

Machine Learning and Geographic Information Systems for large-scale mapping of renewable energy potential

Thèse N° 9376

Présentée le 4 avril 2019

à la Faculté de l'environnement naturel, architectural et construit
Laboratoire d'énergie solaire et physique du bâtiment
Programme doctoral en énergie

pour l'obtention du grade de Docteur ès Sciences

par

Dan ASSOULINE

Acceptée sur proposition du jury

Prof. S. Haussener, présidente du jury
Prof. J.-L. Scartezzini, Dr N. Mohajeri Pour Rayeni, directeurs de thèse
Prof. M. Kanevski, rapporteur
Prof. U. Eicker, rapporteuse
Prof. F. Golay, rapporteur

2019

No idea is original,
there's nothing new under the sun,
it's never what you do,
but how it's done.
— Nasir “Nas” Jones

Acknowledgements

A doctoral thesis is a very enriching experience, an adventure - a long one. So much so that from time to time, most PhD students wonder what they are doing and why they are doing this (yes that includes me), for a second forgetting about the plethora of reasons that originally brought them here. A thesis is paved by a list of choices and turns made during four years (or more) to reach an ultimate and ambitious goal, under the impulsion of advisors, colleagues and loved ones. These are the ones I want to acknowledge.

First and foremost, I want to thank Professor Jean-Louis Scartezzini for greeting me in the LESO-PB lab, for his knowledge and guidance, and for giving me the big picture. Thank you for putting me back on track when my theoretical mind adventures were too far from reality.

Second, I deeply thank Dr. Nahid Mohajeri. This thesis would have been probably very different without the countless discussions we shared about research - and life in general. Thanks for motivating me constantly to do my best during these four years. Lastly, many thanks for introducing me to academia and showing me the ways to the subtle art of writing articles...

I express my gratitude to Prof. Sophia Haussener, Prof. François Golay, Prof. Mikhail Kanevski and Prof. Ursula Eicker for accepting to be a part of the jury for this thesis, and showing interest for my research.

I would like to gratefully acknowledge the financial support of the Swiss Innovation Agency Innosuisse for this thesis, as a part of the Swiss Competence Center for Energy Research SCCER FEEB&D.

My sincere thanks are also due to Dr. Dasaraden Mauree for his precious help regarding wind energy modeling, and Prof. Agust Gudmundsson for his knowledge and support regarding geothermal energy.

I would also like to give my special thanks to Prof. Alexei Pozdnoukhov, for introducing me to the wonderful world of Machine Learning and sparking my interest to explore it for many different applications.

I naturally thank all the LESO-PB team. I cannot make an exhaustive list of people as I will forget some of you and then I'll have to explain myself. Just know that I have all of you in my heart, forever attached to the souvenir of this adventure, for better and for worse - but mostly better of course. A special note goes to Silvia and Alina, who have been my office mates and therefore had to put up with my strange behavior daily. Cheers to both of you.

Friends from Lausanne, Paris, and the rest of the world, thank you for being here; you made the bitter moments sweeter.

Claudia, grazie di esistere! Alright, that was not very sophisticated but I had to, sorry. Thank you for your infinite love and support. You made this thesis possible in many ways.

I obviously cannot end this part without thanking my brother and my parents. For everything.
Love, love, and love.

Lausanne, 25th December 2018

Abstract

A promising pathway to follow in order to reach sustainable development goals is an increased reliance on renewable sources of energy. The optimized use of these energy sources, however, requires the assessment of their potential supply, along with the demand loads in locations of interest. In particular, large-scale supply estimation studies are needed in order to evaluate areas of high potential for each type of energy source for a particular region, and allow for the elaboration of efficient global energy strategies. In Switzerland, the “Energy Strategy 2050”, initiated in 2011 by the Swiss Federal Council, sets an example with the ambitious goal of reaching a 50-80% reduction of CO₂ emissions by the year 2050, with a clear course of action: phasing-out nuclear power, improving energy efficiency, and greatly increasing the use of renewables.

This thesis develops a general data-driven strategy combining Geographic Information Systems and Machine Learning methods to map the large-scale energy potential for three very popular sources of decentralized energy systems: wind energy (using horizontal axis wind turbines), geothermal energy (using very shallow ground source heat pumps) and solar energy (using photovoltaic solar panels over rooftops). For each of the three considered energy sources, an adapted methodology is suggested to assess its large-scale potential, by estimating multiple variables of interest (with a suitable time resolution, e.g. monthly or yearly), using widely available data, and combining these variables into potential values. These latter estimated variables, dictating the potential, include: (i) the monthly wind speed, and rural and urban topographic/obstacle configuration for wind energy, (ii) the ground thermal conductivity, volumetric heat capacity and monthly temperature gradient for geothermal energy, (iii) the monthly solar radiation, available area for PV panels over rooftops, geometrical characteristics of rooftops and monthly shading factors over rooftops for solar energy. The use of Machine Learning algorithms (notably Support Vector Machines and Random Forests) allows, given adequate features and training data (examples for some locations), for the prediction of the latter variables at unknown locations, along with the uncertainty attached to the predictions. In each case, the developed methodology is set-up with an aim to be applied for Switzerland, meaning that it relies on Swiss available energy-related data. Such data, however, including meteorological, topographic, ground/soil-related and building-related data, is becoming progressively available for most countries, making it possible to widely generalize the proposed methodologies.

Results show that Machine Learning is adequate for energy potential estimation, as the multiple required predictions and spatial extrapolations are achieved with reasonable accuracy. In addition, final values are validated with other existing data or studies when possible, and show general agreement. The application of the suggested potential methodologies in Switzerland outline the very significant potential for the considered renewables. In particular, there is a relatively high potential for Rooftop-Mounted solar PV panels, as it is estimated that they could generate a total electricity production of 16.3 TWh per year, which corresponds to 25.3% of the annual electricity demand in 2017.

Keywords: energy potential, spatio-temporal mapping, Machine Learning (ML), Support Vector Machines (SVM), Random Forests (RF), Geographic Information Systems (GIS), Solar energy, Wind energy, Geothermal energy, Switzerland.

Résumé

Face aux enjeux environnementaux actuels, une solution envisageable est de maximiser l'utilisation d'énergies renouvelables. Cependant, afin d'utiliser ces sources d'énergie de façon optimale, il est nécessaire d'estimer leur potentielle contribution énergétique. En particulier, une telle estimation est nécessaire à grande échelle afin d'évaluer les zones à fort potentiel pour chaque type de ressource au sein d'une région, et ainsi de pouvoir mettre en place des stratégies énergétiques efficaces. En Suisse, la "Stratégie Énergétique 2050", instituée en 2011 par le Conseil Fédéral Suisse, se place en exemple en s'imposant l'objectif ambitieux de réduire la consommation de CO₂ de 50-80% d'ici 2050, en suivant trois principales directives: supprimer progressivement l'usage du nucléaire, améliorer l'efficacité énergétique et augmenter drastiquement l'usage d'énergies renouvelables.

Cette thèse développe une stratégie générale, basée sur l'utilisation de données et combinant Systèmes d'Information Géographique et Apprentissage Automatique pour estimer le potentiel de trois énergies alternatives populaires: éolienne (éoliennes à axe horizontal), géothermique (pompes à chaleur très peu profondes) et solaire (Panneaux Photovoltaïques (PV) montés sur les toits des bâtiments). Afin de déterminer le potentiel à grande échelle de chacune des trois énergies, une méthodologie spécifique, estimant plusieurs variables caractéristiques, est proposée, basée sur l'utilisation de données réelles. Les variables estimées sont ensuite combinées pour former le potentiel. Parmi ces variables figurent: (i) la vitesse de vent mensuelle, et les variables reliées à l'occupation du sol et la présence d'obstacles pour l'énergie éolienne, (ii) la conductivité thermique, capacité thermique volumique, ainsi que le gradient de température mensuel du sol pour l'énergie géothermique, et (iii) la radiation solaire mensuelle, l'espace disponible pour l'installation de panneaux solaires sur les toits, les caractéristiques géométriques des toits, et divers facteurs d'ombrage mensuels pour l'énergie solaire. L'utilisation de méthodes d'Apprentissage Aléatoire (notamment de Machines à Vecteurs de Support et Forêts Aléatoires) permet, avec l'aide de prédicteurs adéquates et de données d'entraînement (exemples), la prédiction spatio-temporelle de ces précédentes variables, ainsi que l'incertitude reliée à la prédiction. Dans chaque cas, la méthodologie est appliquée à la Suisse; elle repose donc sur la présence de données disponibles pour le territoire. Néanmoins, les types de données utilisés, notamment des données météorologiques, topographiques, et reliées au sols et au bâtiment, deviennent progressivement disponible dans bon nombre de pays, permettant ainsi la généralisation des méthodologies suggérées.

Les résultats montrent que l'Apprentissage Automatique est adapté à l'estimation de potentiel énergétique, étant donnée la précision raisonnable obtenue pour chaque prédiction et estimation spatiale. Les valeurs finales de potentiel, globalement validées à l'aide de données existantes, mettent en lumière le fort potentiel énergétique renouvelable en Suisse. En particulier, l'utilisation de panneaux solaires PV sur les toits semble être une solution prometteuse: son potentiel est évalué à 16.3 TWh par an, ce qui correspond à 25.3% de la demande annuelle en électricité en 2017.

Mots-clés: potentiel énergétique, estimation spatio-temporelle, Apprentissage Automatique, Machines à Vecteurs de Support, Forêts Aléatoires, Systèmes d'Information Géographique, énergie solaire, énergie éolienne, énergie géothermique, Suisse.

Contents

List of Figures	xv
List of Tables	xxvii
Nomenclature	xxxix
Abbreviations	xxxix
1 Introduction	1
1.1 Motivation	1
1.2 Energy potential estimation	2
1.2.1 Selection of renewable energies	4
1.2.2 Machine Learning for potential estimation	6
1.3 Research goals	7
1.4 Structure of the thesis	7
2 Machine Learning	9
2.1 Learning from data	10
2.1.1 Supervised learning	11
2.1.2 Model selection and model assessment	12
2.2 Support Vector Machines (SVM)	15
2.2.1 The large margin classifier	15
2.2.2 The <i>soft-margin</i> classifier	17
2.2.3 The kernel trick	18
2.2.4 Kernel functions	19
2.2.5 Support Vector Regression	20
2.2.6 SVM in practice	22
2.2.7 Use of SVM in the thesis	22
2.3 Random Forests (RF)	23
2.3.1 Decision Trees	24
2.3.2 Making decision trees better: a short historical note	28
2.3.3 The Random Forests classifier and regressor	30
2.3.4 RF in practice	32
2.3.5 Quantile Regression Forests for Prediction Intervals estimation	35
2.3.6 Use of RF in the thesis	36
2.4 A last note: RF Vs. SVM	37

3	Theory and modeling of renewable energy systems	39
3.1	Wind energy modeling	40
3.1.1	Significant wind variables	41
3.1.2	Vertical structure of the urban atmosphere	41
3.1.3	Vertical wind modeling	42
3.1.4	Rural wind characteristics	43
3.1.5	Urban wind characteristics	44
3.1.6	Wind energy systems	45
3.2	Shallow geothermal energy modeling	47
3.2.1	Significant ground thermal variables	48
3.2.2	Elements of soil structure and texture	49
3.2.3	Inversion of Vertical Electrical Soundings for resistivity estimation	50
3.2.4	Fourier analysis of temperature data for diffusivity estimation	52
3.2.5	Converting electrical to thermal properties	54
3.2.6	Shallow geothermal energy systems	55
3.3	Solar energy modeling	57
3.3.1	Significant solar variables	58
3.3.2	Global solar horizontal radiation models	60
3.3.3	Solar tilted radiation models	61
3.3.4	Solar energy systems	63
4	Wind energy: a theoretical potential estimation	67
4.1	Related literature	69
4.2	Data	71
4.2.1	Data sources	71
4.2.2	Data processing	71
4.3	Wind speed estimation in rural areas	76
4.3.1	Estimation at 10 m	76
4.3.2	Extrapolation at 100 m	82
4.4	Wind speed estimation in urban boundary areas	84
4.4.1	Urban wind characteristics estimation	86
4.4.2	An important assumption	89
4.4.3	Extrapolation above buildings	91
4.5	Results	94
4.5.1	Discussion	96
4.5.2	Preliminary geographical potential estimation	97
4.5.3	Validation with other potential studies	98
4.5.4	Limitations	102
4.6	Summary	103

5	Very shallow geothermal energy: a theoretical potential estimation	107
5.1	Related literature	109
5.2	Data	111
5.2.1	Data sources	111
5.2.2	Data processing	111
5.3	Ground temperature estimation	117
5.4	Thermal conductivity estimation	117
5.4.1	Processing and interpretation of Vertical Electrical Soundings data	121
5.4.2	Estimation and extrapolation of electrical resistivity	122
5.4.3	From electrical resistivity to thermal conductivity	124
5.5	Thermal diffusivity estimation	127
5.5.1	Fourier modeling for thermal diffusivity estimation	127
5.5.2	Extrapolation of diffusivity	129
5.6	Results	130
5.6.1	Discussion	130
5.6.2	Preliminary geographical potential estimation	135
5.6.3	Validation with other potential studies	137
5.6.4	Limitations	138
5.7	Summary	138
6	Solar energy: a technical potential estimation at commune scale	143
6.1	Related literature	144
6.2	Data	146
6.2.1	Data sources	146
6.2.2	Data processing	146
6.3	Theoretical potential estimation	147
6.3.1	Estimation of horizontal solar radiation	147
6.4	Geographical potential estimation	148
6.4.1	Available rooftop area estimation	150
6.4.2	Shading factors estimation	151
6.4.3	Global solar tilted radiation estimation	154
6.4.4	Final geographical potential estimation	157
6.5	Technical potential estimation	158
6.6	Results	159
6.6.1	Discussion	159
6.6.2	Validation with other potential studies	165
6.7	Summary	167

7	Solar energy: an improved potential estimation at pixel scale	169
7.1	Literature Context	171
7.2	Data	172
7.2.1	Data sources	172
7.2.2	Data processing	172
7.3	Theoretical potential estimation	174
7.4	Geographical potential estimation	175
7.4.1	Available rooftop area estimation: an updated methodology	177
7.4.2	Shading factors estimation	180
7.4.3	Rooftop geometrical properties estimation (in OOSG an GEN zones)	183
7.4.4	Global tilted radiation estimation	192
7.4.5	Final geographical potential	193
7.5	Technical potential estimation	195
7.6	Results	196
7.6.1	Discussion	196
7.6.2	Comparison with the first estimation (commune level)	199
7.6.3	Validation with other potential studies	201
7.6.4	Limitations	204
7.7	Summary	206
8	Conclusion	209
8.1	Main findings	209
8.2	Practical implementation	212
8.3	Future outlook	212
Appendices		
A	Data presentation	217
A.1	Time series data	217
A.2	Raster data	220
A.3	Vector polygon data	220
A.4	Vector point data	223
B	Classes in GeoCover data (GK500)	229
C	Extra calculus for chapter 7	233
C.1	Statistical independence of rooftop slope, aspect, and roof type.	233
References		235

List of Figures

1.1	World installed capacity (for electricity production) for various renewable sources of energy, from 2000 to 2017. (a) Cumulative installed capacity; (b) Net addition to installed capacity. Source: IRENA. Data available from http://resourceirena.irena.org/gateway/dashboard/	2
1.2	Installed capacity (for electricity production) for various renewable sources of energy in Switzerland, from 2000 to 2017. (a) Cumulative installed capacity; (b) Net addition to installed capacity. Source: IRENA. Data available from http://resourceirena.irena.org/gateway/dashboard/	3
1.3	Hierarchical potential approach illustration, for solar energy available from photovoltaic panels installed over rooftops. (Note that in this case the geographical potential is called “urban” potential given the installation of PV solar only over building rooftops, and therefore within urban areas.)	4
2.1	K-fold Cross-Validation scheme, here for a number $K = 7$ of folds.	13
2.2	Illustration of SVM for the linearly separable case, in case of a 2D input space. The support vectors are highlighted in red. The slack variable ξ , allowing to tolerate outliers in the soft-margin formulation (as further explained in 2.2.2) is also illustrated for some examples.	17
2.3	Non linear (Kernel) SVM principle. Once mapped (by the implicit mapping ψ induced by the kernel) in a 3D space (the high-dimensional feature space mentioned for Equation 2.18, in (b)), the points are naturally separable by a linear space, here a 2D plane. It translates into a non linear line once mapped back into the original 2D input space, in (a).	19
2.4	The soft margin setting for a linear SVR model, in case of a 2D input space.	21
2.5	Decision tree and partition of the input space. (a) Decision tree example built over a 2D input space defined by two features X_1 and X_2 (light red nodes are leaves, orange nodes are normal nodes, including the root node); (b) Partition of the 2D input space corresponding to the tree shown in a , resulting from recursive binary splitting. S_1 , S_2 etc. are the obtained subspaces after the tree was grown, and h_1 , h_2 etc. are the thresholds extracted by the algorithm at each split.	25

2.6	Example of one of the regression trees trained in an RF model built for wind speed prediction in Switzerland (as presented in chapter 4). Features (predictors) are X (longitude), Y (latitude), altitude, air temperature, cloud cover, precipitation, sunshine duration, air pressure, terrain slope, terrain aspect, terrain plan curvature ("Curve"), terrain transverse curvature ("Curve_Trans"), terrain longitudinal curvature ("Curve_Long"), terrain roughness ("Roughness") and surrounding terrain roughness ("Roughness_neighbors"). In each node are specified: the variable and threshold of the split, the current impurity (as the mean square error, "mse"), the number of input samples in the node ("samples") and the current estimate for the solar irradiance ("value"). Note that the shade of the color of each node is based on the local estimated value: the darker the node, the higher the value.	27
2.7	An illustration of the training and prediction processes with Random Forests, for the regression case. (a) RF training scheme, (b) RF prediction scheme. For readability purposes, the notation is simplified compared to the one adopted in the text: $T_b(x)$ is the prediction performed by tree T_b and is therefore the predicted value for a regression task and the predicted class for a classification task.	32
2.8	Evolution of the OOB score (1–OOB error) with the number of trees B in a RF for different values of m. This example was computed while training a RF for the estimation of the global horizontal radiation in Switzerland (G_h) in July, as presented in chapter 7.	35
3.1	Volume of air received by an area A (area typically spanned by blades of a wind turbine, called A_w in this chapter). Source: [57]	40
3.2	Schematic of typical layering of the atmosphere over a city (by day). Note the height scale is logarithmic, except near the surface. Source: [59].	42
3.3	Wind profile in the RSL (Roughness sublayer) and ISL (Inertial sublayer) measured in a wind tunnel over an array of cubes ($\lambda_f = 0.16$). Values are horizontal spatial averages. z_h is the average building height (called h in this chapter). It can be seen that in this case the Log-law matches the measurements approximately when $z/h \geq 1.5$. Modified from [59]	43
3.4	Calculated power curve for a large commercial horizontal-axis wind turbine, the ENERCON E-101. Figure taken from https://www.enercon.de/en/products/ep-3/e-101/	45
3.5	Wind turbines types. Besides the HAWT, all other turbines are VAWT. (Figure taken from [62])	46
3.6	(a) An example of soil structure arrangement. The soil structure is split between solid soil, water, and air. V_a , V_w , V_s , V_v and V_T are respectively the volume of air, water, solid soil, void ($V_v = V_a + V_w$), and total volume. (b) A soil texture diagram-soil types according to their clay, silt and sand composition, as used by the USDA, redrawn from the USDA webpage: http://soils.usda.gov/education/resources/lessons/texture/ . Source: [79] (Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/A-soil-texture-diagram-soil-types-according-to-their-clay-silt-and-sand-com fig2_235884102 [accessed 5 Dec, 2018].)	51

3.7	Generalized form of electrode configuration in VES resistivity surveys. Source: [82].	52
3.8	Principle scheme of a ground source heat pump. Source: http://nialls.co.uk/ground-source-heat-pumps/	57
3.9	Multiple types of shallow geothermal technologies. Modified from https://www.energieatlas.bayern.de/thema_geothermie/oberflaeche/nutzung.html	58
3.10	(a) Solar radiation basic components (Source: esri.com); (b) : angles to be considered when computing the solar radiation over a tilted PV panel (source: www.urbangreenenergy.com). Note that the sun zenith angle θ_z is not defined here and the altitude angle α_s is preferred. If it was shown, θ_z would be the angle between the vertical direction line and the line linking the sun and the panel, so that $\theta_z + \alpha_s = 90^\circ$	59
4.1	The Swiss territory divided in two zones, treated differently for the estimation of the wind speed in urban areas.	72
4.2	Terrain features extracted from the Digital Elevation Model in Switzerland. (a) DHM 200×200 [m ²], (b) terrain slope, (c) terrain aspect, (d) terrain curvature, (e) terrain longitudinal curvature, (f) terrain transverse curvature.	73
4.3	Prediction of meteorological variables using RF models. (a) Monthly mean yearly sunshine duration (hours), (b) yearly mean air temperature (degree Celsius), (c) monthly mean yearly precipitation (millimetres), (d) yearly mean cloud cover (percentage), (d) yearly mean air pressure (hectoPascals).	75
4.4	Moore neighborhood of a pixel located in (i, j) within a pixel grid of rows indexed by i and columns indexed by j.	76
4.5	Roughness map in rural areas of Switzerland, obtained using the CORINE land cover data and typical roughness values ([60]).	77
4.6	Prediction Intervals (with 95% confidence) extracted from Quantile Regression Forests while training monthly models for rural wind estimation at 10m. We show the PIs for an example of 4 months. (a) and (b) : PIs for February, respectively in the test set and for 30 random unobserved (unknown) points; (c) and (d) : PIs for May, respectively in the test set and for 30 random unobserved (unknown) points; (e) and (f) : PIs for August, respectively in the test set and for 30 random unobserved (unknown) points; (g) and (h) : PIs for November, respectively in the test set and for 30 random unobserved (unknown) points.	79
4.7	Variable Importance of each feature during the training of a Random Forest model over yearly averaged features to estimate the yearly wind speed in rural areas (at a height of 10m). Note that X, Y and Z are respectively the longitude, latitude and altitude.	80
4.8	Wind speed (yearly average) in rural areas, estimated at a height of 10 m. Note that the color map thresholds are chosen abnormally high so that they match the ones used for the extrapolated rural speed map at 100m, later presented in Figure 4.12. They can therefore be compared easily.	81

4.9	Lower and upper uncertainties ($PE_{s,down}$ and $PE_{s,up}$) attached to RF prediction for the yearly wind speed at 10m in rural areas (obtained from the computation of Prediction Intervals with Quantile Regression Forests). (a) Lower prediction error, (a) Upper prediction error.	82
4.10	Prediction error (PE_s , the average of the lower and upper prediction error) attached to RF prediction for the yearly wind speed at 10m in rural areas (obtained from the computation of Prediction Intervals with Quantile Regression Forests).	83
4.11	Monthly maps for wind speed in rural areas, at a height of 10m. Note that lakes are shown in light blue while urban and forest areas are both set to white in order to improve the readability of the maps.	85
4.12	Wind speed (yearly average) in rural areas, estimated at a height of 100m, suitable for wind turbine installation.	86
4.13	Processing scheme for the extraction of the average building height in zone 2, for an example area in Switzerland. (a) TLM3D building footprints, (b) result of the conversion of the TLM3D footprints into 2×2 [m ²] raster cells, (c) DOM, (d) DTM, (e) result of the subtraction DOM-DTM raster computation, (f) final mean building height raster, obtained by clipping the DOM-DTM raster (e) over the TLM3D raster cells (b)	90
4.14	Mean building height map. The small window on the bottom right of the figure zooms in Geneva city.	91
4.15	Urban characteristics estimated in urban pixels. (a) Displacement height estimated in urban boundary pixels, (b) roughness length estimated in urban boundary pixels. The small windows on the bottom right of the two figures zoom in Geneva city.	92
4.16	Example of pixel configuration in Switzerland: white pixels are rural pixels, light red pixel are urban but not in the boundaries of urban areas, dark red pixels are urban pixels in the urban boundaries (meaning neighbors to at least one rural pixel). Three examples of urban boundary pixels are highlighted with their position (i,j) together with their respective rural neighbors considered when computing $u_u(100)$ as explained in section 4.4.2. The rural neighbors are highlighted in light gray with their position with respect to the urban boundary pixel considered at (i,j).	93
4.17	Wind speed (yearly average) in urban boundary pixels, estimated at 5 m above the mean building height (at $z_{hub} = h + 5$).	94
4.18	Wind speed (yearly average) in both rural areas (at a height of 100m) and urban boundaries (at a height of 5 m above buildings). (a) Wind speed map. (b) Zoom in Geneva city, showing the wind speed in rural pixels only (not the urban boundary pixels). (c) Zoom in Geneva city, showing the wind speed in rural and urban boundary pixels.	95

4.19	$C_{p,w}$ /power curves for typical turbines suitable for rural and urban areas. (a) Considered $C_{p,w}$ /power curve for a typical turbine suitable for rural areas. We fit a polynomial over typical power coefficient $C_{p,w}$ values for a large commercial horizontal axis turbine (Reproduced from https://www.enercon.de/en/products/ep-3/e-101/ , for the EN-ERCON E-101 product), and recompute the power curve based on fitted polynomial (degree 6) for $C_{p,w}$. (b) Considered $C_{p,w}$ /power curve for a typical turbine suitable for urban areas. We fit a polynomial over typical power coefficient $C_{p,w}$ values for a small horizontal wind turbine (Reproduced from [75]), and recompute the power curve based on fitted polynomial (degree 3) for $C_{p,w}$	97
4.20	Wind power (yearly average) potential map, computed based on the estimated wind speed and typical turbine characteristics. (a) Wind power map. (b) Zoom in Geneva city, showing the power in rural pixels only (not the urban boundary pixels). (c) Zoom in Geneva city, showing the power in rural and urban boundary pixels.	98
4.21	Estimated wind speed (yearly average) for both rural and urban boundary areas, aggregated within Switzerland cantons. (a) Average canton wind speed, (b) Maximum canton wind speed, (c) Range of wind speed in each canton.	99
4.22	Estimated power (yearly average) aggregated within Switzerland cantons, for rural areas (considering a typical large commercial turbine), along with the estimated wind speed. (a) Average canton wind speed, (b) Maximum canton wind speed	100
4.23	Estimated power (yearly average) aggregated within Switzerland cantons, for urban areas (considering a typical small household turbine), along with the estimated wind speed. (a) Average canton wind speed, (b) Maximum canton wind speed	101
4.24	Comparison between Swisstopo yearly wind estimations at 100m and our estimations for 80 pixels chosen randomly in Switzerland in rural areas. Note that the RMSE and NRMSE given is computed over all rural pixels, not only the 80 shown here.	102
4.25	Comparison between wind measurements and estimated values in 17 urban boundary pixels.	103
4.26	Flow chart of the methodology for the theoretical wind potential estimation at pixel scale.	105
5.1	GK500 vector polygons, offering information on the superficial geological strata in Switzerland. Here we display, the geological period characterizing each polygon. (Creta., Quat., "P-C", and "Tec-mixture" respectively stands for Cretaceous, Quaternary, "Permian-Cretaceous" and "Tectonic-mixture"	112
5.2	Marginal PDFs and all possible 2-joint PDFs for the percentages of sand, silt, clay, and F for one example of rock type ("Silty sands with gravels and blocks").	114
5.3	Yearly soil moisture map estimated from the SMAP data in cm^3/cm^3	116
5.4	Monthly ground temperature in Switzerland. Temperatures are shown at 5, 10, 20, 50 and 100cm, and are averaged through all the measurement stations available in Switzerland.	118
5.5	Monthly ground temperature maps as estimated for Switzerland for a depth of 100cm.	118

5.6	Yearly ground temperature maps as estimated for Switzerland at the depths of 5, 10, 20, 50 and 100cm.	119
5.7	Prediction Intervals (with 95% confidence) from Quantile Random Forests for the monthly ground temperature at a depth of 100 and 50cm for an example of 2 months. (a) and (c): PIs in the test set, in January, respectively for 100cm and 50cm; (b) and (d): PIs for 30 random new points, in January, respectively for 100cm and 50cm; (e) and (g): PIs in the test set, in June, respectively for 100cm and 50cm; (f) and (h): PIs for 30 random unknown points, in June, respectively for 100cm and 50cm.	120
5.8	Vertical Electrical Sounding data point locations in Switzerland. “Int” stands for “Intepreted”.	122
5.9	Vertical Electrical Sounding data point inversion example. The right graph shows the measured electrical resistivities given by Ohm’s law for different distances between electrodes A and B in red and the forward model resulting from the inversion in blue (ρ_a is the apparent resistivity, and the RMSE between the two set of values is specified); the left graph shows the resulting depths and resistivities of the ground layers obtained from the inversion.	123
5.10	Electrical resistivity map as estimated in the study with visualization of PIs (with 95% confidence) both in the test set and for new points. (a) Electrical resistivity (ρ) map; (b) PIs in the test set for the label variable used ($\log(5 + \frac{\rho}{c})$ and not ρ) while training the RF model; (c) PIs for 30 random unknown points for the same label variable. . . .	125
5.11	Prediction error attached to the electrical resistivity estimation (for each GK500 polygon) obtained from the trained RF model. The error is computed as the average of the down and up ($PE_{s,down}$ and $PE_{s,up}$) width of Prediction Intervals computed with Quantile Regression Forests.	126
5.12	Thermal conductivity map as estimated in the study with vizualisation of PIs (with 95% confidence) both in the test set and for new points. (a) Thermal conductivity (λ) map; (b) PIs in the test set for the label variable used ($\log(\rho_t)$ where ρ_t is the thermal resistivity) while training the RF model; (c) PIs for 30 random unknown points for the same label variable.	128
5.13	Fourier analysis of a ground temperature time series example. The figure shows, for depth of 10cm, in Bern, during the year 2013: (a) daily average temperature time series over the year, (b) and (d) amplitude and phase of the 30 first frequencies, (c) harmonics for the 3 dominant frequencies (here $n = 1$, $n = 4$ and $n = 11$) and resulting Fourier approximation of the signal.	129
5.14	Slope fitting for diffusivity estimation example. The linear fit between $\ln(R_n)$ and $z\sqrt{n}$ is shown for the three dominant harmonics ($n = m_1, m_2, m_3$) in Fourier analysis in Bern, in 2013, 2014, 2015 and 2016.	129

5.15	Thermal diffusivity map as estimated in the study with vizualisation of PIs (with 95% confidence) both in the test set and for new points. (a) Thermal diffusivity (α) map; (b) PIs in the test set for the label variable used ($\log(\alpha \times C)$) while training the RF model, with soil texture features considered; (c) PIs for 30 random unknown points for the same label variable, with soil texture features considered; (d) PIs in the test set for the label variable used ($\log(\alpha \times C)$) while training the RF model, without soil texture features; (e) PIs for 30 random unknown points for the same label variable, without soil texture features.	131
5.16	Prediction error attached to the thermal diffusivity estimation (for each GK500 polygon) obtained from the trained RF model. The error is computed as the average of the down and up ($PE_{s,down}$ and $PE_{s,up}$) width of Prediction Intervals computed with Quantile Regression Forests.	132
5.17	Estimated variables aggregated (average) at canton level. (a) Volumetric Heat Capacity and Thermal Conductivity aggregated at canton level, (b) Thermal Diffusivity and Electrical Resistivity aggregated at canton level.	136
5.18	Estimated preliminary geographical potential for very shallow geothermal systems. (a) Heating potential, during one heating season in a year, (b) Cooling potential, during one cooling season in a year.	140
5.19	Flow chart of the methodology for the theoretical shallow geothermal potential estimation at pixel scale.	141
6.1	Prediction of horizontal solar radiation maps using SVR. (a) Yearly mean global horizontal radiation, kWh/m ² , (b) Yearly mean diffuse horizontal radiation, kWh/m ² , (c) Yearly mean extraterrestrial, kWh/m ²	149
6.2	Schematic presentation of different steps to estimate available roof area using ArcGIS. (a) building polygons with detailed roof geometries including superstructure (e.g. chimney, dormers, staircase), (b) removing the superstructures from roof surfaces, (c) creating 1 m ² buffer around each remaining roof surfaces, (d) the final available roof area for PV installation after removing the areas less than 28 m ²	152
6.3	Final available area maps in Switzerland. (a) the ratio C_R of average available roof area for PV solar (A_R) to the average ground floor area (A_f) for each commune, percentage (b) total available roof area in each commune, expressed in km ²	153

6.4	Schematic presentation of calculating shading factors (S_{Sh} and S_{hill}). (a) buildings with detailed roof geometry, (b) dissolve detailed roof geometry into a continuous polygon, keeping the same outer boundaries as the original polygons, (c) reverse vector map of buildings as void and the surroundings as filled polygons, (d) DOM (Digital Orthophoto Map) map with a 2 m by 2 m resolution, (e) “negative” DOM raster map extracted by clipping the DOM map (d) over the “reverse” building polygons (c), (f) a boolean raster map or IsNull raster map, Null cells (void cells) assign to value 1 indicating there are buildings, whereas not Null cells assign to value 0 indicating there are no buildings, (g) Hillshade map for buildings and their landscape surroundings in urban areas, (h) clipped Hillshade map for buildings extracted by clip Hillshade map (g) over IsNull raster map (f), (i) binary raster map showing cells that are non-fully shaded in yellow and cells that are fully shaded in blue.	155
6.5	Yearly solar radiation for one specific slope and azimuth configuration (slope=35° and azimuth=10°) out of 63 possible configurations.	158
6.6	Technical potential of rooftop PV solar electricity production for each commune in Switzerland (in GWh/month)	160
6.7	(a) Technical potential of rooftop PV electricity production (GWh/year) for each communes in Switzerland; (b) Histogram showing the distribution of rooftop PV electricity production (GWh/year) among 1901 communes in Switzerland. The maximum values belong to large cities (e.g. Zurich, Bern, Basel).	161
6.8	A comparison between monthly PV electricity production (GWh/month) for 1901 communes in Switzerland and Swiss electricity consumption in 2015, in GWh/month.	162
6.9	(a) Technical potential of rooftop PV electricity production normalized by the population (MWh/year/capita); (b) Histogram showing the distribution of rooftop PV electricity production (MWh/year/capita) among 1901 communes in Switzerland.	163
6.10	Technical potential of rooftop PV electricity production normalized by the population for each commune in Switzerland, in kWh/month/capita.	164
6.11	(a) Cumulative rooftop PV solar electricity production for each canton in TWh/year and cumulative rooftop PV solar electricity production per capita for each canton in MWh/year/capita, in Switzerland; (b) cumulative rooftop PV solar electricity production for each canton, TWh/year and total available roof area, km ² , for each canton in Switzerland.	165
6.12	Flow chart of the methodology for the solar PV potential estimation at commune scale.	168
7.1	An schematic map of Switzerland showing the location of three zones for different data availability and different data processing (see section 7.2.1). GEN: Geneva canton zone, SON: zone spanned by the “Sonnendach” data, OOSG: the “Out Of SON and GEN” zone, meaning the remaining territory in Switzerland.	173
7.2	Prediction of horizontal solar radiation maps using RFs. (a) Yearly mean global horizontal radiation (G_h), in kWh/m ² , (b) Yearly mean diffuse horizontal radiation (G_D), in kWh/m ²	176

7.3	Prediction Intervals (PIs) from Quantile Random Forests for global horizontal and diffuse horizontal radiations (G_h and G_D). (a) and (c) : PIs for G_h in the test set, respectively in January and March, (b) and (d) : PIs for G_h for 30 random unobserved points, respectively in January and March, (e) and (g) : PIs for G_D in the test set, respectively in January and March, (f) and (h) : PIs for G_D for 30 random unobserved points, respectively in January and March. The test confidence (percentage of observed points within the interval) is given for PIs in the test set.	177
7.4	Available Area labelling process. It is divided in steps (a) Detailed building polygons, containing all roof characteristics, including all superstructures, (b) the superstructures are removed from the roofs, (c) a buffer of 40 cm is created around the roof surfaces, (d) PV panels are virtually installed over the roofs where it is possible, as described in step 3 in section 7.4.1.	179
7.5	(a) Available roof area map. It provides the total available area over rooftops for PV installation in each considered pixel, combining the results obtained in both SON and OOSG regions. (b) and (d) : prediction intervals from Quantile Random Forests for the available area ratios, respectively C_R^c and C_R^s , in the test set, (c) and (e) : prediction intervals from Quantile Random Forests for the available area ratios, respectively C_R^c and C_R^s , for 30 random unobserved points. The test confidence (percentage of observed points within the interval) is given for intervals in the test set.	181
7.6	Prediction Intervals (PIs) from Quantile Random Forests for the two estimated shading variables (S_{Sh} and S_{hill}). (a) and (c) : PIs for S_{Sh} in the test set, respectively in January and March, (b) and (d) : PIs for S_{Sh} for 30 random unobserved points, respectively in January and March, (e) and (g) : PIs for S_{hill} in the test set, respectively in January and March, (f) and (h) : PIs for S_{hill} for 30 random unobserved points, respectively in January and March. The test confidence (percentage of observed points within the interval) is given for PIs in the test set.	183
7.7	Convention used for aspect calculation.	184
7.8	Aspect and Slope reclassified ($0,5 \times 0,5$) [m^2] rasters, along with the building polygons from VECTOR25 (grey thick line) and Sonnendach data (black thin line) for a building with a hipped (a,b,c) and a gabled roof (d,e,f). One can observe the significant delay of position between the two polygons. Also, the different aspect and slope patterns between the two types are clearly shown, specially regarding the amount of roof raster cells showing a flat surface, significantly larger in the gable case.	185
7.9	Roof classification scheme.	186
7.10	(a) Yearly technical potential for rooftop PV electricity production, in MWh/year; (b) Zoom in the Zurich urban area (the zoom location is signified by the yellow window within the map); (c) Further zoom within the Zurich urban area at the pixel level. . .	196
7.11	Monthly technical potential for rooftop PV solar electricity production, in MWh/month.	197

7.12	Monthly demand profile (for 2017), in blue, and rooftop PV solar potential for the Swiss cantons, in yellow, in GWh/month. The abbreviations used for the canton names are the following: AR:Appenzell Ausserrhoden, AI:Appenzell Innerrhoden, BL:Basel-Landschaft, BS:Basel-Stadt, BE:Bern, FR:Fribourg, GE:Geneve, GL:Glarus, GR:Graubunden, JU:Jura, LU:Luzern, NE:Neuchatel, NW:Nidwalden, OW:Obwalden, SH:Schaffhausen, SZ:Schwyz, SO:Solothurn, SG:St. Gallen, TG:Thurgau, TI:Ticino, UR:Uri, VS:Valais, VD:Vaud, ZU:Zug, ZH:Zurich.	199
7.13	Comparison between the electricity consumption (for 2017) and the estimated rooftop PV solar potential at two aggregated levels in Switzerland: (a) at the national level, monthly, and (b) at the cantonal level, yearly. The abbreviations used for the canton names are the following: AR:Appenzell Ausserrhoden, AI:Appenzell Innerrhoden, BL:Basel-Landschaft, BS:Basel-Stadt, BE:Bern, FR:Fribourg, GE:Geneve, GL:Glarus, GR:Graubunden, JU:Jura, LU:Luzern, NE:Neuchatel, NW:Nidwalden, OW:Obwalden, SH:Schaffhausen, SZ:Schwyz, SO:Solothurn, SG:St. Gallen, TG:Thurgau, TI:Ticino, UR:Uri, VS:Valais, VD:Vaud, ZU:Zug, ZH:Zurich.	200
7.14	Yearly technical potential normalized by the population in each pixel for rooftop PV electricity production, in MWh/year/capita.	201
7.15	Variable Importance (VI). The graph shows the importance of each feature during the training of a Random Forest for the estimation of the available area in OOSG/GEN zones (describe in section 7.4.1). Here the categorical features are aggregated to reduce the original number of features for visualization purposes (e.g. instead of considering 12 construction period features, we aggregate them to one feature expressing the most frequent period in a pixel). The features in the x axis are as follows: CatFeature: most frequent residential class, PeriodFeature: most frequent construction period, ClassFeature: most frequent building typology, Number_of_Vertices: average number of vertices of building cluster polygons, vec25_Db_build: number of building clusters, vec25_Db_area: ratio of the total ground floor area of building clusters to the total pixel area, Db_build(build_per_km2): number of buildings, Vec25MeanAreaGeneva: average ground floor area of building clusters, Nombre_logements: average number of building flats, Isopq: average isoperimeter quotient of building cluster polygons, Nombre_niveaux: average number of building floors, footprint_mean: ground floor area, Db_area(build_area_per_km2): ratio of the total ground floor area of individual buildings to the total pixel area.	202
7.16	Comparison between the yearly rooftop PV solar potential estimates from SITG and the current study, for all communes in the Geneva canton in GWh/year.	204
7.17	Flow chart of the methodology for the PV solar potential estimation improved at pixel scale.	207
A.1	Locations of measurement stations for weather data, used for training the weather models. (a) solar radiation, (b) sunshine duration, (c) precipitation, (d) cloud cover, (e) temperature.	218

A.2	Locations of additional measurement stations for data, used for training various models. (a) Wind, (b) Air pressure, (c) Snow cover, (d) Ground temperature.	219
A.3	Illustration for some of the elevation models available for Switzerland. (a) Digital Elevation Model (DEM), at a resolution of 250×250 [m ²], (b) Digital Height Model (DHM), downsampled to a resolution of 200×200 [m ²], (c) Digital Height Model (DHM), at a resolution of 25×25 [m ²],	222
A.4	Illustration for the most precise elevation models available for Switzerland. (a) Digital Orthophoto Map (DOM), at a resolution of 2×2 [m ²], (b) Digital Terrain Model (DEM), at a resolution of 2×2 [m ²], for the same area than (a).	223
A.5	Corine polygons defining the land use in Switzerland. Land use categories are simplified for illustration purposes.	224
A.6	Illustration for some of the vector polygon building data available in Switzerland. (a) Building footprint data (TLM3D), (b) Bulding rooftop data (swissBUILDINGS3D/Sonnendach), (c) Building clusters (VEC25), (d) Digital Orthophoto Map (DOM).	225
A.7	Illustration for some of the building vector polygon data available for the canton of Geneva. (a) Building rooftops data for Geneva canton (SITG rooftops), (b) Building superstructures for Geneva canton (SITG superstructures).	227

List of Tables

1.1	Annual Swiss production for different types of renewable energies, in 2017. (HAWT: Horizontal Axis Wind Turbines, HP: Heat Pumps, PV: PhotoVoltaic. The figures are taken from SFOE [6])	5
2.1	Comparison of tree-based Ensemble Learning classical models. For each strategy, the assessment is based on a benchmark model: CART for trees, bagging CART predictors for bagging and AdaBoost for boosting. (✓ signifies low ability, ✓ signifies medium ability and ✓ signifies high ability)	31
3.1	Roughness length ($z_{0,u}$) values table proposed for the Corine Land Cover (CLC) classes, as described in [60].	44
3.2	Recommnded days from Klein et al. [109] to compute the earth declination for each month in a year.	63
3.3	Diffuse tilted radiation models. H, D, and H&D mean that the model is suitable respectively for hourly, daily and both hourly and daily estimations.	64
4.1	Testing RMSE (E_R , in the same unit as the variable of interest) and NRMSE (E_{NR} , in %) for Random Forest models trained for weather variables.	74
4.2	Errors related to the building of monthly wind speed in rural areas using RF models. Left side of the table: Testing errors for each monthly model, in the form of the RMSE (E_R , in m/s), NRMSE (E_{NR} , in %) and OOB score (between -1 and 1). Right side of the table: Monthly Prediction Errors, computed using Quantile Regression Forests, averaged over a random sample of 1000 unobserved pixels. $PE_{s,down}$ is the average bottom error above the mean predicted value, $PE_{s,up}$ is the upper error above the mean predicted value, PE_s is the average of $PE_{s,down}$ and $PE_{s,up}$	78
5.1	Testing RMSE (E_R) and NRMSE (E_{NR} , in percentage) for Random Forest models trained for monthly ground soil moisture (Volumetric Water Contents).	116
5.2	Testing RMSE (E_R), NRMSE (E_{NR} , in percentage) and OOB score for Random Forest models trained for monthly ground temperature at multiple depths.	121
6.1	Testing RMSE (E_R) and NRMSE (E_{NR} , in percentage) from SVR models trained for weather and solar variables.	148
6.2	Testing errors (RMSE and NRMSE) for G_h , G_D and G_{oh} , while no weather variables are considered (no w.) and weather variables are considered (w.).	150

6.3	Testing errors (RMSE and RRMSE) in SVR training for slope distribution (β) and for the available area ratio (C_R), that is, the ratio of the average available area for PV to the average ground floor area.	151
6.4	Testing errors (RMSE and NRMSE) for shading variables using SVR.	156
7.1	Testing RMSE (E_R) and NRMSE (E_{NR} , in percentage) for Random Forest models trained for weather and solar variables.	175
7.2	Prediction Errors related for G_h and G_D . Monthly Prediction Errors, computed using Quantile Regression Forests, averaged over a random sample of 10000 unobserved pixels. $PE_{s,down}$ is the average bottom error above the mean predicted value, $PE_{s,up}$ is the upper error above the mean predicted value, PE_s is the average of $PE_{s,down}$ and $PE_{s,up}$	176
7.3	Testing RMSE (E_R) and NRMSE (E_{NR} , in percentage) for Random Forest models trained for available area ratios (C_R^c , C_R^s). We provide also the Out Of Bag (OOB) score to highlight the difficulty to estimate C_R^c : although the E_R and E_{NR} errors for C_R^c are very high, the OOB score is larger than 0, which means our model brings improvement over a simple average estimate. PE_s is the average prediction error estimated on a random sample of 10000 unknown pixels.	180
7.4	Left side of the table: Testing RMSE (E_R) and NRMSE (E_{NR} , in percentage) for Random Forest models trained for shading variables (S_{hill} , S_{Sh}); Right side of the table: Monthly Prediction Errors, computed using Quantile Regression Forests, averaged over a sample of 10000 unobserved pixels. PE_s is the average of the lower and upper error (half width of PIs)	182
7.5	Roof shape estimation confusion matrix for Classification 1.	188
7.6	Roof shape estimation confusion matrix for Classification 2.1.	188
7.7	Roof shape estimation confusion matrix for Classification 2.2.	188
7.8	Final accuracy of the overall classifier to detect each roof class.	189
7.9	Aspect estimation confusion matrix.	190
7.10	Slope estimation confusion matrix.	191
7.11	Roof characteristics considered for each roof type. β_l and γ_m are the center value of respectively the main slope (tilt) and the main aspect (direction) class predicted for the roof of interest.	191
7.12	Roof characteristics considered in the spreading function F_{c_n} , depending on the roof type c_n . A_R^c is the average roof available area for PV installation. Note that the available area is used for all sides, besides the special case of Hip roofs, where the area is different in the two main sides and the two "hips" (the two small triangles). See previous Table 7.11 for more details on the considered roof characteristics.	195
7.13	Comparison of the estimated PV solar potential in 20 randomly chosen pixels between Sonnendach project potential study and the present study. Note that the Relative error is the absolute error between the two estimations, in MWh/year, while the Absolute error is the ratio of the Relative error to the sonnendach estimation value, in %	205

A.1	Time series and raster datasets used in the thesis, along with their characteristics and reference.	221
A.2	Vector polygons, points and other datasets used in the thesis, along with their characteristics and reference. (FSO: Federal Statistical Office, SITG: Systeme d'Information du Territoire Genevois.)	226
B.1	Hydrogeology [HYDRO] categories in GK500 dataset. N_{GK500} is the number of polygons for each category and Area is the total area spanned by all polygons of the category. .	229
B.2	Geological period [PERIOD] categories in GK500 dataset. N_{GK500} is the number of polygons for each category and Area is the total area spanned by all polygons of the category.	229
B.3	Aquifer productivity [PRODUCTIV] categories in GK500 dataset. N_{GK500} is the number of polygons for each category and Area is the total area spanned by all polygons of the category.	230
B.4	Rock formation types [TYPE ROCHE] categories in GK500 dataset. N_{GK500} is the number of polygons for each category and Area is the total area spanned by all polygons of the category.	230
B.5	Rock/soil types [LITH PET] categories in GK500 dataset. N_{GK500} is the number of polygons for each category and Area is the total area spanned by all polygons of the category. $N_{\text{soil samples}}$ is the number of NABODAT soil texture points available in each category.	231
C.1	Marginal and conditional probabilities for S (main slope variable) and $S \mid D$ (D being the main aspect variable). Mean Cond. is the average conditional probability $\mathbb{P}(S = \beta \mid D = \gamma)$ for each β value.	233
C.2	Marginal and conditional probabilities for S (main slope variable) and $S \mid T$ (T being the roof type variable). Mean Cond. is the average conditional probability $\mathbb{P}(S = \beta \mid T = c)$ for each β value.	234
C.3	Marginal and conditional probabilities for T (roof type) and $T \mid D$ (D being the main aspect variable). Mean Cond. is the average conditional probability $\mathbb{P}(T = c \mid D = \gamma)$ for each c value.	234

Nomenclature

Machine Learning (Chapter 2)

α_n, α_n^*	[-]	Lagrange multipliers
$\Delta i(v)$	[-]	Impurity decrease between node v and children nodes v_L and v_R
ε	[-]	Margin width (tolerated error) in SVR
ϵ	[-]	Generic random error
ζ_n, ζ_n^*	[-]	Slack variables in SVR
ξ	[-]	Slack variable in SVM classification
σ	[-]	Deviation of gaussian (hyperparameter for SVM with RBF kernel)
φ	[-]	Generic predictive function learned from training process
ψ	[-]	Mapping to the feature space (kernel trick)
$\omega_i(\mathbf{x}, T)$	[-]	Weight stored by tree T for training sample \mathbf{x}_i (within a RF)
b	[-]	Constant (offset) in the SVM solution
B	[-]	Number of trees in a Random Forest
c	[-]	Number of classes considered in a classification task
C	[-]	Hyperparameter of SVC and SVM to control the outliers penalty
d	[-]	Number of features (predictors) in the training data for an ML task
\mathbb{E}	[-]	Expectation operator
E_R	[-]	Root Mean Square Error
E_{NR}	[-]	Normalized Root Mean Square Error
f	[-]	Generic solution function extracted by a ML model
h	[-]	Threshold defining a splitting query in a tree
$i(v)$	[-]	Impurity of node v in a decision tree
I_{95}	[-]	95% Prediction Interval
$K(.,.)$	[-]	Kernel function
$l(\mathbf{x}, T)$	[-]	Leaf reached by data point \mathbf{x} when passed through tree T
\mathcal{L}	[-]	Loss function
m	[-]	Number of variables to consider to split a node v in a decision tree

N	[-]	Number of training points (size of training data)
$N_{l(x,T)}$	[-]	Number of training samples in leaf $l(x, T)$ in tree T
N_{PV}	[-]	Number of PV panels fitting within within a roof polygon
N_{test}	[-]	Number of testing point (Size of testing data)
N_v	[-]	Number of learning samples in node v in a decision tree
N_{v_L}	[-]	Number of learning samples in left child node v_L in a decision tree
N_{v_R}	[-]	Number of learning samples in left child node v_R in a decision tree
\mathbb{P}	[-]	Probability operator
Q_α	[-]	α -quantile
\mathbb{R}	[-]	Set of real numbers
S_v	[-]	Subspace stored in a node v within a
$S_L(j, h)$	[-]	Subspace stored in left child node during a tree splitting
$S_R(j, h)$	[-]	Subspace stored in right child node during a tree splitting
T	[-]	Generic decision tree
v	[-]	Node in a decision tree
v_L	[-]	Left child node of node v in a decision tree
v_R	[-]	Right child node of node v in a decision tree
w	[-]	Vector parametrizing the solution of SVM
x	[var. unit]	Generic input vector (realisation of X_1, \dots, X_d)
X_1, \dots, X_d	[var. unit]	Generic input variables
y	[var. unit]	Generic output value (label, realisation of Y)
Y	[var. unit]	Generic ouput variable
$\mathbb{1}$	[-]	Indicator function

Wind energy (Chapter 4)

β	[-]	Coefficient in Macdonald models, reflecting the obstacle configuration
κ	[-]	Von Karman constant
λ_f	[-]	Frontal area ratio (the ratio of the total facade area of the surface obstacles to the total plan area)
λ_p	[-]	Plan area ratio (the ratio of the total plan area of the surface obstacles to the total plan area)
ρ	[kg/m ³]	Density of air

A	[-]	Coefficient in Macdonald models, reflecting the obstacle configuration
A_w	[m ²]	Area spanned by rotor of wind turbine
C_D	[-]	Drag coefficient of a single obstacle
$C_{p,w}$	[-]	Coefficient of performance of a wind turbine
h	[m]	Mean building height
L	[-]	Number of rural neighbors of an urban pixel
n_l	[-]	l^{th} rural neighbor of an urban pixel
P_w	[W]	Power delivered by a wind turbine
u	[m/s]	Generic wind speed variable
u^*	[m/s]	Friction velocity
u_r	[m/s]	Wind speed in a rural area
u_u	[m/s]	Wind speed in an urban area
z	[m]	Generic altitude variable
z_0	[m]	Generic Roughness length variable
$z_{0,r}$	[m]	Roughness length in a rural area
$z_{0,u}$	[m]	Roughness length in an urban area
z_d	[m]	Displacement height (in an urban area)
$z_{d,r}$	[m]	Displacement height in a rural area
z_{hub}	[m]	Displacement height in a rural area

Geothermal energy (Chapter 5)

α	[m ² s ⁻¹]	Thermal diffusivity
γ_d	[g/cm ³]	Dry (bulk) density
γ_s	[g/cm ³]	Particle density
γ_w	[g/cm ³]	Water density
λ	[WK ⁻¹ m ⁻¹]	Thermal conductivity
ρ	[Ωm]	Electrical resistivity
ω	[s ⁻¹]	Angular frequency of one period in Fourier series
c_v	[Jm ⁻³ K ⁻¹]	Volumetric heat capacity
D	[m]	Damping depth
e	[-]	Void ratio
F	[%]	Percentage sum of sand and gravel fractions in the soil

h_i	[m]	Width of soil strata i
I_1, \dots, I_{10}	[-]	Possible fraction intervals for soil texture variables
M_s	[g]	Mass of solid soil in ground
M_w	[g]	Mass of water in ground
n	[-]	Harmonics index in Fourier series
R_n	[-]	Amplitude of n^{th} harmonic of Fourier series solution for T
S_d, S_t, C_l	[%]	Sand, silt and clay fraction percentage in soil
t	[s]	Time
T	[°C]	Shallow ground temperature
w	[%]	Gravimetric water content
z	[m]	Ground depth

Solar energy (Chapter 6)

β	[°]	Roof tilted angle
γ	[°]	Roof azimuth angle
γ_s	[°]	Sun azimuth angle
δ	[°]	Declination (monthly estimated)
η	[%]	PV panel efficiency
θ	[°]	Incidence angle of the sun rays on a tilted plane
θ_z	[°]	Sun zenith angle
ρ	[-]	Foreground's albedo
ϕ	[°]	Latitude
A_f	[m ²]	Average of building ground floor area in a commune
$A_{f, \text{sum}}$	[m ²]	Total building ground floor area in a commune
b_i	[-]	Number of buildings in urban area i
C_R	[-]	Average portion of ground floor area available for PV over a building in a commune
D_b	[-]	Building density in a commune
D_p	[-]	Population density in a commune
G_h	[kWh/m ²]	Global Horizontal Radiation
$G_{h, \text{ no w.}}$	[kWh/m ²]	Global Horizontal Radiation, estimated without weather variables
$G_{h, \text{ w.}}$	[kWh/m ²]	Global Horizontal Radiation, estimated with weather variables

G_B	[kWh/m ²]	Direct (Beam) Horizontal Radiation
G_{Bt}	[kWh/m ²]	Direct (Beam) Tilted Radiation
G_D	[kWh/m ²]	Diffuse Horizontal Radiation
G_{Dt}	[kWh/m ²]	Diffuse Tilted Radiation
G_{oh}	[kWh/m ²]	Extraterrestrial Horizontal Radiation
G_{Rt}	[kWh/m ²]	Reflected Tilted Radiation
G_t	[kWh/m ²]	Global tilted Radiation
$N_{sh,x}$	[-]	Number of fully shaded 2×2 [m ²] cells over roofs in a commune, at time x
N_{cells}	[-]	Number of 2×2 [m ²] cells over roofs in a commune
p_l	[-]	Prob. of each possible slope value (7 values) to be within a frequency bin l
$P_{j,geo}$	[GWh]	Geographical potential in commune j
$P_{j,tech}$	[GWh]	Technical potential in commune j
q_m	[-]	Prob. of each possible aspect value (9 values) to be within a frequency bin m for one side and –m for the other side
R_b	[-]	Beam Radiation factor
R_d	[-]	Diffuse Radiation factor
R_r	[-]	Reflected Radiation factor
S_{sh}	[-]	Daily ratio of fully shaded cells over rooftops in a commune
S_{hill}	[-]	Hillshade average value over partially shaded cells in a commune
X^j	[-]	variable X in commune j
X_i	[-]	variable X in urban area i (defined by CORINE land use)

Solar energy (Chapter 7)

β	[°]	Roof tilted angle
γ	[°]	Roof azimuth angle
γ_s	[°]	Sun azimuth angle
δ	[°]	Declination (monthly estimated)
θ	[°]	Incidence angle of the sun rays on a tilted plane
θ_z	[°]	Sun zenith angle
η	[%]	PV panel efficiency

ρ	[-]	Foreground's albedo
ϕ	[°]	Latitude
A_t^s	[m ²]	Average tilted area over a roof surface s in a pixel
A_f^c	[m ²]	Average ground floor area for a building cluster c in a pixel
b_j	[-]	Number of buildings in pixel j
c_n	[-]	Roof type of class n
C_R^c	[-]	Average portion of ground floor area available for PV over a building cluster c
C_R^s	[-]	Average portion of tilted area available for PV over a roof surface s
D_b	[-]	Building density
D_p	[-]	Population density
F_{c_n}	[-]	Spreading function distributing aspect and roof area values over a roof based on its type c_n
G_h	[kWh/m ²]	Global Horizontal Radiation
G_D	[kWh/m ²]	Diffuse Horizontal Radiation
G_B	[kWh/m ²]	Direct (Beam) Horizontal Radiation
G_t	[kWh/m ²]	Global tilted Radiation
G_{Dt}	[kWh/m ²]	Diffuse Tilted Radiation
G_{Bt}	[kWh/m ²]	Direct (Beam) Tilted Radiation
G_{Rt}	[kWh/m ²]	Reflected Tilted Radiation
G_{oh}	[kWh/m ²]	Extraterrestrial Horizontal Radiation
h_R	[m ²]	Available area for PV panels over the lateral sides for a hipped roof
$N_{sh,x}$	[-]	Number of fully shaded 2×2 [m ²] cells over roofs in a pixel at time x
N_{cells}	[-]	Number of 2×2 [m ²] cells over roofs in a pixel
p_l	[-]	Prob. for a building cluster roof to have a main slope of β_l in a pixel
$p_{j,geo}^{OOS}$	[GWh]	Geographical potential in pixel j , for OOSG and GEN zones
$p_{j,geo}^{SON}$	[GWh]	Geographical potential in pixel j , for SON zone
$P_{j,tech}$	[GWh]	Technical potential in pixel j
q_m	[-]	Prob. for a building cluster roof to have a main asp. of γ_m in a pixel
r_n	[-]	Probability for a building cluster to have a roof of type c_n in a pixel
R_b	[-]	Beam Radiation factor
R_d	[-]	Diffuse Radiation factor

R_r	[-]	Reflected Radiation factor
S_{sh}	[-]	Daily ratio of fully shaded cells over rooftops
S_{hill}	[-]	Hillshade average value over partially shaded cells
X^j	[-]	variable X in pixel j

Other (used on Chapter 3)

ϕ_n	[°]	Amplitude of general Fourier solution for T
ω_s	[°]	Sunset hour angle
ω_{sr}	[°]	Sunrise hour angle on titlted surface
ω_{ss}	[°]	Sunset hour angle on titlted surface
a_n	[-]	First Fourier coefficient for the n^{th} harmonics
A	[-]	Ratio of beam horizontal to extra-terrestrial horizontal radiation
\mathcal{A}	[-]	Function of angles in solar direct models
\mathcal{B}	[-]	Function of angles in solar direct models
b_n	[-]	Second Fourier coefficient for the n^{th} harmonics
m	[-]	Recommended representative day of the month
n	[-]	Harmonics index in Fourier series
n_p	[-]	Porosity
S_r	[%]	Saturation degree
T_0	[°C]	Average ground surface temperature over a year
V_a	[m ³]	Volume of air in ground
V_s	[m ³]	Volume of solid soil in ground
V_T	[m ³]	Total volume of ground
V_v	[m ³]	Volume of void in ground
V_w	[m ³]	Volume of water in ground

Abbreviations

1-2-3D	One-, two-, three-dimensional.
ABL	Atmospheric Boundary Layer
AE	Accuracy Error
BHE	Borehole Heat Exchanger
CART	Classification And Regression Trees
COP	Coefficient of Performance (also Conference of The Parties)
CV	Cross-Validation
DEM	Digital Elevation Model
DHM	Digital Height Model
DOM	Digital Orthophoto-Map (also called Digital Surface Model - DSM)
DSM	Digital Surface Model
DTM	Digital Terrain Map
FFT	Fast Fourier Transform
GEN	Zone of Switzerland defined by the Geneva canton
GIS	Geographic Information Systems
GSHP	Ground Source Heat Pump
GWC	Gravimetric Water Content (or noted w)
HAWT	Horizontal Axis Wind Turbines
HDF5	Hierarchical Data Format 5
HP	Heat Pump
IEA	International Energy Agency
ISL	Inertial Sublayer
LiDAR	Light Detection And Ranging
ML	Machine Learning
NABODAT	NAtionale BOdenDATenbank (National ground data bank)
NRMSE	Normalized Root Mean Square Error (also noted E_{NR})
OOB	Out-Of-Bag

OOSG	Zone of Switzerland remaining while not considering either SON or GEN zones (Out Of SON and GEN)
PCA	Principal Components Analysis
PDF	Probability Density Function
PE	Prediction Error
PI	Prediction Interval
PR	Performance Ratio
PV	Photovoltaic
QRF	Quantile Random Forests
SFOE	Swiss Federal Office of Energy
SITG	Système d'Information du Territoire Genevois (Information System for the Geneva Territory)
SMAP	Soil Moisture Active Passive (satellite data)
SON	Zone of Switzerland covered by the Sonnendach project data
SVC	Support Vector Classification
SVF	Sky View Factor
SVM	Support Vector Machines
SVR	Support Vector Regression
RBF	Radial Basis Function
RBL	Rural boundary layer
RF	Random Forest
RMSE	Root Mean Square Error (also noted E_R)
RSL	Roughness Sublayer
UBL	Urban boundary layer
UCL	Urban canopy layer
VAWT	Vertical Axis Wind Turbines
VES	Vertical Electrical Soundings
VI	Variable Importance
VSGs	Very Shallow Geothermal systems
vSGP	very Shallow Geothermal Potential
VWC	Volumetric Water Content

1

Introduction

This chapter borrows from the articles:

Assouline, D., Mohajeri, N., and Scartezzini, J-L. (2017). Quantifying rooftop photovoltaic solar energy potential: a machine learning approach, *Solar Energy* 141 278-296.

Assouline, D., Mohajeri, N., and Scartezzini, J-L. (2018). Estimation of Large-Scale Solar Rooftop PV Potential for Smart Grid Integration: A Methodological Review. In *Sustainable Interdependent Networks* (pp. 173-219). Springer, Cham.

In this introducing chapter, Section 1.1 provides the motivation behind the present thesis. Section 1.2 presents the concept of energy potential estimation and the related general state-of-the-art. Section 1.3 expresses the research questions asked in the present work and states the objectives we set up to achieve. Finally, Section 1.4 highlights the structure of the thesis.

1.1 Motivation

On December 12th 2015, the 21st Conference of the Parties (COP21), held in Paris, saw an historic universal agreement between 196 states, in order to shape durable solutions to the critical climate situation we are currently facing. Multiple directions have been discussed, notably allowing countries to adapt themselves to climate change and initiate a transition to a decarbonized economy and society, with the aim to keep the global warming *largely* under 2°C (since pre-industrial era). The COP24, discussing the executing phase of the Paris Agreement, is being held as this thesis is being finalized. This global agreement answers catastrophic environmental issues that require no introduction, as they are fortunately being more and more acknowledged throughout the world.

Switzerland, located in the center of Europe, is embracing an environment-aware path; and has been for a significant period of time. The Swiss Federal Council “Energy Strategy 2050”, initiated in 2011 partly as a consequence of the Fukushima nuclear accident, proposes a phasing-out of nuclear power, currently generating 40% of the national electricity demand, by the year 2035 (<http://www.bfe.admin.ch/>). To compensate for the loss of nuclear energy, the federal Council's Energy Strategy anticipates not only the improvement of energy efficiency, but also the increase in the use

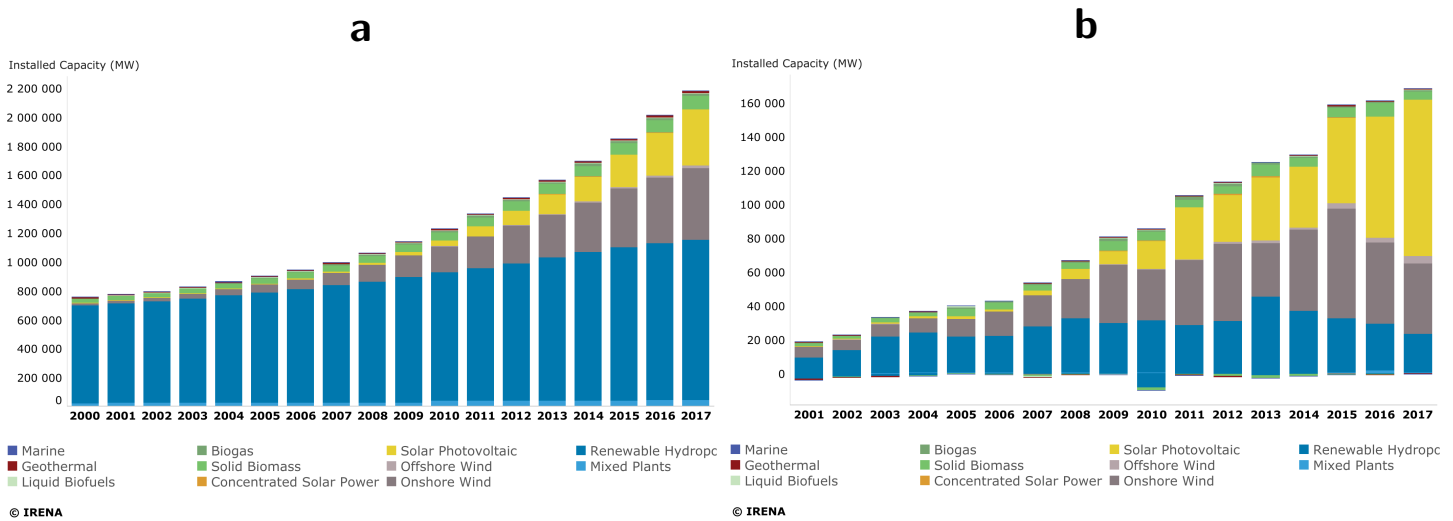


Figure 1.1: World installed capacity (for electricity production) for various renewable sources of energy, from 2000 to 2017. **(a)** Cumulative installed capacity; **(b)** Net addition to installed capacity. Source: IRENA. Data available from <http://resourceirena.irena.org/gateway/dashboard/>.

of renewable energy and associated development of grid and storage capacity. In addition, the Swiss climate policy aims at a drastic reduction of the country's greenhouse gas emissions, including 20-30% reduction of the country's CO₂ emissions from the 1990 level by the year 2020, according to the revised federal CO₂ Act, and a possible 50-80% reduction by 2050. Buildings have the largest share in energy demand in Switzerland: heating, ventilation, and air conditioning account for roughly 40% of the overall energy demand; 32% of the national electricity demand is also caused by building (HVAC, lighting, space heating). Therefore, the goals of the Energy Strategy 2050 and the Swiss climate strategy can only be met when buildings become much more energy efficient compared to today's situation.

To reach these goals, one of the most promising strategies besides the improvement of energy efficiency is to rely on an increased use of renewable energies. Consequently, the installed capacity of renewables is significantly increasing all over the world (Figure 1.1), including Switzerland (Figure 1.2). It can be notably observed from the above figures while hydro still cover a very large fraction of the installed capacity, “decentralizable” energies such as solar (PV) and wind (onshore) are gradually taking more place, and offered the largest additions to the total capacity in 2017 (Figure 1.1b). More specifically, since 2010, solar PV has been the most newly installed type of renewable energy in the world (Figure 1.2b).

The optimized use of decentralized energy systems, however, requires the estimation of both supply and demand values. A supply estimation study, in particular, is critical to make at a large scale of a region or a country, to determine areas of high potential for each type of sustainable energy, and allow for the elaboration of efficient global energy strategies.

1.2 Energy potential estimation

In order to assess the potential for a renewable energy, it is convenient to follow a general approach which provides the structure of the strategy. A *hierarchical approach* to the estimation of renewable

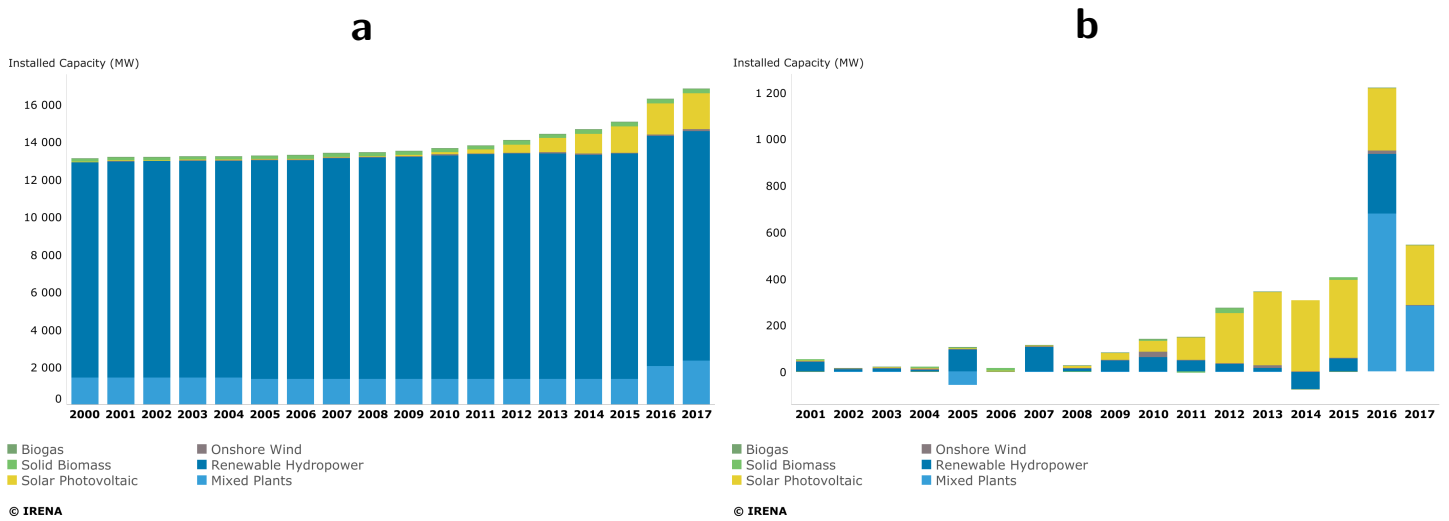


Figure 1.2: Installed capacity (for electricity production) for various renewable sources of energy in Switzerland, from 2000 to 2017. **(a)** Cumulative installed capacity; **(b)** Net addition to installed capacity. Source: IRENA. Data available from <http://resourceirena.irena.org/gateway/dashboard/>.

energy potential has been proposed by van Wijk and Coelingh [1] and became the staple potential estimation strategy. It has been widely used at various scales (region, country, world) [2–4]. This approach consists of a list of consecutive steps, representing different levels of constraints regarding the availability of the energy of interest. The steps are as follows [4]:

- (i) The *theoretical* potential is the theoretical maximum potential provided by the considered energy resource. It is therefore defined by the collection of variables expressing raw energy (e.g. solar radiation or wind speed).
- (ii) The *geographical* potential is the theoretical potential reduced to the amount of energy available from areas which are suitable to the production of the energy.
- (iii) The *technical* potential is the geographical potential reduced by the various losses and practical considerations induced from the conversion of the raw incoming energy into usable energy forms through energy systems (eg. PhotoVoltaic panels or wind turbines).
- (iv) The *economic* potential is the technical potential constrained to cost considerations, making the energy of interest an economically competitive and attractive solution.
- (v) The *market (or implementation)* potential is the final fraction of technical potential which can be implemented in practice. It is notably subject to societal constraints and regulations, including implementation policies, social acceptance, and legal considerations.

An illustration of the full hierarchical approach is shown in case of solar energy harvested with photovoltaic panels in Figure 1.3. Within the framework of this thesis, however, we focus on the technical potential for the renewables of interest. The economic and market potentials require socio-economic analysis, which is beyond the scope of the present study.

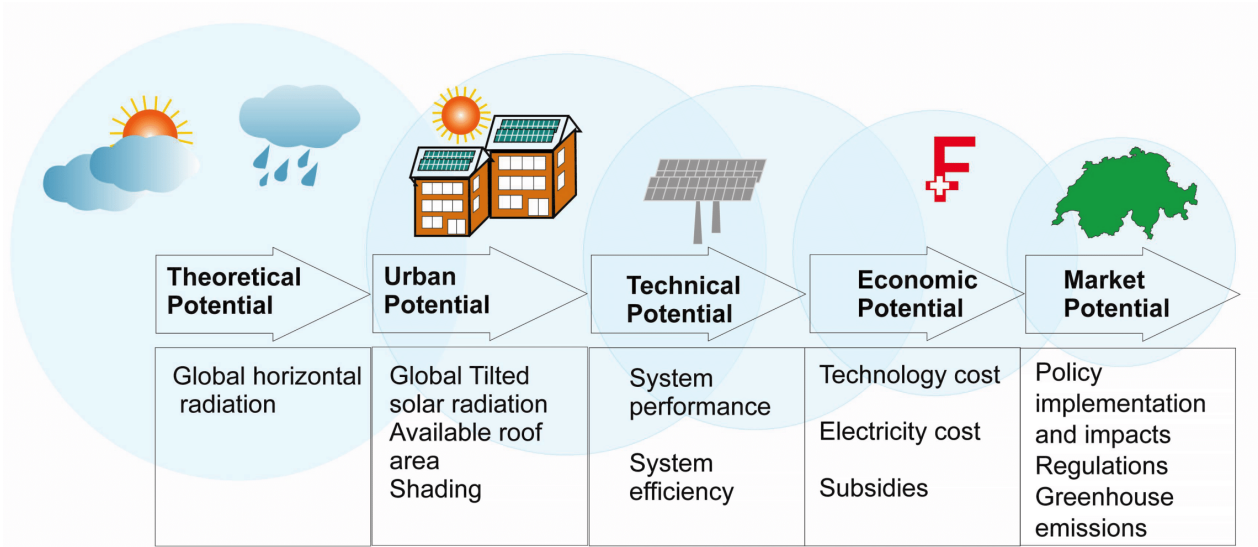


Figure 1.3: Hierarchical potential approach illustration, for solar energy available from photovoltaic panels installed over rooftops. (Note that in this case the geographical potential is called “urban” potential given the installation of PV solar only over building rooftops, and therefore within urban areas.)

1.2.1 Selection of renewable energies

Not all used sustainable energies can be tackled in the present study, for obvious time constraints reasons, and to allow for a thorough adapted strategy for each energy resource. As a result, we focus on renewables which have the most promising for the future: wind energy, geothermal energy and solar energy. The reasons are the following: (i) they are easily deployable at a large scale, including in urban areas, (ii) they have been (and still are) theoretically studied extensively, and therefore benefit from regular technological advances increasing their efficiency, (iii) they have been chosen as some the most popular renewable systems (as seen from Figure 1.1). Note that, from the installed power figures, the geothermal values are not representative of the reality since the figures show only the contribution to electricity production, while geothermal systems are mainly used for space heating and domestic hot water production. It has, however, known a “boom” in Switzerland in the beginning of the 2000s [5], and is therefore a particularly important source of energy within the country; Switzerland is one of the largest users of geothermal energy in the world, with a total installed capacity of 4222 MW for geothermal heat pumps in 2017 [6].

In addition, many different energy conversion systems are available for each form of renewable energy (e.g. solar thermal or solar PV, horizontal or vertical ground source heat pumps, etc.). Since these systems have very different characteristics impacting on the ultimate energy potential, we also have to select systems of preference. We choose some of the most easily deployable systems (which can be grid connected or stand-alone) and currently used, in Switzerland and all over the world: solar photovoltaic panels mounted on building rooftops, horizontal axis wind turbines (both small mounted on buildings and large commercial turbines) and very shallow ground connected heat pumps (installed in the first 1-2m of the ground). The reasons for these specific choices of systems as well as their presentation are discussed in chapter 3, dedicated to energy models and installations. In order to have a view of the current state of sustainable energy in Switzerland, and as a point of comparison for the potential

Table 1.1: Annual Swiss production for different types of renewable energies, in 2017. (HAWT: Horizontal Axis Wind Turbines, HP: Heat Pumps, PV: PhotoVoltaic. The figures are taken from SFOE [6])

<i>System / Energy type</i>	Wind	Geothermal	Solar	
	HAWT / Electricity	HP / Heat	PV / Electricity	Thermal / Heat
Production [GWh]	132.6	6665.0	1683.0	697.0
Installed Power [MW]	75.2	4222.0	1905.8	1177.0
Number of Installations	37	290'000	70'070	1650

studies to be performed in this thesis, Table 1.1 provides Swiss renewable energy production figures, for the systems selected the thesis (along with solar thermal, which has also an important role), in 2017.

Considering the latter energies and conversion systems discussed, let us define the specific levels of potential. Note that the different variables constituting these levels are also presented and discussed in chapter 3, and we therefore only outline them here. The hierarchical levels for each of the tackled energies are as follows:

- Concerning wind turbines: the *theoretical* potential is given mainly by the estimation of wind speed at heights that are suitable for turbine installation; the *geographical* potential is the portion of the theoretical potential available from locations where wind turbines can be installed (both in rural and urban areas, with a reasonable design of arrays in case of multiple installations); the *technical* potential is the portion of the geographical potential obtained after the conversion of wind speed to electricity through the considered turbines (with a specific coefficient of performance).
- Concerning very shallow ground connected heat pumps: the *theoretical* potential is given by the three main variables defining the ground thermal behavior, namely the ground temperature gradient (for the ground depth considered), thermal conductivity and volumetric heat capacity (or the thermal diffusivity, being the ratio of the two latter variables); the *geographical* potential is the fraction of the theoretical potential available from areas suitable for the installation of very shallow ground connected heat pumps, considering the space needed for the latter, the length of the pipes, etc.; the *technical* potential is the fraction of the geographical potential extractable from considered ground source heat pumps and corresponding practical characteristics, including their coefficient of performance.
- Concerning solar rooftop PV: the *theoretical* potential is given by the estimation of solar horizontal radiations (global, direct and diffuse) over rooftops; the *geographical* potential is the fraction of the theoretical potential available from suitable areas for PV systems over rooftops, considering the rooftops'slope and direction, along with shading factors; the *technical* potential is the fraction of the geographical potential obtained when losses induced from the use of PV panels are taken into account, including mainly the efficiency and performance ratio of considered panels.

The assessment of the theoretical, geographical and technical potential for the three latter energies can be performed using various adapted strategies. Accounting for the scale of the study, the considered conversion systems and the data available, many methodologies have been suggested in the literature. A specific literature review for each of the three discussed renewable energies is provided in the corresponding chapters (chapter 4 for wind energy, chapter 5 for geothermal energy, and chapters 6 and 7 for PV solar energy). The main gaps left by the literature regarding the three energies, however, can be summarized as follows: (i) there is a lack of large-scale potential methodologies covering a whole region or country (most suggested studies are for city or small regional scale), (ii) the spatial extrapolation of required variables (e.g. available area for PV) is often performed using qualitative methods such as averaged coefficients rather than quantitative ones, and (iii) the data available is often used for sampling strategies or parameter tuning rather than advanced statistical methods which often take better advantage of the information contained within the data. Regarding the potential estimation in Switzerland in particular, while there exists studies (as detailed in the literature reviews of each chapter), there are often important missing aspects in the assessment (e.g. the consideration of urban areas for wind energy, very shallow depth for geothermal energy, superstructures over the rooftops in case of solar PV energy). Therefore, there is a need for an updated assessment, particularly concerning the most popular sources of sustainable energy.

1.2.2 Machine Learning for potential estimation

In parallel with the need for the increase of sustainable energy use, there has been a tremendous growth in the availability of data, in every domain. So much data, in fact, that it became challenging to handle it adequately. Yet, it is crucial to be able to process, analyze and learn from this data in the most efficient way possible, to extract solutions to incoming problems. Machine Learning (ML), a domain which started its development in the middle of the 20th century, is precisely designed for that aim, and has thus recently known a revival over the last few years. In particular, ML can be used to learn patterns from data and perform prediction/estimation for a given phenomenon described by the data, based on inputs of interest. It is therefore applicable to a very large variety of domains.

ML algorithms have allowed to reach breakthrough results for a number of useful tasks (time series forecasting, speech recognition, image recognition, language processing, video tracking, etc.) applied to many different domains, including biology, energy, medicine, security or social networks. Regarding energy applications and environmental sciences in general, it has been used for a number of tasks, including weather and energy supply forecasting [7–12], spatial interpolation (mapping of environmental variables) [13–16], or natural hazard assessment [17–20]. Besides for the spatial estimation of environmental variables (such as solar radiation), however, ML has been very rarely used in the framework of a renewable energy potential estimation. The latter, often requiring the assessment of multiple variables of various types (meteorological, geological, building-related, etc.), could nonetheless greatly benefit from it, in particular for an estimation over large-scale regions, which often have missing data over a large portion of the territory.

Note that a chapter is dedicated to Machine Learning methods (chapter 2), and that the state of the art concerning ML use for each energy is provided in each corresponding chapter.

1.3 Research goals

With the earlier presented concepts and literature gaps in mind, the following research questions are asked:

- *Is Machine Learning adequate for large scale mapping of renewable energy potential values?*
- *Can Machine Learning be realistically combined with traditional models to predict potential values? How?*
- *What is the spatio-temporal potential, in Switzerland, for the three following promising forms of energy:*
 - *wind energy, using both small (building mounted) and large wind turbines?*
 - *geothermal energy, using very shallow ground source heat pumps?*
 - *solar energy, using PhotoVoltaic panels mounted over building rooftops?*

The ultimate goal of this thesis is twofold: (i) provide data-driven methodologies combining Machine Learning, Geographic Information Systems and traditional modeling to estimate the renewable energy potential of the mentioned energies at the large-scale of a region or a country and (ii) apply the latter methodologies using available data to estimate these renewable energy potentials in Switzerland.

Note that, because of constraints imposed by the framework of a thesis, the full technical potential cannot be considered for all three energies. We therefore focus on one source of energy by extracting its full technical potential, and assess the theoretical potential for the two others. It is rather clear that, even though geothermal heat pumps are currently dominating in Switzerland, solar photovoltaics show a remarkable growth and are overall the most popular of the sustainable energy forms, and possibly the most used in the world in the near future (Figures 1.1 and 1.2). As a result, it was decided to focus on this technology, and develop an estimation strategy to assess its technical potential. Wind and very shallow geothermal energies are therefore studied at the theoretical potential level.

1.4 Structure of the thesis

The thesis will attempt to achieve the discussed objectives, using the following structure:

Chapter 2 provides a gentle introduction to Machine Learning (ML), in the context of the present thesis. In particular, the chapter presents theoretical background and practical considerations related to the principle of learning models from data in general, along with more detailed sections on the two extensively used ML algorithms in this thesis, namely Support Vector Machines (SVM) and Random Forests (RF).

Chapter 3 presents theoretical concepts, modelling strategies and systems characteristics for the three tackled renewable energies in this thesis: solar energy, geothermal energy and wind energy. It focuses on concepts used in the thesis and presents energy conversion systems of interest for each type of energy, highlighting the ones considered in the potential studies suggested.

Chapter 4 provides a methodology to estimate the large-scale theoretical potential for wind energy, with an application to Switzerland, using a pixel grid resolution. It constitutes a first potential study, which focuses on the estimation of wind speed in both rural and urban areas.

Chapter 5 presents a strategy to estimate the large-scale theoretical potential for very shallow geothermal energy, with an application to Switzerland, using a pixel grid resolution. This second study provides steps to estimate the multiple ground variables of interest impacting the latter geothermal potential, including the thermal conductivity, heat capacity and ground temperature gradient.

Chapter 6 suggests a methodology to estimate the large-scale technical potential for solar PhotoVoltaic generated electricity over rooftops, with an application to Switzerland, at the scale of the Swiss communes. The consideration of the geographical and technical aspects brings significant complexity to the estimation, notably regarding the assessment of the suitable area for PV panels over rooftops and building-related variables over the whole territory.

Chapter 7 revisits the methodology provided in chapter 6 in order to improve it and perform the estimation at a finer grid resolution. Several improvements are suggested regarding the methodology steps (available area for PV panels, geometrical characteristics of rooftops etc.) and the accuracy of the estimation of the multiple variables of interest.

2

Machine Learning

This chapter borrows from the book chapter [21]:

Assouline, D., Mohajeri, N., and Scartezzini, J-L. (2018). Estimation of Large-Scale Solar Rooftop PV Potential for Smart Grid Integration: A Methodological Review. In Sustainable Interdependent Networks (pp. 173-219). Springer, Cham.

and the following articles [22, 23]:

Assouline, D., Mohajeri, N., and Scartezzini, J-L. (2017). Quantifying rooftop photovoltaic solar energy potential: a machine learning approach, Solar Energy 141 278-296.

Assouline, D., Mohajeri, N., and Scartezzini, J-L. (2018). Large-scale rooftop solar photovoltaic technical potential estimation using Random Forests, Applied Energy 217 189-211.

This chapter aims at introducing Machine Learning (ML) and some of the popular ML algorithms applied in this thesis. It is meant for the reader to grasp the main theoretical notions applied in this thesis and the processing needed to use them in practice, and particularly in the scope of the present work. Note that the already repetitive use of the expression “applied in this thesis” was chosen on purpose: this chapter is neither an exhaustive review of Machine Learning concepts and algorithms, as that naturally falls outside the scope of the thesis, nor a very thorough and mathematically formal presentation of the concepts and algorithms used in this context, as there are already many textbooks and articles which achieve that goal remarkably well. The chapter is rather a presentation of the concepts and theory behind the algorithms mentioned in the next chapters to prevent the reader from seeing them as “black boxes”, which is unfortunately frequently the case.

The structure of the chapter is as follows. Section 2.1 presents generalities about ML in theory and practice. Section 2.2 presents the concept of kernel methods and in particular Support Vector Machines for classification and regression tasks. Section 2.3 presents the the family of Ensemble Learning methods and in particular Random Forests for classification and regression tasks.

2.1 Learning from data

Machine Learning (ML) methods are based on algorithms that improve their performance with increasing experience. Experience, in this case, is primarily the information provided by examples. Examples in the data expose patterns and dependencies within this data, from which an ML algorithm learns and shapes a model. This model is said to be *trained*. Once trained, the model can be used for prediction tasks. ML methods can be seen as a new, direct way to use data for modelling. Classically, statistics used parametric methods to model a phenomenon: observe data, go through a list of known distributions (models), pick the one that seems to fit best, tune the parameters of the distribution to fit the data in the best way possible. Contrary to parametric methods, ML methods learn a model purely based on data, with no prior distribution knowledge required.

The lack of underlying distribution in the ML paradigm causes some potential *generalization* issues, which is their main intrinsic disadvantage compared to traditional parametric methods: (i) the chosen data needs to be large enough and sampled wisely in order to be representative of the phenomenon, in most of its patterns, (ii) special care is required during the training phase: if the training data is fit too closely, the model will not be general enough to perform a prediction over unseen data (*overfitting*); if the training data is fit too loosely, the model is too simple and has not learned useful patterns (*underfitting*). In practice, these two issues are solved by the use of large training datasets (as much as possible) and particular techniques (called *regularization* techniques) designed to avoid overfitting while learning from the training data. Some regularization techniques will notably be tackled during the presentation of ML algorithms further in the chapter.

Yet, the lack of prior distribution also allows for more flexibility in terms of modelling, notably for the choice of the input variables in the model, which is often motivated by expert knowledge in the domain of interest or pure intuition. More importantly, the current availability of extraordinarily large datasets further motivates to explore data-driven methods, which can possibly extract models that sole theoretical knowledge would not be able to produce. ML algorithms have therefore been explored for a wide variety of tasks and successfully applied in many different domains, as shortly discussed in section 1.2.2 (chapter 1). A good review on the foundations of ML and the different families of algorithm can be found in [24]. A well designed, popular (and heavily used in this thesis) ML library to apply the algorithms in practice is the Scikit-Learn [25] library, implemented in the Python language.

There exists three different paradigms for data-based learning methods: *supervised*, *unsupervised* and *semi-supervised* learning. In supervised learning, models are built based on (inputs, outputs) pairs in order to extract the link between input and output. On the contrary, unsupervised learning aims at building models by learning solely from input data and its inner structure, without the output information. Finally, semi-supervised learning methods are approaches which attempt to extract patterns between inputs and outputs, yet with a very limited number of output values available, and often a large number of input values with no corresponding output. In the framework of the present thesis, solely supervised learning approaches are used.

Let us end this introduction with a semantic note: several other terms, closely related to Machine Learning, are often used, including *pattern recognition*, *artificial intelligence*, *statistical learning* or *data mining*. Even though these terms may often seem to be used in a random fashion, they are domains or subdomains related to the machine learning paradigm, focusing on certain aspects of learning information from data. In the framework of this thesis, we will always refer to ML algorithms,

without making a difference between various practices. Note finally that *Big Data* is another rather trending topic. It is not a subgenre of ML techniques, but rather a fact, a paradigm reflecting the current availability of data in extremely large quantities: if used properly, the almost infinitely large source of data can greatly help ML algorithms; yet, the algorithms need to be (re-)designed intelligently in order to be able to handle this gigantic amount of data.

2.1.1 Supervised learning

Let us define the general problem more formally. We first note X_1, X_2, \dots, X_d the input variables of interest for a studied phenomenon. These variables can be for example environmental variables such as temperature, precipitation, etc. Each data point is one realization of these variables, forming a d -dimensional vector of values taken by the input variables $(x_1, x_2, \dots, x_d) = \mathbf{x}$. As a result, an input dataset of l points can be seen as an $l \times d$ matrix formed by l rows $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l)^T$. Similarly, let us define an output variable Y (a variable to be predicted, for example the solar radiation) that takes a value y for each data point. The collection of output values forms the output data (y_1, y_2, \dots, y_l) . Note that the input variables are sometimes called *features*, *predictors* or *attributes*, data points are *samples* or *instances*, the output variable is the *target*, and the target values are *labels*. In a supervised learning framework, we know both feature and target values for a certain amount of data points. This gathering of observed points is known as the *labeled set*, or sometimes simply the *learning set*, and consists of couples $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_l)$.

In a general sense, a supervised learning task aims at the following: learning a function $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$, where \mathcal{X} and \mathcal{Y} are respectively the input and output spaces (where the input and output variables live), based on the labeled set, so that the predictions $\varphi(\mathbf{x})$, for any \mathbf{x} , are as close as to its actual corresponding target y as possible.

The obtained function φ can then be applied to new input points to perform a prediction of their output labels. The type of the output variable (the space it lives in) defines the type of task. If y is a real number ($\mathcal{Y} = \mathbb{R}$), the task at hand is a *regression task* and φ is a *regressor*; if y is a relative integer (\mathcal{Y} is a finite set of classes), the task at hand is a *classification task* and φ is a *classifier*. Note that for classification tasks the actual values defining the classes do not matter and simply signify the affiliation to a class. Therefore, multiple conventions may be taken. Binary classification problems (with only two considered classes) often use $y \in (-1, 1)$ or $y \in (0, 1)$. Multi-class problems traditionally use $y \in (0, 1, 2, 3, \dots, c)$ where c is the number of considered classes. Many classification and regression algorithms or families of algorithms have been developed and perfected over the years, including kernel methods (notably Support Vector Machines presented in section 2.2), ensemble learning methods (notably Random Forests), neural networks (nowadays referred to as *Deep Learning* methods, based on the newly acquired computational power allowing the networks to be very large), etc. In fact, most of these methods often have a classification and a regression version, as the regression case can be seen as a generalization of the classification case, with an infinite number of classes.

2.1.2 Model selection and model assessment

One of the main concerns in ML is to maximize the performance of a learned model φ . The performance of the model is given through a *loss function* \mathcal{L} that measures the discrepancy between predicted output values and known target values. Many different loss functions are used, based on the model and task of interest. A very commonly used example of loss function (for regression tasks) is the *squared-error* loss: $\mathcal{L}(Y, \varphi(X)) = (Y - \varphi(X))^2$. In a statistical sense, the quantity of interest to minimize is the expected prediction error, over all possible values of $\mathcal{X} \times \mathcal{Y}$. The fact of using all possible values in the product space $\mathcal{X} \times \mathcal{Y}$ is crucial: the finality is not to learn the model perfectly in the labeled set (known set), but to assure that this model *generalizes* well outside of the labeled set, to allow for the prediction of labels for unobserved points. Formally, the previous statement means that one desires to minimize the expected prediction error $\mathbb{E}_{\text{EPE}}(\varphi)$ between prediction performed by φ and the actual output for any possible input point, measured by the loss function \mathcal{L} :

$$\text{Goal: Choose } \varphi \text{ to minimize } \mathbb{E}_{\text{EPE}}(\varphi) = \mathbb{E}_{\mathcal{X}, \mathcal{Y}} [\mathcal{L}(Y, \varphi(X))] \quad (2.1)$$

However, since we generally do not know the distribution of the input and output variables, we usually approximate the expected prediction error with the *test sample estimate* of the error. The principle of the test sample estimate is simple: (i) separate the labeled set into a *training set* and a *test set*, (ii) train a model only using the training set, (iii) predict the output values for points in the test set and compare with the actual labels to compute the test error. The test set serves as a virtual unobserved set, as a part of the data that was not seen by the model during the training process. A typical used proportion is 75% of the labeled set for the training set, and 25% for the test set.

Once the type of algorithm has been chosen for a task (together with the loss function embedded in the algorithm), a very important step is performed in order to maximize the performance of the model and avoid generalization issues: the *model selection* step. Before describing what the model selection steps is practically, note that, as already discussed several times in this chapter, the generalization capabilities of a model is a central issue which gave birth to a theoretical framework known as *statistical learning theory*, or *VC theory*. This theory is named after its developers, Vapnik and Chervonenkis [26–28], in a desire to provide formal mathematical guarantees for algorithms that, more often than not, seemed to “work magically”. VC theory “characterizes properties of learning machines which enable them to generalize well to unseen data” [29]. It is therefore a very important piece in the history of Machine Learning and largely contributed to the development of ML and its regained popularity at the beginning of the 2000's. For the sake of brevity, however, and since this thesis does not have the ambition to provide theoretical developments regarding ML algorithms, we will not present this theory in details. Note, however, that model selection is formalized within the framework of this theory; readers are invited to refer to the suggested literature for more details.

In practice, the model selection concept is rather simple. Most models have global tuning parameters, often called *hyperparameters* to differentiate them with the parameters extracted by the core algorithm within an optimization task, that the user chooses when training the model. Selecting the best model consists in extracting the optimal set of hyperparameters so as to maximize the performance of the model. A portion of the training set can be used as a *validation set* in order to compare models trained with different parameters. Model selection is however often achieved at the same time as the

training itself, through a procedure called *K-fold Cross-Validation (CV)* [30], depicted in Figure 2.1. The procedure goes as follows: The training set is first separated in K equal parts. It will then be used K times (corresponding to K different folds of the data), in each of which one of the K parts is used to test the model trained with the $K-1$ remaining parts. The error between observed and model predicted output (usually in the form of RMSE, presented later in the section) is stored for each fold, and the mean error is computed. This whole process is done for multiple sets of parameters. The “best” set of parameters is then the one offering the lowest mean error.

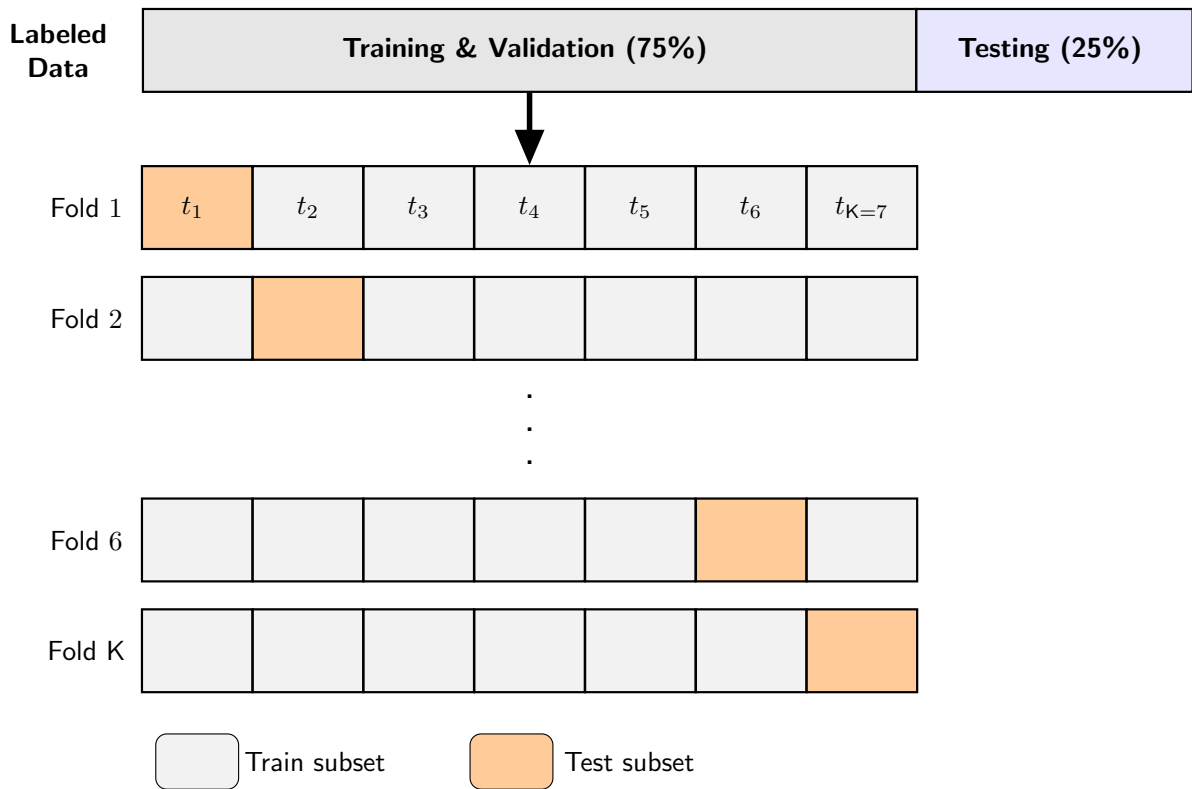


Figure 2.1: K-fold Cross-Validation scheme, here for a number $K = 7$ of folds.

The model selection step is followed by the *model assessment* step, which consists in evaluating the performance of the model by extracting the error resulting from the model prediction in the test set, and therefore estimating the *test error*, also called the *generalization error*. It is in particular differentiated from the *training error*, obtained when using the model in the training set, and the *CV error*, obtained while selecting the best model with cross-validation (it can be seen as the mean K-fold CV error through the K folds). There exist multiple functions to compare known labels and predicted values for the test sample estimate of the error [31]. Although some of these functions are attached to certain problems as the “standard” measures of error, multiple errors can be used to capture the different discrepancies of the model. Some of the most used errors are presented here. In all the following definitions, y^{obs} is the observed (known) output, y^{pred} is the predicted output, and N_{test} is the size of the testing set, meaning the number of data points for testing, used to compute the generalization error.

The *Root Mean Square Error* (RMSE) is the standard error of estimation for regression. It is the square root of the mean square error:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N_{\text{test}}} (y_i^{\text{pred}} - y_i^{\text{obs}})^2}{N_{\text{test}}}} \quad (2.2)$$

The RMSE expresses the average error between observed (i.e. known, measured or validated in some way) and predicted output, in the order of magnitude of the quantities of interest. As a result, it only has sense if compared to typical values of the data.

The *Normalized Root Mean Square Error* (NRMSE) is used to neutralize the relativity of the RMSE, and will be the preferred normalized error within the thesis. It is simply normalizing the RMSE by the average observed value, so that the obtained error is a relative fraction, often given in percentage. In the percentage form, it is :

$$\text{NRMSE} = 100 \times \frac{\text{RMSE}}{\overline{y^{\text{obs}}}} \quad (2.3)$$

where $\overline{y^{\text{obs}}} = (\sum_{i=1}^{N_{\text{test}}} y_i^{\text{obs}}) / N_{\text{test}}$. As a rule of thumb, the “goodness” of the NRMSE can be unformally classified as: [0% - 10%]: “excellent”; [10% - 20%]: “very good”; [20% - 30%]: “good”; [30% - 40%]: “acceptable”. Higher NRMSE values denote a poor accuracy.

The *Mean Absolute Error* (MAE) measures the average discrepancy between observed and predicted values, as RMSE does, but using the absolute value, instead of the root mean square:

$$\text{MAE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |y_i^{\text{obs}} - y_i^{\text{pred}}| \quad (2.4)$$

The *Mean Absolute Percentage Error* (MAPE) is a normalized error, giving a general idea of the model in terms of percentage, and no knowledge of typical data values is needed. An issue of MAPE is that it can be computed only for non-zero positive values. It is given by the following formula:

$$\text{MAPE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \left| \frac{y_i^{\text{obs}} - y_i^{\text{pred}}}{y_i^{\text{obs}}} \right| \quad (2.5)$$

The *Mean Biased Error* (MBE) is a simple mean error between observed and predicted output. The main difference is that the sign of the difference matters, which can be useful in some specific studies to capture how the two quantities compare. It is given by the following formula:

$$\text{MBE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} (y_i^{\text{pred}} - y_i^{\text{obs}}) \quad (2.6)$$

The *Accuracy Error* (AE) is used in a classification task, where the values are discrete, and there is no need for a measure between continuous values. The goal is to compute how many times the predicted label (class number) is the same as the actual label, as a relative fraction or percentage. As a result, the following indicator function is used:

$$\mathbb{1}_{[x=y]} = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases} \quad (2.7)$$

The Accuracy Error is then given by:

$$AE = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} 1_{[y_i^{\text{pred}} \neq y_i^{\text{obs}}]} \quad (2.8)$$

where y_i^{pred} and y_i^{obs} are in this case discrete classes.

2.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) is one of the most popular and efficient machine learning algorithms for classification tasks and the flagship of the so-called *kernel methods* [32]. The latter term comes from a particular “trick” that this family of methods has in common, embedded within the algorithm (the *kernel trick*), which allows to tackle non linear problems with high performance - and will be explained in detail during the presentation of the algorithm. SVM was developed by Cortes and Vapnik in 1995 [33] and improved at various levels through the years. Given the period of the development of the method, it is deeply rooted in the framework of VC theory, and was introduced, in its original form, with strong mathematical guarantees of its performance. It was also successfully extended to a regression setting, under the name of Support Vector Regression (SVR) [34]. In this section, we will first present the algorithm for classification, Support Vector Classification (SVC), since it was the original idea of the algorithm; we will then present its extension, SVR. Finally, we will highlight important aspects to take into account when training an SVM model in practice and more specifically detail how it is used in this thesis.

2.2.1 The large margin classifier

For simplicity, let us first define a binary classification problem - the principle of the algorithm can be easily extended to a multi-class problem. We consider a set of training data with two classes of points C_1 and C_2 . The goal of a binary classification task is to design a function φ that can assign any new point \mathbf{x} to either C_1 or C_2 , with the help of the training data providing us with examples. Instead of extracting a function that analyzes each point one by one to estimate its class, the basic idea of SVM is to find the “best boundary” separating the two classes. Then, a point located on one side of the boundary will be labeled with one class, a point on the other side of the boundary will be labeled with the other class. The best boundary is found by maximizing the distance between the two classes, called the margin. This idea gives the classifier its name: the *large margin classifier*. If the two groups of points are separable by a linear space, we say that they are *linearly separable*. This linear space in 2D is a straight line, a plane in 3D and a so-called *hyperplane* (space of dimension $d - 1$) in an arbitrary dimension d .

The linearly separable case, illustrated in Figure 2.2, is fundamental since it develops the basis of the algorithm. Let us write it more formally. We consider the set of N training data points $(\mathbf{x}_n, y_n)_{n=1, \dots, N}$ where each \mathbf{x}_n is a d -dimension point input vector containing values for the d features of interest and y_n is the corresponding class label, for example -1 for class C_1 and $+1$ for class C_2 . We consider the data to be linearly separable, so we know the form of the solution: it is a hyperplane separating the positive from the negative examples. Since it is a hyperplane, it can be parametrized by a vector normal (perpendicular) to the hyperplane and we call this vector \mathbf{w} . We also consider the offset b , defining the shift between the hyperplane and the origin of the space where the data points live (the dimension

of that space is equal to the number of features). The points that lie on the hyperplane satisfy the equation $\mathbf{w}^\top \mathbf{x} + b = 0$, where $\mathbf{w}^\top \mathbf{x}_n$ is the scalar product between \mathbf{w} and \mathbf{x} , in a matrix formulation. The quantity $\frac{|b|}{\|\mathbf{w}\|}$ is the distance from the hyperplane to the origin, where $\|\mathbf{w}\|$ is the Euclidean norm of \mathbf{w} , a measure of its length. For the linearly separable case, our φ function is then [35]:

$$\varphi(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b) = \begin{cases} +1 & \text{if } \mathbf{x} \in C_1 \\ -1 & \text{if } \mathbf{x} \in C_2 \end{cases} \quad (2.9)$$

where C_1 and C_2 are two considered classes. The solution is the hyperplane with the highest margin and the training points satisfy the following constraints:

$$\mathbf{w}^\top \mathbf{x}_n + b \geq +1, \text{ for } y_n = +1 \text{ with } n = 1, 2, \dots, N. \quad (2.10)$$

$$\mathbf{w}^\top \mathbf{x}_n + b \leq -1, \text{ for } y_n = -1 \text{ with } n = 1, 2, \dots, N. \quad (2.11)$$

These two set of constraints can be combined in one more convenient one:

$$y_n (\mathbf{w}^\top \mathbf{x}_n + b) \geq 1, n = 1, 2, \dots, N. \quad (2.12)$$

Eq. 2.12 will form the constraints of our optimization problem. Now it needs a clear objective function to be optimized. As it was explained earlier, the objective is to maximize the margin, the distance between the two classes. A small derivation leads to a value of $\frac{2}{\|\mathbf{w}\|}$ for this margin, as shown on the right of Figure 2.2, which illustrates a typical linearly-separable classification setting. Since (i) maximizing a quantity is equivalent to minimizing its inverse, and (ii) manipulating the square of a norm is equivalent and usually simpler than manipulating the norm itself, the optimization problem is finally the following constrained problem [35]:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2} \|\mathbf{w}\|^2 \\ &\text{subject to} \quad y_n (\mathbf{w}^\top \mathbf{x}_n + b) \geq 1, n = 1, 2, \dots, N. \end{aligned} \quad (2.13)$$

One can use the Lagrangian formulation to solve this optimization problem, which gives expressions for \mathbf{w} and therefore, by plugging \mathbf{w} in Eq. 2.9, the form of the final SVM decision function φ :

$$\mathbf{w} = \sum_{n=1}^N y_n \alpha_n \mathbf{x}_n \quad (2.14)$$

and

$$\varphi(\mathbf{x}) = \text{sign} \left(\sum_{n=1}^N y_n \alpha_n \mathbf{x}_n^\top \mathbf{x} + b \right) \quad (2.15)$$

Replacing the expression of \mathbf{w} in the original formulation 2.13 yields the so-called *dual problem*:

$$\begin{aligned} &\text{Minimize}_{\alpha} \quad D(\alpha) = \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N y_n y_m \alpha_n \alpha_m \mathbf{x}_n^\top \mathbf{x}_m - \sum_{n=1}^N \alpha_n \\ &\text{subject to} \quad \alpha_n \geq 0, n = 1, 2, \dots, N, \\ &\quad \sum_{n=1}^N \alpha_n y_n = 0. \end{aligned} \quad (2.16)$$

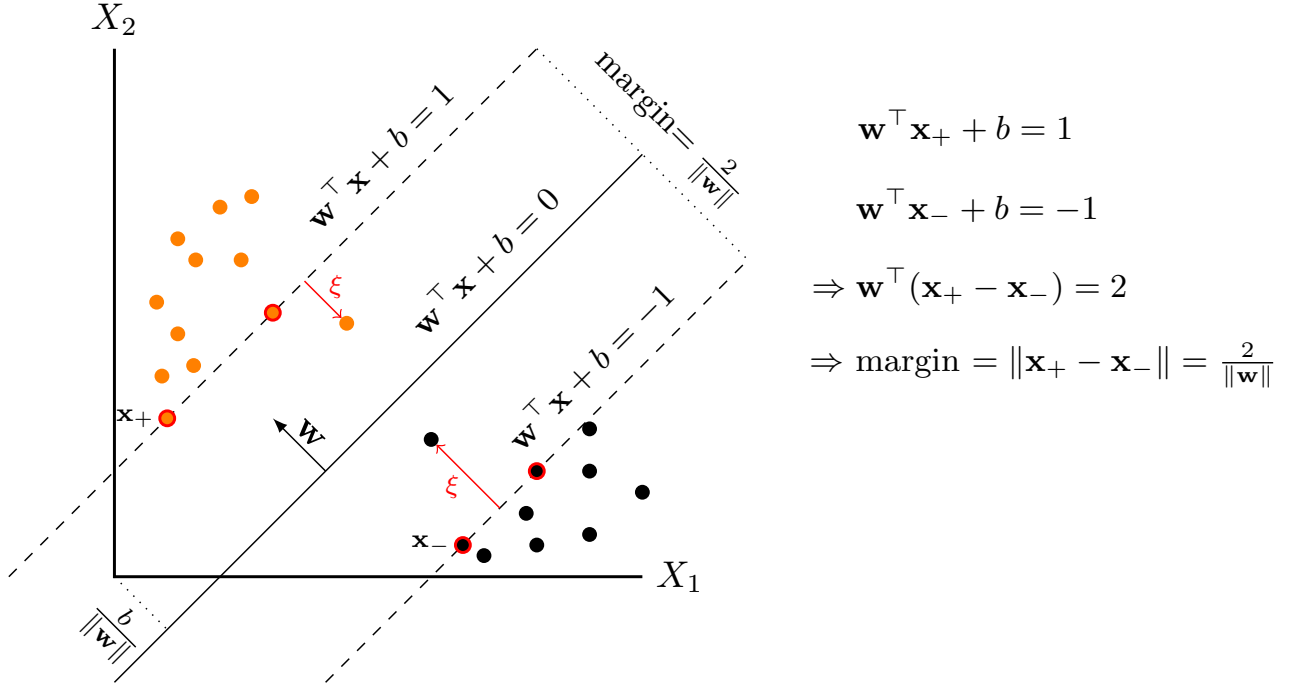


Figure 2.2: Illustration of SVM for the linearly separable case, in case of a 2D input space. The support vectors are highlighted in red. The slack variable ξ , allowing to tolerate outliers in the soft-margin formulation (as further explained in 2.2.2) is also illustrated for some examples.

where α_n are the Lagrange multipliers for the constraints of the problem.

The dual formulation is a quadratic problem that can thus be solved with Quadratic Programming (QP). Solvers for QP extract the α_n , which allows to obtain a unique solution for the final solution φ , by plugging the obtained α_n in Eq. 2.15). Note that there is one α_n for each training point. Only a small portion of the training points are given a non-zero α_n and only these points actually impact the decision function. They lie on one of the two hyperplanes defining the boundaries of the classes and are therefore called the *support vectors*.

Eventually, for a binary classification task, the predicted class label is given by φ (Eq. 2.15): -1 if φ is negative, $+1$ if φ is positive. For a multi-class task, involving c classes with $c > 2$, the strategy consists in combining multiple binary SVM classifiers, using either a *one-vs-one* approach or a *one-vs-all* approach. In the first approach, $\frac{c(c-1)}{2}$ binary classifiers are trained with 2 classes at a time, and the overall predicted class is the one that got a majority of vote, meaning the most frequently predicted class. In the latter approach, c binary classifiers are trained to separate one class from all the others, and the overall predicted class is given using a slightly modified solution for each binary classifier: for each sample, the actual value of $\varphi_c(x) = \left(\sum_{n=1}^N y_n \alpha_n x_n^\top x + b \right)$ is stored instead of solely its sign and the class which has the largest $\varphi_c(x)$ is assigned to x .

2.2.2 The soft-margin classifier

There is one crucial problem brought by the formulation of large-margin classification: if a point is within the margin and therefore in neither class 1 nor class 2, it is completely discarded from the classification. Another formulation has therefore been developed in order for the margin to be

more flexible, by relaxing the constraints. This formulation, called the *soft margin* SVM, allows for points slightly outside of a class (inside the margin) to still be considered in this class, with an added penalty. This penalty is expressed by slack variables (one for each training point) ξ_n that measures how far from the boundaries the points are [33]. In order to control the impact of the created penalty on the objective function, a parameter C is added as a multiplier in front of the sum of the slack variables. C determines the trade-off between the generalization of the classifier and the amount of outliers tolerated (if C is too big the data will be overfitted and the capacity to adapt to a new data will be low, if C is too small it might adapt well but will not have enough memory of the training data to classify well). As a hyperparameter, C will further have to be tuned during model selection. This soft classification is now formulated as follows:

$$\begin{aligned} \underset{\mathbf{w}}{\text{Minimize}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi_n \\ \text{subject to} \quad & y_n (\mathbf{w}^\top \mathbf{x}_n + b) \geq 1 - \xi_n, \quad n = 1, 2, \dots, N. \end{aligned} \quad (2.17)$$

The Lagrangian formulation is similar to the previous one, with the extra slack term added. The obtained dual problem is almost the exact same as in Eq. 2.16 with the only difference that the α_n are bounded by C in the first set of constraints: $0 \leq \alpha_n \leq C$, $n = 1, 2, \dots, N$.

2.2.3 The kernel trick

The main contribution of SVM, which created a whole family of methods (the kernel methods), is that it can also yield very good results in the case of non-linearly separable points. In that case, the data is mapped to a higher dimension to cleverly get back to a linearly-separable situation with the so-called *kernel trick* [36]. This trick consists of replacing scalar products with a *kernel* function $K(.,.)$, applying it to the same input vector points. Due to the fundamental properties of the kernel function, this simple substitution allows to implicitly apply a non-linear mapping ψ to a higher dimensional space (called the reproducing kernel Hilbert space). Once the input points are mapped in this space, the previously presented linear classifier can be applied, as illustrated in Figure 2.3 Note that the non linear mapping ψ is never explicitly chosen. The scalar products are simply replaced by the kernel $K(.,.)$ [33]:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \psi(\mathbf{x}_i)^\top \psi(\mathbf{x}_j) = \mathbf{z}^\top \mathbf{z}' \quad (2.18)$$

where \mathbf{z} and \mathbf{z}' are living in a higher dimensional space called the *feature space*, while the original space is called the *input space*.

A crucial point of the kernel trick is that all the computations are performed in the original input space through the kernel and do not need to be performed in the feature space, where the high dimension would make the calculation potentially intractable. Note that the kernel trick and therefore the generalization of SVM to non-linear settings is allowed by the presence of scalar products in the dual formulation. When compared to the original linear problem, the mathematical formulation and the solution remains the same, with the only difference that the scalar products are substituted with the kernel function (the program is still solvable by QP). Notably, the final decision function extracted by kernel SVM is:

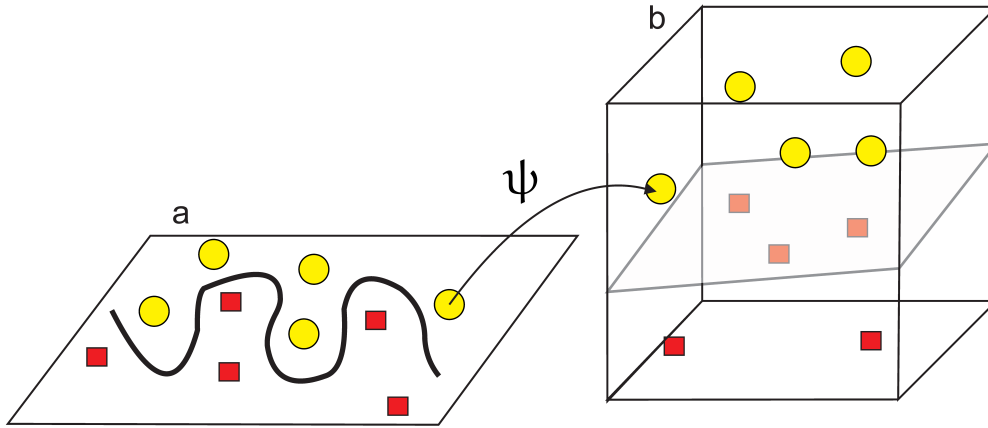


Figure 2.3: Non linear (Kernel) SVM principle. Once mapped (by the implicit mapping ψ induced by the kernel) in a 3D space (the high-dimensional feature space mentioned for Equation 2.18, in (b)), the points are naturally separable by a linear space, here a 2D plane. It translates into a non linear line once mapped back into the original 2D input space, in (a).

$$\varphi(\mathbf{x}) = \text{sign} \left(\sum_{n=1}^N y_n \alpha_n K(\mathbf{x}_n, \mathbf{x}) + b \right) \quad (2.19)$$

2.2.4 Kernel functions

In order for a bivariate function to be a kernel function, it has to satisfy several mathematical properties called *Mercer's conditions* [36, 37]. Notably, the matrix associated to the kernel applications has to be symmetric and positive semidefinite, which significantly constraints the possibilities. Many different kernel functions have been developed in an attempt to improve the performance of SVM classifiers. The most popular kernel functions include:

- Linear kernel: $K(\mathbf{x}_n, \mathbf{x}_m) = \mathbf{x}_n^\top \mathbf{x}_m$
- Polynomial kernel: $K(\mathbf{x}_n, \mathbf{x}_m) = (1 + \mathbf{x}_n^\top \mathbf{x}_m)^\alpha$
- Gaussian RBF (Radial Basis Function) kernel: $K(\mathbf{x}_n, \mathbf{x}_m) = \exp \left(-\frac{\|\mathbf{x}_n - \mathbf{x}_m\|^2}{2\sigma^2} \right)$, $\sigma \in \mathbb{R}^+$
- Laplace kernel: $K(\mathbf{x}_n, \mathbf{x}_m) = \exp \left(-\frac{\|\mathbf{x}_n - \mathbf{x}_m\|}{\sigma} \right)$, $\sigma \in \mathbb{R}^+$

The use of the linear kernel is equivalent to a case for which the kernel trick is not applied. In situations where the data is actually linearly-separable, it is naturally the one to use. The RBF and Laplace kernels, using the euclidian distance between input samples in order to extract the similarity between them, are two common non-linear kernels. The RBF kernel, however, is arguably the most popular one, as it yields the best results in a large number of situations. Note that most kernels add hyperparameters to the algorithm (for example α for the polynomial kernel and σ for the laplace and the RBF kernel), usually tuned together with C by Cross-Validation.

Eventually, the kernel trick is very powerful because it can be (and has been) embedded in all sorts of methods to extend them to a highly non-linear setting and potentially improve their performance. With the help of an adequately chosen kernel, it often yields very good results. That is part of the reason why kernel methods were very popular and extensively studied in the 2000's. Some of the popular kernel versions of famous methods include for example Kernel PCA [38] or Kernel K-Means [39]. Support Vector Regression, perhaps the most natural extension of SVM classification, is also one of the most popular among them.

2.2.5 Support Vector Regression

Support Vector Regression extends all the ideas of Support Vector Machines to a regression framework. We will present the ε -SVR algorithm, one of the most famous and most used versions of the algorithm. In the case of ε -SVR, we want to build a function that deviates not more than ε from the target values y_n , but is as *flat* as possible, meaning it does not follow the fluctuations of the y_n too closely. A lack of flatness would result in a poor generalization to unseen data. Note that ε is therefore the first hyperparameter of SVR. Following the same scheme adopted in the classification case, let us first describe the case of linear functions. Such linear functions can still be parametrized with a normal (or perpendicular) vector \mathbf{w} , as:

$$\varphi(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b, \text{ with } \mathbf{w} \in \mathcal{X}, b \in \mathbb{R} \quad (2.20)$$

The basic idea of the algorithm is to impose the flatness by minimizing the euclidean norm of \mathbf{w} (or its square, which is equivalent), noted $\|\mathbf{w}\|$. This defines the objective function, as a regularization term. Together with the ε constraint, it forms the optimization problem:

$$\begin{aligned} &\text{Minimize} && \frac{1}{2} \|\mathbf{w}\|^2 \\ &\text{subject to} && y_n - \mathbf{w}^\top \mathbf{x}_n - b \leq \varepsilon \\ &&& \mathbf{w}^\top \mathbf{x}_n + b - y_n \leq \varepsilon \end{aligned} \quad (2.21)$$

with $n = 1, 2, \dots, N$.

As in the classification case, this strict optimization problem is not always feasible and the ε boundaries should be relaxed to avoid that unfeasibility issue. It results in the soft margin version of the SVR algorithm. Slack variables ζ_n and ζ_n^* (there is one at each side of the margin) are added to deal with points that are out of the ε margin. An illustration of the soft margin variables in linear SVR is shown in Figure 2.4. The optimization problem becomes:

$$\begin{aligned} &\text{Minimize} && \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N (\zeta_n^* + \zeta_n) \\ &\text{subject to} && y_n - \mathbf{w}^\top \mathbf{x}_n - b \leq \varepsilon + \zeta_n \\ &&& \mathbf{w}^\top \mathbf{x}_n + b - y_n \leq \varepsilon + \zeta_n^* \\ &&& \zeta_n^*, \zeta_n \geq 0 \end{aligned} \quad (2.22)$$

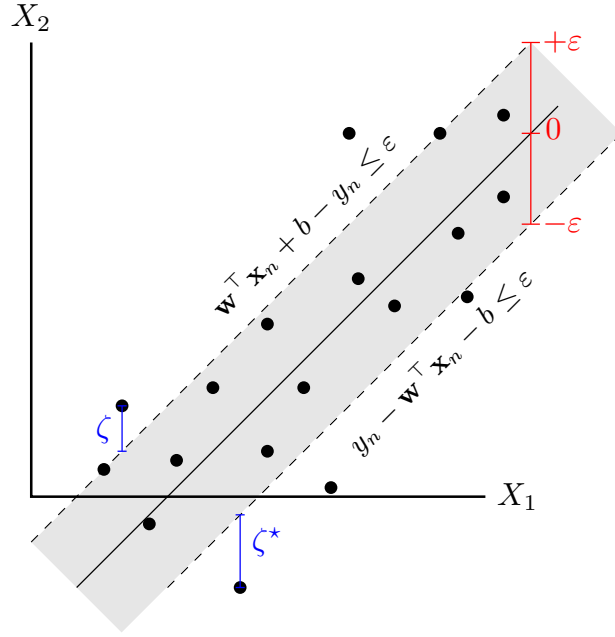


Figure 2.4: The soft margin setting for a linear SVR model, in case of a 2D input space.

In practice, the main modification of this problem is the constant $C > 0$ that will allow one to control the balance between how flat the function is and how far from the margin we tolerate the outliers. This will be tuned as a hyperparameter of the model, with a very similar role to the C constant in SVC.

One can use the Lagrangian formulation to solve this optimization problem, which yields the following dual problem:

$$\begin{aligned}
 \text{Maximize} \quad & -\frac{1}{2} \sum_{n,m=1}^N (\alpha_n - \alpha_n^*) (\alpha_m - \alpha_m^*) \mathbf{x}_n^\top \mathbf{x}_m \\
 & -\varepsilon \sum_{n=1}^N (\alpha_n + \alpha_n^*) + \sum_{n=1}^N y_n (\alpha_n - \alpha_n^*) \\
 \text{subject to} \quad & \sum_{n=1}^N (\alpha_n - \alpha_n^*) = 0 \text{ and } \alpha_n, \alpha_n^* \in [0, C]
 \end{aligned} \tag{2.23}$$

where α_n^* and α_n are the Lagrange multipliers for the two first constraints of the problem. Similar to the classification case, Quadratic Programming solvers for this problem extract the α_n and α_n^* . The few non-zero α_n and α_n^* are of course still called the support vectors, as they lie on the boundaries of the ε region. Finally, it gives us:

$$\mathbf{w} = \sum_{n=1}^N (\alpha_n - \alpha_n^*) \mathbf{x}_n, \text{ and } \varphi(\mathbf{x}) = \sum_{n=1}^N (\alpha_n - \alpha_n^*) \mathbf{x}_n^\top \mathbf{x} + b \tag{2.24}$$

The kernel trick is applied in cases where the data cannot be fit by a linear function. In this case, inner products are substituted by a kernel $K(.,.)$, which yields the general solution provided by SVR:

$$\varphi(x) = \sum_{n=1}^N (\alpha_n - \alpha_n^*) K(\mathbf{x}_n, \mathbf{x}) + b \quad (2.25)$$

All kernels presented in section 2.2.4 are naturally valid here, as for any other kernel method. For more details on SVR, a very popular tutorial on SVR can be found in [29].

2.2.6 SVM in practice

SVM is a very powerful algorithm, which shows great prediction performance (particularly for the classification case). It requires, however, some care in preprocessing and training steps in order to work at its best in practice. We present here the main aspects to take into account when training an SVM model in practice.

Due to its intrinsic learning strategy, SVM requires its user to perform thorough data preprocessing prior to training. First, since SVM requires all input samples to be real-valued vectors, it cannot handle mixed type variables. In particular, *categorical variables have to be transformed into numeric variables*. While there are multiple ways to do so, it is often advised in the literature to use a *one hot encoding approach* ([40]), meaning using c different binary features for c different categories: category 1 is represented by $(1,0,0,\dots,0)$, category 2 by $(0,1,0,\dots,0)$ etc. Second, because SVM relies on the inner products of the feature vectors (in the linear and non-linear case), *the data has to be scaled* prior to the training step. All input vectors are compared two by two and therefore require to be in the same scale. Some practitioners advise to scale the data points to the range $[-1,1]$ or $[0,1]$ [40]. Note that the entire dataset needs to be scaled in the same fashion, including the training data, testing data and any unobserved point to be predicted. Finally, since inner products of all couples of training samples are computed regardless to their values, SVM is highly *sensitive to outliers*. It is therefore advised to discard outliers during the training process.

In addition to the pre-processing, SVM requires a careful model selection process. As presented earlier in the subsection, SVM classification has originally one hyperparameter, C , while SVR has the additional ε to tune. The use of a particular kernel will often further add other hyperparameters. The RFB kernel, for example, adds the gaussian width parameter σ . In practice, the performance of the model is quite sensitive to the choice of hyperparameters and K-fold Cross-Validation is advised to extract the optimal set of hyperparameters, based on a list of possible choices.

Because of some of its intrinsic properties, SVM has additional practical downsides: (i) it is quite slow when the data is relatively large (already with thousands or tens of thousands of samples), specially considering the Cross-Validation process needed, which entails a limited scalability, and (ii) it is difficult to interpret, particularly in high dimensions.

2.2.7 Use of SVM in the thesis

SVM will be used multiple times within this thesis, particularly in chapter 6, for the estimation of various energy-related variables, including meteorological, terrain or urban geometrical variables. As a result, the input features and output labels will naturally be adapted to the task at hand, in content and in form, and described during the presentation of the task. The general pre-processing

and training strategy, however, will remain the same, and will not be specified for every task. The strategy adopted when training a SVM model is the following:

1. Separate (randomly) the data into 75% for the training set and 25% for the test set. (We visualize both sets to assure a certain homogeneity between the two sets and check if the majority of different available labels are represented in the training set)
2. Transform the categorical feature data in numerical values (using a one hot encoding strategy, or another one depending on the task at hand).
3. Scale the data, by centering to the mean and scale to unit variance (of the entire data, including the learning data set as well as unobserved points to be predicted), using the simple `sklearn.preprocessing.scale` function from Scikit-Learn [25].
4. Choose the RBF kernel. Although it is good practice to first try a linear kernel, especially in case of very high dimensions of the input space [40], the RBF kernel always worked better for tasks tackled in this thesis.
5. Perform K-fold Cross-Validation with grid search over the training set to extract the optimal configuration for (C, ϵ, σ) (or (C, σ) for classification), with exponentially increasing values for each of the three hyperparameters ($C = 2^{-5}, 2^{-4}, \dots, 2^{15}$, $\sigma = 2^{-10}, 2^{-9}, \dots, 2^5$), $\epsilon = 2^{-5}, 2^{-9}, \dots, 2^5$). The choice of K in the K-fold Cross-Validation is motivated by a rule of thumb suggested by Friedman et al. [24], which consists in choosing a value for K that offers an overall Cross-Validation error within one standard error of the minimum mean Cross-Validation error (through the folds).
6. Train a SVM model with the obtained optimal hyperparameters over the training set.
7. Test the trained model over the test set, assessing the performance of the model using RMSE and NRMSE for regression task and AE for a classification task.

(Note that because the number of features for most considered data is rather small and to keep some meaning to the features, the dimensionality of the data also remains untouched, unless it is mentioned otherwise in the section of interest (using for example Principal Component Analysis - PCA).) SVM, however, will not be used in this thesis as often as the algorithm presented in the following section, the result of another very clever idea, but with a certain number of practical advantages.

2.3 Random Forests (RF)

Random Forests (RF) [41] is a ML algorithm for classification, regression and related tasks, part of the Ensemble Learning (EL) family of methods. EL aims at combining multiple “weak learners” (simple and fast models with a poor performance) in order to obtain one “strong learner” offering good prediction capabilities. In case of RF, the weak learners are decision trees [42]. What separates RF from the other EL methods using trees, however, is that the combination of these trees is designed to provide a very good tradeoff between speed/easiness to use and prediction performance. In this section, we will (i) present the principles of decision trees for classification and regression, (ii) further

explain their use in Ensemble Learning methods and in particular Random Forests, (iii) give a few guidelines on how to use RF in practice, (iv) present Quantile Regression Forests, a very useful generalization of the regression version of Random Forests allowing to extract the uncertainty attached to RF predictions, (v) finally present specifically how RF is used in this thesis.

2.3.1 Decision Trees

Decision Trees are decision models that have been widely used for many years and in various applications. Their prediction scheme is easy but very efficient: they partition the input space into a set of rectangle subspaces and fit a very simple model (a constant) in each one. The subspaces are defined from the training data using a series of binary splits performed at the nodes of the tree, which successively divide the input space in two. Note that multiple decision tree algorithms have been developed. We will present the CART (Classification And Regression Trees) algorithm, one of the first and most famous decision tree algorithms [42].

Regression trees

Let us consider a regression problem with two input features X_1 and X_2 and a generic output variable Y (the classification version will only require a slight change, as explained later in the section). At each node, a value query is performed on one of the features, for example “Is $X_2 < 3.5$?”. The local input space that gathers the training samples in this node is then split into two subspaces according to this query, and we model the output by the mean of Y in each subspace. The variable and threshold of the query is used in order to have the best fit offered by our very simple mean model. By abuse of language, it is often said that the node itself is split. Following the split, the training samples for which the query is “True” are moved down the tree in a one node (left child node), while the training samples for which the query is “False” are stored in another node (right child node). An illustration of the iterative binary splitting of the input space by a tree is shown in Figure 2.5 for a 2D input space case. In this example, the 2D space is first split at $X_2 = h_1$ (therefore horizontally on the figure, at h_1 level). The two subspaces are stored respectively in the left and right child nodes. Then, they are again split according to the chosen variables and thresholds, vertically with respect to X_1 , horizontally with respect to X_2 . The subspace $X_2 \leq h_1$ is split at $X_2 = h_2$, and the subspace $X_2 > h_1$ is split at $X_1 = h_4$. Finally, the subspace $X_2 > h_2 \cap X_2 \leq h_1$ is split at $X_1 = h_3$. The recursive splits lead to the creation of the five subspaces S_1, S_2, S_3, S_4 and S_5 , which define a partition of the 2D space, and gather, when put together, all the input training samples. Note that terminal nodes are called *leaves* and the first node of the tree (containing all the training samples) is called the *root node* of the tree.

The choice of the feature and the threshold for the splitting query is at the heart of the decision tree algorithm. They are chosen in a way to obtain the *best split*, meaning the splitting couple (variable, threshold) for which the mean of Y in the two subspaces is the closer to the output training values in the subspaces. To measure the *goodness of split*, the notion of *impurity* $i(v)$ of a node v is defined. The impurity of a node measures how far the temporary modelled response (mean of Y) is from all the labels of the points contained in this subspace, and therefore how “spread” the label values are in this node. The best split is therefore the one which reduces the impurity as much as possible, or

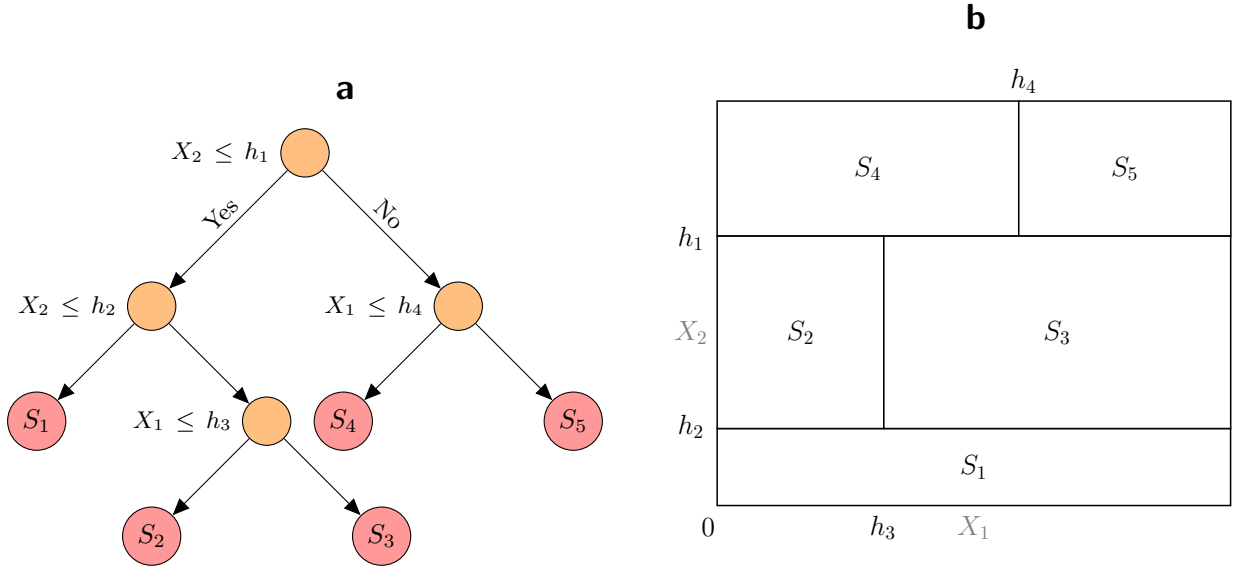


Figure 2.5: Decision tree and partition of the input space. **(a)** Decision tree example built over a 2D input space defined by two features X_1 and X_2 (light red nodes are leaves, orange nodes are normal nodes, including the root node); **(b)** Partition of the 2D input space corresponding to the tree shown in **a**, resulting from recursive binary splitting. S_1, S_2 etc. are the obtained subspaces after the tree was grown, and h_1, h_2 etc. are the thresholds extracted by the algorithm at each split.

equivalently, maximizes the impurity decrease $\Delta i(v)$, between the current node and the two children nodes (left child and right child node). The impurity decrease is defined by Eq. (2.26) [43]:

$$\Delta i(v) = i(v) - \frac{N_{v_L}}{N_v} i(v_L) - \frac{N_{v_R}}{N_v} i(v_R) \quad (2.26)$$

where $i(\cdot)$ is the impurity, v is the current node, v_L is the left child node, v_R is the right child node, and N_v, N_{v_L} , and N_{v_R} are respectively the number of learning samples in the current node, left child node and right child node. Each node corresponds to a particular subspace of the input space obtained from a binary split.

The impurity of the node can in theory be defined by multiple error functions. In the case of a regression problem, the most natural one is the square error function. As mentioned earlier, we consider the mean of the output values as a temporary model in each node. This estimate is called the *local resubstitution estimate*. The impurity function for a regression task is therefore the following, based on the latter node estimate and square error loss: (2.27) [43]:

$$i_R(v) = \frac{1}{N_v} \sum_{x_k \in S_v} (y_k - \bar{y}_v)^2 \quad (2.27)$$

where S_v is the subspace stored in node v (i.e. the y_k are the output values corresponding to the training samples contained in node v), and \bar{y}_v is the mean of the output values in the node.

At each node v , the best split is found by maximizing the impurity decrease, which is equivalent to minimizing the sum of the impurities of the children nodes. If we define the two subspaces resulting from the splitting of node v as follows

$$S_L(j, h) = \{x \mid X_j \leq h\} \text{ and } S_R(j, h) = \{x \mid X_j > h\} \quad (2.28)$$

and use the impurity definition adopted for regression, the splitting variable X_j and threshold h offering the best split are the ones which solve the following optimization problem:

$$\begin{aligned} & \underset{\text{split}}{\text{Maximize}} && \Delta i_R(v) \\ \Leftrightarrow & \underset{\text{split}}{\text{Maximize}} && i_R(v) - \frac{N_{v_L}}{N_v} i_R(v_L) - \frac{N_{v_R}}{N_v} i_R(v_R) \\ \Leftrightarrow & \underset{\text{split}}{\text{Minimize}} && i_R(v_L) + i_R(v_R) \\ \Leftrightarrow & \underset{j, h}{\text{Minimize}} && \sum_{x_k \in S_L(j, h)} (y_k - \bar{y}_{v_L})^2 + \sum_{x_k \in S_R(j, h)} (y_k - \bar{y}_{v_R})^2 \end{aligned} \quad (2.29)$$

Using a greedy algorithm, the threshold h can be extracted rather quickly for each possible splitting variable X_j , and the best couple (X_j, h) is found by going through all the inputs. Algorithms to efficiently solve this problem are notably studied in [43].

One last issue remains: the depth of the tree, meaning the number of times we perform a split, needs to be decided. As for most ML methods, there is here an equilibrium to find between a very large tree (for example fully grown: every leaf contains exactly one sample) which might result in overfitting the data, and a very small tree which would not extract information from the data. The size of the tree is therefore a hyperparameter to tune during the training process. It may be directly chosen or indirectly decided by another parameter: the minimum number of samples per leaf, often called `min_samples_leaf` in implementations of the algorithm. One or the other parameter can be tuned using cross-validation. Some efficient strategies to tune the tree size and even perform additional *pruning* (cutting some leaves of the tree) are discussed in [24]. We will see, however, that within the framework of Random Forests, this parameters does not have a very large impact on the prediction. To see the evolution of a tree in practice, Figure 2.6 shows a regression tree trained to estimate the yearly wind speed in Switzerland (in rural areas and at a height of 10m, as presented in chapter 4).

After the tree has been grown using the training data, new observations are predicted by passing their features through the tree down to a leaf. The leaves are therefore decision nodes giving the predicted value (or the predicted class in case of classification): as for the temporary nodes, a leaf estimates the output value as the mean of its label values. By construction, once a new point falls in a leaf, the estimate given by this leaf is the “last” one and therefore the final prediction of the tree.

Let us define some notations to formally express the solution given by a regression tree (and to prepare for the future section 2.3.5). Formally, to predict the output value for a new point x , we pass x through the tree T until a leaf $l(x, T)$ is reached. We note $N_{l(x, T)}$ the number of training samples in $l(x, T)$. The tree stores weights $\omega_i(x, T)$ for each original training sample x_i in the following fashion: $\omega_i(x, T) = \frac{1}{N_{l(x, T)}}$ if $x_i \in l(x, T)$, else, $\omega_i(x, T) = 0$. The tree prediction is then the weighted average of the training labels y_i , using the discussed weights, which is equivalent to the average of the output labels contained in the leaf containing x . The solution $\varphi_{R, T}$ (the regressor) extracted by the decision tree T is therefore defined by the following equation:

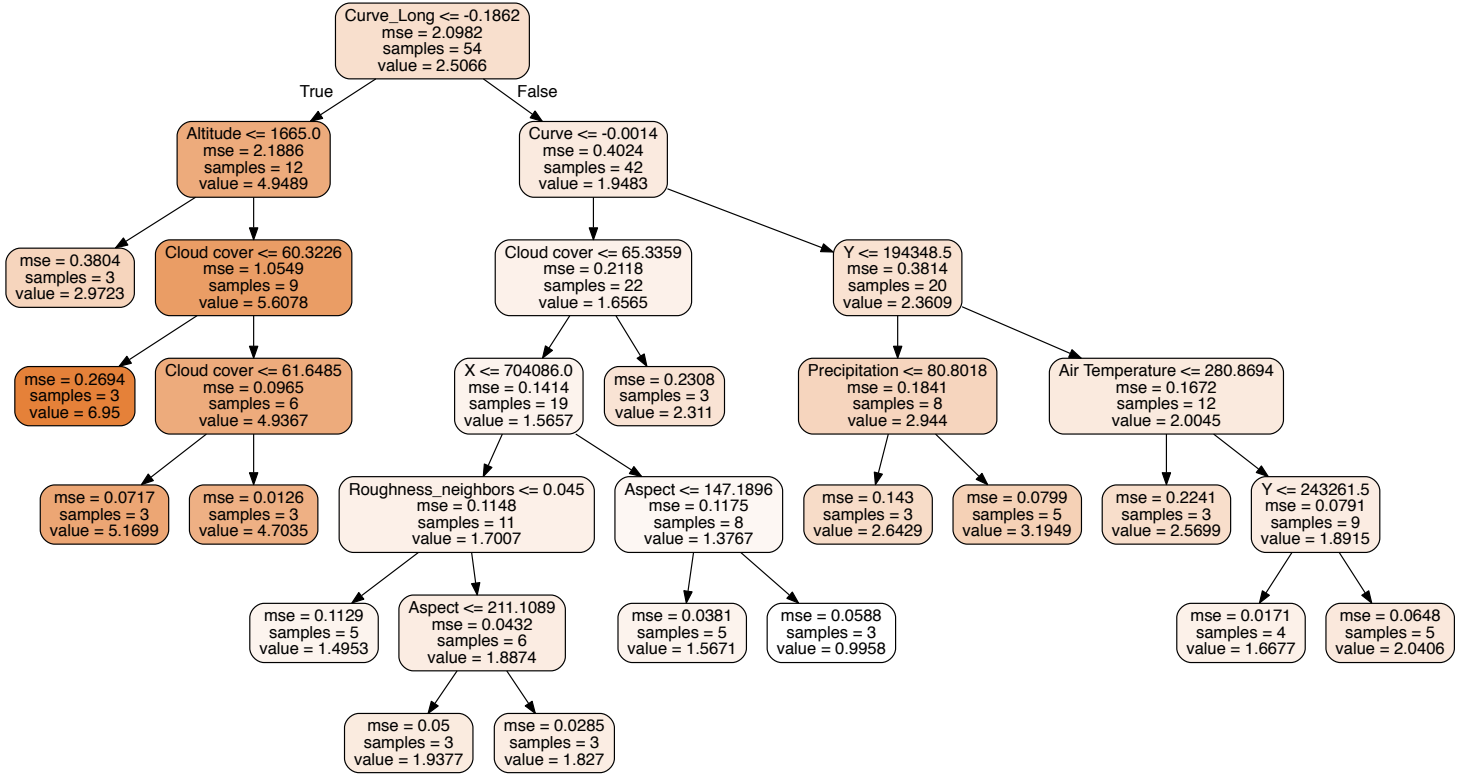


Figure 2.6: Example of one of the regression trees trained in an RF model built for wind speed prediction in Switzerland (as presented in chapter 4). Features (predictors) are X (longitude), Y (latitude), altitude, air temperature, cloud cover, precipitation, sunshine duration, air pressure, terrain slope, terrain aspect, terrain plan curvature (“Curve”), terrain transverse curvature (“Curve_Trans”), terrain longitudinal curvature (“Curve_Long”), terrain roughness (“Roughness”) and surrounding terrain roughness (“Roughness_neighbors”). In each node are specified: the variable and threshold of the split, the current impurity (as the mean square error, “mse”), the number of input samples in the node (“samples”) and the current estimate for the solar irradiance (“value”). Note that the shade of the color of each node is based on the local estimated value: the darker the node, the higher the value.

$$\varphi_{R, T}(x) = \sum_{i=1}^N \omega_i(x, T) y_i = \frac{1}{N_{l(x, T)}} \sum_{x_k \in S_{l(x, T)}} y_k \quad (2.30)$$

Classification trees

In the case of a classification task, the main change is the splitting procedure and therefore the definition of the impurity measure. When predicting classes, the impurity of a node will measure the uncertainty of Y within the node. For example, if the probability for an input sample to be in class C_k is almost the same for all classes (there is almost the same number of input samples in each class), the uncertainty is high because the node cannot decide which class is dominating based on the samples it contains; the impurity is therefore high.

The two most common definitions of impurity used for classification are the ones based on the

Shannon entropy and the *Gini index*. For simplicity, let us define some notations. The probability for a sample point \mathbf{x}_i to belong to class C_k when in the subspace S_v defined by node v is equal to the proportion of sample points in class k within node v . Let us note this probability by the following expression:

$$\mathbb{P}(C_k | v) := \mathbb{P}(\mathbf{x}_i \in C_k | \mathbf{x}_i \in S_v) = \frac{1}{N_v} \sum_{\mathbf{x}_i \in S_v} \mathbb{1}_{\{y_i=k\}} \quad (2.31)$$

where N_v is the number of samples in node v and $\mathbb{1}$ is the indicator function, here returning 1 if $y_i = k$ and 0 otherwise. Following this notation, the possible definitions for classification impurity are:

- based on Shannon entropy:

$$i_S(v) = - \sum_{k=1}^c \mathbb{P}(C_k | v) \log_2 \{\mathbb{P}(C_k | v)\} \quad (2.32)$$

- based on the Gini index:

$$i_G(v) = \sum_{k=1}^c \mathbb{P}(C_k | v) \{1 - \mathbb{P}(C_k | v)\} \quad (2.33)$$

where c is the number of possible classes. Both definitions for the impurity are differentiable and therefore “easily” usable within an optimization framework. Following a similar reasoning than the one presented to extract problem 2.29, and choosing one of the two mentioned impurity definitions, the optimization problem leading to the best split for each node in a classification task can be easily expressed and ultimately solved with a greedy procedure.

The predicted class for a new input point \mathbf{x} is obtained by passing the point through the tree, and given by the *majority vote* in the leaf in which it ultimately falls. The majority vote gives the class with the largest number of representative samples in the leaf, which is equivalent to the class with the largest probability in the leaf. Formally, the solution $\varphi_{C,T}$ (the classifier) extracted by a decision tree T is therefore defined by the following, where $l(\mathbf{x}, T)$ still denotes the leaf containing \mathbf{x} in tree T :

$$\begin{aligned} \varphi_{C,T}(\mathbf{x}) &= \arg \max_{C_k \in \mathcal{Y}} \mathbb{P}[Y = C_k | X = \mathbf{x}] \\ &= \arg \max_{C_k \in \mathcal{Y}} \mathbb{P}[C_k | v = l(\mathbf{x}, T)] \\ &= \arg \max_{C_k \in \mathcal{Y}} \frac{1}{N_{l(\mathbf{x}, T)}} \sum_{\mathbf{x}_i \in S_{l(\mathbf{x}, T)}} \mathbb{1}_{\{y_i=k\}} \end{aligned} \quad (2.34)$$

2.3.2 Making decision trees better: a short historical note

Decision trees have the advantage of being simple, easy to understand and most importantly fast models. Their performance, however, is poor, for both classification and regression tasks. In fact, their error rate is very high (only slightly better than random guessing [24]) which make them a part of the “weak learners” family. More specifically, the main default of decision trees can be better identified through the classical *bias-variance* decomposition, particularly used for regression models. Let us assume that $Y = f(X) + \epsilon$, where the random error ϵ is independent of X and such that $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) = \delta_\epsilon^2$. Using the squared-error loss function (by far the most common and convenient loss function), we can express the expected prediction error of a fitted model φ at an input point $X = \mathbf{x}$ [24]:

$$\begin{aligned}
\mathbb{E}_{\text{EFE}}(\varphi(\mathbf{x})) &= \mathbb{E} \left\{ (Y - \varphi(\mathbf{x}))^2 \mid X = \mathbf{x} \right\} \\
&= \delta_{\epsilon}^2 + \{\mathbb{E}\varphi(\mathbf{x}) - f(\mathbf{x})\}^2 + \mathbb{E}\{\varphi(\mathbf{x}) - \mathbb{E}\varphi(\mathbf{x})\}^2 \\
&= \text{Irreducible error} + \text{Bias}_{\varphi(\mathbf{x})}^2 + \text{Variance}_{\varphi(\mathbf{x})}
\end{aligned} \tag{2.35}$$

By construction, decision trees have relatively low bias but suffer from high variance [24]. Generally speaking, it means that they manage to extract the right distribution from the data, but their predictions are scattered around the mean - instead of being focused on it, which would happen in case of low variance -, resulting in large discrepancies with the actual labels. It would be therefore desirable to reduce this variance in order to obtain better predictors.

An interesting idea was formulated in the middle of the 1990's in order to reduce the variance of noisy predictors: *bootstrap aggregating* or *bagging* [44]. The idea is based on the bootstrap, a technique consisting in creating multiple datasets from one training data by sampling N times from the training data with replacement. Repeating this process B times results in B different datasets, yet following the same distribution as the original training data. These datasets are called *bootstrap-sampled versions* of the training set, or simply *bootstrap samples*. Note that this denomination can be confusing, since the input training points of the original data are also called *samples*. It is however a standard denomination, so we will keep it throughout the chapter. To avoid confusion as much as possible, bootstrap-sampled versions of the training set will always be called “bootstrap samples”, and never solely “samples”, which will always refer to input training points.

The bootstrap has many interesting applications related to statistical accuracy assessment, notably to extract statistics along with predictions. Bagging is a simple idea: averaging the prediction from a collection of models trained over bootstrap samples of the training data will provide a prediction with reduced variance. Since each model is identically distributed, the bias is the same as the one from an individual model (the average of B models is the same as the expectation of any one of them), but the variance is reduced as a result of the averaging. Bagging became the founding idea of the Ensemble Learning family of methods.

Given the properties of decision trees, they were considered as promising weak learners to be used within a bagging framework. Indeed, as pointed out by Hastie et al. [24]: “*Trees are ideal candidates for bagging, since they can capture complex interaction structures in the data, and if grown sufficiently deep, have relatively low bias. Since trees are notoriously noisy, they greatly benefit from the averaging.*” For regression tasks, the prediction is the average of the predictions from the B bootstrap trees. For classification tasks, the predicted class is then the one with the majority of votes from the bootstrap trees, a vote being the class predicted by one of the trees. As a result of bagging, the variance of trees is reduced, therefore offering a good performance for bagged trees. More specifically, the variance decreases with the number of trees, which makes the B hyperparameter a very easy one to tune: the larger the B , the better the prediction.

Observing the bias-variance decomposition and the performance of bagging, another natural thought comes to mind: What if it was possible to also reduce the bias of trees? By definition, bagging cannot improve the bias, but solely the variance, as discussed previously. The method therefore needs a slight improvement in order to achieve that goal. This improvement was soon found with the development of *boosting* [45]. The idea, still relatively simple, is to rethink the combination of the weak learners to extract a “powerful committee”: instead of building models on bootstrap samples of the training

data, boosting considers weighted versions of the training data. The most famous boosting algorithm is AdaBoost.M1 [45], originally designed for classification. For each weighted version, different weights w_i are applied to each of the training samples, and a model G_k is trained based on this weighted version. The weights are tuned iteratively for each new weighted version so that the focus is put on the observations that were previously misclassified. The training steps are as follows: (i) The first model is trained on the original training data, meaning the weights are set to $w_i = \frac{1}{N}$. (i) The k^{th} weighted version is defined as follows: the input points misclassified by the previous model G_{k-1} see their weight increase, while the well classified observations see their weight decrease. Eventually, the predicted class is given a weighted majority of votes $\sum_{k=1}^W v_k G_k(\mathbf{x})$, where W is the total number of models trained and v_k are weights extracted by the algorithm. As a result, the models are iteratively trained in an adaptive way to reduce the bias (the weighted versions of the training data are not identically distributed), which allows for a better performance to be obtained when compared to a single model.

Boosting (notably AdaBoost) was tried with decision trees, to obtain “boosted trees”. It was a large success, since it appears to significantly outperform bagging in most cases. A main issue, however, is that boosting is slower than bagging and more difficult to tune (it is the case for many algorithms based on boosting, notably gradient tree-boosting [46], an algorithm also became very popular), given its higher level of complexity within the core of the algorithm.

By the end of the 1990's, combining weak learners in clever ways had proven to be very efficient in improving their performance significantly. However, there was still a desire to design an ideal methodology, combining the simplicity and relative speed of bagging with the newly discovered performance of boosting. *Random Forests*, presented by Breiman in 2001 [41], (almost) achieved that goal.

2.3.3 The Random Forests classifier and regressor

Random Forests [41] is a modified version of bagging, which improves the variance reduction of bagging by de-correlating the trees trained on the bootstrapped versions of the training data. In fact, the correlation between the trees is a serious issue of the bagging procedure: they are identically distributed but not independent, and the variance of the average of the trees is proportional to the correlation. Considering B non independent but identically distributed random variables with variance δ^2 and pairwise correlation ρ , the variance of their average is given by [24]:

$$\rho\delta^2 + \frac{1-\rho}{B}\delta^2 \quad (2.36)$$

The first term is therefore a problem for the overall variance of the ensemble of trees (the second term vanishes with increasing B). RF de-correlates the trees by slightly changing the core of the growing process of each tree, during the split at each node:

Instead of using the best split among all d input variables, $m \leq d$ variables are selected at random out of all possible ones, and the best split among these m variables is used.

This added layer of randomness reduces the correlation between the trees and therefore further reduces the variance of the model. Note that it may slightly increase the bias of the model, but this increase in bias is greatly compensated by the variance reduction, resulting in an overall better error. This small modification considerably increases the performance of the algorithm compared to

Table 2.1: Comparison of tree-based Ensemble Learning classical models. For each strategy, the assessment is based on a benchmark model: CART for trees, bagging CART predictors for bagging and AdaBoost for boosting. (✓ signifies low ability, ✓ signifies medium ability and ✓ signifies high ability)

	Trees	Bagging	Boosting	Random Forests
Speed	✓	✓	✓	✓
Simple to tune	✓	✓	✓	✓
Prediction accuracy	✓	✓	✓	✓

bagging, making it comparable to boosting for many problems. It is, however, easier to train and tune (with only two important hyperparameters, m the number of candidates for splitting and B the number of bootstrap trees) and ultimately faster due to the smaller number of candidates for splitting at each tree node. As a result, RF has become very popular and is preferred to boosting for various problems; the technical characteristics of bagging, boosting, and Random Forests can be summarized in the simplified comparison shown in Table 2.1.

(For the sake of exhaustivity, one should note that when tuned properly, gradient boosting, AdaBoost's big brother, may yield a slightly better performance than RF. Also, RF recently saw an even fiercer competitor in the new implementation of gradient boosting offered by the now popular XGBoost library [47], which further improves its speed and scalability. It is however not included in popular ML libraries such as Scikit-Learn, and is independently maintained, making it less accessible to the large public.)

The RF classifier and regressor are identical to the ones defined by bagged trees, besides the splitting difference when building each individual tree. In a regression task, the RF prediction $\varphi_{R,RF}(\mathbf{x})$ of the label corresponding to the unseen input point \mathbf{x} is the average of the predictions from the trees. Following the notations adopted in section 2.3.1, the RF regressor computes $w_i(\mathbf{x})$, that is, the average weight over the collection of bootstrap trees and estimates the prediction using the following expressions [48]:

$$\varphi_{R,RF}(\mathbf{x}) = \sum_{i=1}^N \omega_i(\mathbf{x}) y_i = \frac{1}{B} \sum_{b=1}^B \varphi_{R,T_b}(\mathbf{x}) \quad (2.37)$$

with

$$\omega_i(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \omega_i(\mathbf{x}, T_b) \quad (2.38)$$

where T_1, \dots, T_B are the trees built respectively with each bootstrapped version of the training data, and $\varphi_{R,T_b}(\mathbf{x})$ is the value predicted by bootstrap tree T_b . In a classification task, the RF prediction $\varphi_{C,RF}(\mathbf{x})$ of the class corresponding to the unseen input point \mathbf{x} is the one cast by a majority vote through the bootstrap trees (the most predicted class):

$$\varphi_{C,RF}(\mathbf{x}) = \text{majority vote} \{ \varphi_{C,T_b}(\mathbf{x}) \}_1^B \quad (2.39)$$

where $\varphi_{C,T_b}(\mathbf{x})$ is the class predicted for \mathbf{x} by tree T_b .

Illustrations for the training and prediction schemes of RF are provided in Figure 2.7.

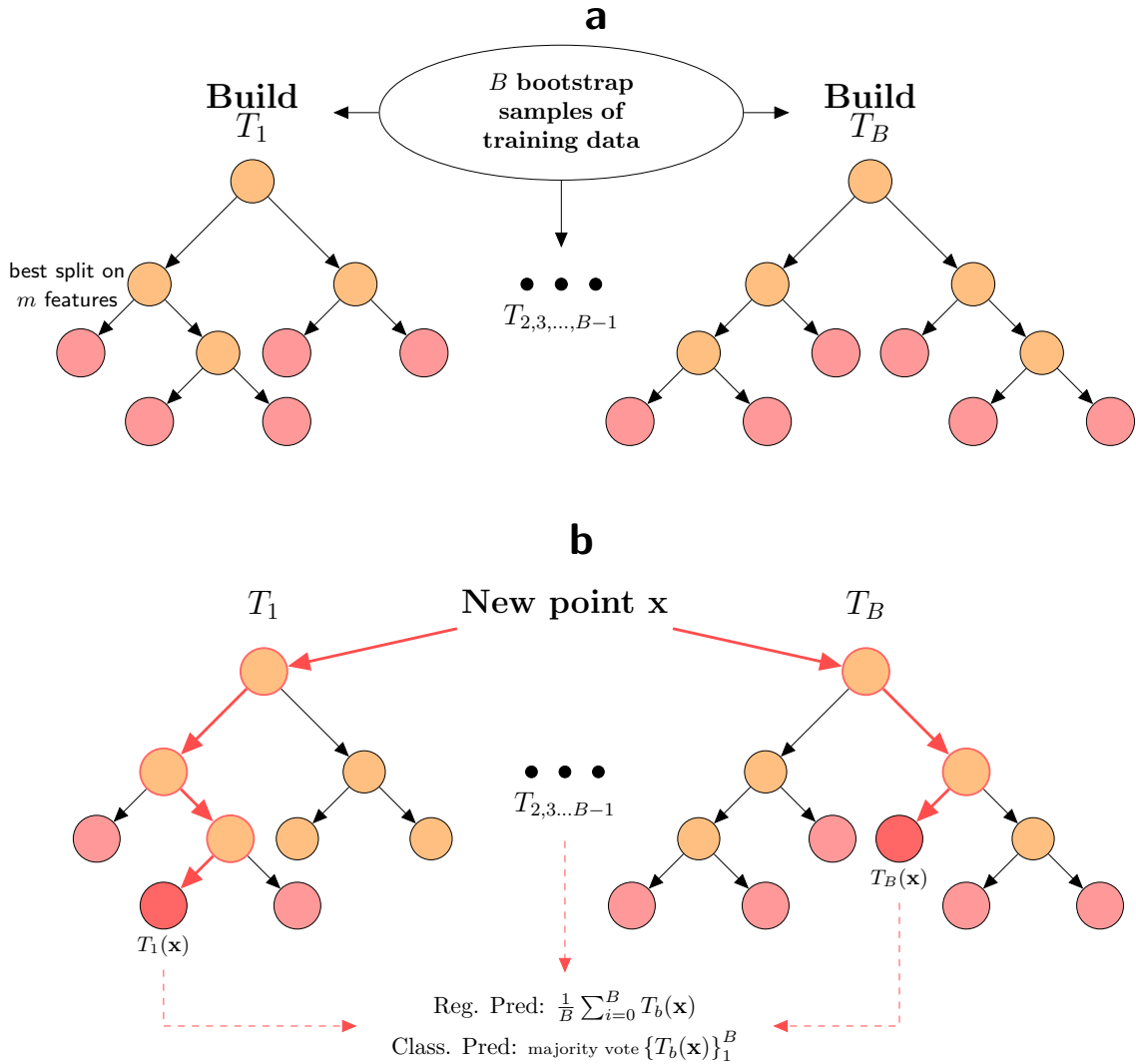


Figure 2.7: An illustration of the training and prediction processes with Random Forests, for the regression case. **(a)** RF training scheme, **(b)** RF prediction scheme. For readability purposes, the notation is simplified compared to the one adopted in the text: $T_b(\mathbf{x})$ is the prediction performed by tree T_b and is therefore the predicted value for a regression task and the predicted class for a classification task.

2.3.4 RF in practice

Besides its high training speed and easiness to tune, the Random Forests algorithm has numerous advantages compared with other machine learning algorithms (e.g. SVM) [41, 49], which contributed to its great success. These advantages can be categorized in two families: (1) useful tools embedded in the algorithm and (2) practical advantages allowing for easier pre-processing and training (as a consequence of certain properties of the algorithm).

Useful RF embedded tools

Out-Of-Bag samples. The first “by-product” of Random Forests is the *Out-Of-Bag* (OOB) error provided when training a forest [41]. The term OOB is quite explicit and expresses how RF can use

the bootstrap trees to perform predictions within the training data. Note that by definition, in each bootstrap sample, some of the original training samples may not be included since the sampling process is done with replacement; these training samples are called *Out-Of-Bag samples*, the “bag” being the considered bootstrap sample. Every time a bootstrap tree is grown, it can then be used to predict the output values corresponding to the OOB samples. The discrepancy between predictions and actual labels therefore provides a validation error for the bootstrap tree. The OOB error of all trees can then be averaged to form the OOB estimate of error rate (or simply the “OOB error”) for the RF (on average, each data point would be out-of-bag around 36% of the times [49].) As a result, the OOB error can serve as a validation error, very similar to K-fold Cross-Validation error, which can be obtained while training the forest. Note that implementations of RF often give the OOB score rather than the OOB error. The OOB score is simply given by $1 - \text{OOB error}$ and is defined between -1 and 1 (the closer it is to 1 , the better the generalisation capabilities of the model). Besides the OOB error, the OOB samples can be used for multiple tasks, including the ones presented in the following paragraphs.

Variable Importance. The second additional information provided by a trained RF is the *Variable Importance* (VI) measure [41, 42]. The VI seeks to estimate the impact of each feature when building the trees, and ultimately a measure of how important each feature is in the prediction of the output of interest. Within the framework of Random Forests, there are two possible ways to extract a VI measure.

The first VI definition is based on the building strategy of a decision tree: at each split in each tree, the variable chosen for the split is the one which maximizes the impurity decrease of the children nodes, meaning the improvement in squared error (for regression) when fitting a constant model in the children subspaces. For each variable, one can sum the improvements (impurity decreases) over all internal nodes for which this variable was chosen for the split, therefore providing an importance measure of the variable within the tree. For bagged trees, and particularly a RF, the importances of each variable can be averaged over the trees to obtain an overall importance measure of each variable within the RF. This definition of VI is called the *mean decrease impurity* or the *Gini importance*.

RF can provide another Variable Importance measure using the OOB samples [41]. The idea is to monitor the change in performance of the forest if the variable of interest is wrongly used, for example with altered values. For each tree grown in the forest, the OOB labels are predicted using that tree, and then predicted again using the same bootstrap sample but with the values for one variable randomly permuted. The decrease in accuracy resulting from the permutation of the variable values provides the importance of the variable. It can be averaged over all trees to extract the importance of the variable over the overall tree. This definition of VI is called *OOB randomization* [24].

Proximity measure. A last interesting additional measure one can extract from RF is the *proximity measure* [41]. The fraction of trees for which input samples x_i and x_j fall in the same leaf defines the (i, j) element of the proximity matrix. The original idea was to attempt to extract a similarity measure between samples to perform clustering or related tasks, based on the intuition that samples of the same “type” should fall in the same leaf more often than unrelated samples. The shape of the resulting proximity plot, however, seems to be quite constant across different tasks and data making it difficult to interpret. It is as a result perhaps the least popular aspect of RF.

Practical advantages of RF

The properties of RF entail a certain number of practical advantages regarding the training and the pre-processing of the data prior to the model training. These include [24, 41]:

- Feature selection is embedded in the core of the algorithm by choosing the variable offering the best split at each node. There is therefore no need to carefully select variables prior to the training of a model. In particular, all variables intuitively having an impact on the output variables may be used. Unimportant or redundant variables are not an issue as they will be automatically discarded when splitting the nodes.
- It is robust to outliers.
- It can handle very large datasets.
- It can handle mixed features, meaning quantitative or qualitative (for example categorical).
- Missing data can also be handled by RF in various ways [50].
- There is no need to scale or normalize the data prior to fitting (one feature is never actually compared to another one).
- Because of the OOB error, RF does not necessarily need additional cross-validation (as the OOB error serves as a validation error).

Note that all the previously mentioned advantages (besides the last one) are actually properties inherited from decision trees. The main contribution of RF (as well as bagging and boosting) is that it improves the main weakness of trees being the prediction performance.

In addition to the above advantages, it appears that the performance of an RF model is not highly sensitive to the parameters and it is relatively straight forward to obtain a near-optimal forest for the task in hand, with very little tuning required [24, 49]. Also, the accuracy increases with the number of trees (B) [24], so there is no need to fine-tune it. It is current practice to fix m and simply try to increase the value of B until an accuracy plateau is reached. While a high number of variables calls for a large number of trees, $B = 500$ trees appears to be sufficient to reach the plateau in most cases. Fig. 2.8 shows an example of the evolution of the OOB score ($1 - \text{OOB error}$) with an increasing number of trees. The number of variables considered for split m can be tuned using OOB estimates or K-fold Cross-Validation. It is however advised to simply select the m value which gives the best results from a list of values advised by Breiman in practice. The list of values consists in a default value, twice the default value and half this value. The default m value is $\lfloor \frac{d}{3} \rfloor$ for regression and $\lfloor \sqrt{d} \rfloor$ for classification where d is the total number of features in the training data [41, 49]. If not included in the previous list, the choice of $m = 1$ also often gives good results. Finally the `min_samples_leaf` does not dramatically change the performance of the model. While it could in theory lead to overfitting to set it to a low number, experience show that a small value (higher than 1) leads to near-optimal results.

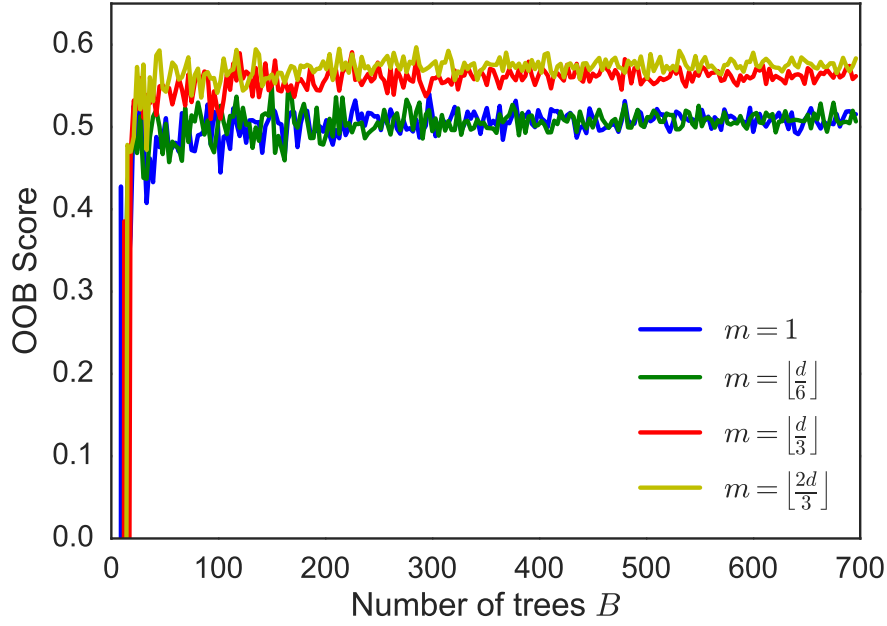


Figure 2.8: Evolution of the OOB score (1–OOB error) with the number of trees B in a RF for different values of m . This example was computed while training a RF for the estimation of the global horizontal radiation in Switzerland (G_H) in July, as presented in chapter 7.

2.3.5 Quantile Regression Forests for Prediction Intervals estimation

Random Forests can be generalized in order to provide information on the probability distribution of the output variable. This generalization is called Quantile Regression Forests (QRFs) [48]. QRF compute informative statistics (quantiles) over all samples in the trees' leaves instead of only aggregating them. It notably allows to extract Prediction Intervals (PIs) when predicting the output value of a new point.

Following the notation used in sections 2.3.1 and 2.3.3, QRF allow to compute the full conditional distribution of the output variable (given this new input point \mathbf{x}) $\mathbb{P}(Y \leq y \mid X = \mathbf{x})$, instead of extracting only the conditional expectation $\mathbb{E}(Y \mid X = \mathbf{x})$ of the variable, which is the quantity estimated by traditional RF (note that we denote the vector of input variables as one d -dimension input variable X to lighten the notation). The computation of the full conditional distribution can be easily derived from the RF estimation using the following expression [48]:

$$\mathbb{P}(Y \leq y \mid X = \mathbf{x}) = \mathbb{E} \left(\mathbb{1}_{\{Y \leq y\}} \mid X = \mathbf{x} \right) \quad (2.40)$$

where $\mathbb{1}_{\{Y \leq y\}}$ is an indicator function, returning 1 if $Y \leq y$ and 0 otherwise. The conditional distribution for any $y \in \mathbb{R}$ can be approximated in a similar way as the conditional mean $\mathbb{E}(Y \mid X = \mathbf{x})$ in traditional RFs, that is, by replacing the observed Y_i outputs by $\mathbb{1}_{\{Y_i \leq y\}}$. Hence the expression is given by [48]:

$$\mathbb{P}(Y \leq y \mid X = \mathbf{x}) = \sum_{i=1}^N \omega_i(\mathbf{x}) \mathbb{1}_{\{Y_i \leq y\}} \quad (2.41)$$

where the weights $\omega_i(\mathbf{x})$ are the same weights used in the original RF in Eq. (2.38).

Given the conditional distribution, PIs can be easily estimated as they are based on α -quantiles. The α -quantile Q_α is defined as the y value such that the probability for Y to be smaller than y is equal to α , for a given $X = x$. The α -quantile is given by [48]:

$$Q_\alpha = \inf \{y : \mathbb{P}(Y \leq y | X = x) \geq \alpha\} \quad (2.42)$$

For any input point x , a prediction interval can then be computed at any confidence level CL with $\alpha = 1 - 0.01 \times \text{CL}$. For instance, a 95% PI is given by:

$$I_{95}(x) = [Q_{\frac{\alpha}{2}}, Q_{1-\frac{\alpha}{2}}] = [Q_{0.025}, Q_{0.975}] \quad (2.43)$$

Sadly, QRF are not available in many libraries despite their straight-forward theoretical computation. In this study, a python machine learning library called Scikit-Learn [25] is used. As it includes an RF implementation but not a QRF one, a small script is added to extend the original RF class. Using equations 2.41, 2.42 and 2.43, the added script allows us to estimate the PIs. The main requirement for the computation of the PIs is the knowledge of the training samples in the tree leaves, which is not provided by Scikit-Learn. This information, however, can be acquired by passing the training input data back in the built trees of the forest using the same bootstrap data sets as the ones used to build the trees. In practice, in order to keep track of the bootstrapped versions of the training data one can access the `random_state` used by the algorithm for each tree of the forest. The trees are stored in the `estimators_` list of the forest.

PIs can naturally be derived for both observed points (in the test set) and unseen points. The test set serves as a basis to cross-validate the intervals by observing the percentage of known points characterized with an output value within the built interval; this percentage is called the test confidence. For a PI to be reliable, the test confidence should be close to the confidence level chosen while building the interval. For new points, the PIs give a measure of the uncertainty of the prediction. In addition, they are a good indication of how difficult the variable is to be predicted at this point. The narrower the interval, the smaller the variability of the estimation (and the easier it seems to predict the variable), and the more one can be confident about the prediction.

2.3.6 Use of RF in the thesis

Random Forests will be used multiple times within this thesis, notably in chapters 4, 6 and 7. As a result, the input features and output labels will naturally be adapted to the task at hand, in content and in form, and described during the presentation of the task. The general pre-processing and training strategy, however, will remain the same. In particular, the steps are as follows:

1. Separate (randomly) the data into 75% for the training set and 25% for the test set. (We visualize both sets to assure a certain homogeneity between the two sets and check if the majority of different available labels are represented in the training set)
2. All unnecessary pre-processing steps described in the advantages are avoided. In particular the data is not scaled, and outliers remain untouched. Also, as explained in the case of SVM section 2.2.7, the dimensionality of the data generally remains untouched, unless it is mentionned otherwise in the section of interest.

3. Choose $B = 500$ trees (or 1000 trees in case of very large and high dimensional data)
4. Perform K-fold Cross-Validation to pick the best value for m in the list of advised values for m , as discussed previously in section 2.3.4 (for regression: $1, \lfloor \frac{d}{6} \rfloor, \lfloor \frac{d}{3} \rfloor, \lfloor \frac{2d}{3} \rfloor$, for classification: $1, \lfloor \frac{\sqrt{d}}{2} \rfloor, \lfloor \sqrt{d} \rfloor, \lfloor 2\sqrt{d} \rfloor$). The choice for K is based on the rule of thumb explained in section 2.2.7.
5. Set `min_samples_leaf=3`.
6. Train an RF with previously chosen m , B and `min_samples_leaf` over the training data.
7. Test the trained model over testing data, and assess its performance, based on the RMSE, NRMSE and OOB score.
8. Extract the Variable Importance of the chosen features for the model using the `scikit-learn` `feature_importances_` attribute for the trained RF model. It is based on the Gini definition. Since the VI resulting from each RF slightly changes based on the distribution of the bootstrap samples, the VI is computed 100 times (based on 100 trained RF models) and averaged over the 100 times to obtain an overall VI.
9. In case of a regression task, compute Prediction Intervals using QRFs, both in the test set and for unobserved points to be predicted.

2.4 A last note: RF Vs. SVM

SVM and Random Forests are both the results of brilliant ideas, which are perhaps two of the most powerful ideas in Machine Learning (along with Deep Learning, which is not discussed here). Interestingly, they were developed around the same period (Bagging and RF are slightly younger than SVM) and, since they show similar prediction performances, they were therefore compared many times for various tasks [51–55]. Naturally, it is unreasonable (and in any case impossible) to extract a clear “winner” and elect the method which would provide the best performance regardless of the task at hand. A few conclusions, however, can be drawn from their multiple comparisons and in the light of their respective properties:

- They seem to have comparable prediction capabilities. SVM, however, may provide slightly better results in some cases, especially when the training data is relatively small and when the dimensionality is high compared to the size of the training set.
- In terms of speed and simplicity to use and tune, RF is the clear winner. It is faster, easier to train and tune, more robust to outliers and more scalable.
- RF have many additional useful tools, including a better interpretability (from the intrinsic structure of trees through Variable Importances) and the possibility to easily extract uncertainty measures based on its intrinsic “bootstrapped” structure

As a result of the latter points, RF will be generally preferred in this thesis, unless the size of the training set is very small and the dimensionality is higher than the number of samples.

Finally, let us conclude this chapter with two additional notes.

(i) While the systematic use of one algorithm for multiple tasks should be avoided, the choice of SVM and RF to build many models in the thesis was motivated by all the advantages previously mentioned, particularly for RF, which provide remarkable results with relatively small processing efforts, therefore facilitating the use of ML in multiple modeling steps. Even though in some cases extensive testing of several methods could yield slightly better results for each task at hand, the aim of this thesis is to show that ML, with the help some of its more powerful methods, can achieve good estimations for energy-related variables.

(ii) While the two presented methods were extremely popular in the end of the 1990's and beginning of 2000's, the current domination of Deep Learning methods in the state-of-the-art ML strategies (since the beginning of the 2010's, really), cannot be denied. As shortly mentioned earlier, the regained popularity is notably due to the new availability of tremendous computational power, which, coupled with complex network structures (Recurrent Neural Networks, Convolutional Neural Networks) proved to yield spectacular results. There are two main reasons why these methods were not chosen within this thesis. The first one is that while these methods showed great results for traditional ML tasks (for example tasks related to audio, image, video and text data), they were not extensively employed for energy-related tasks. The second reason is the difficulty and requirements needed to train such methods. They cannot be trained with a systematic strategy, often require many blind try-outs to extract the best structure of the network and may take several days to train in case of large networks. They are therefore not adequate for multiple predictions of many different variables often showing different patterns, as performed in this thesis. More importantly, they require very large amounts of data to be trained properly and provide good performance. As most of the datasets used in this thesis are relatively small (hundreds to thousands samples), these methods could not be appropriately used.

3

Theory and modeling of renewable energy systems

This chapter borrows from the book chapter:

Assouline, D., Mohajeri, N., and Scartezzini, J-L. (2018). Estimation of Large-Scale Solar Rooftop PV Potential for Smart Grid Integration: A Methodological Review. In Sustainable Interdependent Networks (pp. 173-219). Springer, Cham.

This chapter presents fundamental theoretical and practical concepts related to each of the three renewable energies explored in this thesis: wind energy, shallow geothermal energy and solar energy. More specifically, the chapter, for each energy form, presents:

1. The significant physical variables related to the type of resource, allowing to express its potential,
2. The models and methods used in the thesis to model the mentioned variables and therefore the behaviour of the renewable energy system,
3. The systems and power plants related to the energy and their fundamental differences. The choice of the considered system(s) within the potential studies performed in this thesis will also be justified.

The chapter presents a summary of the models and concepts that we use in the rest of the thesis. For more details on each of them, we provide several references. While we may mention models and concepts which are not used in the thesis (e.g. because they are fundamental or popular), their presentation will therefore remain succinct.

The structure of the chapter is as follows. Section 3.1 presents concepts and models concerning the wind energy, as a theoretical background for chapter 4. Section 3.2 tackles theory related to shallow geothermal energy, as a theoretical background for chapter 5. Section 3.3 finally presents solar energy fundamentals, specifically for PhotoVoltaic panels, as a theoretical background for chapters 6 and 7.

3.1 Wind energy modeling

Wind is a consequence of differences in atmospheric pressure, mainly caused by the unequal heating of the earth's surface by the sun; as such, wind can be interpreted as a indirect portion of solar energy, which is nonetheless non negligible. Clean, safe, affordable, and available in the long-term, it is considered to be one of the most promising sources of renewable energy.

The physical principles behind the power delivered by wind turbines (WT) is rather simple, and similar for all types of turbines, from the traditional large turbine installed in rural areas to a micro turbine installed over buildings. The turbine converts the kinetic energy from the moving air to mechanical energy through the rotor, and then to electrical energy through the generator. The power delivered by the turbine with a wind with speed v is originally the kinetic energy $E_k = mv^2/2$, extracted during an amount of time t . It receives a cylindrical volume $V = A_w vt$ of air through its blades spanning an area A_w , as shown in Figure 3.1. Thus, the mass of air is $m = \rho V = \rho A_w vt$, with ρ the density of air. Plugging this in the kinetic energy term, and writing that $P = E_c/t$, we obtain the output electrical power P of the turbine:

$$P = \frac{C_p \rho A_w v^3}{2} \quad (3.1)$$

with C_p the coefficient of performance of the turbine (capturing the various losses in the conversion process). It should be noted that, while the theoretical maximum value for C_p is 59% (Betz limit [56]), it usually fluctuates between 20% and 50% (which is generally larger than typical PV panels) and varies with the wind speed. More details on the behavior of C_p will be given in section 3.1.6.

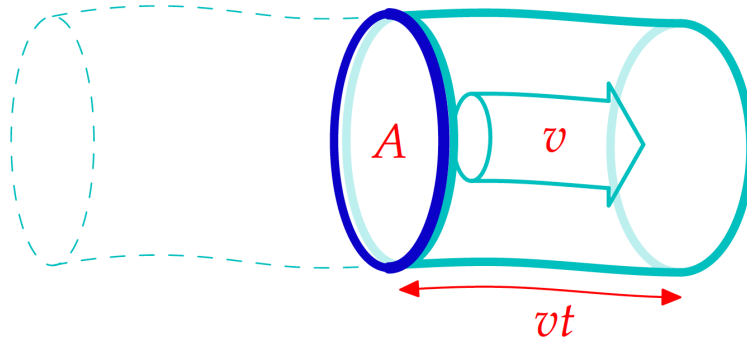


Figure 3.1: Volume of air received by an area A (area typically spanned by blades of a wind turbine, called A_w in this chapter). Source: [57]

The wind speed is the fundamental theoretical variable to express the wind potential; there are, however, multiple wind-specific variables impacting directly on the latter one. In the rest of this section, we will therefore present these variables and various laws used to model them as well as wind speed. We will notably differentiate rural areas from urban areas, as the wind behavior varies significantly from one to the other setting.

The notions presented in this section will be applied in chapter 4.

3.1.1 Significant wind variables

Based on the expression of the wind power delivered by a turbine, it is clear that **wind speed** is the most impactful variable on the wind potential. The **wind direction**, although less important, is also significant since the electricity generated (through the C_p coefficient) by a wind turbine depends in practice on the angle of incidence between the wind direction and the turbine blades. Typically, the optimal performance is obtained when the wind is perpendicular to the area spanned by the turbine blades.

In addition, there are variables which directly impact the wind speed and therefore require attention: the **roughness length** and **displacement height**. These variables generally express, respectively, the “smoothness” of the terrain (or surface over which the wind flows) and the typical height of local obstacles. In particular, they are the two main parameters of a logarithmic-law, or *log wind profile*, which describes the vertical variations of wind speed (further presented in section 3.1.3). Note that in the present thesis, 2D (“plane”) extrapolation of variables is often performed with means of Machine Learning (ML) methods, as presented in the previous chapter. Vertical extrapolation, however, is more challenging to perform with ML methods since they rely on the availability of data, which is often measured at the same height for wind speed (usually 10m). Thus, the variables and models related to the vertical behavior of wind speed are crucial and the focus of the present section. In order to introduce these latter concepts, we first need to shortly present the vertical structure of the atmosphere, and in particular in urban areas.

3.1.2 Vertical structure of the urban atmosphere

The structure of the atmosphere, particularly within the framework of wind modeling, is described by a thorough theory [58, 59]. We here present some of the concepts of this theory, needed for the development of the wind potential estimation presented in chapter 4.

The lowest part of the Earth’s atmosphere, called the *Atmospheric Boundary Layer (ABL)*, is divided in multiple layers defining domains for which the wind has a certain behavior. Above an urban area, the ABL has a particular structure; typically, a particular layer, the *Urban Boundary Layer (UBL)*, is formed over an urban area, starting from the rural-urban border. It is therefore differentiated from the *Rural Boundary Layer (RBL)*, characteristic of rural areas. The UBL is divided into four layers [59], depicted in Figure 3.2:

- *Urban canopy layer (UCL)*. Defines the “envelope” of buildings, from the ground up to the mean height of buildings/trees.
- *Roughness sublayer (RSL)*. Defines the domain strongly perturbed by the presence of buildings/trees, from the ground up to two to five times the mean height of buildings/trees (including the UCL).
- *Inertial sublayer (ISL)*. Above the RSL, and characterized by (i) a logarithmic speed vertical profile and (ii) small variation of turbulent fluxes with height. The first point signifies that the wind speed can be treated in the sole vertical direction using a log-law; the second point has for a consequence that, using the latter log-law, one can express the wind speed at a particular height as a function of the wind speed at another height.

- *Mixed layer*. Above the ISL, where atmospheric properties are uniformly mixed by thermal turbulence.

The typical height for each layer is illustrated in Figure 3.2. Within the present thesis, we are neither interested in the Mixed layer, typically too high for wind turbine installations, nor in the UCL, where the wind flow behavior can only be modeled with thorough 3D analysis and Computational Fluid Dynamics (CFD), intractable at a large scale. Thus, the two main layers of significance for the present study are the ISL and the RSL. The ISL is theoretically convenient because of its two mentioned properties, which will allow to consider some assumptions discussed in chapter 4. Its typical height is also adapted for traditional large turbines, installable in rural areas. The RSL is in theory within a perturbed domain subject to specific flow conditions based on individual obstacles; it therefore requires thorough 3D approaches for wind modeling. It is, however, the layer in which building mounted turbines can be installed within urban areas.

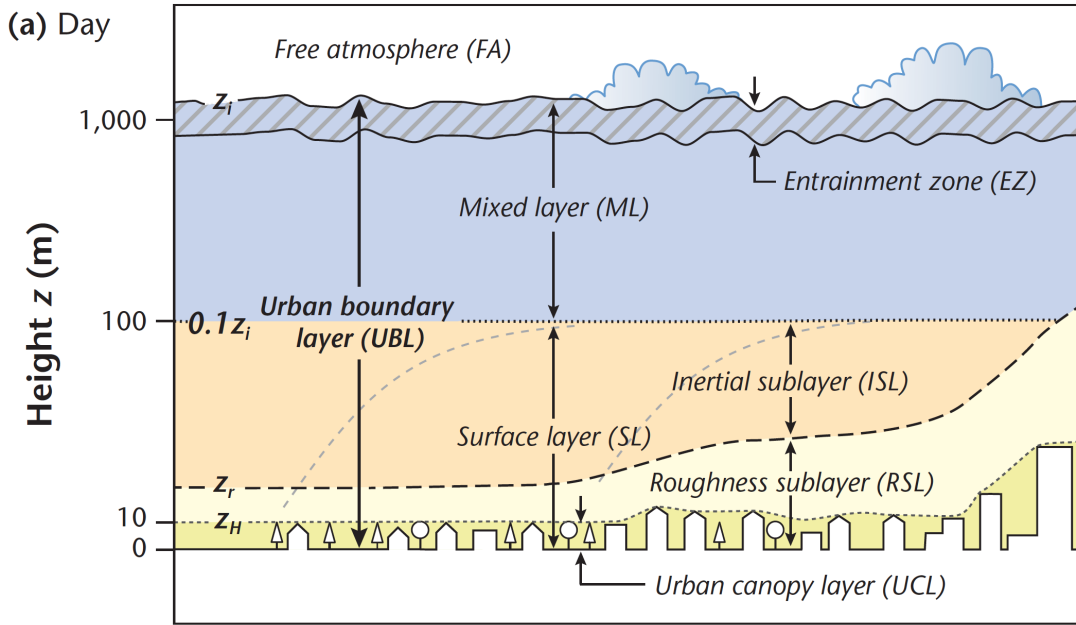


Figure 3.2: Schematic of typical layering of the atmosphere over a city (by day). Note the height scale is logarithmic, except near the surface. Source: [59].

3.1.3 Vertical wind modeling

As previously mentioned, it is possible to express the vertical behavior of wind speed with a logarithmic law (sometimes called log wind profile), in the boundary layer [58], as follows:

$$u(z) = \frac{u^*}{\kappa} \ln \frac{z - z_d}{z_0}, \text{ for } z > z_0 \text{ and } z \in \{ \text{Rural BL} \cup \text{ISL} \} \quad (3.2)$$

where $u(z)$ is the wind speed at a height of z , u^* is the friction velocity, κ is the Von Karman constant, and z_0 and z_d are respectively the roughness length and displacement height of the considered location. Note that the thorough definition of z_0 is defines by a boundary condition of this log-law: the roughness length is the height under which the speed $u(z)$ is theoretically considered to be equal to 0.

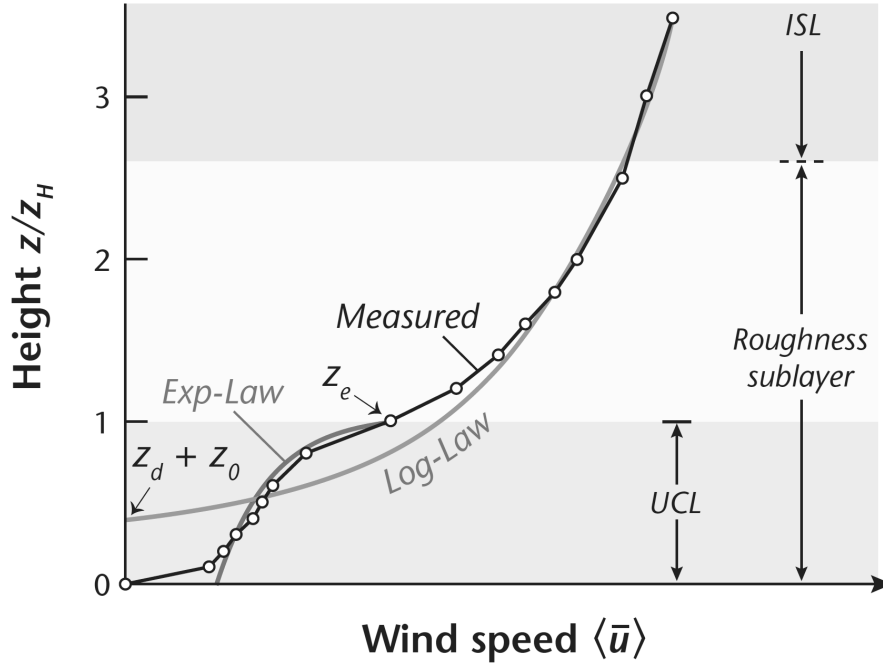


Figure 3.3: Wind profile in the RSL (Roughness sublayer) and ISL (Inertial sublayer) measured in a wind tunnel over an array of cubes ($\lambda_f = 0.16$). Values are horizontal spatial averages. z_h is the average building height (called h in this chapter). It can be seen that in this case the Log-law matches the measurements approximately when $z/h \geq 1.5$. Modified from [59]

This law is generally valid in rural areas (within the Rural Boundary Layer), and in the ISL when in urban areas. However, it is shown [59] that the log-law often provides a good estimation starting approximately from a height of $1.5h$, where h is the mean height of buildings, even though it is still located within the RSL, as shown from Figure 3.3; this approximation will notably be used in chapter 4.

One has to mention that the friction velocity u^* is a parameter related to turbulence. The ISL property stating that turbulent fluxes show small variations with height signifies that the friction velocity is considered constant with height. It is therefore possible, in one particular location, to link wind speed values at two different heights. To do so, (i) Eq. 3.2 is evaluated at these two different heights z_1 and z_2 , which gives two equations; (ii) by taking the ratio of the two resulting equations, it allows to make u^* and κ vanish, and express the ratio of $u(z_1)/u(z_2)$ as a function of z_1 , z_2 , z_d and z_0 (as used in section 4.3.2 of chapter 4).

3.1.4 Rural wind characteristics

In rural areas, the presented log-law can be used to model the vertical behavior of wind speed. The two parameters of the log-law (roughness length and displacement height) are often computed based on land use data, defining the different types of terrain and obstacles present in a particular location. Notably, tables values for z_0 are suggested by [60] based on land cover types, as shown in Table 3.1. In Switzerland, the CORINE database (described in annex Table A.2) can be used as to assess the land cover types in the country, and therefore the corresponding roughness length values. The displacement height, on the other hand, is mainly related to the height of trees within rural areas, or other related obstacles. It

Table 3.1: Roughness length ($z_{0,u}$) values table proposed for the Corine Land Cover (CLC) classes, as described in [60].

CLC classes	CLC codes	CLC roughness	
		Range	Most likely value
Continuous urban fabric	111	1.1 - 1.3	1.2
Broad-leaved forest; Coniferous forest; Mixed forest	311;312;313	0.6 - 1.2	0.75
Green urban areas; Transitional woodland; Burnt area	141;324;334	0.5 - 0.6	1.1
Discontinuous urban fabric; Construction sites; Industrial or commercial units; Sport and leisure facilities; Port areas	112;133;121; 142;123	0.3 - 0.5	0.5
Agro-forestry areas; Complex cultivation patterns; Land principally occupied by agriculture, with significant areas of natural vegetation	242;243;244	0.1 - 0.5	0.3
Annual crops associated with permanent crops; Fruit trees and berry plantations; Vineyard; Olive groves	241;221;222; 223	0.1 - 0.3	0.1
Road and rail networks and associated land; Non-irrigated arable land; Permanently irrigated; land; Rice fields; Salt marshes	122 211;212;213; 411;421	0.05 - 0.1	0.075 0.05
Sclerophyllous vegetation; Moors and heathland; Natural grassland; Pastures	321;322;323; 231	0.03 - 0.1	0.03
Dump sites; Mineral extraction sites; Airports; Bare rock; Sparsely vegetated areas	131;132;124; 332;333		0.005
Glaciers and perpetual snow	335		0.001
Peatbogs; Salines; Intertidal flats	422;412;423		0.0005
Beaches, dunes, and sand plains	331		0.0003
Water courses; Water bodies; Coastal lagoons; Estuaries; Sea and ocean	511;512;523; 522;521		0

can be for example assess based on Digital Elevation Data. It will be however discarded in the wind potential study of chapter 4, for environmental and suitability reasons discussed in section 4.3.1.

3.1.5 Urban wind characteristics

In urban areas, it is still possible to use the log-law within the ISL and above a height of $1.5 \times$ (mean building height) with a reasonable approximation. The parameters of the law, however, have to be computed to reflect the particular characteristics of the urban area. In particular, the displacement length z_d is in this case related to the height of the buildings, and the roughness length z_0 (renamed $z_{0,u}$ to specify the urban setting), also requires a computation above the urban canopy layer. Models exist to determine the two parameters based on various characteristics of buildings. The expressions given by by Macdonald et al. [61] are notably popular, and are as follows:

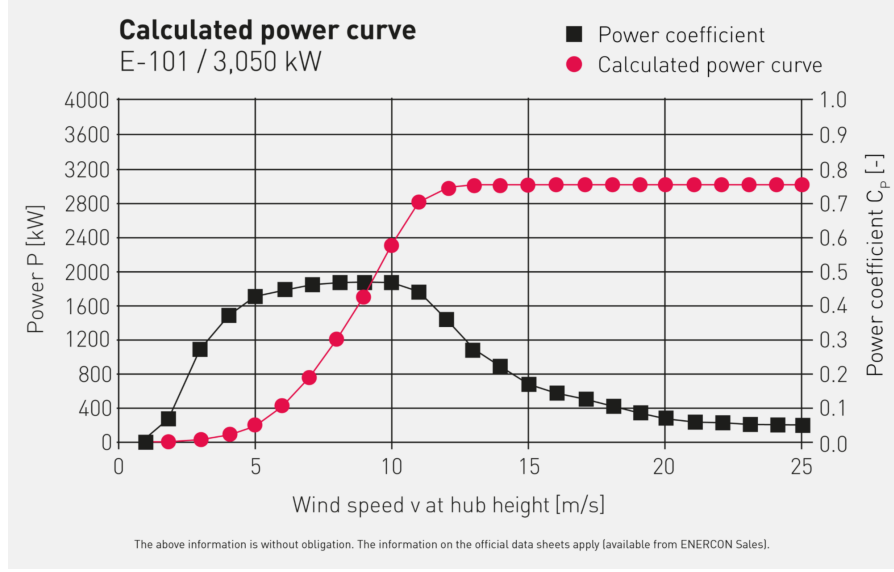


Figure 3.4: Calculated power curve for a large commercial horizontal-axis wind turbine, the ENERCON E-101. Figure taken from <https://www.enercon.de/en/products/ep-3/e-101/>.

$$\frac{z_d}{h} = 1 + A^{-\lambda_p} (\lambda_p - 1) \quad (3.3)$$

$$\frac{z_{u,0}}{h} = \left(1 - \frac{z_d}{h}\right) \exp \left(- \left[0.5\beta \frac{C_D}{\kappa^2} \left(1 - \frac{z_d}{h}\right) \lambda_f \right]^{-0.5} \right) \quad (3.4)$$

where h is the average height of buildings, λ_p the plan area ratio (the ratio of the total plan/footprint area of the surface obstacles to the total plan area), λ_f the frontal area ratio (the ratio of the total facade area of the surface obstacles to the total plan area), C_D is a drag coefficient of a single obstacle, A is an experimental coefficient and β a parameter. Note that λ_f is a function of the frontal area and is therefore subject to a particular direction we have to consider and for which we compute the facade area of obstacles. As a result, $z_{u,0}$ is also different for each considered direction. It is a common practice to consider multiple directions (for example North-South and East-West) and compute λ_f and $z_{u,0}$, and ultimately the wind speed for each of these directions.

3.1.6 Wind energy systems

As mentioned earlier, electrical power can be generated from the wind through a turbine. The efficiency of the turbine is expressed by the C_p in Eq. 3.1. As C_p is a function of the wind speed, each turbine is characterized by a C_p curve and a corresponding *power curve*; an example of such a curve is shown in Figure 3.4. Depending on the type and size of turbine, the two curves naturally vary. One of the main feature of the turbine is its *nominal power*, defined by the maximum power the turbine can produce. For example, the turbine characterized by the curves shown in Fig. 3.4 has a nominal power of 3MW.

Wind turbines are commonly separated in two types, based on their axis direction: *Horizontal Axis Wind Turbines (HAWT)* and *Vertical Axis Wind Turbines (VAWT)*. While the HAWT exist mainly in one type, multiple kinds of VAWT exist, including notably the Savonius and Darrieus types, as shown in Fig. 3.5 [62, 63]; we will not discuss them here. Instead, we will shortly review the advantages and

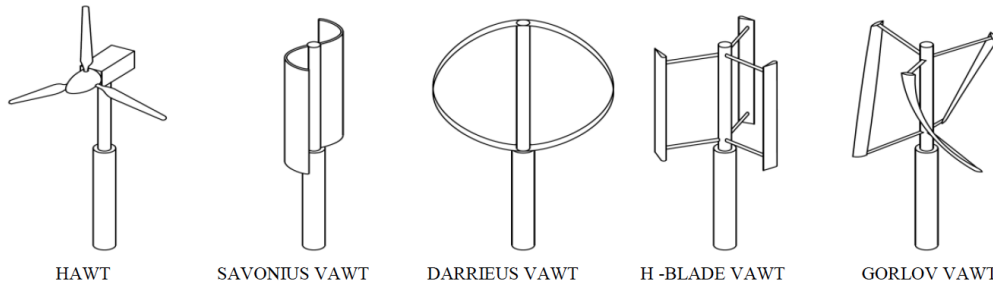


Figure 3.5: Wind turbines types. Besides the HAWT, all other turbines are VAWT. (Figure taken from [62])

disadvantages of the two types, and further discuss their use in different conditions, and in particular in rural versus urban areas. Note that a general comparison of the two systems can be found notably in [64].

The main advantages (✓) and disadvantages (♣) of HAWT and VAWT are as follows:

- Horizontal Axis Wind Turbines:
 - ✓ High efficiency. The blades are perpendicular to the wind and generate more power.
 - ✓ Because the tower of a HAWT is high, it can benefit from high altitude stronger wind.
 - ♣ Complex installation and maintenance. It is difficult to handle heavy blades, gearbox, and generator; it needs a massive tower their support.
 - ♣ High visibility. It can disrupt the appearance of the landscape.
 - ♣ Do not handle turbulence well. HAWT work best with a steady wind behavior.
 - ♣ Do not handle high winds well. HAWT require a braking or yawn device in high winds to stop the turbine from being damaged.
 - ♣ Do not benefit from multiple wind directions. HAWT may require a system to rotate the rotor axis.
- Vertical Axis Wind Turbines:
 - ✓ Benefit from the wind in every direction and are therefore more robust to change of wind behavior. In particular, no yawn system is required.
 - ✓ Are located closer to the ground; meaning an easier maintenance.
 - ✓ Benefit from lower startup wind speeds.
 - ✓ Can be built where taller structures are prohibited.
 - ✓ Are less noisy.
 - ♣ Lower efficiency. Their blades imply additional drag, which result in a lower generated power compared to HAWT.
 - ♣ Do not benefit from high altitude wind.
 - ♣ Need a “starting” system in most cases (e.g. Darrieus VAWT).

As a result, VAWT have a lot of practical advantages. A few studies attempted to rehabilitate their use [64], pleading for these multiple advantages. Nevertheless, the majority of research dedicated to wind turbines has been focusing on HAWT, and in practice, HAWT still benefit from a much larger popularity, given their much higher efficiency. A great majority of large commercial turbines installed in Europe, and particularly in Switzerland, are indeed HAWT (as observed from the Swiss wind atlas https://www.uvek-gis.admin.ch/BFE/storymaps/EE_WEA/index.php?lang=en). Besides their overall popularity, it is significant to differentiate the particular type of turbine suitable for installation within rural and urban areas, specially within the framework of a potential study. For rural area, where large scale commercial turbines are dominating, it is clear HAWT are more suitable; for urban areas, however, the use of HAWT or VAWT is debated in the literature.

Although small wind turbines are significantly less efficient than large ones, wind turbines have been recently seriously considered as an alternative energy source for urban areas, both VAWT and HAWT [63, 65, 66]. Recent interest, however, have been shown for VAWT for small/micro turbines installed over buildings, since their multiple advantages (as discussed earlier in the section) make them attractive for this use [64, 67]. Their ability to handle slow and turbulent wind, in particular, as well as changes in wind direction, could result in their outperforming HAWT in urban areas [68]. Many studies have therefore considered the integration of small VAWT in urban areas, above buildings notably [69, 70]. Yet, small wind turbines installed over buildings are still in majority HAWT, notably in the U.K. [66, 71], one of the largest users of wind energy in Europe. In addition, when HAWT and VAWT are compared within a potential study with respect to performance, it appears that HAWT yield better results [72]. Despite their numerous practical advantages in urban areas, further developments are therefore needed to make them viable for large scale deployment [73]. On the other hand, small scale HAWT have been already considered for a relatively large scale deployment, notably within the framework of several projects, including the SWIP project (<http://swipproject.eu>).

In the scope of the present thesis, notably for chapter 4, we will therefore consider HAWT, for both rural (with large-scale commercial turbines) and urban areas (with traditional small scale HAWT). In practice, another problem related to the consideration of VAWT (besides the issues mentioned previously) is their C_p curve. Indeed, it is often given with respect to the blade tip speed ratio rather than wind speed, which makes it difficult to consider C_p when assessing the potential power extracted by the turbines within a potential study. In chapter 4, typical C_p /wind speed curves will be considered, for both small (for urban areas) and large (for rural areas) scale HAWT. Notably, the general shape of C_p curves are similar for both sizes, but their respective maximum are different, reaching typically 0.2 to 0.3 for small turbines, instead of 0.45 for large ones [63, 74, 75].

3.2 Shallow geothermal energy modeling

Geothermal energy originally consists of the available heat coming from the earth's core, where the decay of radioactive elements occurs. It can be captured in many different ways, using various installations. The crucial point, however, is the depth at which one decides to extract heat from. It is naturally related to the scale of the installation, greatly varying in power capacity. On the one hand, deep geothermal boreholes (≥ 500 m depth) look for high temperature resources, and can provide tremendous amounts of energy (both for heating and electricity), but are naturally not suitable for

small decentralized systems. On the other hand, shallow (depth ≤ 500 m) and very shallow (depth ≤ 10 m) geothermal resources are characterized by lower temperature profiles and are more suitable for small scale and single homes installations. Ground source heat pumps (GSHPs), perhaps the most popular shallow geothermal system, and the chosen system in this thesis, is in the latter category and can be installed at shallow or very shallow depths. Also note that GSHPs are particularly popular in Switzerland and have known a drastic increase in number of installations, with notably a total number of 290'000 heat pumps installed in 2017 [6]. This increase was notably supported by an attractive feed-in tariff instituted in Switzerland (0.30 to 0.40 CHF/kWh depending on the installed power).

The potential assessments for high or low temperature geothermal resources are different in nature, because of their heat usage. While deep resources consist in very high temperature available heat to be used directly (through an energy system) or to generate power, shallow resources act like low temperature heat reservoirs from which heat can be extracted for heating or deposited for cooling. This is possible because of the temperature of the subsurface, consistent below a depth of 10 to 20 m (4° to 13° for most places [76]). In the case of very shallow installations, the heat can be extracted from the surface which is in direct contact to solar radiation. The temperature at these low depths is therefore not constant throughout the year and the efficiency of very shallow GSHPs is in general lower than traditional shallow GSHPs. They can, however, offer an easy installable, less expensive and still efficient solution, particularly with recently developed very shallow systems (section 3.2.6).

Consequently, the energy available from the development of GSHPs is available almost everywhere, and the main challenge in assessing their potential in specific locations is to estimate the local thermal characteristics of the ground. These characteristics include notably the ground temperature gradient, the presence and properties of local ground water, and significant ground thermal physical properties (thermal conductivity, heat capacity).

We will present concepts, models and strategies useful to extract the mentioned significant ground variables, and particularly how it was performed in this thesis. We also shortly present the current, existing systems allowing the extraction of the shallow geothermal potential, and discuss the significance of very shallow geothermal installations as efficient renewable energy systems. Note that, within the framework of this thesis, we focus on the theoretical potential of geothermal energy, and therefore do not discuss system design (length of pipes for the installations etc.) and heat modeling which would be required in order to estimate the corresponding technical potential [77].

The notions presented in this section will be applied in chapter 5.

3.2.1 Significant ground thermal variables

We focus here on the fundamental ground thermal (or if not possible, electrical) properties of the ground, which define the theoretical shallow geothermal potential. (The presence of ground water, notably, is an important variable - discussed in chapter 5 - yet not the subject of models presented in this section.) Since these ground properties can be defined in various forms, and for clarity of units and definitions, let us present briefly all electrical and thermal variables of interest:

- **Electrical Resistivity** (ρ) [$\Omega \cdot m$]: measures the ability of a material to oppose a flow of electrical current.

- **Thermal Resistivity** (ρ_t) [K.m.W^{-1}]: measures the ability of a material to oppose the flow of heat.
- **Thermal conductivity** (λ) [$\text{W.K}^{-1}.\text{m}^{-1}$]: measures the ability of a material to conduct heat. It is the inverse of thermal resistivity. Naturally, the electrical conductivity can also be defined as the ability of a material to conduct electricity, or the inverse of electrical resistivity.
- **Volumetric Heat Capacity** (c_v) [$\text{J.m}^{-3}.\text{K}^{-1}$]: measures the ability of a given volume of a material to store internal energy while experiencing a given temperature change, yet without a phase transition. Note that the specific heat capacity is also used, for which the storage ability is measured for a given mass rather than a given volume.
- **Thermal Diffusivity** (α) [$\text{m}^2.\text{s}^{-1}$]: measures the rate of transfer of heat within a material. It is by definition the ratio of the thermal conductivity to the volumetric heat capacity: $\alpha = \frac{\lambda}{c_v}$.
- **Ground temperature** [T] [K]: temperature of the ground at various depths, defining the temperature gradient, meaning the change of temperature with depth.

The traditional key variables of interest considered in a theoretical geothermal potential study (besides groundwater conditions) are the thermal conductivity λ , the volumetric heat capacity c_v and the ground temperature T. Note, however, that (i) the conductivity is easily obtainable from the resistivity, (ii) the heat capacity can be obtained from the knowledge of the conductivity and the diffusivity, and (iii) thermal properties should be extractable from the electrical properties, based on the specific soil conditions. The geothermal potential is therefore equivalently extracted from the electrical resistivity ρ , thermal diffusivity α and ground temperature T. Based on the availability of data in Switzerland, the three latter variables are easier to estimate and are therefore the focus of the potential estimation in chapter 5. As a result, they are also the focus of the models and strategies to be discussed in this section. We will first present soil structure and texture notions impacting on these variables, then present modeling strategies allowing to respectively extract electrical resistivity and thermal diffusivity values from various data, and finally offer a discussion on the link between electrical and thermal properties of the ground, which will be useful in the thesis to achieve the conversion of electrical resistivity into thermal resistivity values).

3.2.2 Elements of soil structure and texture

The configuration of the soil in a particular location has a large impact on the thermal and electrical properties of the ground. In particular, the soil structure and texture, which define two very important characteristics to differentiate multiple types of soil, are variables of interest for the geothermal energy potential (and will therefore be used in chapter 5).

The soil structure, on the one hand, consists of the arrangement of solid, liquid and void parts within the soil. Figure 3.6a shows an illustration of a soil structure example, including the used variables to describe the different volumes: V_a , V_w , V_s , V_v and V_T are respectively the volume of air, water, solid soil, void, and total volume. The void volume can be filled with air and water, so that $V_v = V_a + V_w$, and the total volume is given by $V_T = V_v + V_s = V_a + V_w + V_s$. The masses of the different parts are also of use and are noted M_s and M_w respectively for the mass of solid soil and the mass of water.

Various soil structure quantities are often used to describe the amount of water or air within the soil, including:

- Volumetric Water Content (VWC) $:= V_w/V_T$
- Gravimetric Water Content (GWC or w) $:= M_w/M_s$
- Porosity (n_p) $:= V_v/V_T = e/(1 + e)$
- Void ratio (e) $:= V_v/V_s = n/(1 - n)$
- Saturation degree (S_r) $:= V_w/V_v$
- Particle density (γ_s) $:= M_s/V_s$
- Dry (bulk) density (γ_d) $:= M_s/V_T$
- Water density (γ_w) $:= M_w/V_w \approx 1\text{g/cm}^3$

Note that formulas can be derived to link some of these quantities, using their respective definitions. One of these formulas, expressing VWC as a function of the GWC, will notably be used in chapter 5) of the thesis. It states that:

$$\text{VWC} = \text{GWC} \frac{\gamma_d}{\gamma_w} \quad (3.5)$$

The soil texture, on the other hand, differentiates soil types based on the repartition of minerals given their particle size, defined by their diameter (\varnothing). Often, the very coarse minerals with $\varnothing > 2\text{mm}$ (block, rocks and gravels) are excluded from the texture classification and the soil texture is defined by the respective percentage of the three fine soil elements, namely *sands* ($50\mu\text{m} < \varnothing < 2\text{mm}$), *silts* ($2\mu\text{m} < \varnothing < 50\mu\text{m}$) and *clays* ($\varnothing < 2\mu\text{m}$). Note that the percentages of sand, silt and clay are often given independently of the coarse minerals, meaning that the sum of the three percentages is 100%. In order to create a finite set of typical soil textures, soil texture classes can be extracted. One of the most common classification is the one created by the USDA, as shown in Figure 3.6b, which is followed by the American Society of Agronomy and used in the United States [78]. In addition to fine soil classes, another variable is often used in soil-related studies to specify the portion of coarse minerals: F, which is defined by the percentage sum of the sand and gravel fractions (where the sand portion is computed over all the possible soil classes, not only within fine soils like it is the case in the USDA classification)

3.2.3 Inversion of Vertical Electrical Soundings for resistivity estimation

Although it would be desirable to obtain thermal conductivity information from real data, it is rather challenging to perform in-situ measurements for this variable (even though some strategies exist, for examples suggested by [80]), and particularly for many locations. It is relatively straight forward, however, to conduct an electrical resistivity study, using a simple setup which induces a flow of electrical current in the ground. This is known as Vertical Electrical Sounding (VES).

VES is a classical geophysical method aiming at estimating the electrical resistivity or conductivity of a medium. As it is one of the oldest and trusted methods for extracting resistivity values and one of the least expensive to perform per unit depth, it is a very commonly conducted type of study.

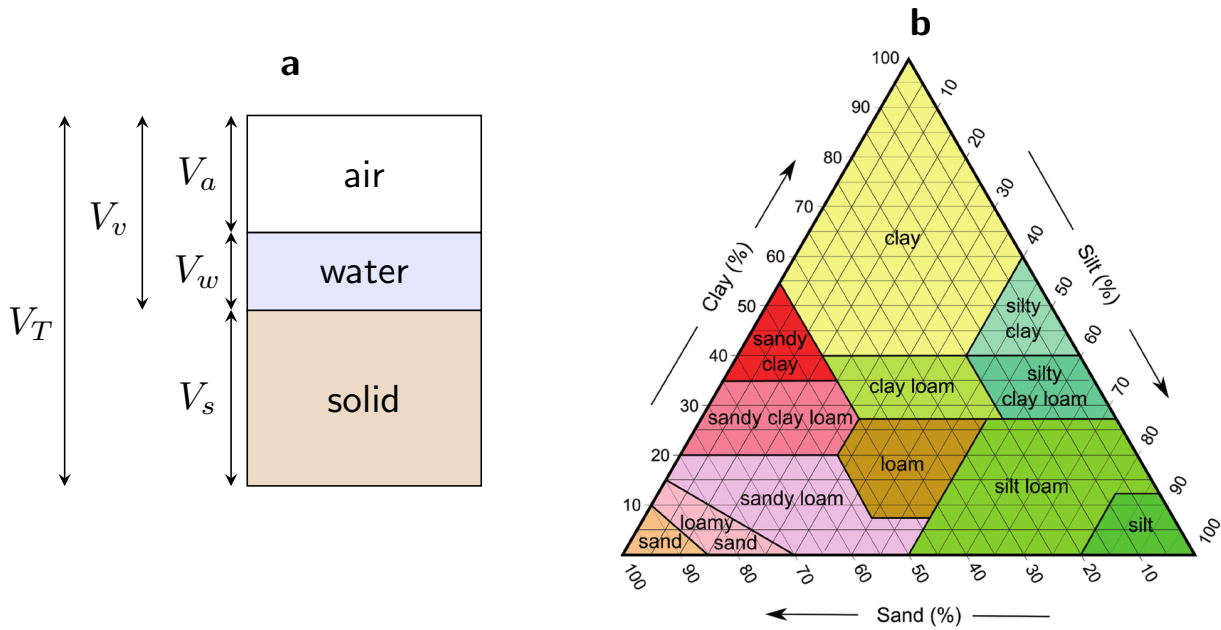


Figure 3.6: **(a)** An example of soil structure arrangement. The soil structure is split between solid soil, water, and air. V_a , V_w , V_s , V_v and V_T are respectively the volume of air, water, solid soil, void ($V_v = V_a + V_w$), and total volume. **(b)** A soil texture diagram-soil types according to their clay, silt and sand composition, as used by the USDA, redrawn from the USDA webpage: <http://soils.usda.gov/education/resources/lessons/texture/>. Source: [79] (Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/A-soil-texture-diagram-soil-types-according-to-their-clay-silt-and-sand-composition-as_fig2_235884102 [accessed 5 Dec, 2018].)

It is based on the measurement of the voltage between grounded electrodes that are installed at multiples distances from each other in order to reach various depths of the ground. The apparent electrical resistivity/conductivity is then given by Ohm's law as a function of the induced current, the measured voltage, and the geometry of the installation setup.

Although there exists multiple configurations for the implementation of a VES study, the geometry of the electrode array can be generalized in order to extract generic formulas. A generalized form of array is shown in Figure 3.7. Following the figure, a current is induced between points A and B, and the difference of potential δV is measured between points M and N. The value of δV and therefore the resistivity depends on the distance between A and B and the larger this distance the deeper in the ground the current flows between the two electrodes. As a result, the AB distance is gradually increased in practice in order to extract the apparent resistivity at increasing depths of the ground. The list of AB distance (or AB/2) with the corresponding measured apparent resistivities is what is often called a VES curve, or in the chapter 5 VES data.

VES curves are said to be interpreted (or inversed) when the depth and corresponding resistivity of the different ground strata are extracted from the AB and apparent resistivity measurements. There are multiple methods to interpret VES curves, including fitting simple known curve shapes, graphical modeling or numerical modeling. The latter is, however, the most rigorous and up to date general method. Many algorithms are available for automatic inversion of VES curves. In the present study, a 1D inversion function from a C++/python library (pyGIMLi [81]) is used. pyGIMLi generally uses

regularization methods to perform inversion, with different schemes, including the popular Marquardt scheme. For more details on pyGIMLi and details about the inversion algorithms, please see [81].

For more details on Vertical Electrical Sounding studies and geophysical exploration methods in general, the reader is invited to see [82].

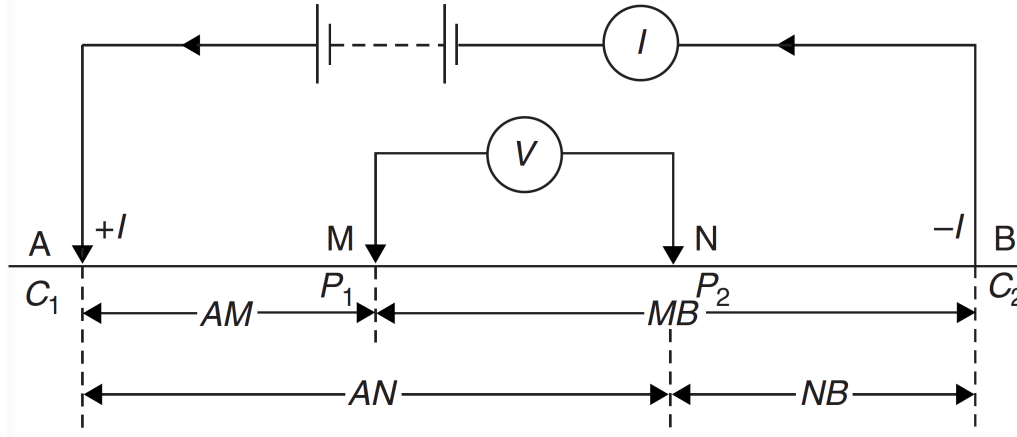


Figure 3.7: Generalized form of electrode configuration in VES resistivity surveys. Source: [82].

3.2.4 Fourier analysis of temperature data for diffusivity estimation

The second variable of interest, namely the thermal diffusivity, is also challenging to measure in practice. It is, however, closely related to another ground variable which is easily measurable: the ground temperature. Indeed, the thermal diffusivity is a crucial parameter of the classical heat equation, for which we have a good understanding of the solution. As a result, multiple methods have been used to estimate the apparent thermal diffusivity using ground temperature time series. The simplest ones include the amplitude and phase equations, modeling the ground temperature as a pure harmonic function (using only the fundamental frequency of the Fourier series) [83, 84] and the arctangent equation (using the first two Fourier frequencies to model the temperature) [83]. Numerical modeling has also been used to extract the diffusivity [83]. Several other studies used a full Fourier series to model the variations of the ground temperature. These latter studies consider the dominant frequencies and estimate the gradients of the amplitudes or phase vs. depth [85–87]. It appears that this last method give the best estimations [83]. It is therefore this strategy used in chapter 5, and the one we shortly present here.

As previously mentioned, the key to the link between diffusivity and temperature is the heat equation. At very shallow depths ($< 10\text{m}$ of depth), the average daily temperature can indeed be described by a heat conduction model, due to its annual cyclic variations. While the temperature generally varies during the day, the average daily temperature at one location shows a period of one year, and can therefore be modeled using a 1D heat equation. The general (3D) heat equation allows to describe the behavior of ground temperature $T(x, y, z, t)$ as a function of the time variable t and space variables (x, y, z) :

$$\frac{\partial T}{\partial t} - \alpha \nabla^2 T = 0 \quad (3.6)$$

where ∇^2 is the laplace operator and α is the ground thermal diffusivity. The thermal diffusivity is in general also a function of space and time. In the assumption of uniform physical properties of the soil, however, both variables can be considered constant. In addition, we are here interested in the variations of the daily average temperature, only with depth, at a specific location (where measurements were conducted). We therefore use the 1D version of the heat equation, along the z component only. Considering the two mentioned assumptions, the 1D heat equation gives, at any fixed location (x, y) :

$$\frac{\partial T(z, t)}{\partial t} = \alpha \frac{\partial^2 T(z, t)}{\partial z^2} \quad (3.7)$$

where $T(z, t)$ is the soil temperature at depth z and time t and $\alpha = \frac{\lambda}{c_v}$ the apparent thermal diffusivity, also considered constant by the uniformity assumption, throughout the year. The apparent thermal diffusivity is representative of the ground at one location, and controls the change of temperature near the surface, independently of time and space.

Following the assumptions that: (i) the ground surface temperature (boundary condition at $z = 0$) is sinusoidal, (ii) the ground temperature is constant at infinite depth and equal to the average ground temperature, and (iii) the apparent thermal diffusivity is, as mentioned, constant with depth and throughout the year; the solution of Eq. 3.7 can be given by a Fourier series, defining an infinite sum of harmonic functions which can be fit to experimental data. The boundary condition (ground surface temperature $T(0, t)$), in particular, can be captured by the following Eq. 3.8 [85]:

$$T(0, t) = T_0 + \sum_{n=1}^{\infty} T_{Sn} \sin(n\omega t + C_n) \quad (3.8)$$

where T_0 is the average ground surface temperature over a year (the period of the Fourier series), T_{Sn} and C_n are the amplitude and phase of the harmonics defined by n , and $\omega = \frac{2\pi}{P}$ is the angular frequency of one period P (a year, or 365.24 days). The general Fourier solution of Eq. 3.7 is then given by [85]:

$$T(z, t) = T_{0,z} + \sum_{n=1}^{\infty} R_n(z) \sin(n\omega t + \phi_n(z)) \quad (3.9)$$

where

$$R_n(z) = T_{Sn} \exp\left(\frac{-z\sqrt{n}}{d}\right) \quad \text{and} \quad \phi_n = \frac{-z\sqrt{n}}{d} + C_n \quad (3.10)$$

and $T_{0,z}$ is the average value of $T(z, t)$ over one year period (and also the constant of the Fourier series, usually noted c_0) and D is damping depth, which traduces the decrease of the temperature amplitude when the depth increases [86, 88]. D is given by:

$$D = \sqrt{\frac{2\alpha}{\omega}} \quad (3.11)$$

The coefficients $\phi_n(z)$ and $R_n(z)$ are the phase and amplitude of the harmonics of the solution given by Eq. 3.9 and can therefore be computed using Fourier analysis based on a set of temperature measurements at various depths and times [85]. In particular, they are given by the classical expressions of Fourier coefficients $a_n(z)$ and $b_n(z)$:

$$R_n(z) = \sqrt{a_n^2(z) + b_n^2(z)} \quad (3.12)$$

$$\phi_n(z) = \arctan\left(\frac{a_n(z)}{b_n(z)}\right) \quad (3.13)$$

where

$$a_n(z) = \frac{\omega}{\pi} \int_0^{\frac{2\pi}{\omega}} T(z, t) \cos(n\omega t) dt \quad (3.14)$$

and

$$b_n(z) = \frac{\omega}{\pi} \int_0^{\frac{2\pi}{\omega}} T(z, t) \sin(n\omega t) dt \quad (3.15)$$

In practice, in order to use the discrete temperature data efficiently, $R_n(z)$ and $\phi_n(z)$ are computed with the Fast Fourier Transform (FFT). The FFT approximates the Fourier series by computing N frequencies rather than an infinity of frequencies. The Nyquist sampling theorem gives the value for N as $N = \frac{N_s}{2}$, where $N_s = 365$ (days) is the number of sampling points (one sample temperature value for each day). In this thesis, the FFT is performed using the implementation from the NumPy Python library,

The damping depth D and therefore the apparent thermal diffusivity can be computed using Eq. 3.10 together with the previously estimated values for $R_n(z)$ and $\phi_n(z)$. An exact computation, however, requires another Fourier analysis to extract T_{sn} and C_n from Eq. 3.8. In practice, it has been shown that the slope of $\ln(R_n)$ vs. $z\sqrt{n}$ curves provides a reliable estimate of the damping depth [86, 87], which allows for an easier estimation of the apparent thermal diffusivity α ; that latter strategy is therefore the one used in this thesis. Note that α can be computed based on multiple chosen harmonics n , and for every year offered by the data. In the framework of the present thesis, the use of the harmonics and the aggregation of the estimates obtained for each year are detailed in chapter 5 within the section of interest.

3.2.5 Converting electrical to thermal properties

The conversion of electrical resistivity into thermal resistivity is not trivial and requires information on the texture and the structure of the soil. In particular, multiple studies have shown that the most important parameters to describe the relationship between electrical and thermal ground properties are the particle size, water content, dry density and saturation [89–91] (all presented in section 3.2.2).

These studies also developed models for the conversion. Sreedee et al. [89] suggested a parametric model to estimate the thermal resistivity ρ_t based on electrical resistivity ρ and percentage sum of the sand and gravel fractions F :

$$\log(\rho) = K_R \times \log(\rho_t) \quad (3.16)$$

where K_R is a constant defined by:

$$K_R = 1.34 + 0.0085 \times F \quad (3.17)$$

This model was then improved by including the saturation degree S_r [90] in the K_R constant, which is re-defined as follows:

$$K_R = X + Y \exp(-S_r \times Z) \quad (3.18)$$

where

$$X = [1.1 + 0.01 \times F] \quad (3.19)$$

$$Y = [0.9 - 0.01 \times F] \quad (3.20)$$

$$Z = [0.02 + 0.0006 \times e^{F/25}] \quad (3.21)$$

In order to extract more complex interdependencies between the soil variables, Erzin et al. [91] attempted a non-parametric approach by training artificial neural networks based on similar variables than the ones used by the two studies by Sreedeeep et al. It ultimately yielded better results than the latter studies.

In the framework of this thesis, the discussed models cannot not be used, as the saturation degree, one of the most important parameters, is not available over the Swiss territory. Instead, and in the light of the good results achieved by Erzin et al. [91] the strategy suggested in chapter 5 is to extract our own non-parametric conversion model. Using the data collected by the mentioned studies, a conversion model (using Random Forests) is trained to predict the thermal resistivity from the electrical resistivity based on similar soil-related variables which are available in Switzerland. Details of this model are given within the section of interest in chapter 5.

3.2.6 Shallow geothermal energy systems

There are multiple systems to extract the shallow geothermal energy available the ground. Note that there are two ways to use shallow thermal energy: ground source heat pumps (GSHPs) and *underground thermal storage*. The latter one, used to store energy throughout the year, is not tackled here, and we focus on GSHPs.

Multiple types of GSHP exist, with different characteristics, relative notably to the installation, the maintenance requirements, and the performance. The performance is the ability of a GSHP to provide heating energy to a house of a flat (the warm source) by pumping it from the environment (the cold source), relatively to the energy needed operation. This feature is usually expressed by the *Coefficient of Performance* (COP):

$$\text{COP} = \frac{\text{Heating energy delivered by the heat pump to the warm source}}{\text{Energy needed to operate the heat pump}} \quad (3.22)$$

The energy needed in the form of electricity (to feed a compressor) is rather small, which ultimately lead to a high COP for a typical GSHP, generally ranging from 2 to 5.

In practice, two general types of GHSPs can be defined, namely open and closed systems, which can further be separated in multiple categories of systems [92]:

- *Open systems* extract heat from groundwater using two or more separate wells (one for extracting the groundwater, one for re-injecting the groundwater). They are sometimes called *open loop* systems, or simply *ground water wells*. Their main advantage is their very high performance (around eight times more than closed systems [77]). Nevertheless, their need for an aquifer restrains the possible locations of installation and cause a high maintenance cost and many environmental restrictions.
- *Closed systems*, on the other hand, extract heat directly from the ground soil using heat exchangers. The heat itself is carried by a fluid circulating within the exchanger, and then passed to a heat pump connected to the heating system of a building. Their advantage is that they can be installed almost anywhere and with relatively low maintenance needs. Closed systems can be traditionally divided into two categories:
 - Vertical installations. The heat is extracted by *Borehole Heat Exchangers (BHE)*, which are mainly built from plastic pipes fixed within a borehole with grout. The fluid circulates within the pipes, which are usually installed at a depth of 100 to 250m. BHEs take advantage of the undisturbed temperature of the ground at this depth, and are not subject to meteorological changes. Different types and arrangements of pipes exist in the market.
 - Horizontal installations. Traditional horizontal systems, called *horizontal collectors*, are installed near the surface of the ground at a depth of 1 to 4m, and extract heat stored in the surface directly exposed to the sun. As opposed to BHEs, the performance of these systems is therefore driven by meteorological conditions (solar radiation, air temperature, precipitation, etc), together with near-ground thermal characteristics. They can be installed in various ways (series, parallel, spiral). Recently, *Slinky horizontal collectors*, flattened and overlapped circular coiled types of horizontal collectors became more popular because of their higher heat extraction rate.

A principle scheme of the operation of closed GSHP is shown in Figure 3.8.

Unless the local soil shows good aquifer conditions, closed systems are in general preferred to open systems. The main choice is then the one offered by the possible closed systems. While traditional Borehole Heat Exchangers have shown they can be of great use for urban settings and single family houses given their very high COP (typically 3 to 5) [93], very shallow geothermal systems (VSGs) such as horizontal and Slinky collectors can often offer a viable alternative solution. Despite their generally lower COP (typically 1 to 3), VSGs offer multiple advantages when compared to BHEs, including [94]: (i) easy maintenance, (ii) low-cost installation, (iii) fewer legal constraints than for deeper installations, (iv) possibility for technical improvements, (v) a potential for installation almost everywhere. For the heat to be easily replenished around the collector, however, steady groundwater flow in the surface layer (normally soil or sediment) is the ideal condition. Thus, water-saturated surface layers offer much better heat sources for the heat-pump collectors than dry and non-cohesive soils and sediments such as sand.

There have been recent efforts to develop, in addition to traditional horizontal loops or heat collectors, new VSGs using less space and with increased performance such as Slinky loops. These include notably *helical heat exchangers*, *stacked tubes*, and *geothermal baskets*. The latter geothermal baskets are particularly interesting since they show relatively stable temperature levels even at the

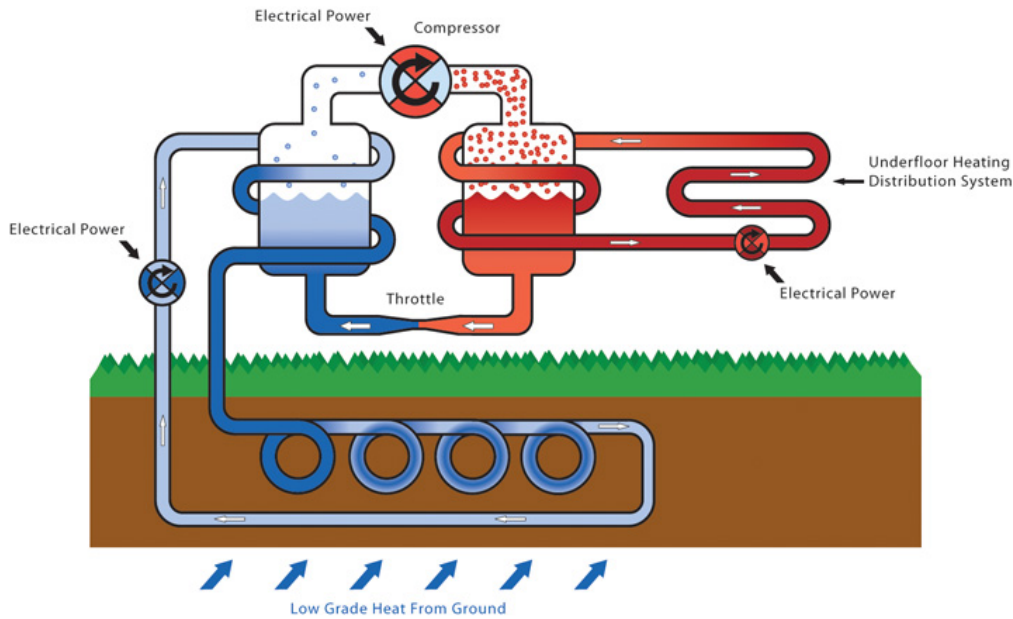


Figure 3.8: Principle scheme of a ground source heat pump. Source: <http://nialls.co.uk/ground-source-heat-pumps/>

extremely shallow depths of 1-2m [95]. Several studies have focused on the main factors influencing their efficiency [94] and to provide design guidance and models for multiple VSGs [96–101]. Thus, when the heat potential exists at a very shallow depth, it is now possible to model and size appropriate VSGs at any location. It is necessary, however, to conduct a large study of the thermal characteristics of the ground/surface layers at such depth to acknowledge adequate locations for such systems.

This recent interest for VSGs is one of the main motivations behind the focus on very shallow geothermal potential (vSGP) in the present thesis (chapter 5). Some of the common discussed systems are illustrated in Figure 3.9, summarizing the main types of GSHPs.

3.3 Solar energy modeling

Solar energy, produced by the sun in the form of electromagnetic radiation, is undoubtedly one of the most socially accepted and perhaps the most popular sustainable source of energy. Indeed, Solar panels, particularly PhotoVoltaic ones, have seen their popularity rising over the last decade and PV power plants are being built all over the world at a rapid rate [102]. This popularity is partly due to the fact that solar energy may very well be the most promising clean energy we have access to, given its continuous and almost unlimited supply. Physicists naturally had that intuition early on and solar energy has been (and still is) studied extensively in a desire to model its power and ultimately manage to use it in the most efficient way possible. We will present in this section the main variables expressing solar energy in its theoretical potential and some of the models developed to define these variables in multiple geometrical cases.

Before beginning the section, let us add a short terminology note. The raw solar energy delivered by the sun is generally referred to as the *solar radiation*. There are, however, multiple related terms

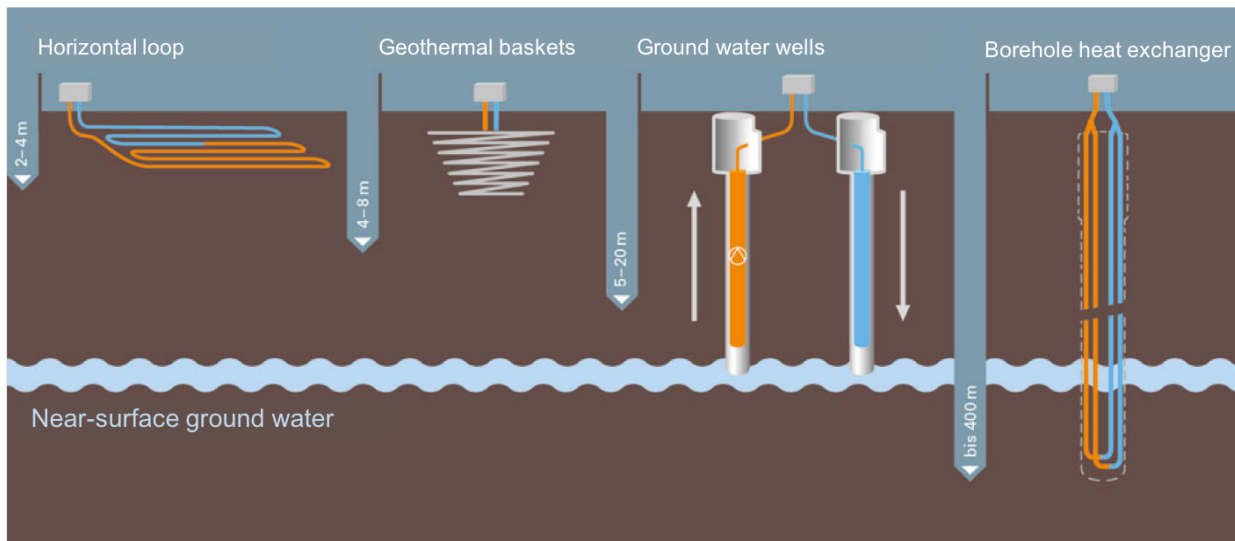


Figure 3.9: Multiple types of shallow geothermal technologies. Modified from https://www.energieatlas.bayern.de/thema_geothermie/oberflaeche/nutzung.html

used in the literature. In fact, two different measurements of solar power are commonly used: solar *irradiance* and solar *irradiation/radiation*. The difference between these two lies in their physical status. Solar irradiance is the solar power, a rate of energy per unit of time and area, usually given in W/m^2 . Solar irradiation, also called more simply radiation is the energy itself over a specific area, usually expressed in Wh/m^2 . Naturally, an amount of energy alone does not offer very precise information about the system without the period of time for which it was considered, received or consumed. As a result, when irradiation values are used, the period of time is usually specified. Some widely used units include kWh/day over a specific area or $\text{kWh}/\text{m}^2\text{year}$. Note also that the conversion between the two consists in a simple factor multiplication. We will use both, depending on the study of interest (when presenting models, we will use radiations).

The notions presented in this section will be applied in chapters 6 and 7.

3.3.1 Significant solar variables

The solar radiation comprises three different components when received by solar panel or a solar thermal collector, as illustrated in Figure 3.10a: the *direct* (or beam), *diffuse*, and *reflected* solar radiations. The sum of the three components is called the *global solar radiation*. An important point is the inclination of the surface on which the solar radiation is evaluated. The components of the radiation over a horizontal plane are simply called *horizontal radiations* (horizontal direct, horizontal diffuse, horizontal reflected). In case of a tilted plane (e.g. a PV panel or a rooftop), the three components of the radiation need to be re-computed to account for the tilt. They are simply called *tilted solar radiations*, and sum to the *global tilted solar radiation*. Note that, in order to express these multiple components of the solar radiation, it is necessary to define several angles to parametrize the positions of the sun and the considered tilted surface. Let us fix some definitions and notations that will serve for the rest of the chapter and other related chapters in the thesis. Regarding solar energy variables:

- G_h , G_B , and G_D are the global horizontal, direct horizontal, and diffuse horizontal radiations.

- G_t , G_{Bt} , and G_{Dt} are the global tilted, direct tilted, and diffuse tilted radiations.
- G_{Bn} is the direct (or beam) normal radiation, sometimes measured instead of the direct horizontal radiation.
- G_{oh} and G_{on} are the extra-terrestrial horizontal and normal radiations, also called the top of atmosphere radiation. It is the radiation coming from the sun before it reaches the Earth's atmosphere.
- G_{Rt} is the ground reflected radiation over a tilted plane, called the reflected tilted radiation.

Regarding solar related angles, illustrated in Figure 3.10b:

- β and γ are the *tilt* (or *slope*) and *azimuth* (direction of the perpendicular direction to the plan, in cylindrical coordinates) of the considered surface.
- θ is the *angle of incidence* on considered surface.
- θ_z , α_s and γ_s are respectively the *sun zenith angle* (angle between the vertical direction and the sun), *sun altitude angle* (angle between the horizontal direction and the sun) and the *sun azimuth angle* (horizontal position of the sun in cylindrical coordinates). Note that either the altitude or the zenith angle can be used to vertically parametrize the position of the sun. As is not necessary to use both, we will only use the zenith angle, as it is the most commonly used one.
- ϕ is the latitude of the location of the considered surface.
- δ is the earth *declination angle* at the time when the solar radiation is considered.

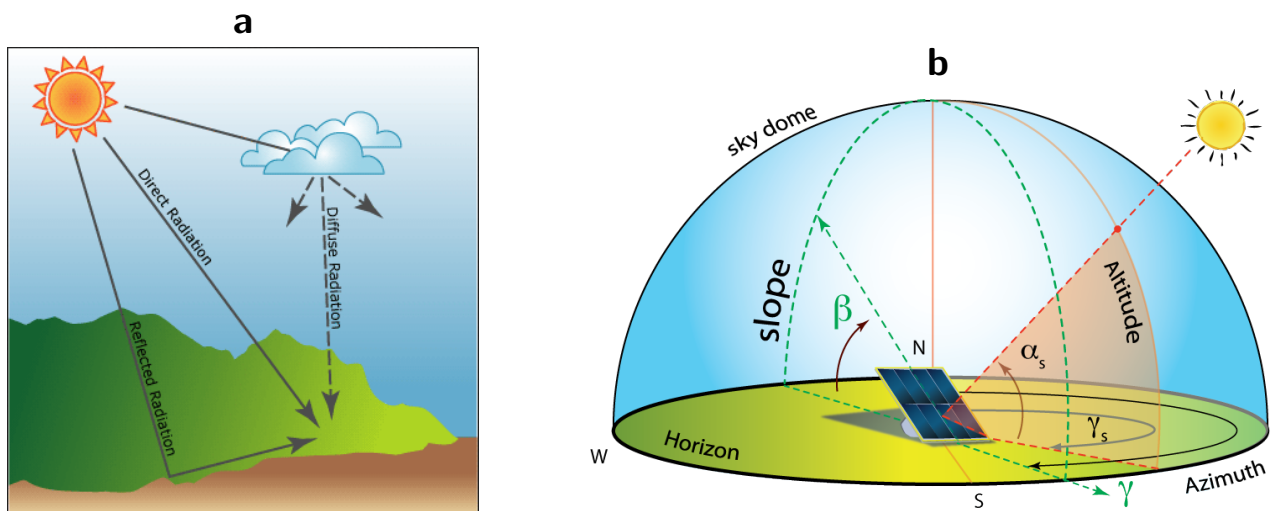


Figure 3.10: (a) Solar radiation basic components (Source: esri.com); **(b):** angles to be considered when computing the solar radiation over a tilted PV panel (source: www.urbangreenenergy.com). Note that the sun zenith angle θ_z is not defined here and the altitude angle α_s is preferred. If it was shown, θ_z would be the angle between the vertical direction line and the line linking the sun and the panel, so that $\theta_z + \alpha_s = 90^\circ$

We will first present the classical physical and empirical models to compute the three components of the solar horizontal radiation, and then further present some of the models developed for the components of the tilted solar radiation.

3.3.2 Global solar horizontal radiation models

As explained earlier, the horizontal radiation is the sum of the horizontal direct, diffuse, and reflected components of the solar radiation. The reflected component, however, is negligible over a horizontal plane, compared to the direct and diffuse radiations. Therefore, it is often discarded in the global horizontal radiation, that is given by the sum of direct and diffuse horizontal radiations:

$$G_h = G_B + G_D = G_{Bh} \cos(\theta_z) + G_D \quad (3.23)$$

where θ_z is the solar zenith angle, the angle between the zenith (vertical direction) and the sun, indicating the position of the sun. The smaller the solar zenith angle, the higher the sun is in the sky. As the sun rises, the angle gradually decreases until midday.

To obtain values for G_h , measurements are preferred to any other kind of model or approximation. However, measurement data is not always available. Consequently, parametric empirical models were developed to estimate G_h based on multiple variables that may be available, including extra-terrestrial radiation, ambient temperature, shining hours, relative humidity etc.

The first family of empirical models for global horizontal radiation is the linear family, among which the most famous is the one developed by Angstrom [103]. It expresses the ratio between G_h and G_{oh} as follows:

$$\frac{G_h}{G_{oh}} = a + b \frac{l_d}{S} \quad (3.24)$$

where l_d and S are respectively the day length and number of shining hours. The model parameters a and b are determined by fitting the model on some solar radiation data at the location of interest. This model is one of the first empirical models developed. Many more complex linear models were used in the literature, as presented for instance in [104].

A variety of non linear models can also be found in the literature for global solar radiation: polynomial, logarithmic, exponential, etc. However, one commonly used is the quadratic model [105], which simply adds a second order non linear term to the previous linear model:

$$\frac{G_h}{G_{oh}} = a + b \frac{l_d}{S} + c \left(\frac{l_d}{S} \right)^2 \quad (3.25)$$

where the variables are defined as in the previous model. Other more sophisticated models add terms to account for ambient temperature or relative humidity.

Since it is more delicate to monitor diffuse and direct horizontal radiations than it is to measure global horizontal solar radiation, models have been developed through the years to separate the diffuse and the direct from the global solar radiations. Some of these models were derived by Spitters et al. [106]. More recently, Yao et al. [107] suggested new models for this separation.

3.3.3 Solar tilted radiation models

Over a tilted plane, the global solar radiation is given by the sum of the three tilted components:

$$G_t = G_{Bt} + G_{Dt} + G_{Rt} \quad (3.26)$$

Since it can be troublesome to directly measure tilted components of the solar radiation, models were developed to derive them based on horizontal solar components values. Each of the tilted components is therefore expressed as a simple function of the horizontal components, so that the the global tilted radiation can be expressed as follows:

$$G_t = R_b G_B + R_d G_D + R_r G_h \quad (3.27)$$

where R_b , R_d and R_r are the direct, diffuse and reflected factors, defined as the ratios between the respective horizontal and tilted radiation. Various solar transposition models have been developed in the literature to compute these three ratios, in order to calculate G_t .

The direct (or beam) radiation factor R_b is analytically computable given the geometric properties of the direct radiation. It must however be treated differently depending on the temporal resolution considered for the radiation. In case of an hourly computation, R_b is given by:

$$R_b^{\text{hourly}} = \max \left(0, \frac{\cos(\theta)}{\cos(\theta_Z)} \right) \quad (3.28)$$

with

$$\cos(\theta) = \sin(\beta) \sin(\theta_Z) \cos(\gamma_s - \gamma) + \cos(\beta) \cos(\theta_Z) \quad (3.29)$$

where θ and θ_Z are respectively the angle of incidence on the tilted plane and the sun zenith angle, γ_s and γ are the sun azimuth angle and the aspect of the tilted plane (azimuth of the perpendicular direction to the plan), and β is the tilt angle of the plane. Note that if θ_Z is not directly available, it can be calculated using the following:

$$\cos(\theta_Z) = \sin(\phi) \sin(\delta) + \cos(\phi) \cos(\delta) \cos(\omega) \quad (3.30)$$

where ϕ , δ and ω are respectively the latitude, the declination of the location, and the solar hour angle, expressing the time of the day [108]. In case of daily computation, R_b is given by Klein [109], and corrected by Andersen [110]:

$$R_b^{\text{daily}} = \frac{\int_{\omega_{sr}}^{\omega_{ss}} \cos \theta(\omega) d\omega}{\int_{\omega_r}^{\omega_s} \cos \theta_Z(\omega) d\omega} = \frac{R_b^1}{R_b^2} \quad (3.31)$$

where ω_r and ω_s are the sunrise and sunset hour angle, and ω_{sr} and ω_{ss} are the sunrise and sunset hour angles on the tilted surface. The expressions for R_b^1 and R_b^2 are as follows [109]:

$$\begin{aligned}
R_b^1 = & \{ \cos(\beta) \sin(\delta) \sin(\phi) \left(\frac{\pi}{180} \right) (\omega_{ss} - \omega_{sr}) \\
& - \sin(\delta) \cos(\phi) \sin(\beta) \cos(\gamma) \left(\frac{\pi}{180} \right) (\omega_{ss} - \omega_{sr}) \\
& + \cos(\phi) \cos(\delta) \cos(\beta) (\sin \omega_{ss} - \sin \omega_{sr}) \\
& + \cos(\delta) \cos(\gamma) \sin(\phi) \sin(\beta) (\sin \omega_{ss} - \sin \omega_{sr}) \\
& + \cos(\delta) \sin(\beta) \sin(\gamma) (\cos \omega_{ss} - \cos \omega_{sr}) \}
\end{aligned} \tag{3.32}$$

and:

$$R_b^2 = 2 \left[\cos(\phi) \cos(\delta) \sin(\omega_s) + \left(\frac{\pi}{180} \right) \omega_s \sin(\phi) \sin(\delta) \right] \tag{3.33}$$

where ω_s is the sunset hour angle, ω_{sr} and ω_{ss} are the sunrise and sunset hour angles on the tilted surface, and β , ϕ and γ are respectively the surface tilt angle, the latitude, and surface azimuth angle.

The sunset and sunrise hour angles are given by the following expressions:

$$\cos \omega_s = -\tan(\phi) \tan(\delta) \tag{3.34}$$

$$\omega_{sr} = \begin{cases} -\min \left[\omega_s, \arccos \left(\frac{AB + \sqrt{A^2 - B^2 + 1}}{A^2 + 1} \right) \right], & \text{if } \gamma < 0 \\ -\min \left[\omega_s, \arccos \left(\frac{AB - \sqrt{A^2 - B^2 + 1}}{A^2 + 1} \right) \right], & \text{if } \gamma > 0 \end{cases} \tag{3.35}$$

$$\omega_{ss} = \begin{cases} \min \left[\omega_s, \arccos \left(\frac{AB - \sqrt{A^2 - B^2 + 1}}{A^2 + 1} \right) \right], & \text{if } \gamma < 0 \\ \min \left[\omega_s, \arccos \left(\frac{AB + \sqrt{A^2 - B^2 + 1}}{A^2 + 1} \right) \right], & \text{if } \gamma > 0 \end{cases} \tag{3.36}$$

where \mathcal{A} and \mathcal{B} are defined as:

$$\mathcal{A} = \frac{\cos(\phi)}{\sin(\gamma) \tan(\beta)} + \frac{\sin(\phi)}{\tan(\gamma)} \tag{3.37}$$

$$\mathcal{B} = \tan(\delta) \left[\frac{\cos(\phi)}{\tan(\gamma)} - \frac{\sin(\phi)}{\sin(\gamma) \tan(\beta)} \right] \tag{3.38}$$

and δ is the monthly declination angle and is computed using the following expression:

$$\delta = 23.45^\circ \sin \left(\frac{360(284 + m)}{365} \right) \tag{3.39}$$

where m is the recommended day to represent each month, as given by Klein et al. [109] in Table 3.2.

The above daily equation 3.31 for R_b is also suitable for monthly mean daily estimations, which are widely used in solar energy potential studies.

The diffuse radiation factor R_d suffers from the stochastic behavior of the diffuse radiation (as it depends notably on cloud presence), and is computed empirically. Therefore, many diffuse models have been suggested in the literature, and tested for various locations and under multiple conditions. It is convenient to classify the models based on their temporal resolution, but also their isotropic or anisotropic assumption. We provide a list of suggested diffuse models in Table 3.3. Since presentations of models have been made in many different studies [111–115], our list is not exhaustive, and only a selection of the most used models is proposed. As there exists a great variety of models, a few studies

Table 3.2: Recommended days from Klein et al. [109] to compute the earth declination for each month in a year.

Month	Day of the year	Date
Jan.	17	17 Jan.
Feb.	47	16 Feb.
Mar.	75	16 Mar.
Apr.	105	15 Apr.
May.	135	15 May.
Jun.	162	11 Jun.
Jul.	198	17 Jul.
Aug.	228	16 Aug.
Sep.	258	15 Sep.
Oct.	288	15 Oct.
Nov.	318	14 Nov.
Dec.	344	10 Dec.

attempted to evaluate their performance in multiple locations and for various sky conditions [111–116]. As expected, the model offering the optimal performance varies based on the particular case study. The best results, however, are often given by Reindl and Perez models. In particular, Reindl model appears to offer, for European countries, the best performance in bright overcast sky conditions [116], a rather common condition in central-western Europe. Another advantage of Reindl model is its relative simplicity compared to Perez model, and its need of fewer parameters. The Reindl model is therefore chosen as the preferred model for the solar potential study conducted in the thesis (chapters 6 and 7).

The reflected radiation factor R_r is often computed using the hypothesis that the ground reflected radiation is diffuse (meaning that the ground does not act like a specular surface, e.g. a mirror, but rather reflects the incoming radiation as Lambertian surface, e.g. following an hemisphere for the luminance), which results in the following isotropic model, given by [130]:

$$R_r = \rho \left(\frac{1 - \cos \beta}{2} \right) \quad (3.40)$$

where ρ is the ground reflectance (or *albedo*), and β is the tilt angle of the surface. This expression was only rarely challenged by anisotropic reflection models [122, 131], that ultimately were not validated enough to be fully accepted in the domain. Therefore, the isotropic expression used in Equation 3.40 is widely accepted to be a reasonable estimation of R_r .

3.3.4 Solar energy systems

There are two main active solar energy systems available to convert raw solar radiation into useful energy:

- *Solar thermal collectors*, which convert the solar energy into heat usable for space heating or domestic hot water.
- *Solar Photo Voltaic (PV) panels*, which convert solar radiation into electricity.

Table 3.3: Diffuse tilted radiation models. H, D, and H&D mean that the model is suitable respectively for hourly, daily and both hourly and daily estimations.

Model	Year	Type	Time res.	R_d
Liu & Jordan [117]	1961	Isotropic	H&D	$\frac{1+\cos\beta}{2}$
Koronakis [118]	1986	Isotropic	H&D	$\frac{2+\cos\beta}{3}$
Tian et al. [119]	2001	Isotropic	H&D	$1 - \frac{\beta}{180}$
Badescu [120]	2002	Isotropic	H&D	$\frac{3+\cos 2\beta}{4}$
Bugler [121]	1977	Anisotropic	H	$\left[\frac{1+\cos\beta}{2}\right] + 0.05 \frac{G_{Bt}}{G_D} \left(\cos\theta - \frac{1}{\cos\theta_z} \left(\frac{1+\cos\beta}{2}\right)\right)$
Temps-Coulson [122]	1977	Anisotropic	H	$\left[\frac{1+\cos\beta}{2}\right] \left[1 + \sin^3\left(\frac{\beta}{2}\right)\right] \left[1 + \cos^2\theta \sin^3\theta_z\right]$
Klucher [123]	1979	Anisotropic	H	$\left[\frac{1+\cos\beta}{2}\right] \left[1 + F' \sin^3\left(\frac{\beta}{2}\right)\right] \left[1 + F' \cos^2\theta \sin^3\theta_z\right]$ with $F' = 1 - \left(\frac{G_D}{G_h}\right)$
Perez [124]	1987	Anisotropic	H	$F_1 \frac{a}{b} + (1 - F_1) \frac{1+\cos(\beta)}{2} + F_2 \sin(\beta)$ with F_1, F_2, a, b defined in [124]
Wilmott [125]	1982	Anisotropic	H&D	$R_b \frac{G_{Bn}}{1367} + C_\varphi \frac{1367 - G_{Bn}}{1367}$ with $C_\varphi = 1.0115 - 0.20293\beta - 0.080823\beta^2$
Hay-Davies [126]	1979	Anisotropic	H&D	$AR_b + (1 - A) \left(\frac{1+\cos\beta}{2}\right)$ with $A = \frac{G_B}{G_{oh}} = \frac{G_h - G_D}{G_{oh}}$
Skartveit-Olseth [127, 128]	1986	Anisotropic	H&D	$AR_b + \Omega \cos\beta + (1 - A - \Omega) \left(\frac{1+\cos\beta}{2}\right)$ with $A = \frac{G_B}{G_{oh}} = \frac{G_h - G_D}{G_{oh}}$ and $\Omega = \max(0, [0.3 - 2A])$
Reindl [129]	1990	Anisotropic	H&D	$AR_b + (1 - A) \left[\frac{1+\cos\beta}{2}\right] \left[1 + \sqrt{\frac{G_B}{G_h}} \sin^3\left(\frac{\beta}{2}\right)\right]$ with $A = \frac{G_B}{G_{oh}} = \frac{G_h - G_D}{G_{oh}}$

These two systems are widely known and have been studied for many years. As such, this section is not meant to explain the principle of these systems or propose a list of the multiple sub-types of systems, as it has been done in many exhaustive books and articles ([132] is for example a good reference), but rather discuss their main characteristics and differences in practice, as well as their respective popularity as a result of those differences.

Solar thermal systems are based on the natural idea of using the heat directly gathered from the solar radiation. Many different types of solar thermal installations exist, with various practical characteristics and energy efficiency, including evacuated tube solar thermal collectors, flat plate solar thermal collectors, and more recently thermodynamic solar panels. Their main appeal resides in their energy efficiency (expressing the fraction of incoming solar radiation the system transforms in usable heat), which is on average around 50%, and sometimes more, e.g. up to 70% in the case of evacuated tube thermal systems. They can be installed in many different locations, notably over rooftops, similarly to a PV solar panels. Their main issue, however, is that, due to their low to

middle temperature range, they can be used solely (for most common systems) for space heating and domestic hot water production. As a result, they have greatly suffered from the comparison with PV panels. The annual installation rate of solar thermal collectors has indeed stagnated or decreased in many European countries [133] over the past few years, and shrunk dramatically in the case of Switzerland [133], suffering a decrease of 14% between 2012 and 2013.

PhotoVoltaic panels, on the other hand, attempt to use the energy from the sun radiation through its photonic form in order to transform it into electricity, usable for any desired purpose. This sole point makes PV panels a more desirable technology than thermal panels. The crucial disadvantage of PV panels is their efficiency, which is in general significantly lower than the one offered by solar thermal collectors. In the case of a PV panel, the power generated $P_{PV \text{ panel}}$ can be expressed by the following:

$$P_{PV \text{ panel}} = PR \times \eta \times A_{\text{panel}} G_{t,\text{panel}} \quad (3.41)$$

where PR is the Performance Ratio, η is the efficiency of the panel, A_{panel} is the area of the panel (in m^2) and $G_{t,\text{panel}}$ is the global radiation received by the (tilted) panel (in kWh/m^2). The efficiency η here traduces the relative fractions of solar energy transformed by the panels into electricity. The PR expresses the difference in output between the “standard test conditions” ($1000 \text{ W}/\text{m}^2$, Air Mass 1.5 spectrum, panel temperature 25°C) and the actual output of the panel. Note that in practice, varying additional factors can have an effect on the actual power generated by a panel, including converter losses etc. These factors, however, are often considered ideal within the framework of potential studies (i.e equal to 1); we therefore consider them ideal as well.

Naturally, the energy efficiency and Performance Ratio depends on the type of panel. The three main types of common PV solar panels include polycrystalline silicon, monocrystalline panels and thin film solar panels. Each type has advantages and disadvantages in terms of installation, maintenance and more importantly efficiency (the monocrystalline panels are notably the more efficient). On average, however, common values are $PR = 80\%$ and $\eta = 17\%$ [134]. Yet, considerable research is currently devoted to PV panels to make them more efficient and the average provided PV solar efficiency gradually increases; performant solar PV panels in the market are now able to reach an efficiency of 25% and recent developments showed a theoretical efficiency of up to 45% using a new PV technology [135]. Furthermore, as the efficiency increases, the average price of the PV technology is decreasing [136]. As a result, PV panels have recently seen their popularity growing very fast. The total world installed PV capacity is constantly increasing, reaching 98 GW [102] in 2017. It can be notably seen from Figure 1.1 in the introduction of the thesis (section 1.1) that solar PV is currently the most popular renewable energy source indeed, as it showed the largest addition to the global installed capacity in 2017. In Switzerland in particular, it has been growing for many years and showed a 2017 added installed capacity of solar PV of 260 MW, leading a cumulative capacity of 1924 MW [102, 137]. The feed-in tariff instituted in Switzerland since 2008 (0.158 CHF/kWh for integrated PV, as for 1 October 2017) has notably helped the growth of PV solar within the country.

When the two technologies are compared, it is clear that PV panels have a more promising future, and will most probably become one of the leading renewable energies in the future, particularly in Switzerland. It is thus quite naturally that we chose the latter as the preferred solar system in the thesis (and for the computation of the solar potential).

4

Wind energy: a theoretical potential estimation

This chapter attempts to assess the theoretical potential for wind energy over the whole Swiss territory. As presented in the introduction of the thesis, the theoretical potential is defined by the estimation of the fundamental physical variables impacting the energy resource. In the case of wind energy, these physical variables can be reduced to a sole one: the wind speed. Therefore, the theoretical wind potential is expressed by the wind speed at locations and heights of possible turbine installation. To give a better measure of the actual energy output potential, the power delivered by a turbine can also be very simply extracted from the wind speed value (along with turbine characteristics). Note that the other considered renewable energies in the present thesis, namely geothermal and solar energy, are respectively subject to multiple ground thermal characteristics and solar components. Wind energy, at the theoretical potential level, is therefore simpler to tackle for a first attempt in potential estimation.

The thorough modeling of wind behavior, however, is relatively challenging and is often tackled with CFD (Computational Fluid Dynamics) simulations, particularly within urban areas, where complex turbulent patterns are often observed. Such simulations are naturally highly computationally extensive and therefore barely scalable. Consequently, a certain number of assumptions and simplifications will be considered based on to the large scale nature of this study. While it is unreasonable to attempt a CFD computation at the scale of Switzerland, it is nonetheless still desirable to use a mesh of points which can impact one another. Consequently, and given the availability of data and the typical resolution of the data sources, we consider a grid of pixels of size (200×200) [m²] which cover the entire Swiss territory. This grid is in particular designed to match the resolution of demand data, which in Switzerland is constrained by privacy issues related to the energy consumption of districts. Most estimations will therefore be aggregated within each of these pixels. The considered assumptions for the rest of the chapter are as follows:

- (i) Wind speed. All considered values (measured and estimated) are average horizontal wind speed values. The vertical component of wind speed is not taken into account (given the lack for vertical wind speed data and the much larger importance of the horizontal component for wind turbines, both with vertical and horizontal axis). Moreover, the handled wind values are always averaged values, either on a hourly, daily or a monthly time resolution. Therefore, to lighten the text of the chapter, we will simply refer to wind speed rather than average horizontal wind speed, even though the latter denomination is the correct one.

- (ii) *Horizontal homogeneity.* When dealing with average meteorological variables such as average wind speed, the effects of horizontal advection can be neglected over a block of 1-2 km size[58], which means the horizontal variations of wind speed are considered nul. Following assumption (i), we can consider that the considered average wind speed stays constant horizontally anywhere within a (200×200) [m²], and can further be considered constant (at the same altitude) between two (or even more) neighbor pixels. These assumptions will notably play an important role in the estimation of the wind speed in urban areas.
- (iii) *Wind direction.* The variations of the wind direction are discarded in the study and all wind speed values are estimated on average in all directions, within each pixel. While the direction of the wind has an impact on the electricity output of a wind turbine in practice, it is not as significant in a theoretical potential study: we are interested in the identification of the optimal locations for a turbine installation rather than the choice of its rotor direction.

Based on these assumptions, the goal of the chapter is to use the theoretical background presented in the previous one and apply it to the practical estimation of wind energy potential. Physical models (as presented in chapter 3) are combined with GIS processing and Machine Learning (Random Forests in the present case, as presented in chapter 3) to estimate monthly values for wind speed all over the country, at the (200×200) [m²] pixels resolution. Notably, rural and urban areas are treated separately, given the different behavior of wind related variables within their respective environment. The details of the multiple steps leading to the computation of wind related variables in both rural and urban areas, including the roughness length and displacement height, are presented in the different sections of the chapter, along with the strategy combining these variables with wind measurements in order to compute average monthly wind speed maps for Switzerland. Note that the estimation of the wind potential in urban areas will be performed only for the periphery of urban areas (zones spanned by urban pixels which are adjacent to rural areas, pixels called “urban boundary” pixels), due to the complexity of wind patterns within the center of urban areas, as will be further explained in section 4.4.3.

The chapter is organized as follows. Section 4.1 offers a literature review on wind potential studies and places the present chapter within its context. Section 4.2 presents the data sources used in the chapter and some of the processing performed to extract significant features, including weather, topographic and other wind-related variables. Section 4.3 explains the computation of the theoretical wind potential in rural areas, including an estimation of monthly wind speed maps at an altitude 10m using Random Forests and a vertical extrapolation to reach a realistic height for rural turbines. Section 4.4 details the computation of the theoretical wind potential in urban areas, including the computation of the urban roughness length and displacement height in urban pixels and the computation of the average wind speed above buildings. Section 4.5 provides last results on the potential estimation and a discussion on the obtained results. Finally, section 4.6 concludes the chapter and summaries the proposed methodology.

4.1 Related literature

Given its very clean nature and the efficiency of modern wind turbines, wind energy has become very popular over the last decade. It can notably be seen from the increase of installed capacity in many countries, including particularly China, the US, Germany and India [138]. As a result, many wind potential studies have been conducted in the literature in order to exploit the electricity generated by the wind in regions and countries. They often do not consider, however, the opportunity of using wind turbines in urban areas. Instead, existing studies focus on generating potential wind speed maps based on rural measurements which discard cities, often considered not suitable or with not enough potential for wind turbines.

Wind potential mapping studies (in rural areas) have been conducted using multiple strategies. A popular strategy consists in performing some statistical analysis, and most notably parameter estimation of the Weibull Probability Distribution Function (PDF), which is commonly considered to describe the distribution of wind speed. Cerulla et al. [139] applied this strategy in Sicily and estimated the Weibull parameters based on wind data in order to map the wind speed over the whole country. The study is further continued in [140] where the use of MultiLayer Perceptron (a simple type of Neural Network) and kriging (a classical geospatial mapping method) are explored to improve the results of the previous study. More recently, Mentis et al. [141] extracted the wind power potential in Africa based on the estimation of wind speed at the hub altitude of 80m, which they extrapolated based on a fitted Weibull PDF. Additional coefficients are considered in order to obtain an estimation of the geographical wind potential. In [142], the potential for wind turbine generated electricity is extracted for a city in Chad using wind measurements to estimate Weibull parameters at multiple heights and eventually assess the potential power generated by three wind turbine models. Note that some others PDFs have also been used in the literature to model the wind distribution, as listed by Carta et al. [143]. A few studies use a multi-criteria decision methodology together with GIS processing, considering multiple constraints that are aggregated, often qualitatively, in order to compute a potential estimation. Such studies include [144–146] and more recently [147]. Note that the two previously mentioned strategies can be combined to obtain a technical potential estimation, as suggested by, Sliz et al. [148], where they applied that methodology by combining Weibull analysis for wind estimation and decision making in order to extract the economical potential of wind energy in a region in Poland.

Machine Learning methods have been explored for multiple environmental variables, including naturally wind speed. The focus of ML methods in this domain, however, has been on forecasting rather than mapping tasks [149–153]. Regarding mapping applications, neural networks became popular in the 2000's for energy-related estimations [7], and several mapping studies were proposed using Neural Networks trained using wind data [14, 154, 155]. Other ML methods were more recently used for wind speed spatial estimation, such as ensemble learning and kernel methods. In [15], Foresti et al. propose a ML method, Multiple Kernel Learning regression, to spatially estimate the wind speed in Switzerland and interpret the impact of each feature while training the model. Notably, multiple terrain variables are extracted to serve as features including aspect, slopes at multiple scales, directional derivatives of the terrain along multiple directions, and differences of DEM surfaces at multiple smoothing bandwidths, called Differences of Gaussians. Robert et al. [16] present the general regression neural network method to also perform spatial prediction of monthly wind speed, using similar complex terrain features. Veronesi et al. used a combination of Random Forests and

Weibull fitting in order to map wind speed and direction values in England, along with uncertainty estimates. In particular, Weibull parameters are estimated using Random Forests. In [156], a similar methodology is used to obtain a mean wind speed map of Switzerland.

Concerning the (rural) wind potential in Switzerland specifically, the main studies have been mentioned in the previous paragraph. The data itself, however, is often not publicly available. The main source of available information regarding the wind potential in Switzerland is included in the Swiss Wind Atlas, of the Swiss Federal Office of Energy (SFOE). Particularly, annual wind speed maps are available for multiple heights, based on a grid width of 100 metres (Wind atlas of Switzerland, web link: https://www.uvek-gis.admin.ch/BFE/storymaps/EE_Windatlas/?lang=en). The SFOE maps are computed using Weibull distributions fitted with parameters estimated based on wind measurements. Nevertheless, as mentioned earlier, all these studies completely discard urban areas in their estimations.

Regarding the wind in urban areas, there were some recent advances allowing to extract a significant potential [63, 65]. Some urban potential studies have been performed with the aim to determine the wind speed over buildings, yet often solely for a rather small region such as a city. Several studies use a log-law over urban areas together with estimations of the roughness length and displacement height within cities based on various urban characteristics. In [72], expressions from Macdonald [61] are used in order to compute roughness and displacement over Greater London and the energy potential of small wind turbines is estimated on top of buildings based on the NOABL wind database. The impact of the estimation of the building surface (roughness, etc.) is further discussed in [157]. Sunderland et al. [158] extracted the urban wind resource by using a physically-based empirical model to link the wind behavior in traditional rural measurement sites with its behavior over urban areas. A few studies focused on extracting multiple morphology measures in order to understand the urban surface in particular locations and link it to the wind potential in these urban areas. Such studies include [159, 160]. Given the relatively small size of the urban region often studied for wind potential extraction, several studies have explored the use of Computational Fluid Dynamics to model the wind behavior for small zones within urban areas and eventually extract its potential. Heath et al. [161] proposed a methodology using CFD to model the wind behavior over pitched-roof houses, using a semi-logarithmic inflow profile. The average wind speed is extracted over roofs of West London. In [162], the wind power resource of the MIT campus is explored, using CFD modeling. Simoes et al. [163] proposed a methodology using CFD to extract the wind speed over buildings by constructing the surface of the building roofs, considering this surface as a complex terrain. Yang et al. [164] proposed CFD-based evaluation procedures to extract potential mounting sites of wind turbines and estimate the wind power over these sites. On-sites measurements are used to validate the simulations obtained with CFD. Note that while most studies focus on mounted turbines on top of buildings, some studies also explore the potential for wind turbines in between buildings [165]. A review of methodologies aiming at extracting the wind energy potential in urban areas is given in [166]. No study, however, to the best of our knowledge, has proposed a methodology to extract the wind potential of the urban areas of a large region or country.

This chapter is motivated by the lack of urban wind potential studies at the scale of a large region or a country, and particularly in Switzerland. While many methodologies aiming at extracting rural wind maps have been explored, very few attempts have been made to include the wind potential from the urban parts of a region, which is often considered too small and negligible, or economically unprofitable, together with the rural one. Consequently, the goals of this chapter are threefold: (i) propose a machine

learning strategy (using Random Forests mainly) to extract the wind speed in rural areas (excluding the urban areas) of Switzerland based on measurements and various meteorological, topographic and other features, (ii) suggest a new methodology to extract the wind speed potential specifically in urban areas over a large region, (iii) extract the wind speed potential of the urban areas of Switzerland in particular, which has never been estimated to the best of our knowledge, and discuss its significance. It results in the estimation of the total theoretical wind potential (through the wind speed and ultimately the wind power) in Switzerland, in both rural and urban areas, over monthly estimated maps.

4.2 Data

4.2.1 Data sources

All data sources used within this chapter are presented in Appendix A and signified by a ✓ for the present chapter within Tables A.1 and A.2. They include monitored meteorological data, wind speed data, digital surface and terrain models, land cover data (CORINE) and building footprint and facade data. Given the availability of data, particularly for building facades data, Switzerland is divided, for this chapter, in two zones shown in Figure 4.1. Zone 1 is the area of Switzerland where building facade data (from the Sonnendach project, originally from swissBUILDING3D) is available at the time of the study; zone 2 is the area of Switzerland for which this information is not available. Note that the zone 1 reflects the state of availability of data from the Sonnendach project at the time at which this study was performed. In the future, the whole Swiss territory will be tackled. The two zones will be treated differently when estimating the wind speed in urban areas. A clear distinction is made between urban areas and rural areas for the entirety of the chapter, as the behaviour of the wind is obviously significantly dependent on the presence or the absence of obstacles such as buildings. Note that even though areas containing only a few buildings are not commonly seen as actual “urban areas”, the presence of buildings nonetheless impacts the wind behavior and entails the use of urban wind characteristics (see section 3.1, chapter 4). As a result, urban areas are here defined by the collection of 200×200 [m²] pixels that contain at least one building (as defined by the TLM3D building data). All the remaining pixels, which do not contain any building, define the rural areas. Note that, based on our wind-specific definition of urban areas, we could not use the CORINE land cover data to differentiate between urban and rural areas as it follows a more conservative urban definition (the “urban fabric”) and does not take into account areas with a few isolated buildings as urban areas.

4.2.2 Data processing

Before training a Machine Learning model, the first step is to select significant features (input variables) to include when gathering the training data. As explained in chapter 2, these features must be chosen for their impact on the output variable of interest, and are therefore usually selected based on expert knowledge on the matter. Also, they must be observed (known) for the entirety of the points to be predicted. In the present chapter, we are interested in features that impact the wind speed. Concerning the wind speed over rural areas, which will be the focus of a machine learning model further in the chapter, significant features include topographic variables describing the terrain, weather variables and wind specific variables (e.g. roughness length).

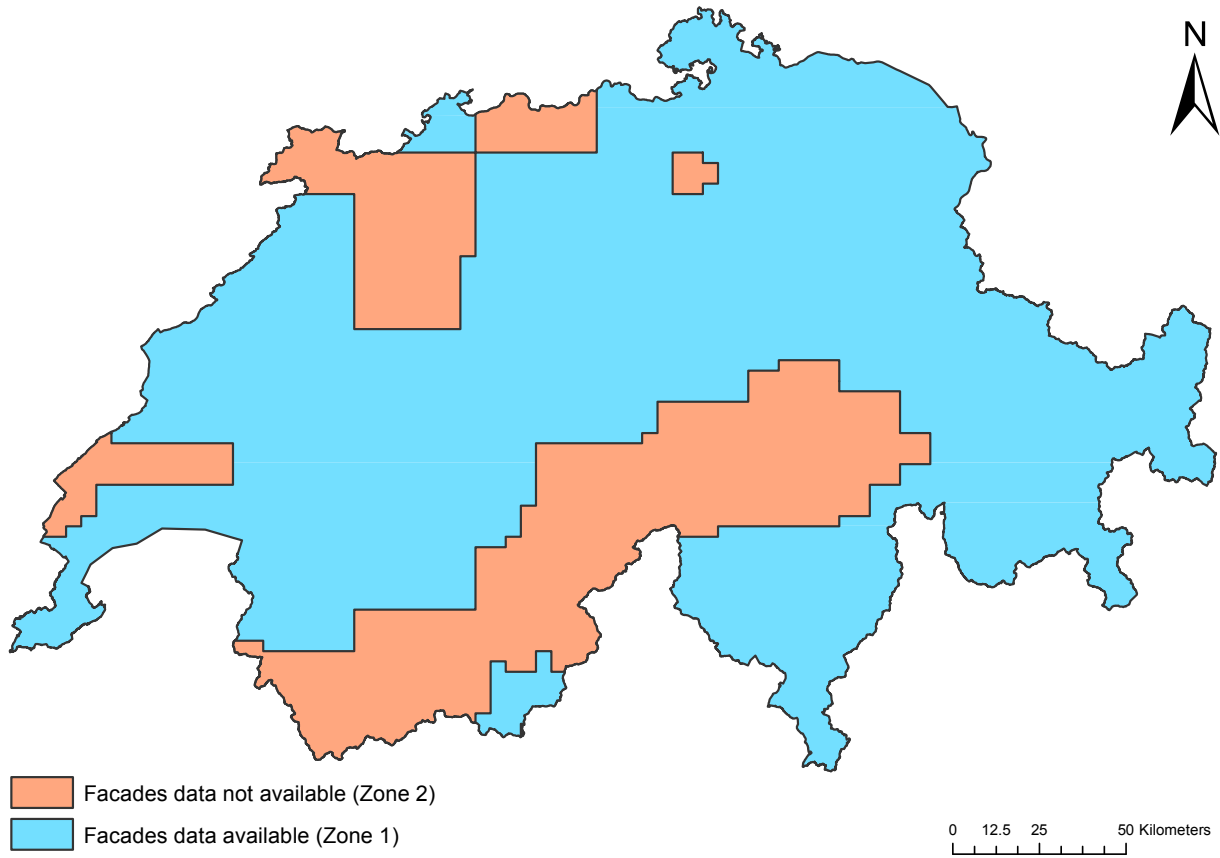


Figure 4.1: The Swiss territory divided in two zones, treated differently for the estimation of the wind speed in urban areas.

To extract features for the whole Switzerland, a Digital Elevation Model (DEM) available for the entire territory (called DHM, as presented in Table A.1) with a resolution of $25 \times 25 \text{ [m}^2\text{]}$ is used. It is down-sampled to a resolution of $200 \times 200 \text{ [m}^2\text{]}$ in order to offer a good precision at the scale of the country while being easily manageable computationally. This DHM forms the base of the mesh grid that will be used for the estimation of the wind speed and most of the estimated variables in the study. Each grid cell will be called a *pixel*, to avoid a confusion with raster cells. The grid defined contains 1140 (rows) \times 1921 (columns) pixels for a total number of 2194500 points indicating the centroid of each pixel, and covers the entire Switzerland. For each pixel, if data is available, we simply allocate the value of a data point to the closest pixel identified by its centroid. Given the rather large number of data points, a binary data format called Hierarchical Data Format 5 (HDF5) is used to process the data within the Python programming language; H5py, an interface for HDF5 in Python, is used to manipulate the data.

Topographic features can be extracted from the DHM using raster processing tools of the Spatial Analyst toolbox from ArcGIS. In addition to longitude, latitude and altitude offered originally by the DHM, the additional topographic features obtained within ArcGIS are: the terrain slope, the terrain aspect (also known as exposition or direction of the terrain), and three types of curvature, namely the plan, longitudinal and transverse curvature of the terrain. All topographic features are shown at the scale of the country in Figure 4.2.

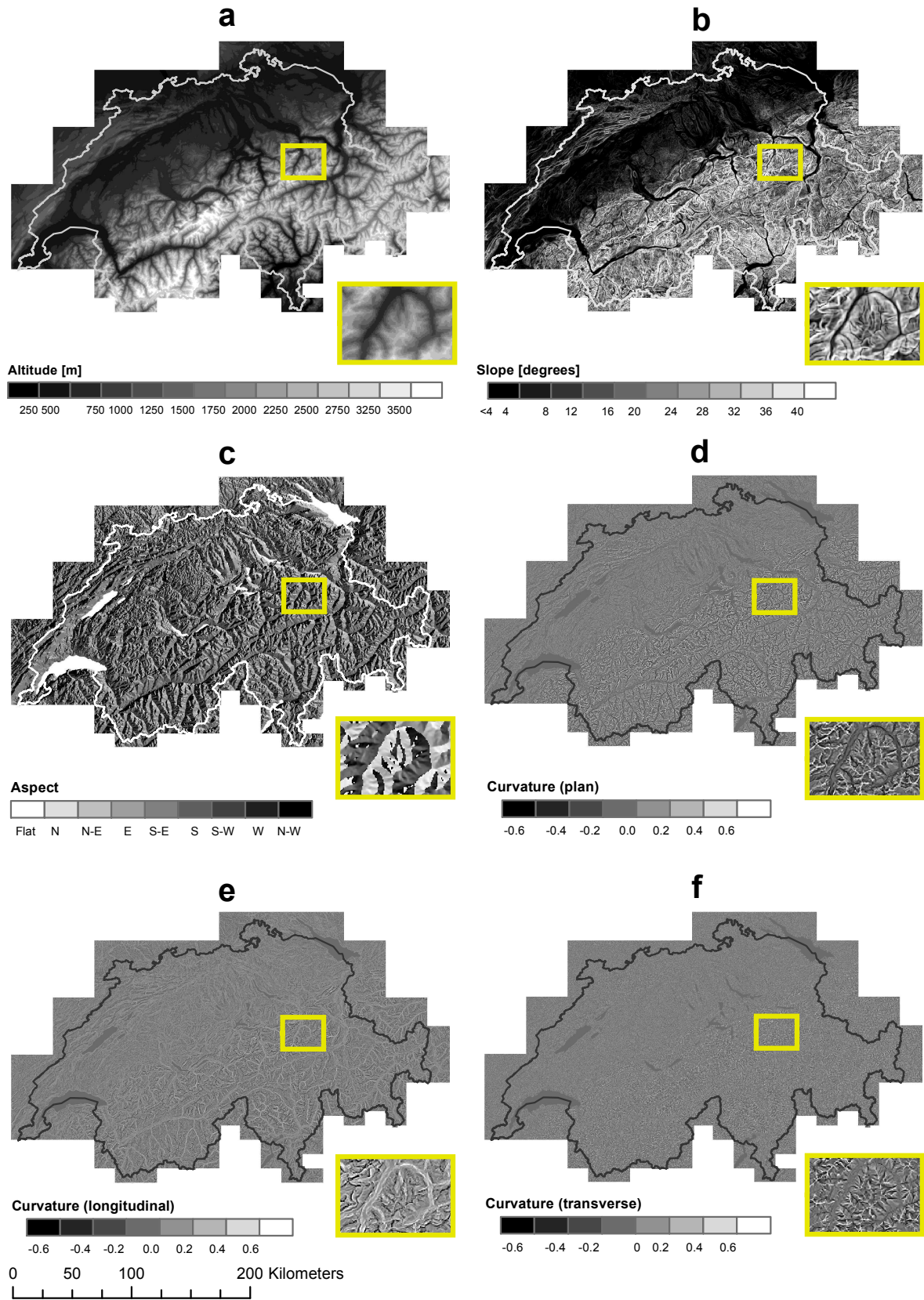


Figure 4.2: Terrain features extracted from the Digital Elevation Model in Switzerland. (a) DHM 200×200 [m²], (b) terrain slope, (c) terrain aspect, (d) terrain curvature, (e) terrain longitudinal curvature, (f) terrain transverse curvature.

Table 4.1: Testing RMSE (E_R , in the same unit as the variable of interest) and NRMSE (E_{NR} , in %) for Random Forest models trained for weather variables.

Month	Temperature		Cloud cover		Precipitation		Sunshine duration		Pressure	
	E_R [°C]	E_{NR} [%]	E_R [%]	E_{NR} [%]	E_R [mm]	E_{NR} [%]	E_R [hours]	E_{NR} [%]	E_R [hPa]	E_{NR} [%]
Jan.	1.63	0.60	9.49	16.68	21.58	27.42	12.95	16.29	4.09	0.46
Feb.	1.40	0.52	8.21	14.67	20.06	27.55	9.18	9.07	4.28	0.48
Mar.	0.88	0.32	9.17	15.13	21.46	24.28	11.22	8.05	3.88	0.44
Apr.	0.82	0.29	8.37	12.87	19.59	20.82	13.32	8.61	2.62	0.30
May.	0.78	0.27	6.61	10.03	18.88	14.94	15.41	8.78	3.05	0.34
Jun.	1.03	0.36	8.41	13.24	19.50	14.44	16.41	8.66	3.76	0.42
Jul.	0.99	0.34	6.84	11.82	22.65	16.23	19.09	8.83	2.77	0.31
Aug.	0.84	0.29	5.42	9.26	20.71	14.56	15.58	7.87	2.38	0.27
Sep.	0.73	0.26	4.24	7.09	19.08	16.27	12.99	8.24	2.98	0.33
Oct.	0.71	0.25	5.08	8.51	16.86	17.36	12.72	10.43	2.47	0.28
Nov.	1.16	0.42	6.34	10.36	20.00	20.65	12.01	15.41	3.28	0.37
Dec.	1.65	0.61	9.13	15.02	21.20	22.96	11.18	17.30	2.90	0.33

Meteorological features are also desirable when attempting to predict an environmental variable such as wind speed. We consider the following weather variables: sunshine duration, air temperature, precipitation, cloud cover and air pressure using monthly measured data described in Table A.1, collected from Meteoswiss [167]. This monitored data for each of the weather variables solely offers local information as the measurement stations vary for each variable and are limited to a number of points across Switzerland: 66 point locations for sunshine duration, 91 different point locations for temperature, 417 points for precipitation, 23 points for cloud cover and 257 points for pressure. As a result, we train monthly Random Forests models, in order to interpolate the training data and estimate monthly maps for each of the five variables. In the training process, the input features for each meteorological variable are latitude, longitude and altitude, while the labels are meteorological variables. Given the small dimensionality of the feature data ($d = 3$), no feature selection or dimensionality reduction technique is performed. Note that a RF model is trained separately for each weather variables so as to estimate values at unknown points and produce monthly weather maps, as shown in Fig. 4.3. Testing errors are obtained for the prediction of weather variables and shown in Table 4.1. The resulting weather maps are stored in HDF5 format in order to be used for later variable estimation, including wind speed and other variables to be estimated in the thesis.

Finally, wind specific variables such as roughness length and displacement (as presented in chapter 3) seem to be natural features to estimate the wind speed spatially. Since the roughness length in rural areas ($z_{0,r}$) is traditionally extracted based on land cover, we use tabled values suggested by [60] offering average roughness values (Table 3.1 presented in chapter 3) for the various land cover types defined by the CORINE Swiss database (described in annex Table A.2). Note that in case of a pixel with multiple land cover types, the one covering the largest area in the pixel is considered. The obtained rural roughness map of Switzerland is shown in Figure 4.5. In addition to the roughness of the considered pixel, we would like to capture the dynamical behavior of wind speed through the analysis of the adjacent pixels' characteristics: if these adjacent pixels show high roughness levels or contain obstacles, it should reflect on the wind speed value of the considered pixel. Classically (notably in cellular automata theory), a pixel adjacent to another pixel can be defined as a pixel in the

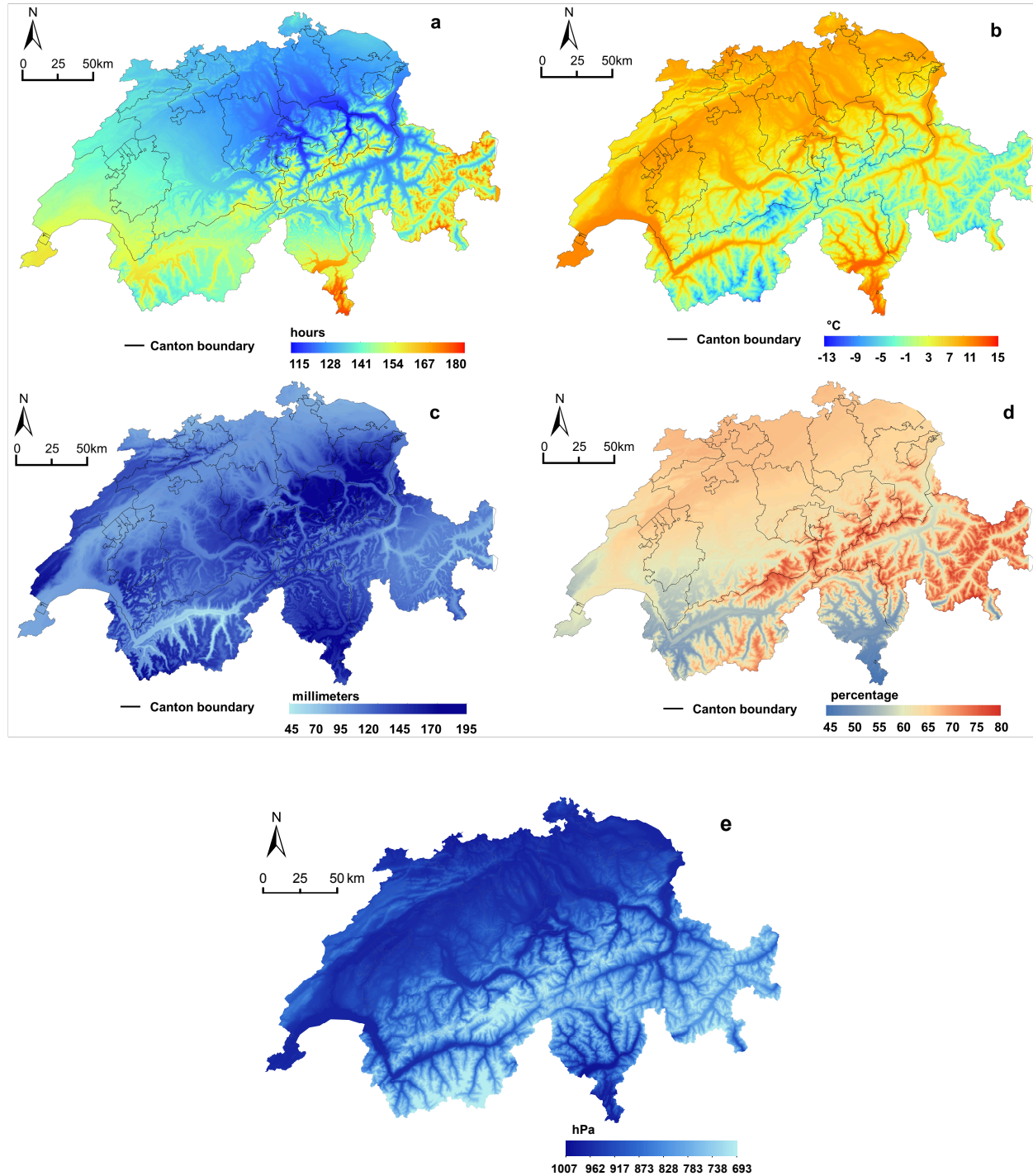


Figure 4.3: Prediction of meteorological variables using RF models. **(a)** Monthly mean yearly sunshine duration (hours), **(b)** yearly mean air temperature (degree Celsius), **(c)** monthly mean yearly precipitation (millimetres), **(d)** yearly mean cloud cover (percentage), **(e)** yearly mean air pressure (hectoPascals).

$(i - 1, j + 1)$	$(i, j + 1)$	$(i + 1, j + 1)$
$(i - 1, j)$	(i, j)	$(i + 1, j)$
$(i - 1, j - 1)$	$(i, j - 1)$	$(i + 1, j - 1)$

Figure 4.4: Moore neighborhood of a pixel located in (i, j) within a pixel grid of rows indexed by i and columns indexed by j .

neighborhood of the latter, and is therefore called a neighbor pixel. While there are multiple ways to define the neighborhood of a pixel, we consider in this study all the pixels adjacent to a given pixel, in all directions, as defined by Moore [168]. The Moore neighborhood is illustrated in Figure 4.4. For each pixel in rural areas, we compute the average roughness of all its neighbor pixels, which defines an additional neighbor roughness feature for the pixel when estimating the wind speed. Note that some rural pixels adjacent to urban areas might have urban pixels as neighbors. It is therefore necessary to determine the roughness in urban pixels ($z_{0,u}$) in order to obtain the neighbor roughness feature for all rural pixels. For details of the urban roughness computation, the reader is invited to forward to section 4.4.1, which focuses on the computation of urban characteristics estimation.

4.3 Wind speed estimation in rural areas

The first stage in the assessment of the wind potential in Switzerland is the estimation of the wind speed in all rural areas of the country (defined in section 4.2.1), for the considered 200×200 [m²] pixels and at a monthly time resolution. It is first estimated at a height of 10 m, using available wind speed measurements at this height and Random Forests models together with the features extracted in the previous section. It is then vertically extrapolated at a height of 100 m using a classical log-law, in order to estimate the wind potential at a adequate height for a rural wind turbine installation (for a large commercial horizontal axis wind turbine, considered as the chosen system in this study. See section 3.1.6 in chapter 3). Also, the height of 100m will be useful for several assumptions allowing the further estimation of the wind speed in urban areas.

4.3.1 Estimation at 10 m

The computation of the wind speed at 10 m is based on the training of Random Forests (RF) models together with daily wind speed measurement data from MeteoSwiss, available for 197 stations (as described in Annex A). The measurement data is processed in order to aggregate the values monthly, averaging the speed for each month from 2000 to 2017. The measurement values are then aggregated

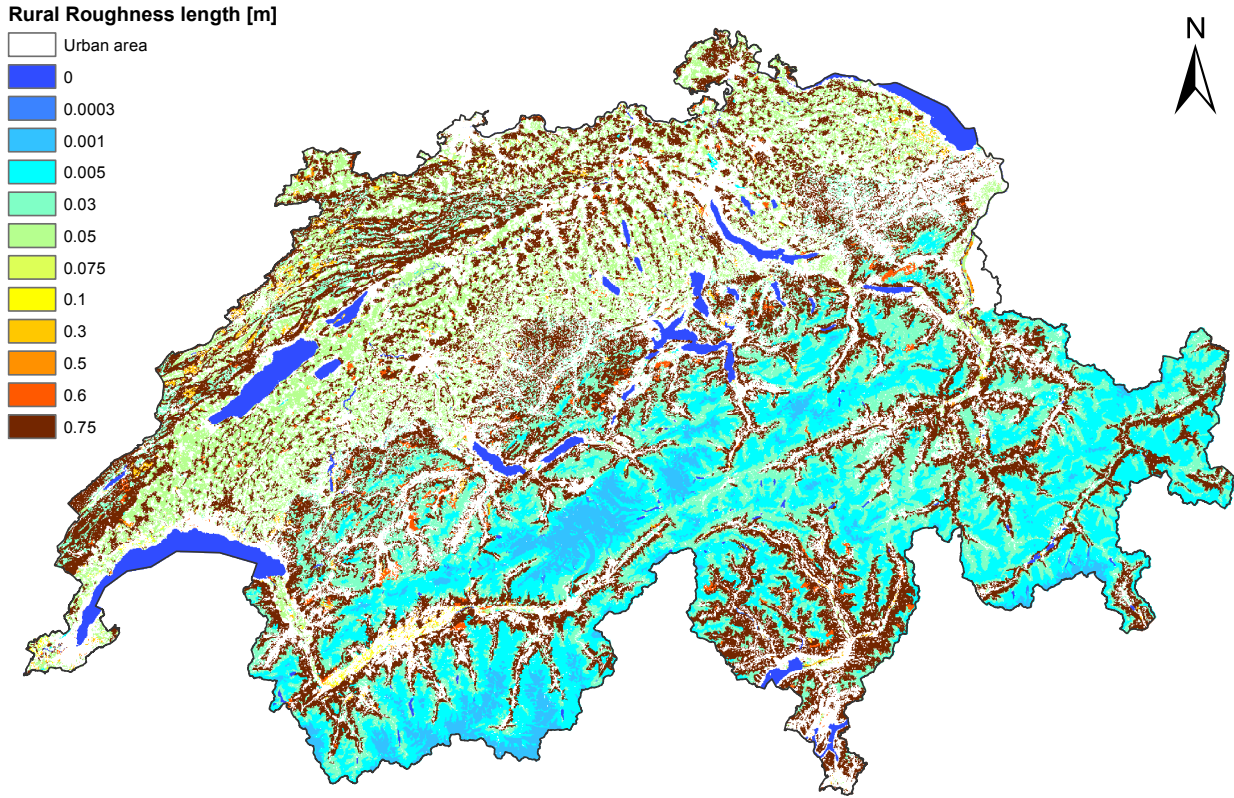


Figure 4.5: Roughness map in rural areas of Switzerland, obtained using the CORINE land cover data and typical roughness values ([60]).

to match the 200×200 [m²] pixel resolution: if a station is alone within a pixel, the monthly measurements of this sole station are considered to represent the whole pixel; if multiple stations are located in one pixel, the monthly measurements are averaged through the stations to obtain the considered pixel value. We thus obtain 159 pixels with monthly wind measurements. The latter pixels are then separated between rural and urban pixels: the rural pixels are selected for the training process, the urban pixels are not used at this stage and are stored for further validation purposes. We obtain 118 rural (training) pixels and 41 urban pixels with wind measurements. Note that, in the following paragraph, all machine learning and RF-related concepts (RF, Variable Importances, Prediction Intervals, Quantile Regression Forests etc.) are presented in theory and practice in chapter 2.

The training data for the building of RF models for rural wind estimation is therefore composed of 118 points, for which the labels are the monitored monthly wind speed, and the features are the ones extracted in section 4.2.2: latitude, longitude, altitude, aspect, slope, longitudinal curvature, transverse curvature, plan curvature, monthly air temperature, monthly sunshine duration, monthly precipitation, monthly cloud cover, monthly air pressure, roughness length and neighbor roughness length. Each feature is obtained for each training pixel by sampling the feature maps (available or estimated in section 4.2.2) at the location of the training pixels. Random Forests models are trained for each month separately (using the monthly features). The testing errors resulting from the training

Table 4.2: Errors related to the building of monthly wind speed in rural areas using RF models. Left side of the table: Testing errors for each monthly model, in the form of the RMSE (E_R , in m/s), NRMSE (E_{NR} , in %) and OOB score (between -1 and 1). Right side of the table: Monthly Prediction Errors, computed using Quantile Regression Forests, averaged over a random sample of 1000 unobserved pixels. $PE_{s,down}$ is the average bottom error above the mean predicted value, $PE_{s,up}$ is the upper error above the mean predicted value, PE_s is the average of $PE_{s,down}$ and $PE_{s,up}$.

Month	Wind speed at 10m [$u_r(10)$]					
	Testing errors			Mean Prediction Errors		
	E_R [m/s]	E_{NR} [%]	OOB [-]	$PE_{s,down}$ [m/s]	$PE_{s,up}$ [m/s]	PE_s [m/s]
Jan.	0.74	23.13	0.57	2.09	1.53	1.81
Feb.	0.80	24.61	0.55	1.34	1.48	1.41
Mar.	0.68	20.57	0.51	0.70	1.63	1.16
Apr.	0.70	22.21	0.42	1.12	1.34	1.23
May.	0.62	20.73	0.42	0.40	1.58	0.99
Jun.	0.60	21.19	0.41	0.65	1.37	1.01
Jul.	0.59	21.12	0.43	0.28	1.42	0.85
Aug.	0.59	22.97	0.46	0.82	1.19	1.00
Sep.	0.57	22.04	0.43	1.09	1.10	1.10
Oct.	0.66	24.26	0.50	0.89	1.29	1.09
Nov.	0.77	24.92	0.56	1.13	1.50	1.31
Dec.	0.80	24.53	0.62	2.18	1.52	1.85

of the monthly RF models are shown in Table 4.2, in the form of the RMSE, NRMSE, and OOB score obtained for each month. Prediction Intervals (PIs) are computed for each monthly model, both in the test set and for unobserved points, using Quantile Regression Forests (as presented in section 2.3.5), in order to extract a measure of uncertainty attached to the estimations. PIs are illustrated for some months on Figure 4.6. Also, prediction errors (width of PIs, below and above the predicted value) are computed over a sample of 1000 unobserved points for monthly predictions, to have an estimation of the uncertainty variations with time through the months. Finally, the prediction errors are computed for the yearly prediction of all rural pixels and shown in Figures 4.9 and 4.10, to have an estimation of the uncertainty variations with space across the country. If we note $PE_{s,down}$ the bottom error and $PE_{s,up}$ the top error, we have, with an approximately 95% confidence, schematically:

$$\text{prediction} - PE_{s,down} \leq \text{actual wind speed value} \leq \text{prediction} + PE_{s,up}$$

The Variable Importances embedded in the RF algorithm is also computed for each monthly RF model and averaged through the month in order to extract the importance of each predictor (feature) in the overall training process, as shown in Figure 4.7. The monthly RF models are then used in unobserved pixels in order to estimate the monthly wind speed in all rural areas of Switzerland. Note that we exclude the forest areas, defined by a roughness length of 0.75, in the estimation, as we do not consider them as suitable areas for wind turbines installation, for natural environmental and social acceptance reasons. The obtained rural monthly wind maps are averaged through the year in order to obtain a rural yearly wind map at 10m, shown in Figure 4.8.

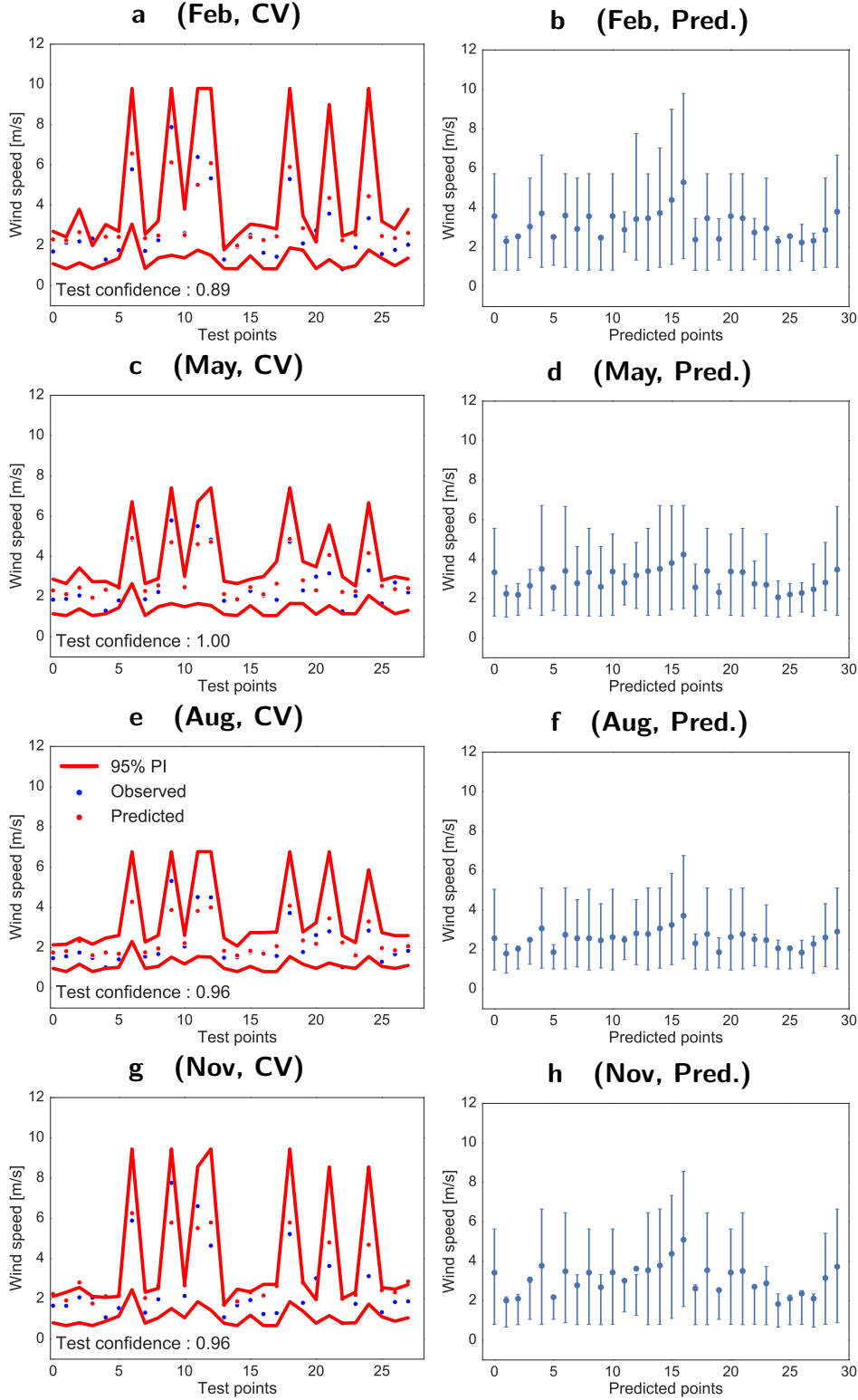


Figure 4.6: Prediction Intervals (with 95% confidence) extracted from Quantile Regression Forests while training monthly models for rural wind estimation at 10m. We show the PIs for an example of 4 months. **(a)** and **(b)**: PIs for February, respectively in the test set and for 30 random unobserved (unknown) points; **(c)** and **(d)**: PIs for May, respectively in the test set and for 30 random unobserved (unknown) points; **(e)** and **(f)**: PIs for August, respectively in the test set and for 30 random unobserved (unknown) points; **(g)** and **(h)**: PIs for November, respectively in the test set and for 30 random unobserved (unknown) points.

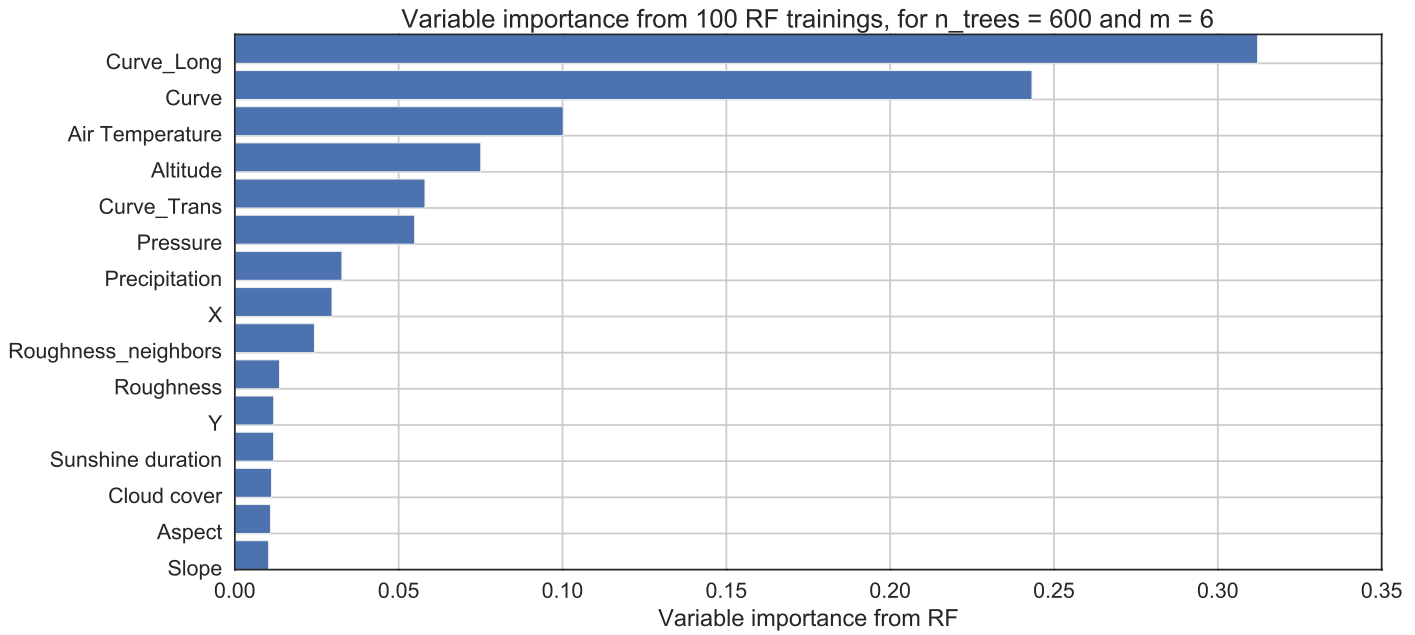


Figure 4.7: Variable Importance of each feature during the training of a Random Forest model over yearly averaged features to estimate the yearly wind speed in rural areas (at a height of 10m). Note that X, Y and Z are respectively the longitude, latitude and altitude.

A few comments can be made regarding the results obtained from the trained monthly wind speed RF models. From the error table (Table 4.2), it can be observed that the overall year NRMSE is around 22%, which is a rather acceptable error given the often very unstable behavior of wind speed through the months. Furthermore, the slightly smaller errors during the warmest months show that the models perform better in the summer than in the winter, which is intuitively explained by a larger presence of wind in the winter, with generally higher values. It is therefore harder to predict the wind speed during the winter, as it can fluctuate and possibly reach unexpected high values very quickly. This seasonal trend is confirmed on the Prediction Intervals shown in Figure 4.6: while the test confidence stays acceptable for all months, it is better in May and August, with respectively 100% and 96% of observed points in the PIs, than in February, when the wind reaches higher values and the test confidence is only of 89%. It is also worth noting on Figure 4.6 that some points, for all months, seem harder to predict than others, as shown by a large PI, for example for points 6, 9, 11 and 12. This is partly explained by a larger range of measurement wind values for these points, therefore making it harder for the RF to predict the correct value. The seasonal trend is also confirmed through the average sample prediction errors shown in Table 4.2, particularly for the bottom error. An additional piece of information can be extracted from the observation of the bottom error: it seems that in April and September, in between clear winter and summer periods, the prediction error has a sudden jump. This jump can be explained by the often unpredictable weather that characterizes the beginning of spring and autumn: depending on the year, the temperature and the wind vary significantly. The spatial trend of the uncertainty attached to the prediction can be shown on the yearly error maps on Figures 4.9 and 4.10. As expected, the uncertainty is larger (± 1.5 to 3 m/s) in locations of high altitude, particularly mountain areas, where wind values can reach higher levels and tend to have a larger range of values. High uncertainties can also be the result of a lower density of measurement stations, which is for example the case in

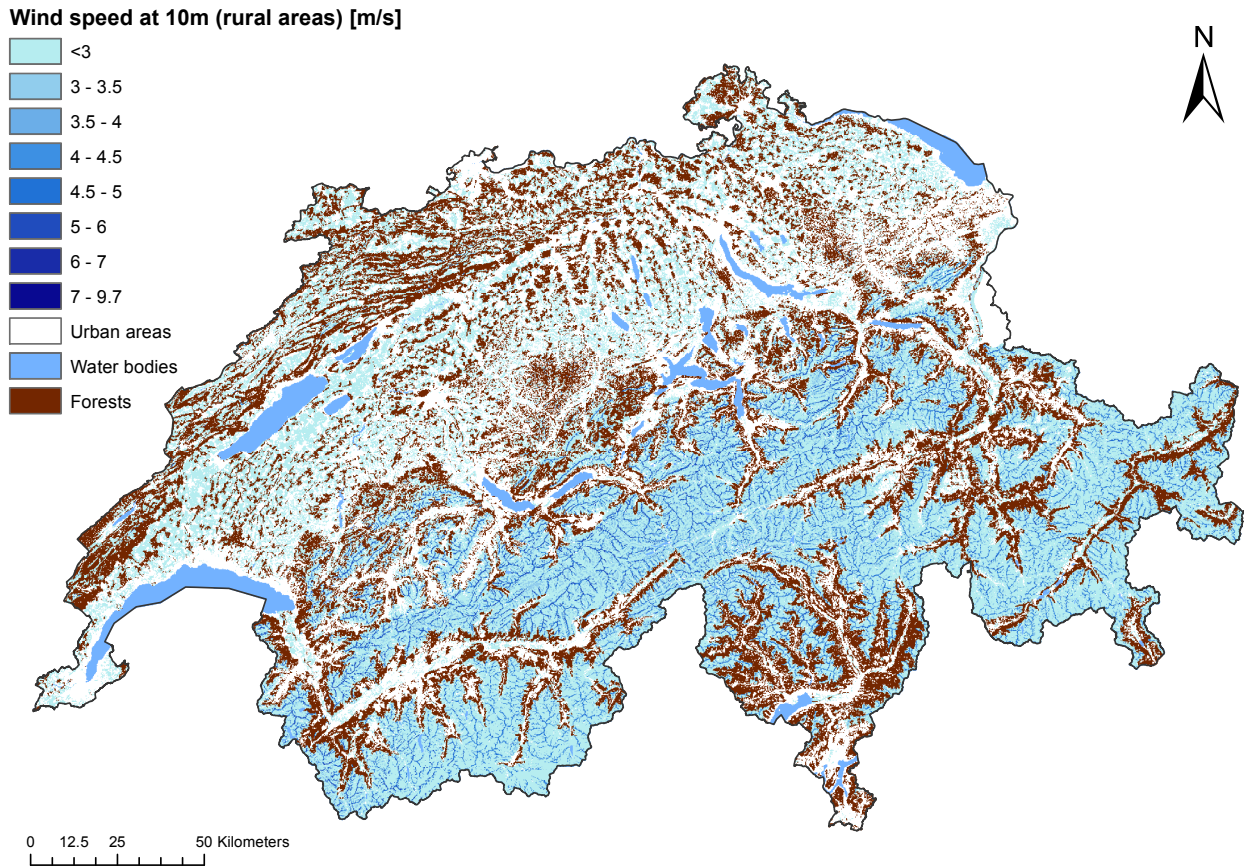


Figure 4.8: Wind speed (yearly average) in rural areas, estimated at a height of 10 m. Note that the color map thresholds are chosen abnormally high so that they match the ones used for the extrapolated rural speed map at 100m, later presented in Figure 4.12. They can therefore be compared easily.

the far east of the country (Graubunden canton), as seen from stations locations in Figure A.2 in Appendix A. In lower altitude locations, notably the Swiss Plateau and in valleys between mountains, the uncertainty range from ± 0.5 to ± 1.5 m/s. Note that the difference in uncertainty between the plateau and the alps is particularly clear from the upper error map, in Fig. 4.9b.

Other useful pieces of information are given by the Variable Importances, on Figure 4.7. As expected from intuition, the terrain features are the most important in the training of the model, along with altitude, air temperature and air pressure. Note that while it is intuitive that the terrain curve has a very important impact on the wind speed, it is worth observing that the longitudinal curve is significantly more impactful than the transverse one, since it actually shows the highest VI. Besides, the roughness, both in each pixel and from the neighbors do not have a tremendous importance. This is partly explained by the fact that, even though the roughness has an impact on the wind speed at 10m, it has a bigger impact on the vertical extrapolation of the wind speed rather than the horizontal one, which is the one explored at this step.

Note that even though, as presented in chapter 2, feature engineering is a popular practice and an important part of ML (consider functions of the features, combining features, etc.), here we don't manipulate the features in order to keep the physical meaning of each feature and because we don't have specific knowledge motivating such manipulations. Also, we do not perform dimensionality reduction

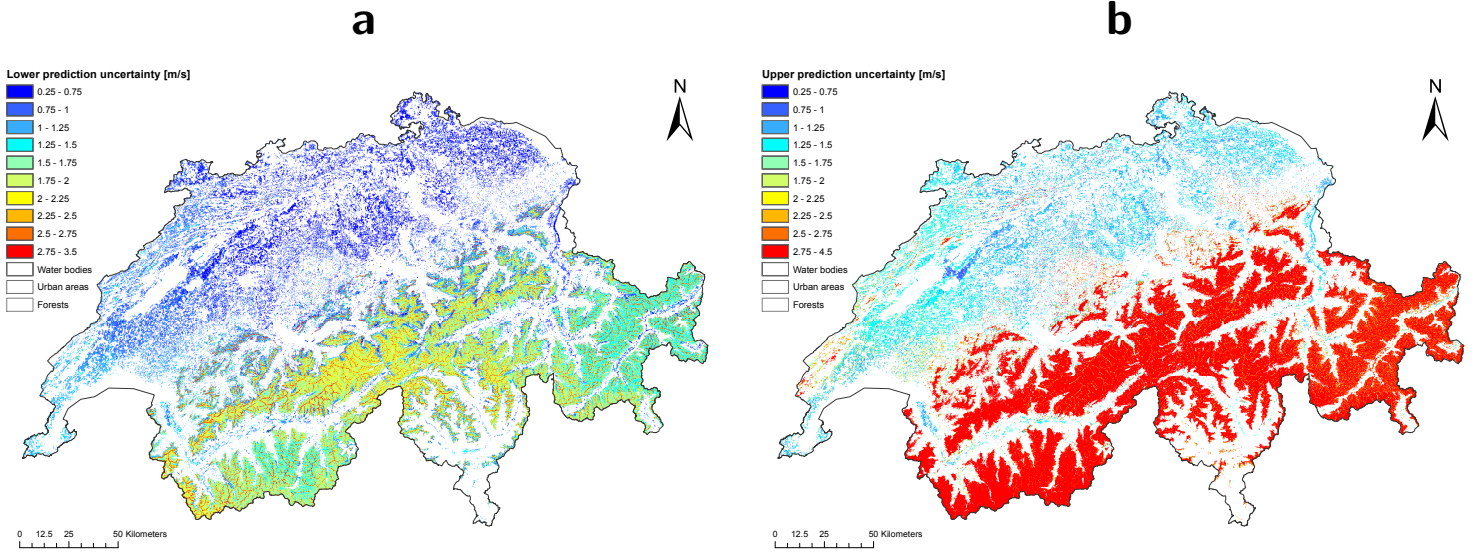


Figure 4.9: Lower and upper uncertainties ($PE_{s,down}$ and $PE_{s,up}$) attached to RF prediction for the yearly wind speed at 10m in rural areas (obtained from the computation of Prediction Intervals with Quantile Regression Forests). (a) Lower prediction error, (b) Upper prediction error.

since the number of considered predictors is rather small. Note that in environmental mapping [13], the feature engineering is minimal and predictors are often used in their original form.

4.3.2 Extrapolation at 100 m

While wind turbines can be installed at various height in rural areas (10-20m for small turbines, 30-50m for medium turbines, 100-200 for large turbines), commercial turbines are often installed at a height of 50 to 100m, where the wind is stronger and more stable, in order to achieve the maximum potential at the location of interest. Therefore, we wish to extrapolate our wind estimation at 10m to a height of 100m to obtain a realistic commercial wind potential in each rural pixel. Given the estimated wind speed in rural areas at 10m, it is possible, in the boundary layer [58], to vertically extrapolate the rural wind speed to another height using a log-law profile (presented in section 3.1.3):

$$u_r(z) = \frac{u_r^*}{\kappa} \ln \frac{z - z_{d,r}}{z_{0,r}} \approx \frac{u_r^*}{\kappa} \ln \frac{z}{z_{0,r}}, \text{ for } z > z_{0,r} \quad (4.1)$$

where $u_r(z)$ is the rural wind speed at a height of z , u_r^* is the rural friction velocity, κ is the Von Karman constant, and $z_{0,r}$ and $z_{d,r}$ the rural roughness length and displacement of the considered rural area. Since we exclude forest areas in the estimation and therefore only consider rural areas with rather small displacement values (generally < 0.7 [59]), we consider the impact of $z_{d,r}$ negligible in the log-law. This equation is considered true in any rural area, and will be therefore used to characterize the wind velocity in any rural pixel j .

We adopt the assumption that the friction velocity remains constant along the profile above rural areas (above the roughness length) at any height within the Inertial Sublayer (ISL) (presented in section 3.1.1), since the ISL is characterized by a small variation of turbulent fluxes with height ($< 5\%$)

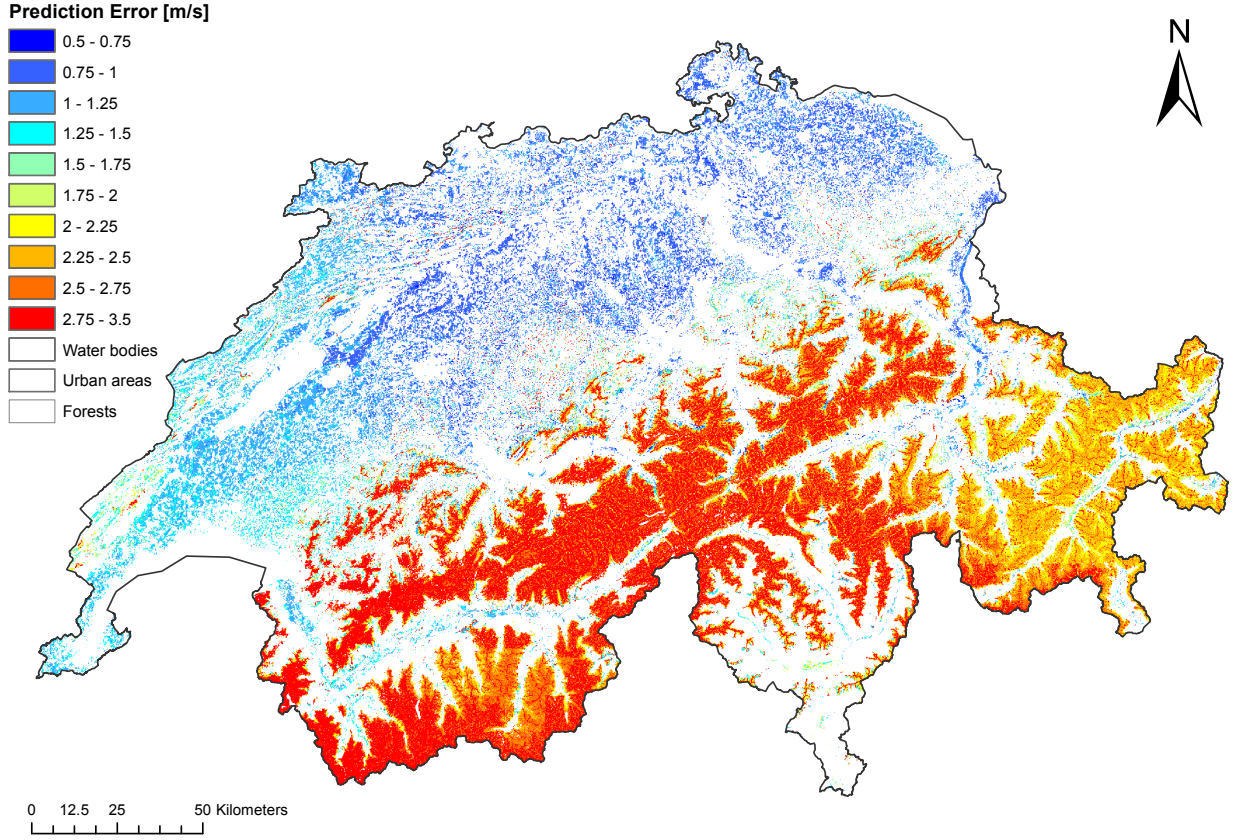


Figure 4.10: Prediction error (PE_s , the average of the lower and upper prediction error) attached to RF prediction for the yearly wind speed at 10m in rural areas (obtained from the computation of Prediction Intervals with Quantile Regression Forests).

and often considered as a “constant flux layer” [59]. Therefore, we can take the ratio of Eq. 4.1 to the same equation evaluated at $z = 10$ and see the friction velocity and the Von Karman constant getting simplified to obtain the rural wind speed at a height z (within the ISL) as a function of the rural wind speed at a height of 10m, for any rural pixel j :

$$u_r^j(z) = u_r^j(10) \frac{\ln(z/z_{0,r}^j)}{\ln(10/z_{0,r}^j)}, \text{ for } z \in \text{ISL} \quad (4.2)$$

where $z_{0,r}^j$ is the rural roughness length in pixel j estimated in the pre-processing section 4.2.2, $u_r^j(10)$ is the previously estimated wind speed in a rural pixel j at a height of 10m, and $u_r^j(z)$ the wind speed in pixel j at a height of z within the ISL. Note that even though the width of the ISL naturally varies depending on the location, it is often admitted that its upper boundary is approximately located at $z = 100$ m, as shown in the illustration by Oke et al. [59] presented in chapter 3, in Figure 3.2. Therefore, Eq. 4.2 can be evaluated at $z = 100$ and the wind speed above rural areas at 100 m ($u_r^j(100)$) can be estimated for the considered rural areas of Switzerland.

The height of 100 m is a key altitude for our estimation: (i) it provides a realistic height for wind turbine installation and is therefore suitable for potential estimation, (ii) it is within the ISL and

therefore: the log-law can be used with constant friction velocity and the wind behavior can be treated vertically as horizontal changes are minimal, (iii) it is “high enough” from the roughness layer to adopt a second assumption: the periphery of the urban areas are characterized by a wind speed (at this altitude of 100 m) that can be considered equal to the wind speed estimated in “near-by” (adjacent) rural areas (at the same altitude of 100 m). Note that the assumption (ii) is strengthened by the horizontal homogeneity assumption discussed in the introduction of the chapter [58]. Eventually, the wind speed above rural areas at 100 m ($u_r^i(100)$) is computed over the whole country to extract the theoretical wind potential of rural areas. The obtained monthly wind speed maps at 100m are shown in Figure 4.11. They are averaged yearly in order to obtain a yearly mean map for wind speed at 100m, as shown in Figure 4.12. The obtained values are stored for further use in the estimation over urban areas, which is the focus of the next section.

4.4 Wind speed estimation in urban boundary areas

We wish to compute the wind speed in urban areas based on an urban log-law and near by rural wind speed estimations performed in the previous section. To obtain such wind values in urban areas, we need the assumptions mentioned previously for a height of 100m and most importantly extract urban characteristics which will be used as inputs for the “urban log-law”, the same log-law defined in Eq. 4.1 but with the variables adapted for urban areas, and a displacement height that cannot be neglected given the large height of urban buildings. As a result, the structure of this subsection will be the following: (i) extract urban wind characteristics (mean building height, urban roughness length, displacement), (ii) make the assumption that the wind velocity stays constant around the boundaries of urban areas at a high altitude, and express it using pixels, (iii) use the estimated rural wind speed at 100m and the urban log-law to compute the wind velocity over urban pixels at the boundaries of urban areas. Note that all of wind characteristics are aggregated for each pixel, as in the rural wind section.

The “urban” log-law, used throughout this section, expresses wind velocity in an urban pixel for any height z within the ISL layer:

$$u_u(z) = \frac{u_u^*}{\kappa} \ln \frac{z - z_d}{z_{0,u}}, \text{ for } z \in \text{ISL} \quad (4.3)$$

where u_u^* is the friction velocity in the urban pixel, and $z_{0,u}$ and z_d are the urban roughness length and urban displacement height, previously computed for each urban boundary pixel. Note that, in order to be coherent with the notations of the chapter, the urban displacement height should be written $z_{d,u}$ to specify that we are in urban areas. However, since we do not mention the displacement in rural areas in the rest of the chapter, there is no confusion in continuing to note z_d the urban displacement height, in order to lighten the notations. Note that this log-law is not theoretically valid in the RSL (Roughness sublayer), located under the ISL and above the UCL (Urban Canopy Layer), meaning between the buildings’ height and the ISL.

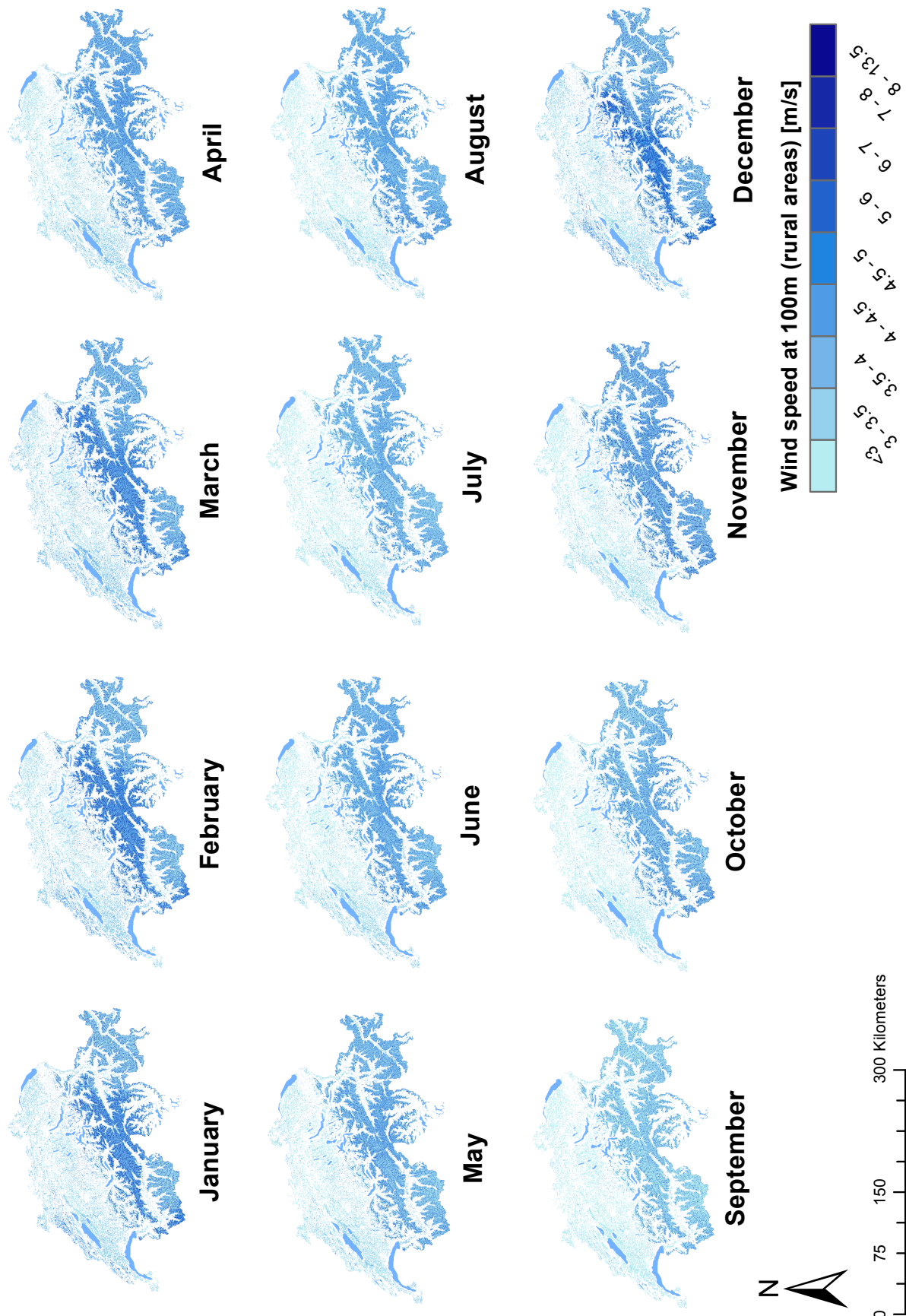


Figure 4.11: Monthly maps for wind speed in rural areas, at a height of 10m. Note that lakes are shown in light blue while urban and forest areas are both set to white in order to improve the readability of the maps.

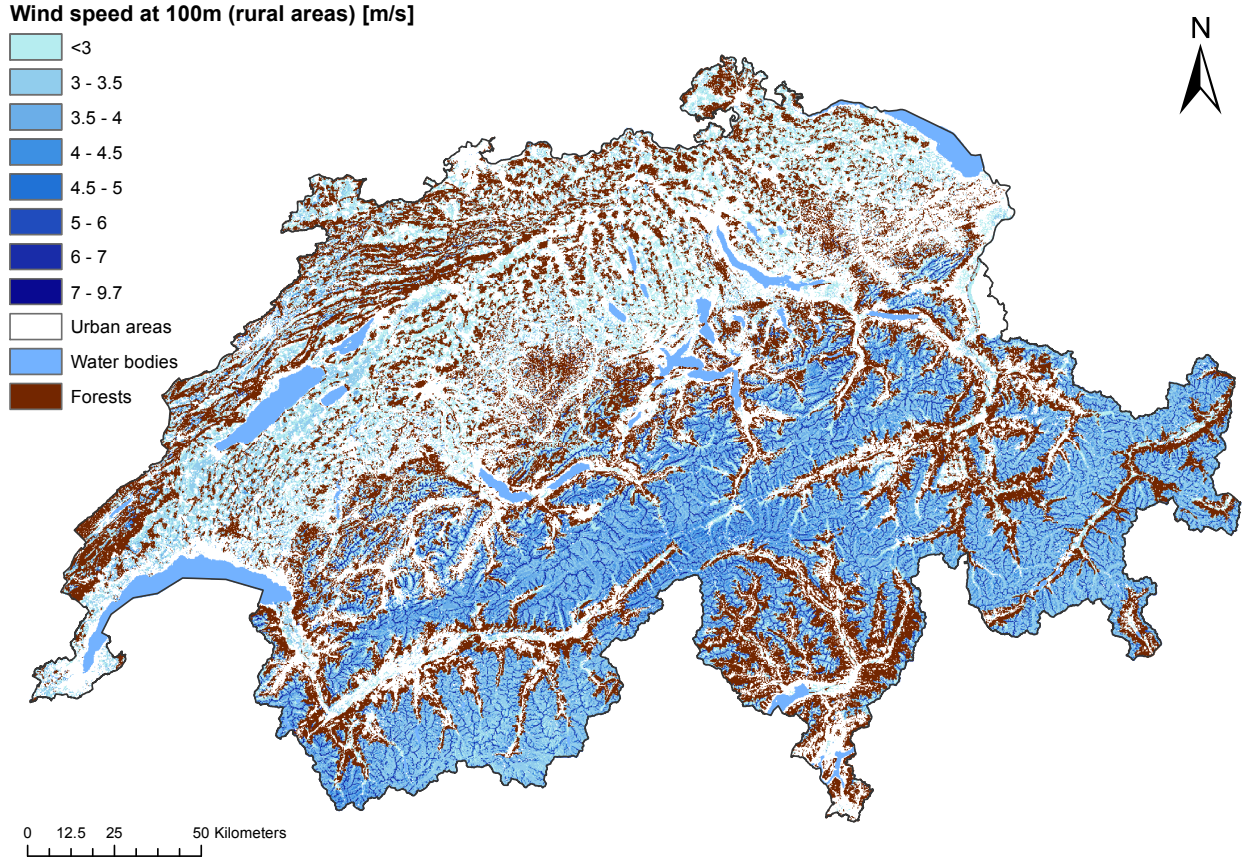


Figure 4.12: Wind speed (yearly average) in rural areas, estimated at a height of 100m, suitable for wind turbine installation.

4.4.1 Urban wind characteristics estimation

In order to use the urban log-law, $z_{0,u}$ and z_d (namely the urban roughness length and displacement height) must be estimated in each urban pixel. These estimations are performed using expressions by Macdonald et al. [61], presented in section 3.1.5 (chapter 3).

$$\frac{z_{u,0}}{h} = \left(1 - \frac{z_d}{h}\right) \exp \left(- \left[0.5\beta \frac{C_D}{\kappa^2} \left(1 - \frac{z_d}{h}\right) \lambda_F \right]^{-0.5} \right) \quad (4.4)$$

$$\frac{z_d}{h} = 1 + A^{-\lambda_p} (\lambda_p - 1) \quad (4.5)$$

where h is the average height of buildings, λ_p and λ_f are the plan area ratio (ratio of total plan/footprint area of obstacles to the total plan area) and frontal area ratio (ratio of total facade area of obstacles to the total plan area), $C_D = 1.2$ is a drag coefficient, A an experimental coefficient and β a parameter. We consider a staggered obstacle array configuration, leading to a choice of $A = 4.43$ and $\beta = 1$ (see section 3.1.5 for more details on Macdonald models and their parameters). Note that λ_f varies based on the wind direction considered when computing the facade area of obstacles. a function of the frontal area and is therefore subject to a direction we consider and for which we compute the facade area of obstacles. Since this chapter aims at computing the average monthly

potential (computed over historical data) for wind at a large scale, it was decided to avoid directional considerations of the wind (as explained in assumption (iii) in the introduction of the chapter). As a result, λ_f and therefore $z_{u,0}$ will be computed in an average fashion, over all directions. The details of the computation are given in the following section.

The use of Macdonald models (Eq. 4.4 and 4.5) requires the computation of the urban roughness $z_{u,0}$ and displacement z_d for each pixel and therefore the computation of h , λ_p and λ_f . It will be the focus of rest of the subsection 4.4.1, using multiple GIS vector and raster processing steps. Notably, the total plan and facade area of obstacles are to be computed in each urban pixel. In addition, the two different zones in Switzerland, as presented in section 4.2.1 (Fig. 4.1), will be treated differently based on the availability of the data.

Extraction of building height and area: Zone 1

In zone 1, two useful building data are available in order to compute the building area ratios and mean heights(details are given in annex A): a precise building footprint polygon data (TLM3D) and a building facade data (swissBUILDINGS3D/Sonnendach), defined as polylines defining each facade. These two datasets will allow for a relatively easy computation of the mentioned urban characteristics. Regarding the computation of the plan area ratio λ_p for each pixel in zone 1, the steps are the following:

- Select the portion of the TLM3D data located in zone 1.
- Split the TLM3D footprint data in 5 separated parts in order to make each part computationally tractable.
- Perform a Joint Spatial operation (using ArcGIS) between the TLM3D vector polygons and the vector pixels in order to extract the building footprints in each pixel. Note that the “HAVE THEIR CENTER IN” option is used while performing the Joint Spatial to consider each building in the pixel containing its centroid and avoid redundancy issues (62 buildings were found to have their center exactly in between pixels and were discarded).
- Sum the footprint area of all buildings within each pixel to obtain the obstacle plan area.
- Compute, for each pixel, $\lambda_p^{\text{zone 1}} = \frac{\text{building plan area}}{\text{total plan area}}$, where the total plan area is the area of the considered pixel, meaning $200 \times 200 = 40000 \text{ [m}^2\text{]}$.

Regarding the computation of the frontal area ratio λ_f for each pixel in zone 1, the steps are the following:

- Cut the building facade data (swissBUILDINGS3D/Sonnendach) in 5 pieces to make each part easily computationally tractable. Each vector line in the data contains the facade area of the corresponding wall and its length.
- Perform a Joint Spatial operation (using ArcGIS) between the facade vector lines and the vector pixels in order to extract the building facade information in each pixel, and average the areas to obtain the average facade area of a wall in each pixel. Note that the “CONTAINS CLEMENTINI” option is used while performing the Joint Spatial, to consider solely the facades which are fully englobed by pixels.

- Compute, for each pixel, $\lambda_f^{\text{zone } 1} = \frac{\text{obstacle facade area}}{\text{total plan area}} = \frac{\text{average wall facade area} \times \text{number of buildings}}{\text{total plan area}}$.

The computation of the mean building height h for each pixel in zone 1 is based on the previously extracted facade area and the length of the walls given by the facade area. Given the average wall facade length and area in each pixel, and assuming the walls are rectangular, the average building height is simply given by $h^{\text{zone } 1} = \frac{\text{average wall facade area}}{\text{average wall length}}$.

Extraction of building height and area: Zone 2

In zone 2, while the building footprint data (TLM3D) is still available, the building facade data used for zone 1 is not available. It is therefore required to adopt a modified strategy to extract the average building facade area in each pixel. This strategy involves the use of the DOM (Digital Orthophoto Map), offering an altitude raster with all obstacles considered (buildings, trees etc.), together with the DTM (Digital Terrain Map), offering an altitude raster for the terrain alone, which are both available for the whole country. Processing steps based on the subtraction of the DTM from the DOM allows an estimation of the average building height in each pixel, which, combined with the average wall length, gives an estimation of the average facade area.

The computation of the plan area ratio $\lambda_p^{\text{zone } 2}$ for each pixel in zone 2 is identical to its estimation in zone 1, with the only difference that the building polygons from TLM3D data are selected in zone 2 (first point in the estimation of $\lambda_p^{\text{zone } 1}$).

The extraction of the mean building height h for each pixel in zone 2 precedes the computation of the frontal area ratio. The computation steps are the following:

- Extract the DTM and DOM for pixels located in zone 2. They both have a 2×2 [m²] resolution.
- Compute the subtraction DOM-DTM to obtain a 2×2 [m²] obstacle height raster. Note, however, that it considers all obstacles at this stage and not only the buildings. The next following steps are performed in order to extract the height raster values solely over the buildings.
- Select TLM3D building polygons located in zone 2.
- Convert the TLM3D buildings into a 2×2 [m²] raster (using the “From polygon to raster” conversion tool from ArcGIS).
- Adjust the TLM3D raster so that it matches the raster cells of the DOM-DTM height raster (using the “Raster clipping” tool from ArcGIS, with the same clip extent as the DOM-DTM raster).
- Extract the DOM-DTM raster values at the TLM3D raster locations by combining them in the same raster calculation (using the Raster calculator from ArcGIS). For example, the raster calculator input $(\text{TLM3D} * 0 + 1) * (\text{DTM} - \text{DOM})$ will result in the height raster extracted over the buildings. Finally, we obtain a 2×2 [m²] building height raster.
- Average the building height raster within each 200×200 [m²] pixel (using the Zonal Statistics tool from ArcGIS) to obtain the average building height $h^{\text{zone } 2}$.

- To avoid mismatch issues between the DOM and the TLM3D buildings (some buildings present in the TLM3D data are not depicted in the DOM, which was not updated recently) that could cause a negative or abnormally small height value, we apply a $h > 1$ filter. Therefore, we do not consider pixels for which the estimated mean height is lower than 1m, as this configuration happens mostly in case of mismatch.

The computation steps for the extraction of the mean building height h raster in zone 2 are summarized and illustrated in Figure 4.13.

The computation of the frontal area ratio λ_f for each pixel in zone 2 is based on the previous estimation of h and the average wall length extracted from the TLM3D data. The computation steps are the following:

- Extract the length of building walls from the TLM3D buildings selected in zone 2 (using the “Split Line At Vertices” tool from ArcGIS over the TLM3D footprints, tool which splits polygons at their vertices and automatically computes the length of the obtained lines, which are by construction the walls of the buildings).
- Perform a Joint spatial operation (using ArcGIS) between the TLM3D walls with the pixels in order to extract the average wall length in each pixel.
- Compute for each pixel the average facade area as the product of the average building height $h^{\text{zone 2}}$ and the average wall length.
- Compute, for each pixel, $\lambda_f^{\text{zone 2}} = \frac{\text{average facade area}}{\text{total plan area}}$

Computation of wind urban characteristics

Using Macdonald models presented in Eq. 4.4 and 4.5, and h , λ_f and λ_p previously estimated over the whole country, we compute the roughness length and displacement $z_{0,u}$ and z_d in all urban pixels. Note that, as mentioned in section 4.4.1, the urban roughness is computed as an average in all directions. Also, the estimation lead, in some rare cases, to values of z_d higher than the average height h , a situation which should not happen in theory; the few pixels showing this issue are therefore discarded. The maps obtained for these two variables are shown in Figure 4.15.

4.4.2 An important assumption

As mentioned earlier, we assume that for a certain altitude significantly high above the buildings and within the ISL, the wind velocity observed in a rural area is equal to the wind velocity in an *adjacent* urban area. As explained at the end of section 4.3.2, the height of $z = 100\text{m}$ is suitable for such an assumption and the horizontal homogeneity assumption further validates that the horizontal changes can be neglected between two adjacent pixels. Therefore, we can write:

$$\text{For adjacent rural and urban areas, } u_r(100) = u_u(100) \quad (4.6)$$

We wish to express the statement defined in Eq. 4.6 using pixels. A pixel *adjacent* to a second pixel can be defined as a pixel in the neighborhood of the second pixel. Using the definition of the Moore neighborhood, previously used in section 4.2.2 (and illustrated in Figure 4.4) and defining (i) a

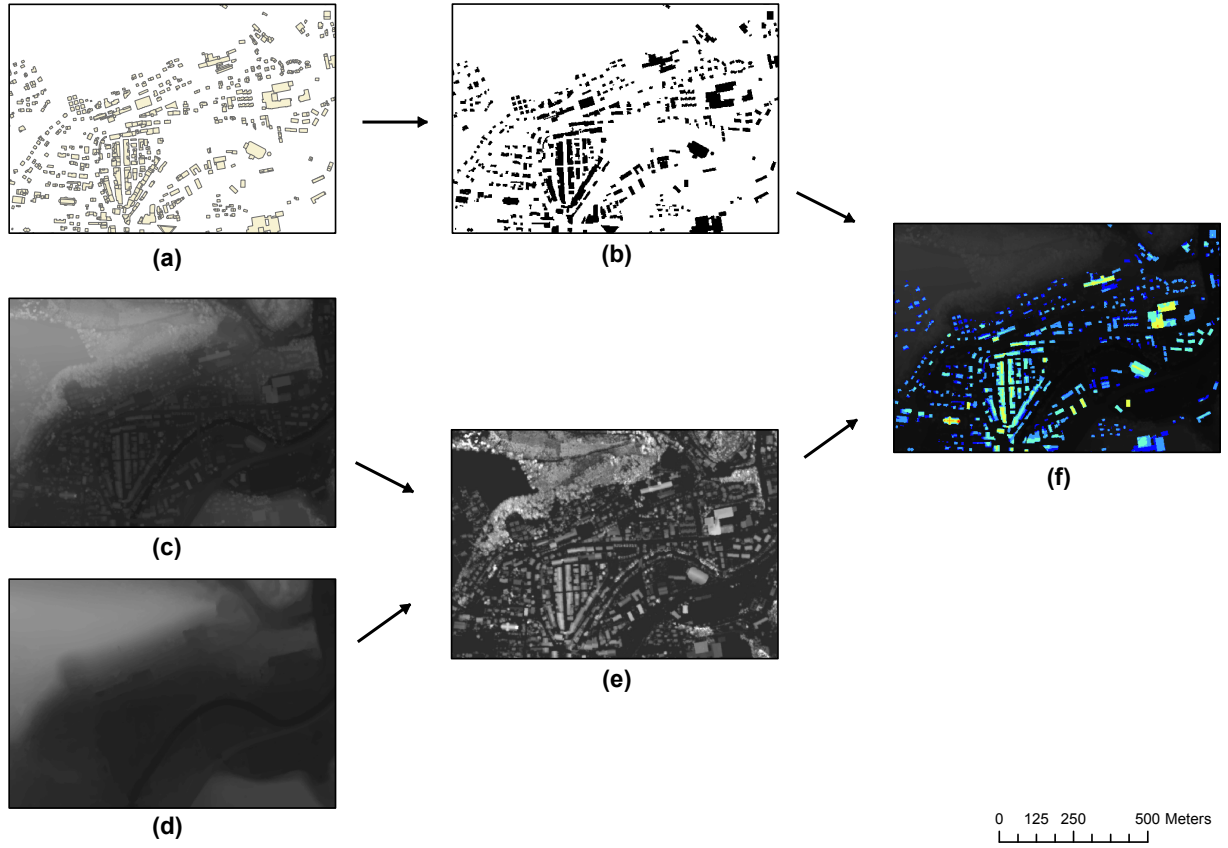


Figure 4.13: Processing scheme for the extraction of the average building height in zone 2, for an example area in Switzerland. **(a)** TLM3D building footprints, **(b)** result of the conversion of the TLM3D footprints into 2×2 [m²] raster cells, **(c)** DOM, **(d)** DTM, **(e)** result of the subtraction DOM-DTM raster computation, **(f)** final mean building height raster, obtained by clipping the DOM-DTM raster **(e)** over the TLM3D raster cells **(b)**.

rural area as a rural pixel, and (ii) an urban area adjacent to a rural area as an urban pixel which has at least one rural pixel in its neighborhood, we can assume that urban pixels have the same wind velocity at 100m than rural pixel in their neighborhood. These neighbor rural pixels, however, may be multiple. Therefore, the most natural way to extract the wind speed for such an urban pixel is to average the wind speed over all its neighbor rural pixels. For simplicity, let us call such an urban pixel with at least one rural neighbor an *urban-boundary* pixel. Following our reasoning, we can write:

$$\text{For an } \textit{urban-boundary} \text{ pixel, } u_u(100) = \frac{\sum_{l=1}^L u_r^{n_l}(100)}{L} \quad (4.7)$$

where L is the number of rural neighbors of the urban boundary pixel, n_l is its l^{th} rural neighbor and $u_r^{n_l}(100)$ is the wind speed estimated in the rural neighbor n_l . Note that L can be any natural number from 1 to 9, depending on the configuration of the urban and rural pixels. An example of urban and rural pixels possible configuration is given in Figure 4.16. The urban wind velocity at 100 m ($u_u(100)$) is therefore computed for all urban-boundary pixels in Switzerland using Eq. 4.7.

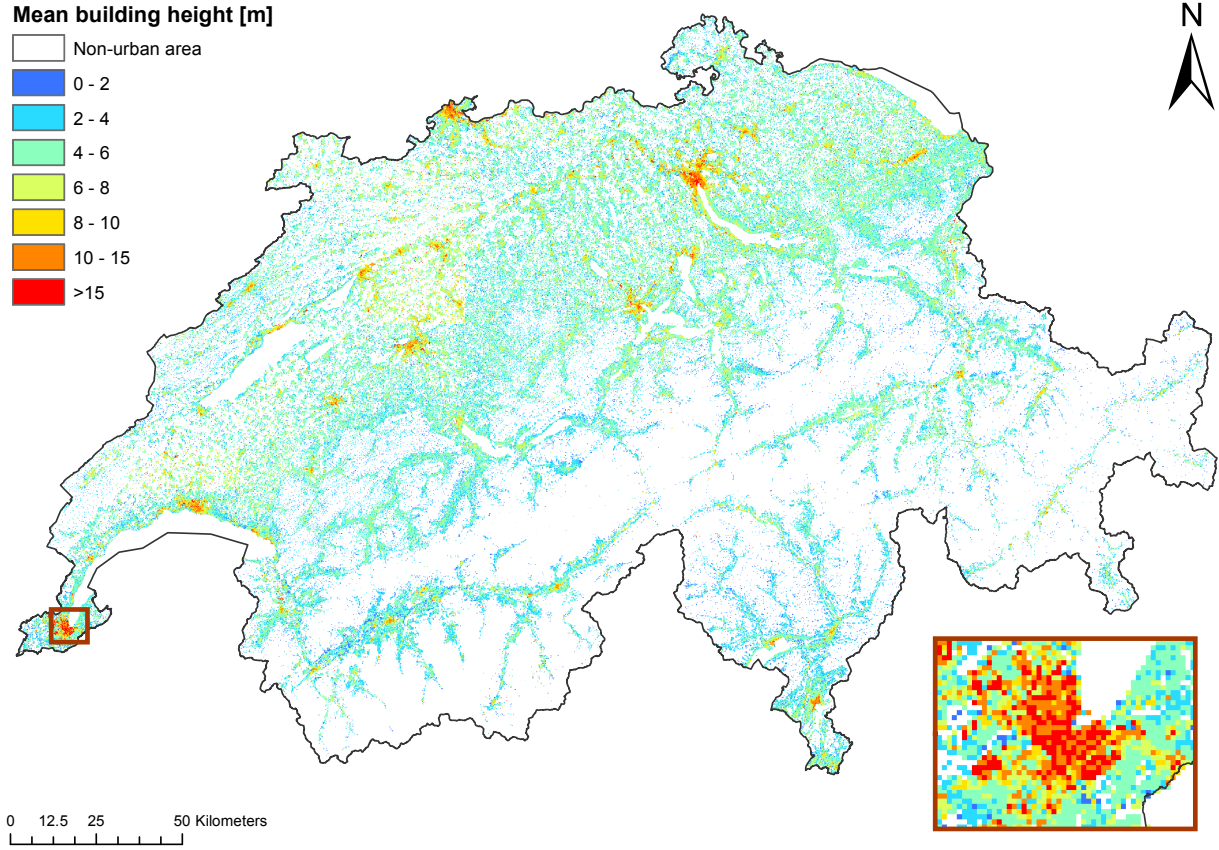


Figure 4.14: Mean building height map. The small window on the bottom right of the figure zooms in Geneva city.

4.4.3 Extrapolation above buildings

To obtain the wind velocity above buildings in urban boundary pixels, the wind velocity at 100 m is used together with the urban log-law, defined in Eq. 4.3. The same way we obtained Eq. 4.2, taking the ratio of Eq. 4.3 with the same log-law evaluated at $z = 100$ gives $u_u(z)$ as a function of $u_u(100)$, for any urban pixel j :

$$u_u(z) = u_u(100) \frac{\ln\{(z - z_d)/z_{0,u}\}}{\ln\{(100 - z_d)/z_{0,u}\}}, \text{ for } z \in \text{ISL} \quad (4.8)$$

where we omitted the pixel indexes j in order to simplify the notations.

Note that while Eq. 4.8 theoretically gives the wind speed at any height z within the ISL (and such that $z > z_d + z_{0,u}$ by construction), it is nonetheless possible to consider heights relatively close to buildings. Oke et al. [59] showed that often the log-law offers a good estimation approximately above $1.5h$ (as seen in Figure 3.3 within section 3.1.2, in chapter 3) even though we are still located within the RSL. Nevertheless, for pixels with a high average building height ($h > 10\text{m}$) it is unrealistic to install wind turbines at $z = 1.5h$ as it becomes quickly too high for turbine installation. Therefore, we compute the speed above buildings at a height of $z_{\text{hub}} = h + 5\text{m}$ in order to be sufficiently far from a possible turbulent behavior (98% of urban pixels in Switzerland show $h < 10$ and therefore $h + 5 > 1.5h$), while

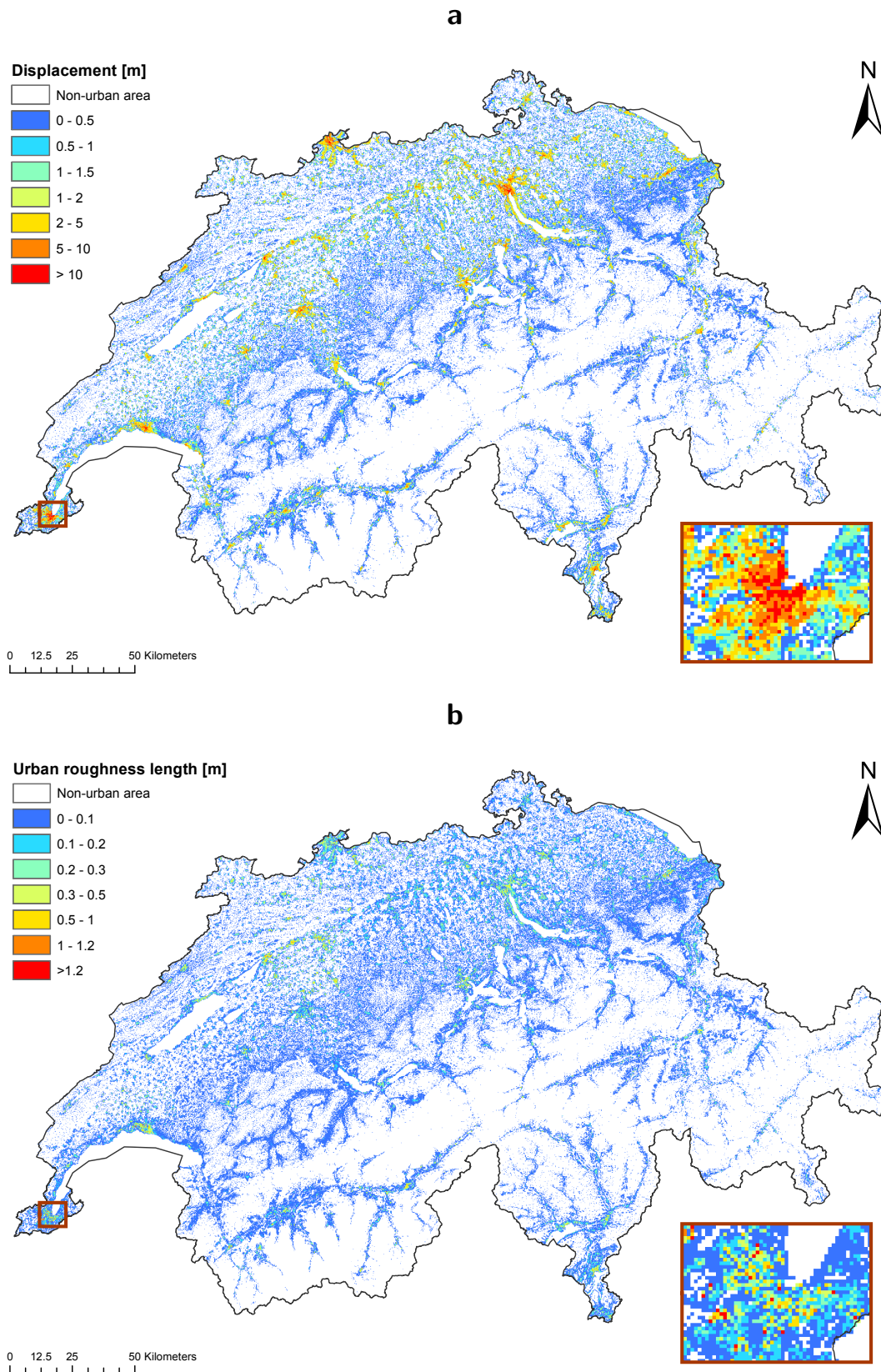


Figure 4.15: Urban characteristics estimated in urban pixels. **(a)** Displacement height estimated in urban boundary pixels, **(b)** roughness length estimated in urban boundary pixels. The small windows on the bottom right of the two figures zoom in Geneva city.

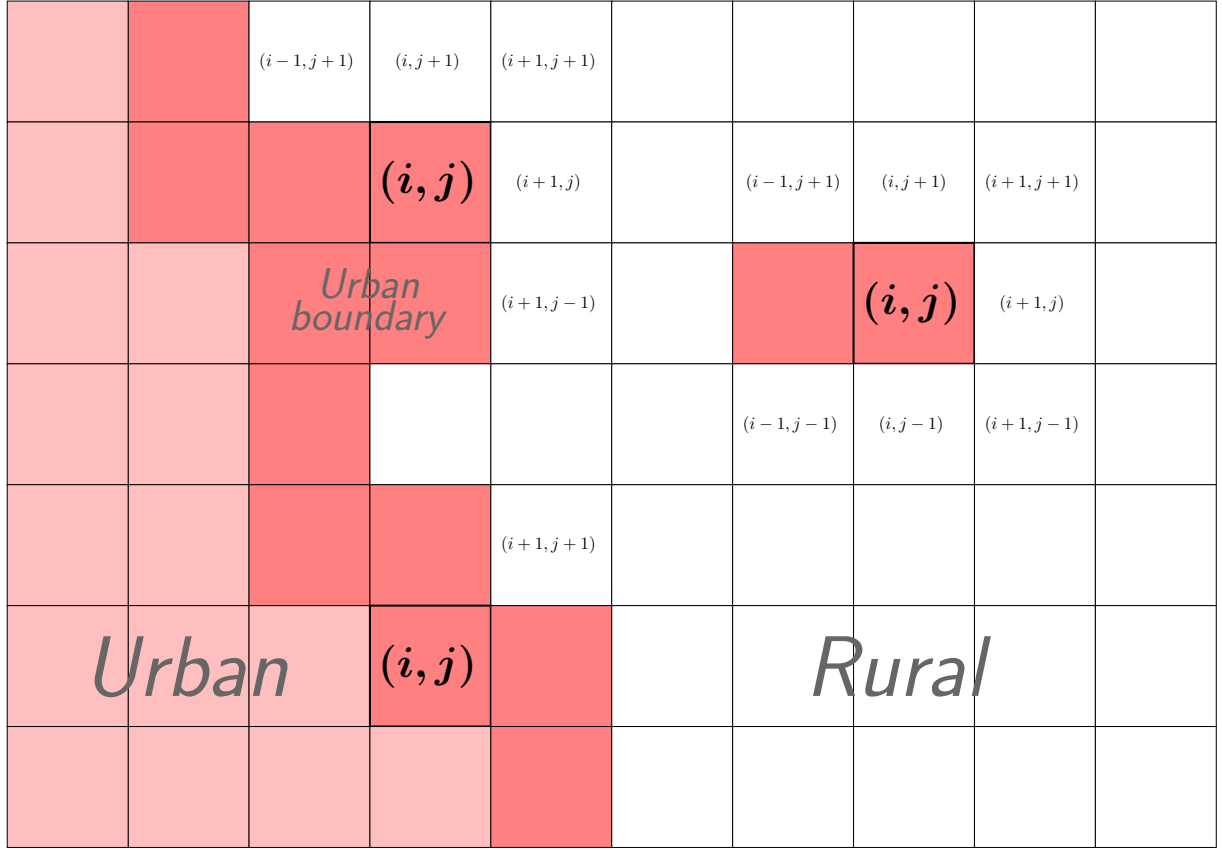


Figure 4.16: Example of pixel configuration in Switzerland: white pixels are rural pixels, light red pixel are urban but not in the boundaries of urban areas, dark red pixels are urban pixels in the urban boundaries (meaning neighbors to at least one rural pixel). Three examples of urban boundary pixels are highlighted with their position (i, j) together with their respective rural neighbors considered when computing $u_{ul}(100)$ as explained in section 4.4.2. The rural neighbors are highlighted in light gray with their position with respect to the urban boundary pixel considered at (i, j) .

considering a realistic height for wind turbines installation. Note that $z_{hub} = h + 5m$ is a standard height considered to estimate the potential of small wind turbines in the literature (e.g. in [72]), as it is generally considered to be high enough to avoid turbulence issues. The obtained map for wind speed above buildings (at a altitude of z_{hub}) in urban boundary pixels is shown in Figure 4.17.

It should finally be noted that the properties of the ISL layer and the horizontal homogeneity assumption, used in the present chapter to deduct the urban boundary wind speed from the neighbor rural wind speed, can only be used locally and does not allow for a reliable estimation of the wind speed within the urban centers. Theoretically, one could invoke these assumptions to recursively estimate the wind speed of any urban pixel by averaging the wind speed of its neighbor pixels previously determined and therefore explore the wind speed behavior within the centers of urban areas. It would however raise the question of the order of the considered initial conditions: since it is a recursive process, the order in which the wind is estimated in the different pixels will change the results, which would then need a physical justification in order to be validated. Eventually, using the similar strategy to the one we use to extract the wind potential in urban centers raises an interesting propagation problem which falls within the domain of cellular automata theory and ultimately outside of the scope of this thesis. The estimation of the wind speed above buildings is here therefore limited to

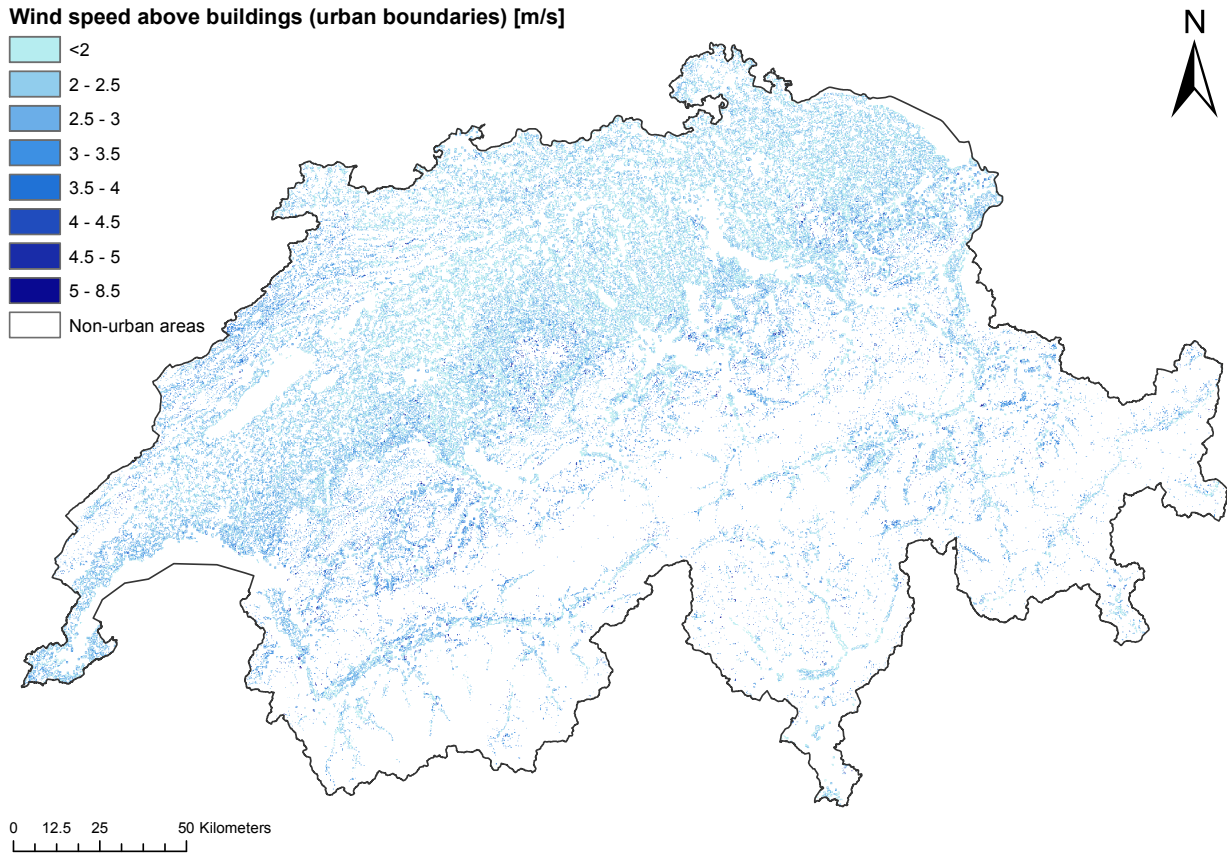


Figure 4.17: Wind speed (yearly average) in urban boundary pixels, estimated at 5 m above the mean building height (at $z_{\text{hub}} = h + 5$).

urban boundaries. The extension of the potential study to the center of urban areas is nevertheless an important topic that would require further research. To tackle this topic, one possible strategy of interest for future work is the use of deterministic meteorological models which are valid in urban areas, allowing to obtain large amounts of data (through simulations) which could later be used as training data for a machine learning model to learn wind speed patterns within urban centers. Such meteorological models include for example the WRF model [169, 170].

4.5 Results

The theoretical wind energy potential represented by the estimation of wind speed both in rural areas and in boundary urban areas, and at reasonable height for wind turbines installation ($z = 100\text{m}$ for rural areas, and $z = z_{\text{hub}} = \text{mean building height} + 5\text{m}$ for urban areas), has been assessed over Switzerland. The yearly wind speed map (obtained by averaging the monthly maps) for both rural and urban areas is illustrated in Figure 4.18. Note that the determination of wind potential based on the sole wind speed is useful because it gives the the flexibility of the choice of the turbine to be installed to extract the actual electricity from the wind. It should be remembered, however, that the speed is naturally not the only factor impacting the electricity generated from a turbine (as

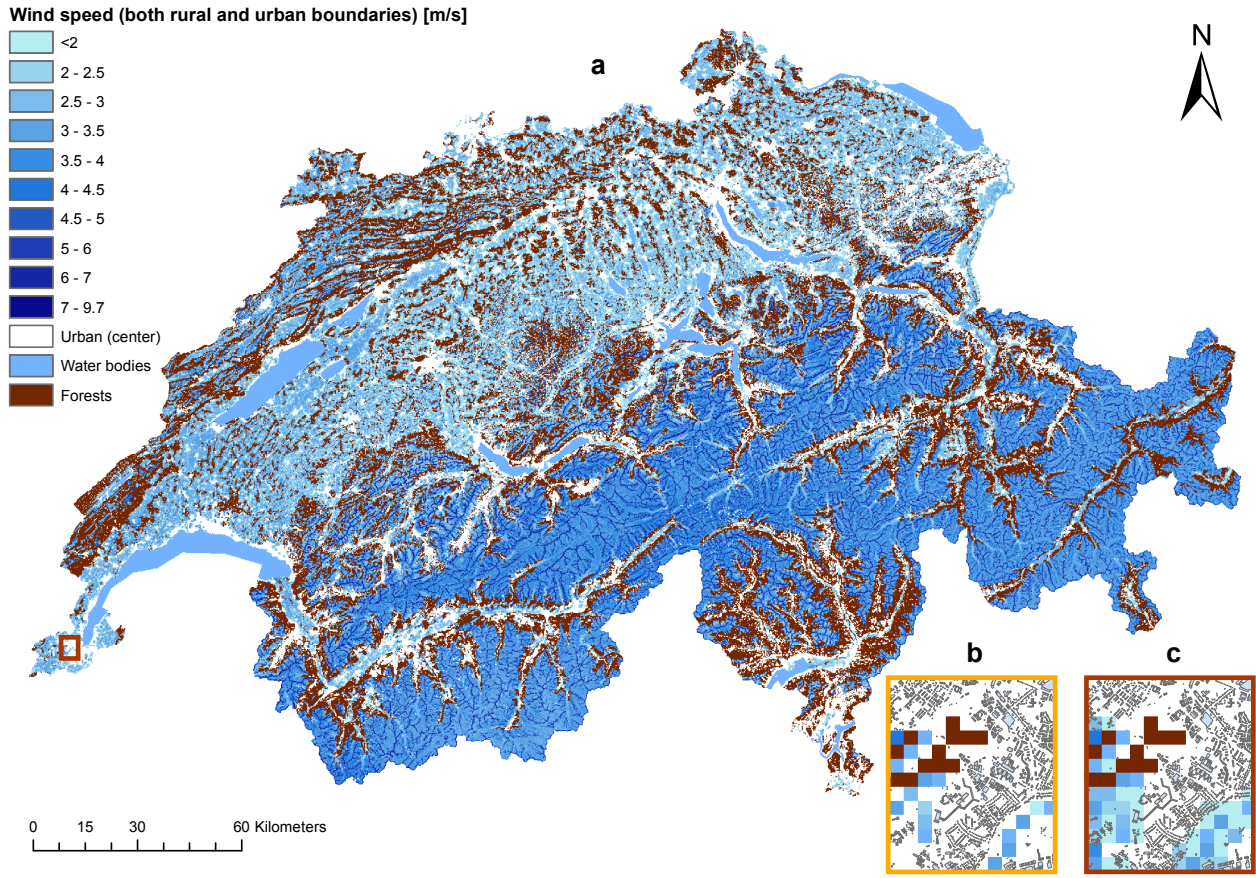


Figure 4.18: Wind speed (yearly average) in both rural areas (at a height of 100m) and urban boundaries (at a height of 5 m above buildings). **(a)** Wind speed map. **(b)** Zoom in Geneva city, showing the wind speed in rural pixels only (not the urban boundary pixels). **(c)** Zoom in Geneva city, showing the wind speed in rural and urban boundary pixels.

explained in chapter 3, section 3.1). Another very important factor is the diameter of the rotor of the turbine, which dramatically increases the gap between the rural and urban potentials for wind, given the difference of turbine size typically installed in the two settings. While a 100m high commercial turbine often has a rotor with an average diameter of 100m, an urban small turbine above buildings will have a much smaller diameter, ranging from 1 to 10 meters. The general urban potential is therefore considerably smaller than that of the rural areas.

To be able to assess a measure of the output electricity potential from wind turbines in Switzerland, the power generated by typical wind turbines is estimated, for both rural and urban areas, using the expression presented in section 3.1:

$$P_w = C_{p,w} \frac{\rho A_w u^3}{2} \quad (4.9)$$

where P_w is the wind turbine power (in W), $C_{p,w}$ is the coefficient of performance of the turbine, expressing its efficiency, ρ is the air density, A_w is the area spanned by the rotor of the turbine, and u is the wind speed, supposedly in a perpendicular direction with respect to the rotor. Eq. 4.9 can be used to determine the average yearly power produced by a wind turbine in rural and urban

areas, based on the wind speed values we estimated all over Switzerland. Let us consider typical turbine characteristics for rural and urban areas [63]: (i) a HAWT (horizontal axis wind turbine) rotor diameter of 100m for rural areas (large commercial turbine); and a small HAWT with a diameter of 3m for urban areas (household turbine) and (ii) an air density considered at 1.2 kg/m^3 at both 100m height and above buildings since it does not vary significantly from sea-level to an altitude of 100m. The coefficient of performance $C_{p,w}$ requires more care, as it changes dramatically with the wind speed, depending on the type and quality of the wind turbine. Note that the choice of a horizontal axis wind turbine rather than a vertical one for urban areas, even though the latter are getting more attention, is discussed in section 3.1.6 of chapter 3.

For $C_{p,w}$ reference in rural areas, let us consider a common wind turbine available in the industry, the ENERCON E-101, with a rotor height of 101m, a rotor diameter of 100m, and a power/ $C_{p,w}$ curve shown in Figure 3.4 (as presented in the following website: <https://www.enercon.de/en/products/ep-3/e-101/>). In the other hand, for urban areas, let us consider the curve offered by a typical small HAWT. As the $C_{p,w}$ curve is rarely given for such small turbines (the sole power curve is given in most cases), we consider the curves obtained within a study on small turbines for lows winds, by Singh and Ahmed [75]. It appears that the typical shape of the $C_{p,w}$ curve is relatively close to large HAWT curves, but with an optimal achieved $C_{p,w}$ of 0.25 to 0.3 instead of 0.45). This is notably confirmed in [63, 171]. A reproduced $C_{p,w}$ curve for small HAWT is shown in 4.19b.

In order to extract the discussed $C_{p,w}$ for any wind speed, however, a continuous function is required rather than discrete values. Given the shape of the $C_{p,w}$ curves, we fit the curves with a polynomial (degree 6 for the rural C_p curve, degree 3 for the urban C_p curve), as shown in 4.19a and 4.19b. Note that, for the rural case, even though the relatively high degree of the fitted polynomial results in a fluctuating power curve for high wind speed values ($\geq 14 \text{ m/s}$), it does not impact the power calculation since our wind speed estimated values are maximized by 10 m/s across the country. The obtained polynomials $C_{p,w}$, functions of wind speed, are then both used together with the discussed turbines characteristics for rural and urban areas in order to compute the yearly average potential wind power in each pixel of Switzerland. The obtained wind power map is shown in Figure 4.20.

4.5.1 Discussion

The yearly wind speed potential has been aggregated within the 26 cantons in order to have a global view of the differences of potential through the country, in urban and rural areas. Figure 4.21 shows the aggregated estimated speed values for three aggregation levels: (i) average the pixel values in each canton (Figure 4.21a), (ii) consider the pixel with the maximum estimated speed in each canton (Figure 4.21b), (iii) consider the range offered by the pixels in each canton (Figure 4.21c). The overall largest potential is located in the Alps region, in the Valais and Ticino canton, as well as in Uri and Glarus. In the average case, the obtained speed values are significantly higher in rural areas, with a general average of 3.5 m/s , as the average building height is naturally lower than 100m in most urban pixels, where the average speed is around 2.5 m/s . In the maximum case, the difference slightly diminishes, with average maximum of 8 m/s and 6 m/s respectively for rural and urban areas; the corresponding range of speed is relatively similar for both settings. As expected, although there are some high potential and low potential pixels in each canton, for both rural and urban areas, the speed potential is naturally higher in rural areas, where turbines can be installed systematically higher than in urban areas.

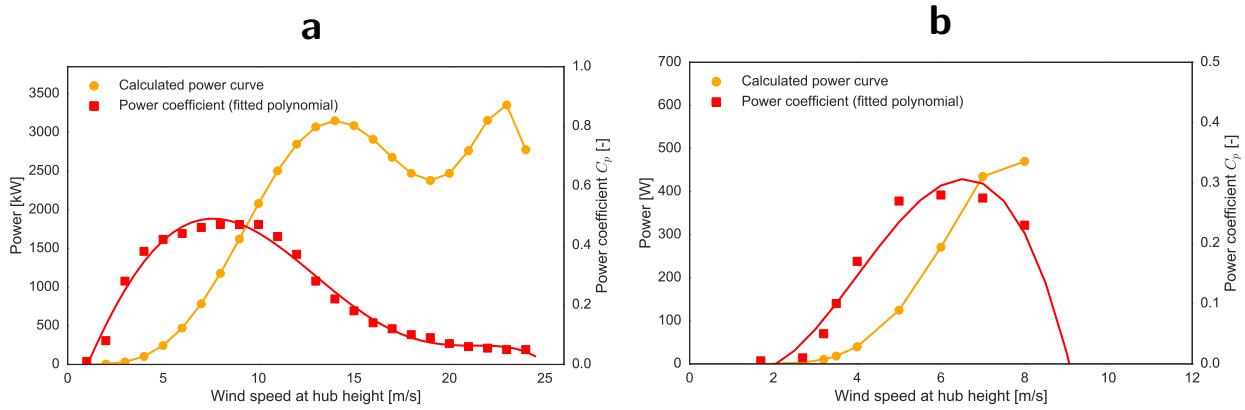


Figure 4.19: $C_{p,w}$ /power curves for typical turbines suitable for rural and urban areas. **(a)** Considered $C_{p,w}$ /power curve for a typical turbine suitable for rural areas. We fit a polynomial over typical power coefficient $C_{p,w}$ values for a large commercial horizontal axis turbine (Reproduced from <https://www.enercon.de/en/products/ep-3/e-101/>, for the ENERCON E-101 product), and recompute the power curve based on fitted polynomial (degree 6) for $C_{p,w}$. **(b)** Considered $C_{p,w}$ /power curve for a typical turbine suitable for urban areas. We fit a polynomial over typical power coefficient $C_{p,w}$ values for a small horizontal wind turbine (Reproduced from [75]), and recompute the power curve based on fitted polynomial (degree 3) for $C_{p,w}$.

The previously computed wind power potential, for typical urban and rural turbines, is also aggregated within cantons, for both the average and maximum potential available in each canton; the result is shown in Figures 4.22 and 4.23. It can be observed that in case of urban pixels, the order of magnitude of the estimated power is significantly smaller than that of the rural pixels, given the difference in size and efficiency for large and small turbines. While the rural power potential can reach 1.6 MW in a pixel (on average 80 kW), the urban power cannot go further than 1.1 kW (on average 15W). Note, however, that the potential power in urban areas is not negligible. The average time for a wind turbine operation over the year is approximately 1600 hours (computed by averaging the ratio of Swiss generated electricity to the installed power from 2010 to 2017 [6]). Therefore, an average rural installation of 80 kW can provide 128 MWh over the year. Assuming the same time of operation in urban areas (even though it may be higher), a small household turbine can also provide a non-negligible amount of renewable energy throughout the year: a highly potential pixel with an installed power of 1 kW and a yearly production time of 1600 hours amounts to a yearly generation of 1600 kWh for one turbine, which corresponds to 31% of the average household yearly electricity demand in Switzerland in 2016 [172].

4.5.2 Preliminary geographical potential estimation

In order to have a preliminary estimation of the overall geographical wind potential over Switzerland, we can further estimate the number of potential turbines installation. Regarding rural areas, we can use two rules of thumb [173–175]: (i) turbines must be installed at a distance of 150m from urban areas and (ii) for wind farm installation, the turbines must be at seven rotor diameters away from each other, meaning 700m within our study. The first rule is embedded within ArcGIS using a Buffer of 150m around urban boundary pixels, which allowed to select 146'350 rural pixels for turbine installation. The second rule corresponds to a requirement of at least three pixels between two turbines, in all directions,

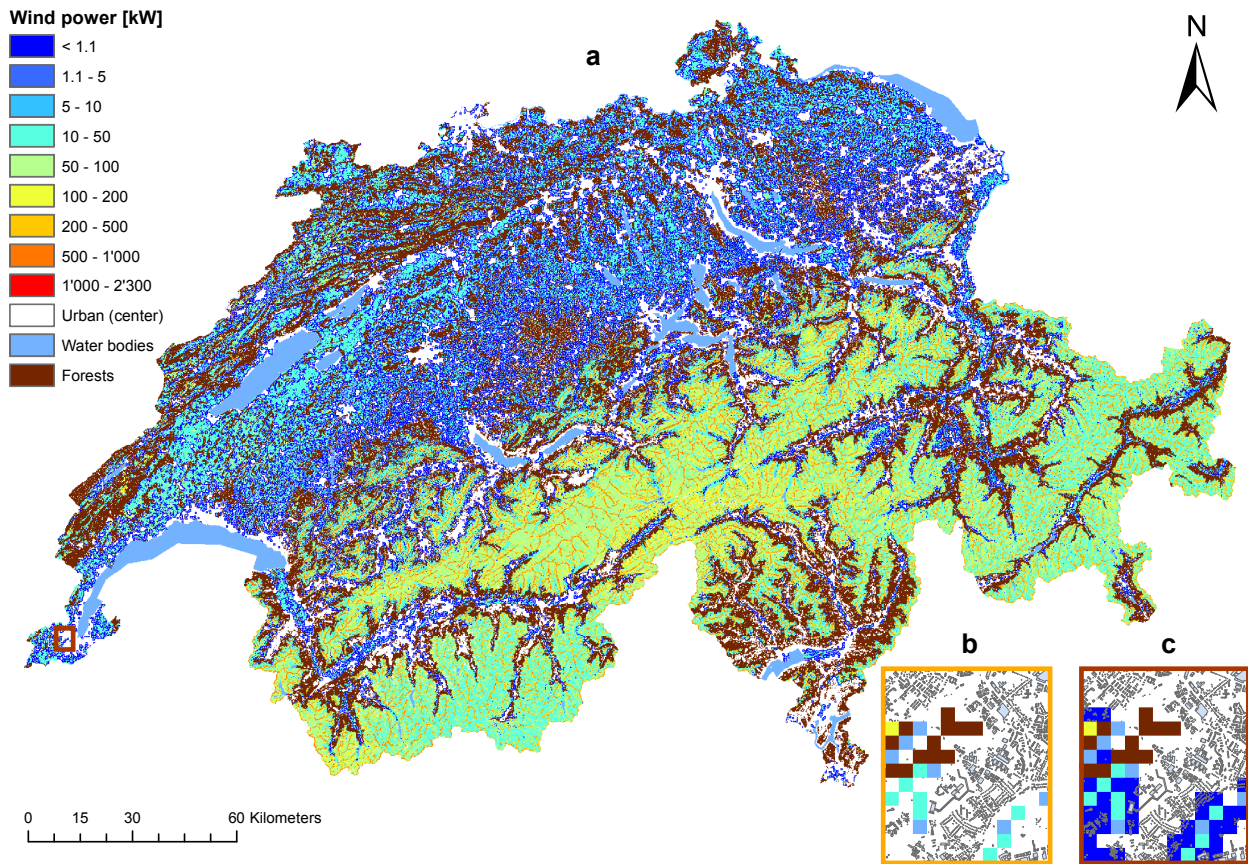


Figure 4.20: Wind power (yearly average) potential map, computed based on the estimated wind speed and typical turbine characteristics. **(a)** Wind power map. **(b)** Zoom in Geneva city, showing the power in rural pixels only (not the urban boundary pixels). **(c)** Zoom in Geneva city, showing the power in rural and urban boundary pixels.

or an average of one turbine for sixteen pixels, which amounts of a total number of 9147 available rural pixels. With an average generation of 128 MWh per rural turbine per year, it results in a crude technical rural potential of 1.17 TWh. The installed Wind capacity of 133 GWh in 2017 (as given by SFOE [6], presented in the chapter introduction, section 1.2.1), therefore corresponds to 11 % of the latter estimated technical potential. Regarding urban areas, the average calculated yearly generation of 24 kWh per small turbine can be multiplied by the 174'757 urban boundary pixels to obtain a conservative estimation of the potential generation within urban boundary areas, with only one small roof-mounted turbine per pixel. It amounts to a “urban boundary potential” of 4.2 GWh per year.

4.5.3 Validation with other potential studies

To validate our estimations, a comparison with existing wind values or wind maps would be desirable. While no wind speed map exists for urban areas in Switzerland, such maps are available from Swisstopo for rural areas, at multiple heights. In particular, we are interested in the yearly wind estimation at a height of 100m (available freely from the following link: https://data.geo.admin.ch/ch.bfe.windenergie-geschwindigkeit_h100/). The estimation is based on existing monitored data to tune

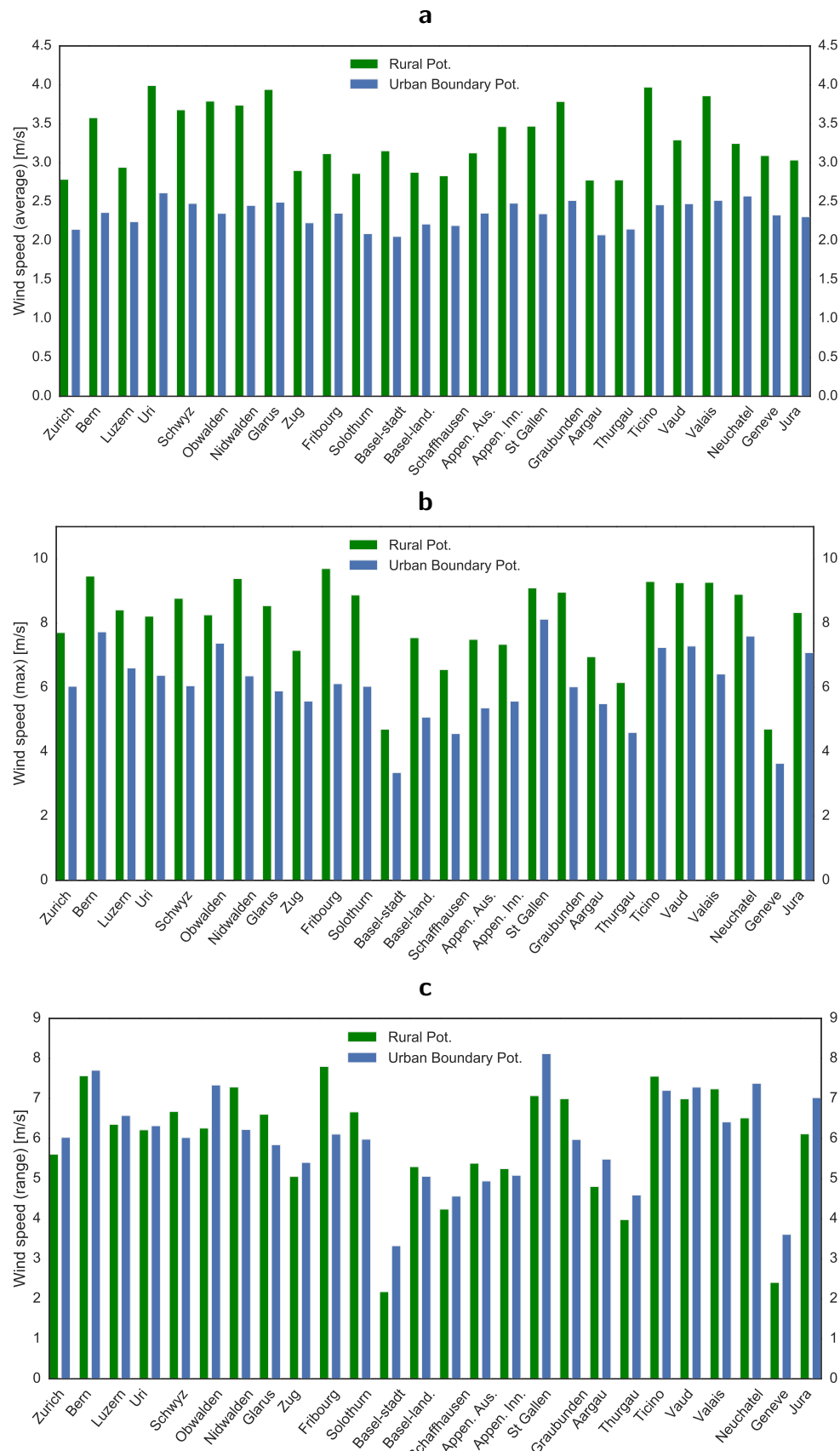


Figure 4.21: Estimated wind speed (yearly average) for both rural and urban boundary areas, aggregated within Switzerland cantons. **(a)** Average canton wind speed, **(b)** Maximum canton wind speed, **(c)** Range of wind speed in each canton.

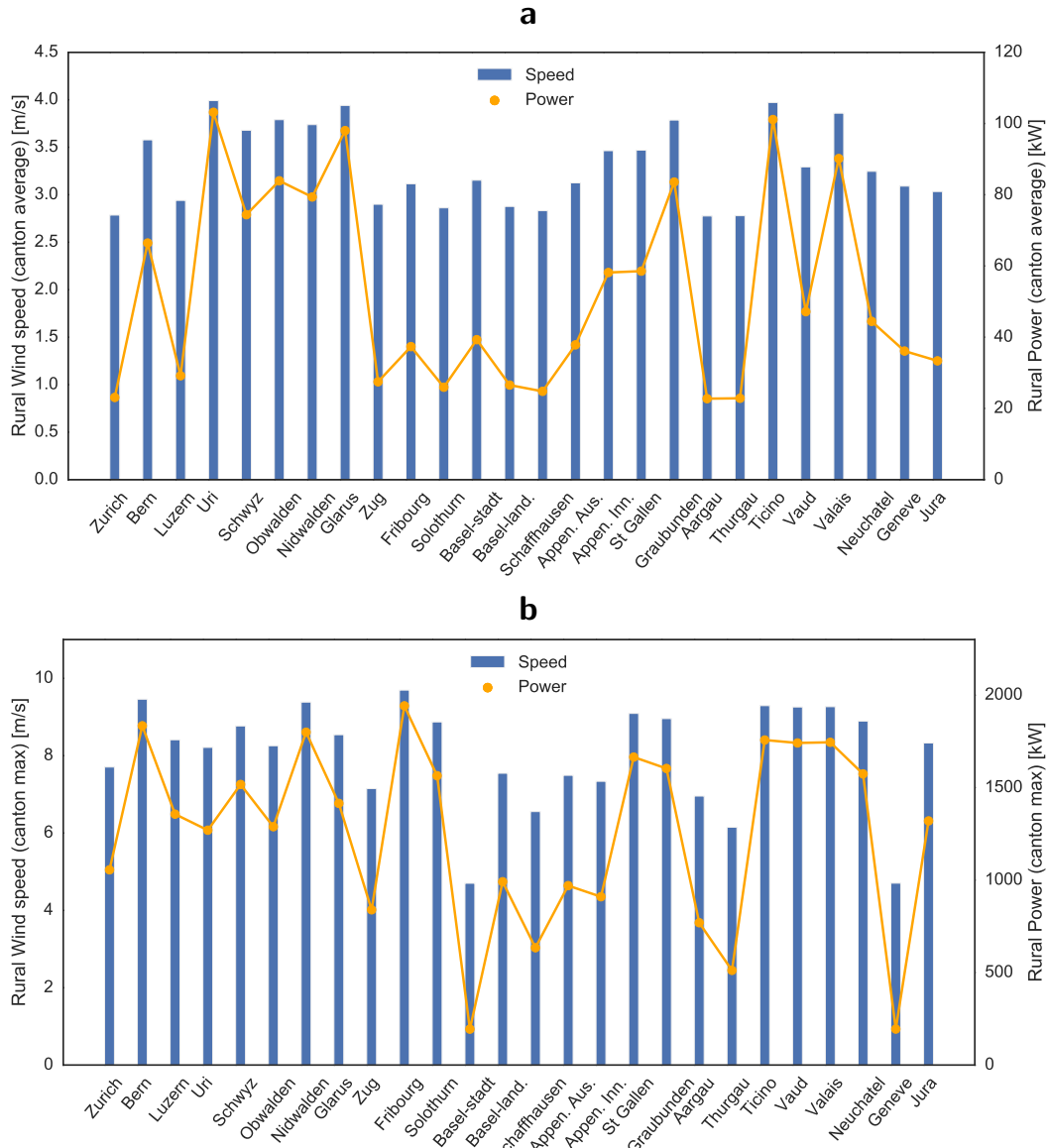


Figure 4.22: Estimated power (yearly average) aggregated within Switzerland cantons, for rural areas (considering a typical large commercial turbine), along with the estimated wind speed. **(a)** Average canton wind speed, **(b)** Maximum canton wind speed

the parameters of the Weibull probability distribution, known to provide a good fit for a typical wind distribution. Although the estimated values from Swisstopo are given for 100×100 [m²] mesh cells which do not match the grid of 200×200 [m²] pixels considered in the present study, it is possible to aggregate the Swisstopo values within the latter pixels by simply averaging the wind speed estimated in the 4 Swisstopo cells whose centroids are the closest to the each pixel. Furthermore, the available wind speed values from Swisstopo at 100m are attached with an error which varies from ± 0.5 m/s to ± 1.5 m/s, depending on the location. A comparison between the wind speed estimated at 100m from Swisstopo and our estimation, across Switzerland, leads to an RMSE error of 1.6 m/s, and an NRMSE of 35% (considering the Swisstopo estimations as ground truth). Given the errors attached

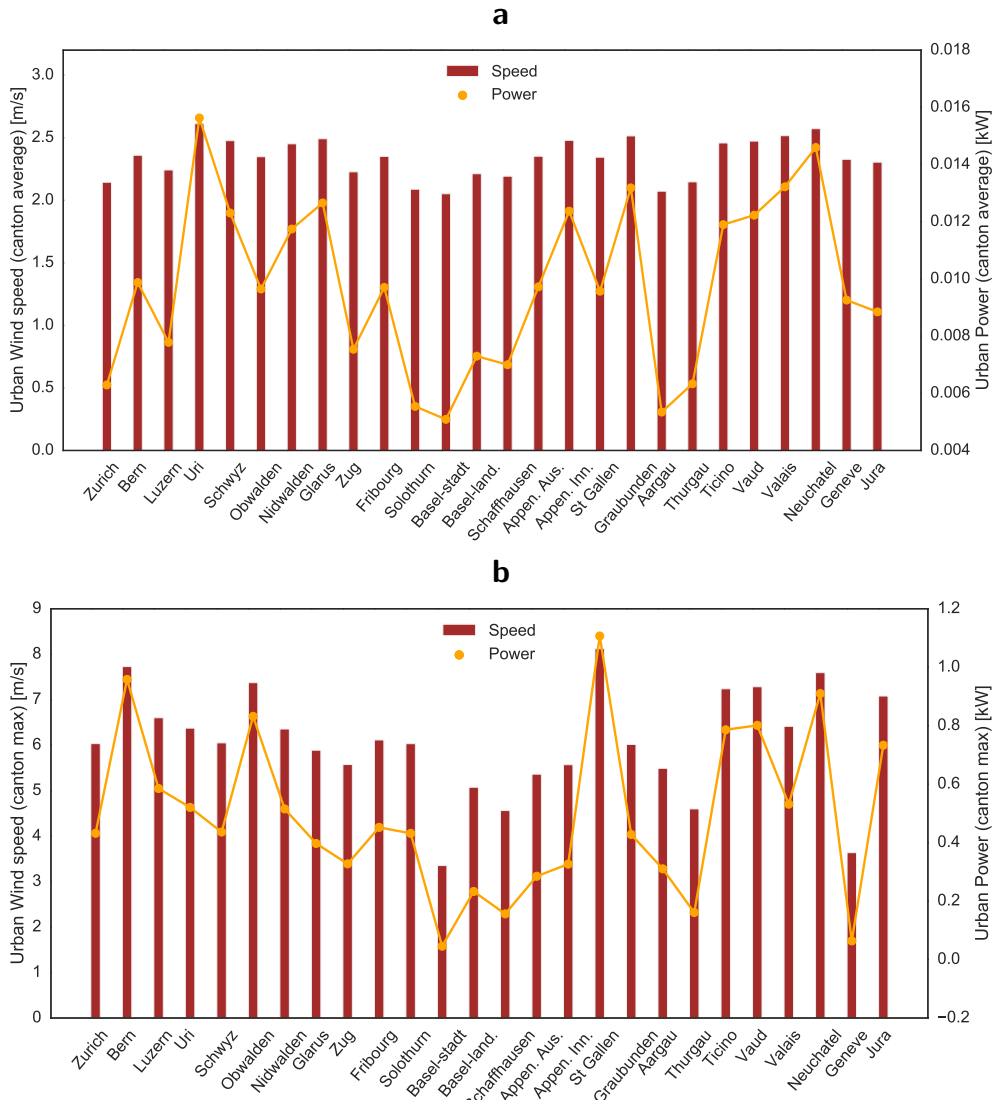


Figure 4.23: Estimated power (yearly average) aggregated within Switzerland cantons, for urban areas (considering a typical small household turbine), along with the estimated wind speed. **(a)** Average canton wind speed, **(b)** Maximum canton wind speed

to the Swisstopo estimations, the obtained RMSE seems reasonable. An illustration of the difference between the two estimations is shown in Figure 4.24, where the wind speed estimations are plotted for 80 pixels chosen randomly in Switzerland. It can be seen from the figure that, even though the values generally are in the same order of magnitude, our estimations tend to slightly overestimate the wind speed compared to Swisstopo. Given the relatively high uncertainty attached to Swisstopo estimations, however, it is rather challenging to draw formal conclusions from the comparison.

Regarding estimations in urban areas, it is possible to validate some of the wind speed values estimated in section 4.4.3 with wind measurements available in urban pixels. As explained in section 4.3.1, out of the 159 observed pixels with available wind aggregated measurements, 41 are urban pixels and were therefore discarded in the training of the wind speed monthly models for rural areas at 10m. Out of these urban pixels, 17 are urban boundary pixels and can, as a result, be used to

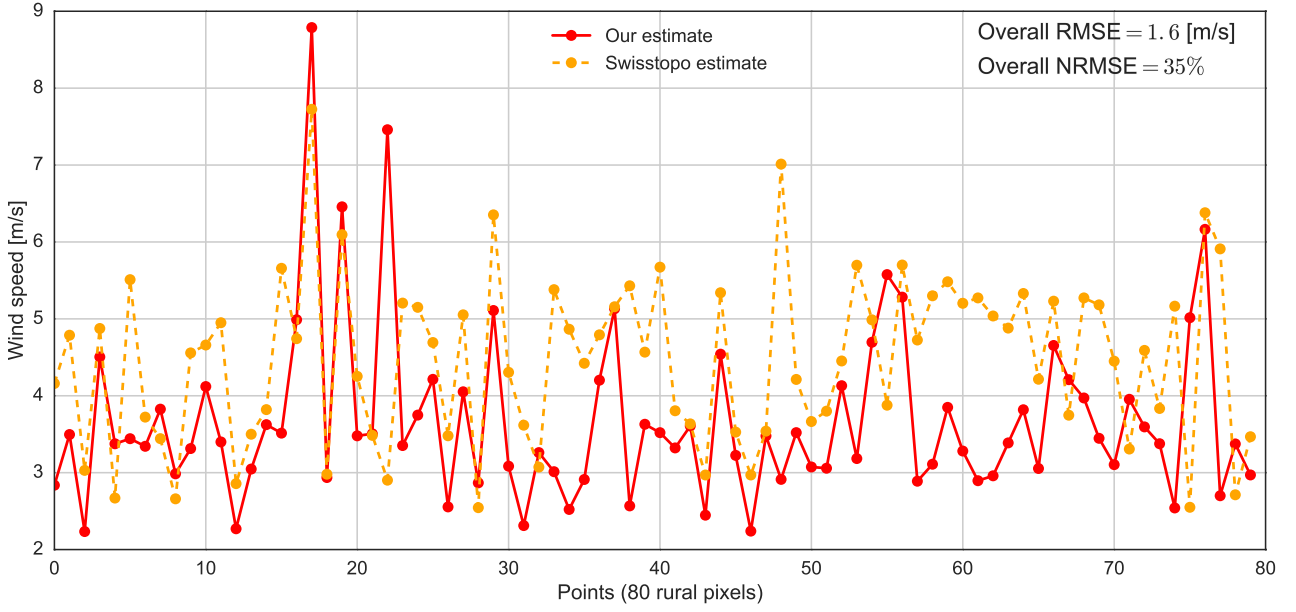


Figure 4.24: Comparison between Swisstopo yearly wind estimations at 100m and our estimations for 80 pixels chosen randomly in Switzerland in rural areas. Note that the RMSE and NRMSE given is computed over all rural pixels, not only the 80 shown here.

validate the estimated urban wind speed obtained with the strategy developed in section 4.4. Note that there are therefore 24 urban center pixels with speed measurement available, which could potentially be used to learn some information within urban centers by training an ML model (to relate with the note at the end of section 4.4.3). The number of 24 points, however, is simply too small to provide a good generalization of the model, and a decent training data to learn from. These points are therefore not used (yet could be used to validate the data provided by meteorological models, as mentioned previously). Concerning the 17 urban boundary pixels, the compared values can be seen in Figure 4.25. The RMSE and NRMSE errors, computed between the observed and estimated values are respectively 0.57 m/s and 31.2%. Note that in addition to our estimates, computed at a height of z_{hub} , the speed is also estimated at a height of 10m to match the original height of the measurements. It can be seen from the figure that in most of the 17 pixels, the estimated speed at z_{hub} is very close to the estimated speed at 10m (which shows that the average building height of these pixels was incidentally around 5m). The RMSE and NRMSE for the speed estimated at 10m are therefore very similar and do not need an extra computation. While the error between the estimated and observed values are in most cases of less than 0.5 m/s, the general error is raised by a few points which extent be located in very dense urban areas. Eventually, even though the very small size of the validation set prevents from drawing definitive conclusions, the related error is rather acceptable for an urban wind estimation, subject in practice to many unknown parameters.

4.5.4 Limitations

The present chapter suggests a novel methodology to assess the large-scale potential for generated electricity from wind turbines in both rural and urban areas. In particular, it shows a non-negligible potential within urban areas in Switzerland, which are unfortunately often discarded from such studies.

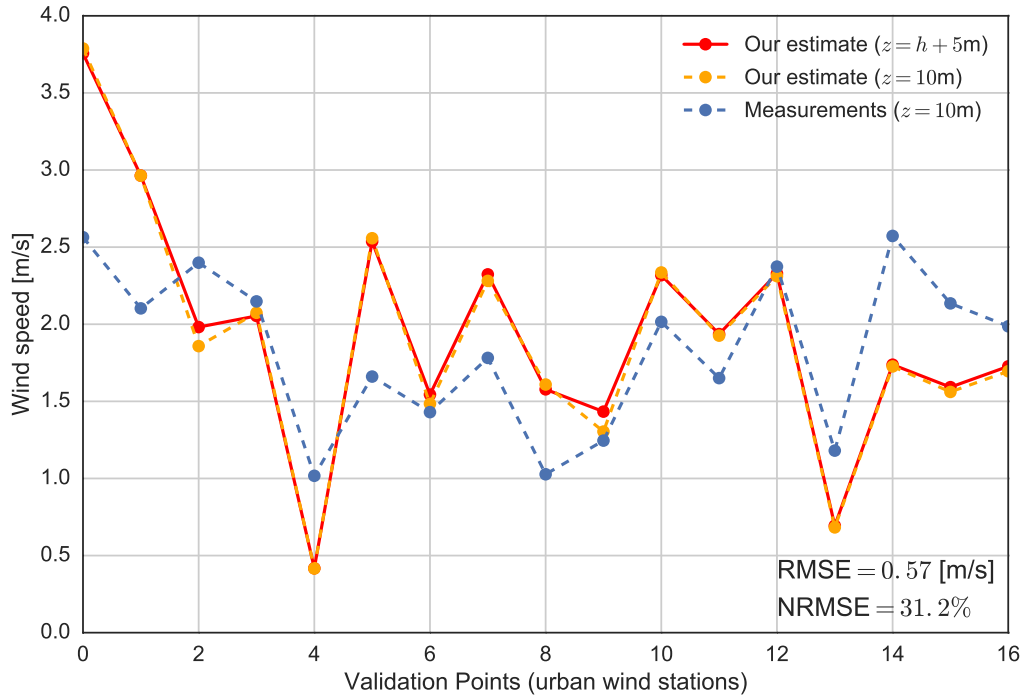


Figure 4.25: Comparison between wind measurements and estimated values in 17 urban boundary pixels.

There are, however, several notable limitations, which could lead to future improvements: (i) **Validation strategy.** Monitored wind speed data within urban areas is challenging to gather, as urban wind measurement stations are rare. This may evolve in the future with the increased use of small integrated wind turbines in urban setting. Nowadays, however, it makes it difficult to validate urban wind speed estimations. Another strategy to be considered in the future is the use of precise simulation results (using Computational Fluid Dynamics) to validate the estimated values. (ii) **Consideration of other variables than wind speed.** Although the average horizontal wind speed is the most important variable in the assessment of the wind potential, other variables shall be considered to complete the study. These include the wind speed variations in multiple horizontal directions and the probability of wind gusts, which can have an effect on the ultimate performance of turbines (in particular horizontal axis turbines). They however require, for a large-scale study, additional data (historical gusts data and high resolution building geometry data), and may not have a very significant impact on the average monthly wind energy potential. Also, additional features could be considered to account for the direct impact of neighbor obstacles/topography on the wind speed within a pixel.

4.6 Summary

This chapter proposes a methodology using a combination of Machine Learning, GIS and wind models to estimate the theoretical wind speed potential over Switzerland, that is, the monthly wind speed over rural and urban areas, for each 200×200 [m²] pixel covering Switzerland. In order to extract values which reflect the wind energy potential in practice, the wind speed is estimated at typical heights for wind turbine installation: $z = 100m$ for rural areas (height of a large commercial

wind turbine), $z = z_{\text{hub}} = \text{mean building height} + 5\text{m}$ for urban areas (height of a small turbine mounted on top of buildings or households).

The steps of the methodology are as follows: (i) extract significant features (predictors) impacting the behavior of wind speed in Switzerland (weather, topographic, and roughness-related variables), (ii) estimate monthly wind speed maps at $z = 10\text{m}$ in rural areas, using Random Forests models together with wind monitored data at 10m and the features previously extracted, (iii) vertically extrapolate the wind speed in rural areas to a height of 100m (height of large commercial wind turbine), using a log-law and the estimated rural roughness length (based on land use), (iv) assume that, for urban pixels neighbor to rural pixels (these pixels are called *urban-boundary pixels*), their wind speed at 100m are equal (invoking the constant flux properties of the Inertial Sublayer and the horizontal homogeneity assumption), (v) compute the roughness length and displacement height in urban boundary pixels, using MacDonald expressions (Eq. 4.4 and 4.5), and various urban characteristics (building facade and footprint area etc.) and (vi) Extrapolate the wind speed in urban pixels to z_{hub} using a log-law and the previously estimated roughness and displacement length computed. Using the estimated wind speed and typical characteristics of wind turbines suitable for rural and urban areas with expression 4.9, the potential wind power is also extracted over the whole country, providing a measure of the potential electricity output generated by installed wind turbines.

The results show a significant potential for wind energy in Switzerland, particularly in the south part of Switzerland (where, unfortunately, the population density is the lowest). Even though it may not be the renewable energy resource with the largest potential in Switzerland, our estimations show that it could approximately represent an installed capacity of, for each $200 \times 200 \text{ [m}^2\text{]}$ pixel and for each turbine installation, on average 80 kW in rural areas (up to 1.6 MW in high potential rural pixels), and 15 W in urban areas (up to 1.1 kW in high potential urban pixels), which is rather not negligible.

A flowchart summarizing the entire methodology proposed in the chapter is shown in Figure 4.26.

Chapter 5 will tackle another promising renewable energy at the theoretical level: very shallow geothermal energy.

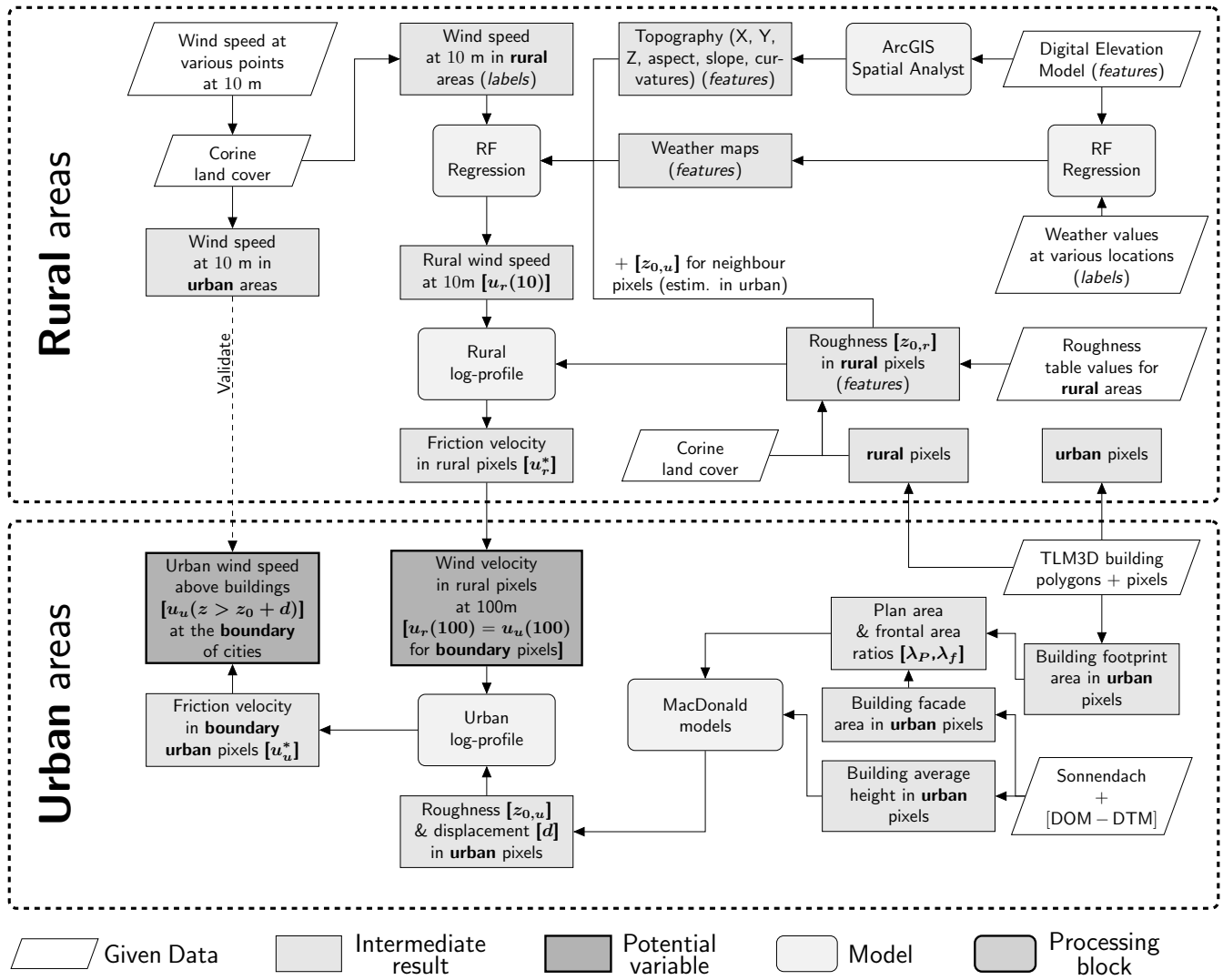


Figure 4.26: Flow chart of the methodology for the theoretical wind potential estimation at pixel scale.

5

Very shallow geothermal energy: a theoretical potential estimation

This chapter borrows from the article:

Assouline, D., Mohajeri, N., Gudmundsson, A. and Scartezzini, J-L. (2018). Combining Fourier Analysis and Machine Learning to Estimate the Shallow-Ground Thermal Diffusivity in Switzerland. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium (pp. 1144-1147). IEEE.

This chapter aims at estimating the theoretical potential for very shallow geothermal energy, or *theoretical vSGP* (very shallow geothermal potential), over the whole of Switzerland. While the theoretical wind potential is mainly expressed by the wind speed, it is significantly more complex for geothermal energy since many different ground variables require consideration within the potential estimation (as presented in chapter 3). In addition, these variables are in general dependent of each other, which makes their individual computation more difficult. It is nonetheless possible to use an estimation strategy similar to the one adopted within the wind potential estimation, under a certain number of choices and assumptions. These are notably related to the main ground variables which are impacting the potential, and the typical ground depth at which the study is performed. They are as follows:

- We consider the *monthly ground temperature*, the *ground thermal conductivity* (λ) and the *ground thermal diffusivity* (α) as the three main ground thermal variables impacting the theoretical vSGP. Note that, however, these variables are not the only ones impacting on the potential. The local groundwater behavior is notably another important factor, which is not discussed here because of a lack of data (but could be added in a future study to complete the potential estimation). Note also that the thermal heat capacity, another important factor, can be derived from the conductivity and the diffusivity.
- The considered ground thermal variables are supposed independent and are estimated separately, even though they are theoretically correlated. It notably allows to avoid uncertainty propagation issues, which would arise from using the estimation of one variable as an input for another one.

- The thermal conductivity and diffusivity are considered to be constant throughout the year, and are therefore computed on an average yearly basis. Even though both variables are in general functions of time (since they are functions of ground temperature, notably), the availability of data and the adopted methodologies for their estimation do not allow for a monthly estimation. Besides, it is a reasonable and common assumption adopted in most potential studies [83, 176, 177].
- We consider a very shallow ground depth of 1m (the uppermost meter of the ground) for the vSGP. This choice is mainly motivated by (i) the lack of attention in the literature for the vSGP, which is a non negligible part of the total potential, and (ii) the lack of large-scale data for higher depths, which in practice makes it challenging to estimate the potential at traditional shallow depths based on real data. There are promising ground source heat pump technologies which are typically installed at this depth (1-2m), as presented in chapter 3, section 3.2.6.

The strategy adopted for estimating each of the three discussed variables is similar to the one adopted for the wind potential estimation in previous chapter. The general steps leading to the estimation are, for each variable, as follows: (i) collect significant data related to the variable (weather, topographic and geological/soil data), (ii) if not existing, extract values for the variable at available locations with the help of traditional models and part of the data as input for these models, (iii) train a ML model (with the Random Forests algorithm) using the extracted variable values as examples (training output labels) and related information contained in the data as features (training input samples), (iv) use the trained ML model to estimate the variable in unknown locations, (v) estimate the uncertainty attached to the estimations. The methodology assesses the variables within the (200×200) [m²] pixels forming the grid presented in the previous chapter 4. The three mentioned ground thermal variables are estimated within polygons defining different categories of surface geology, and the obtained estimations are then re-aggregated within the grid pixels. The chapter details of the multiple processing and estimation steps leading to the estimation of each of the three potential variables.

The chapter is organized as follows. Section 5.1 offers a literature review on shallow geothermal potential studies and places the present chapter within its context. Section 5.2 presents the data sources used in the chapter and some of the processing performed to extract significant features, including weather, topographic, geological and other soil-related variables. Section 5.3 presents the computation of monthly ground temperature maps for Switzerland at different depths. Section 5.4 details the strategy to estimate the shallow thermal conductivity over the country, including the estimation of electrical resistivity values in several locations in Switzerland, its extrapolation over the Swiss territory and its conversion into thermal conductivity. Section 5.5 explains the computation of the apparent thermal diffusivity, including the estimation of thermal diffusivity values in several locations in Switzerland and its extrapolation over the territory. Section 5.6 provides a discussion on the obtained results. Finally, section 5.7 concludes the chapter and summaries the proposed methodology.

5.1 Related literature

There have been several studies proposing large-scale (meaning at regional or national level) methodologies for shallow (50 to 200 m depth) geothermal potential estimation. In particular, the potential for vertical Borehole Heat Exchangers (BHEs) has been studied often, since their high Coefficient of Performance (usually between 3 and 6) make them one of the most attractive means of exploiting shallow geothermal energy [93, 178].

Some studies extract the theoretical potential for very shallow geothermy estimating a thermal ground-related variable at a large scale. In most studies, the estimated variable is the thermal conductivity, which is understandable given its high impact on the potential. Beamish [179] made a GIS study on the thermal conductivity all over the UK using statistical sampling with airborne electromagnetic data together with an available geological database for the country. Di Sipio et al. [180] made a GIS study based on sampling to extract thermal conductivity values for the Calabria region in Italy. Kalogirou et al. [181] explored the use of Machine Learning, more specifically Neural Networks, in order to estimate a thermal conductivity contour map for Cyprus, based on a measurement training data of 41 points at different locations in the island. The features used as inputs for the models for each point were: the lithology class, the elevation, air temperature statistics, rainfall and the x and y coordinates. These studies, however, lack the estimation of other variables to provide a full geothermal potential.

Many studies have developed large-scale methodologies to extract the complete technical potential for shallow geothermal systems. Ondreka et al. [182] proposes a methodology using GIS together with geological, hydrogeological, and lithological ground information, based on the German VDI (Verein Deutscher Ingenieure) guideline 4640, to extract the technical geothermal for two study areas in Germany. The VDI guideline [183] is a table that provides heat extraction values depending on the type of soil or surface rocks/sediments of a particular location, the number of operating hours of the heat pump and other factors. Despite its being comparatively crude, the VDI is very practical to obtain first potential estimations, and is particularly suitable for large-scale studies, where only very general geological information may be available. VDI has been used in many later studies, including a study by Garcia-Gil et al. [184], focusing on the groundwater flow to extract the technical potential, and in another recent study by Schiel et al. [185], using the VDI guideline and demand values to extract the very shallow geothermal potential in the urban area of a city in Germany. Another strategy to extract the technical potential consists in first estimating significant ground thermal variables (commonly the thermal conductivity, ground temperature, and heat capacity) using various methods, and combining these variables with heat conduction models in order to extract the technical potential. Galgaro et al. [186] use multiple models to estimate the annual air and ground temperature, thermal conductivity and monthly energy loads to finally extract technical potential values, using empirical modeling. Casasso and Sethi [187] propose a quantitative method called G.POT to map the shallow geothermal potential. After having extracted thermal heat capacity, thermal conductivity, and ground temperature values from typical values (based on the type of rock/sediment) and empirical models, the G.POT method uses analytical models to simulate the heat transfer in the ground and in the borehole with varying thermal properties as well as operational and design parameters of the system. It results in the technical potential estimation in the form of heat extraction values per year. In a later study [176], the G.POT method was revised so as to extract the potential for open-loop installations such as groundwater heat pumps, in addition to BHEs. Multiple studies have considered the geothermal potential of shallow aquifers

[176, 184]. A notable recent study estimates the geothermal potential of shallow aquifers in Finland [188]. Based on the heat flux, temperature, thermal heat capacity of groundwater, and the design of buildings, the heating capabilities of groundwater for buildings were extracted all over the country.

Since very shallow geothermal resources or systems are particularly suitable for cities, several studies attempted to account for specific urban conditions, and most notably the impact of urban heat islands. In particular, studies by Allen et al. [189], Zhu et al. [190], and more recently Arola and Korkka-Niemi [191] and Rivera et al. [192] all show that the urban heat island effect has a very significant positive impact on the geothermal potential for BHEs installed in urban areas, since the ground temperature in the islands is significantly higher than that of nearby rural areas. These results are, in theory, also valid for shallower geothermal installations.

Although recent very shallow geothermal systems have shown promising results (as discussed in section 3.2.6), few studies have been conducted on large-scale vSGP estimation. Some recent articles tested multiple methods to compute the apparent (very shallow) thermal diffusivity from ground temperature time series, using various analytical methods and numerical models based on the 1D heat equation. Such studies include the works of Busby [84], Rajeev and Kodikara [83], and Andujar Marquez et al. [193]. These studies, however, aim at providing estimations for specific study areas rather than large regions. The main large-scale vSGP study is the one initiated by the ThermoMap project [177, 194, 195] in the past few years. This project aims at extracting and illustrating online the very shallow (top 10 m) geothermal potential in Europe as a whole, including specific case studies. Values are estimated for heat capacity and thermal conductivity based on near surface geology and hydrogeology information, the USDA (United States Department of Agriculture) soil texture classification, and equations from Kersten [196] and Dehner [196]. It should be noted, however, that a large-scale estimation generally refers to regional or national scale. For the ThermoMap project, the continental scale is used, which naturally reduces the resolution of the study. Eventually, a reliable methodology for vSGP at large (regional/national) scale is still to be developed.

Regarding the estimation of energy values and environmental modeling in general, Machine Learning (ML) methods have recently become widely used. In particular, many different ML algorithms have been explored for geospatial modeling of multiple environmental variables, including solar radiation and wind speed [13, 197], forecasting of solar radiation over horizontal and tilted surfaces [8, 198–201], and short-term forecasting of wind speed and wind power prediction [202–204]. Also, Joshi et al. [205] used ML algorithms to perform a rooftop classification and provide a solar potential estimation over rooftops, and Assouline et al. [22, 23] used a combination of GIS and ML methods (Support Vector Machines and Random Forests) to map the technical solar rooftop potential in Switzerland. It has been very rarely used, however, for ground-related variables estimations. One of the main related study is the work of Kalogirou et al. [206], which used neural networks to extract ground temperature maps at various depths in Cyprus, based on measurement data from 41 boreholes. In addition, Beardsmore et al. [207, 208], in the framework of the National ICT Australia, developed a Bayesian inference (which can be seen as an ML sub-family of methods) software tool for geophysical joint inversions, helping for the detection of promising locations for geothermal energy exploration. Nevertheless, ML has never, to the best of our knowledge, been used for a geothermal potential mapping study.

This chapter is motivated by the lack of theoretical vSGP studies at a large scale, and in particular in Switzerland. Most existing large scale potential studies either consider the deep (>200m) or the

shallow (between 50 to 200m) potential. In addition, it is rare to observe more than one significant ground thermal variable estimated, and the adopted methodologies for the estimation are in most cases qualitative, based on general geology data. Thus, the goal of the chapter is to fill in the gaps left by literature concerning geothermal potential estimation, by: (i) exploring the capabilities of Machine Learning algorithms to estimate (very shallow) ground thermal properties, (ii) proposing a novel large scale theoretical vSGP methodology, offering multiple ground variable estimations, and (iii) assessing the vSGP specifically in Switzerland.

5.2 Data

5.2.1 Data sources

All data sources used within this chapter are presented in Appendix A and signified by a ✓ for the present chapter within Tables A.1 and A.2. They include meteorological data, ground temperature data, digital surface models, soil moisture satellite data, geology surface data, soil texture data, Vertical Electrical Soundings data, and electrical/thermal resistivity tabled data. The main source of information on Swiss geology at the national level is the GK500 (or GeoCover500) dataset, provided by Swisstopo, which gives information on surface geological formations and materials for the whole of Switzerland. The data is available in a GIS vector polygon format. Each polygon represents the boundaries of a surface geological formation and includes various related pieces of information. As it is reasonable to assume the ground properties to generally remain similar within one formation polygon, all additional geological features are aggregated within the GK500 polygons (if not specified otherwise). Therefore, the GK500 polygons (Figure 5.1) form the original resolution of the estimation for the thermal conductivity and thermal diffusivity. A posteriori, the estimations will be re-aggregated within the (200×200) [m²] pixels, as used in the wind potential study (chapter 4).

5.2.2 Data processing

As mentioned in previous chapters, feature selection is one of the most important steps when building a machine learning model. In the present chapter, we investigate features that have an impact on electrical and thermal properties of ground rocks and soils.

The GK500 geology data presented in the latter section is first used to extract the first family of features that will be used within this chapter, *geology features*. The properties available from the GK500 data are all categorical (class-based) features, and include:

- Geological period [code: PERIOD] (classes include quaternary, tertiary etc.)
- Main types/classes of rocks [code: TYPE ROCHE] (classes are sedimentary, igneous and metamorphic),
- Detailed rock-type classification [code: LITH PET] (rock type classes include sand, silt, clay, limestone, gneiss, gabbro, basalt, andesite, etc.),
- Hydrogeological characteristics [code: HYDRO] (classes include surface water, presence or absence of aquifers etc.),

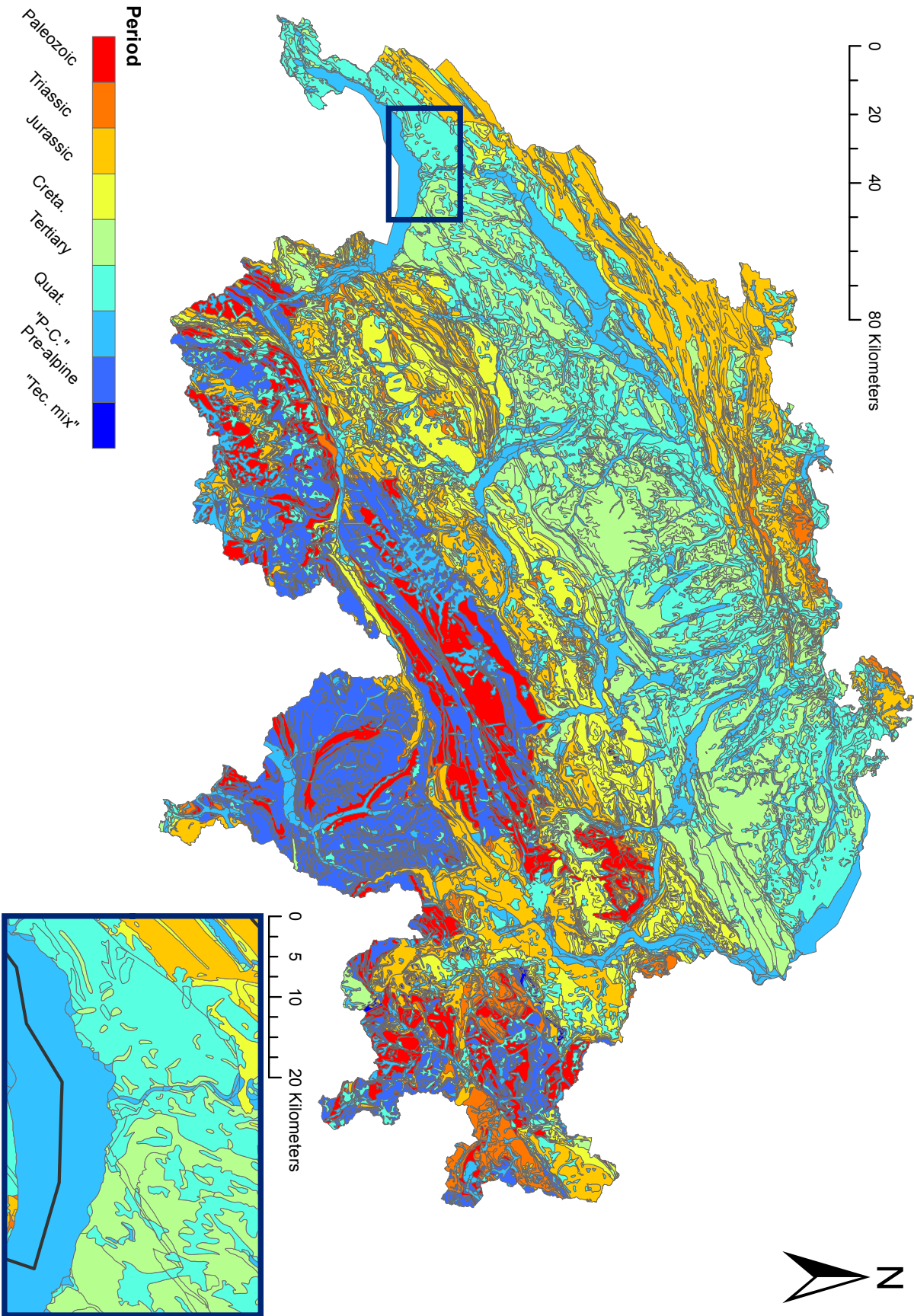


Figure 5.1: GK500 vector polygons, offering information on the superficial geological strata in Switzerland. Here we display, the geological period characterizing each polygon. (Creta., Quat., "P-C.", and "Tec-mixture" respectively stands for Cretaceous, Quaternary, "Permian-Cretaceous" and "Tectonic-mixture")

- Productivity of aquifers [code: PRODUCTIV] (classes include saturated from 2 to 10m, saturated from 10 to 20m etc.).

The previously mentioned features need to be converted into real values. A “one hot encoding” approach is used in order to obtain real numbers from the categorical features: for each original variable, we create as many binary features as there are classes, defining the binary features as “variable=class1”, “variable=class2”, etc. and label with 0 or 1 each feature (1 if the point belongs to the class for this variable, 0 if it does not). All possible classes of each GK500 variable are listed in Annex B. It results in 107 geology features.

Additional *soil/sediment texture features* are gathered for the study. Soil texture information (in the first meter of the ground) was extracted from the NABODAT (NAtionale BOdenDATenbank) dataset from the Swiss Federal Office for the Environment. The dataset contains various soil textural information for an array of 6212 measurements at various locations in Switzerland (mostly in the Swiss plateau), including the sand, silt and clay content, as well as gravel and stones content, in fractional values. The percentage sum of the sand and gravel fractions in the soil, denoted by F , is also computed at each measurement location. As the measurements are often available at multiple depths, all the content values were aggregated at each location, using a weighted average (weighting each measurement by the thickness of the stratum where that measurement was taken). Therefore, the aggregated content value (for sand or silt etc.) at a location l is computed as \bar{q}_l defined as follows:

$$\bar{q}_l = \frac{\sum_{i=1}^{n_l} q_{i,l} h_{i,l}}{\sum_{i=1}^{n_l} h_{i,l}} \quad (5.1)$$

where $q_{i,l}$, $h_{i,l}$, and n_l are respectively the measured content in the rock of interest in stratum i , the thickness (depth) of stratum i , and the number of strata considered in location l . Given that the soil texture depends more on the type of rock (LITH_PET) rather than on spatial location of the measurement, it was decided to aggregate the soil texture information by type of rock rather than within each GK500 geology polygon. For example, for the rock type “silts a sables avec graviers et blocs” (“Silty sands with gravels and blocks”), 1614 NABODAT points spanning over 1962 different GK500 polygons were recorded and considered to extract statistical information. This information is then considered correct for each of these 1962 polygons containing this kind of rock. The number of polygons of each rock type and the number of NABODAT samples for these rock types are specified in Annex B. Note that, for statistical sampling validity, rock types covered by only three or less than three NABODAT sample points were not considered in the study, independently of the number of GK500 polygons typical for these types of rock. For example, the 291 GK500 polygons of one type of gneiss (more specifically “Gneiss a schistes sericito-chloriteux”) could not be considered, as only two NABODAT samples are of this rock type. Eventually, 10812 polygons are characterized by soil texture information, out of the 13320 GK500 polygons.

The statistical information for the soil texture is extracted in two forms: (i) classical summary statistics, and (ii) Probability Density Functions (PDFs). The first form of statistical information is extracted to serve directly as features. Statistics for sand, silt and clay content, as well as F , are computed for each rock type, using the ArcGIS Joint Spatial tool: the minimum, maximum, range, standard deviation, median, 25th and 75th percentiles. It results in 28 soil texture statistical features. The second form of statistical information is extracted to serve as weights (in section 5.4.3).

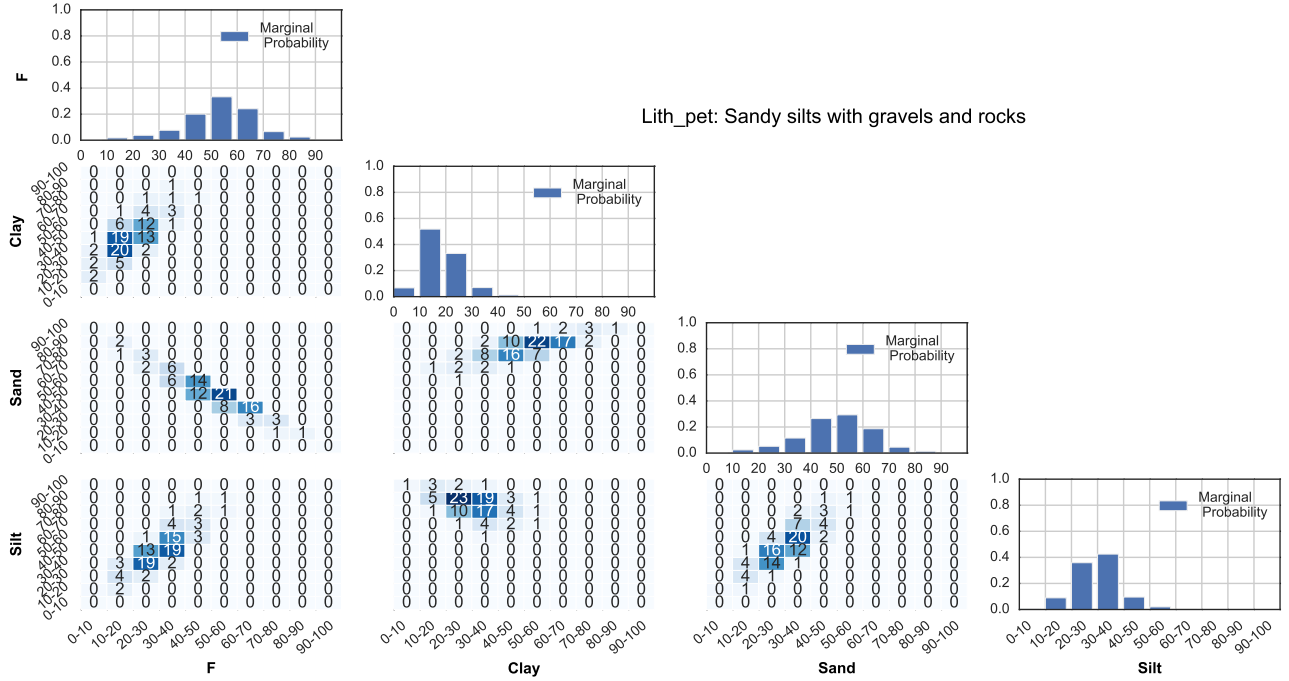


Figure 5.2: Marginal PDFs and all possible 2-joint PDFs for the percentages of sand, silt, clay, and F for one example of rock type ("Silty sands with gravels and blocks").

PDFs for percentages of sand, silt, clay and F, are computed for each rock type using 10% wide bins (for example we compute the probability that the percentage of sand is within $[0\% - 10\%]$, etc.). We note the random variables for sand, silt, and clay content respectively S_d , S_t and C_l , and the possible intervals I_1, I_2, \dots, I_{10} , for $[0\% - 10\%]$, $[10\% - 20\%]$, ..., $[90\% - 100\%]$. In order to store all possible marginal and joint PDFs for the four structure variables efficiently, we compute the full joint PDF $\mathbb{P}(S_d \in I_i, S_t \in I_j, C_l \in I_k, F \in I_l)$ where $(i, j, k, l) \in \{1, 2, \dots, 10\}^4$. Since the four structure variables are considered non-independent random variables, the full joint PDF is expressed as a function of conditional probabilities using a chain rule:

$$\mathbb{P}(S_d, S_t, C_l, F) = \mathbb{P}(S_d | S_t, C_l, F) \mathbb{P}(S_t | C_l, F) \mathbb{P}(C_l | F) \mathbb{P}(F) \quad (5.2)$$

where we omitted the I intervals to lighten the notation. Any marginal or joint PDF (with two or three variables out of the 4) can easily be extracted from the full joint by summing the probabilities over the unconsidered variables. As an example, Figure 5.2 shows the marginals and the 2-joint PDFs calculated from the NABODAT samples for one example of rock type ("silty sands with gravels and blocks"). Note that the conditional probabilities were computed within each rock type in a frequentist fashion, counting the number of samples with particular sand, silt, clay, and F percentages.

Further terrain and weather features are aggregated within the GK500 polygons to bring more additional general information to the machine learning models. These features include:

- **Space features:** latitude, longitude and altitude. Originally available for the whole Switzerland at a 25×25 [m²] resolution from the DHM25 digital elevation model (DEM), they are resampled at a 200×200 [m²] resolution. The latitude and longitude of each GK500 polygon are computed by extracting its centroid using the Graphics and Shapes toolbox [209]. Statistics are computed

for the altitude within each GK500 polygons (minimum, maximum, range, mean, standard deviation and sum). It results in 8 space features.

- **Terrain features:** ground surface slope and aspect. They are computed from the resampled 200×200 [m²] DHM25 DEM using the Spatial Analyst toolbox from ArcGIS. To condense the terrain information, slope and aspect classes are created, as it is often done in the literature. We consider 9 aspect classes: 1=flat, 2=North, 3=North-East, 4=East, 5=South-East, 6=South, 7=South-West, 8=West, 9=North-West; and 12 slope classes: 1=[0-5°], 2=[5-10°], ..., 9=[40-45°], 10=[45-50°], 11=[50-60°] and 12=[60-70°]. Statistics are computed for both features within each GK500 polygon based on classes: variety (number of different classes), majority (most frequent class), minority (least frequent class), "mean" class (the mean value of all registered classes using the class labels, even though it does not have any physical meaning) and median class. It results in 10 terrain features.
- **Weather features:** monthly mean air temperature, mean sunshine duration, mean precipitation, and cumulative snow depth. For each of the previously mentioned weather variables, except snow depth, monthly rasters derived in [23] are used. The rasters are built based on MeteoSwiss measurement data and the DHM25 DEM, using Random Forests. The monthly cumulative snow depth rasters are estimated in a similar fashion, using snow depth measurement data from MeteoSwiss. Statistics are computed for each of the weather variables within each GK500 polygons (minimum, maximum, range, mean, standard deviation and sum). It results in 24 weather features.

Lastly, *soil moisture features* are extracted for the study. Soil moisture (in the top 5cm of the ground) was extracted from the recently available SMAP (Soil Moisture Active Passive) satellite data from NASA, in the form of Volumetric Water Content (WVC). As the SMAP data was very recently collected by NASA (from 2015), it should be noted that our estimation of the soil moisture in Switzerland will only reflect its behavior during the past 3 years. Furthermore, the specificity of the SMAP mission is that the data is collected during 6:00 a.m. descending or 6:00 p.m. ascending half orbits (see <https://smap.jpl.nasa.gov/data/> for more information). It therefore does not reflect the fluctuations of the soil moisture during the entire day. Lastly, the data was only available for 6 months in Switzerland during the last three years (January, February, March, October, November, December), and yearly average values (considering the six mentioned months) were therefore used. The data does not, as a result, allow for precise information about the monthly moisture fluctuations during summer. The spatial variations of the moisture, however, are nonetheless captured by the data. Originally with a resolution of $3\text{km} \times 3\text{km}$, and resampled to a resolution of $1\text{km} \times 1\text{km}$ by NASA, the data was further resampled to follow the $200\text{m} \times 200\text{m}$ pixel grid. A grid of SMAP polygon cells was built in ArcGIS based on the original SMAP points defining the centroids of the cells, using Thiessen polygons (polygons generated based on the centroids, so that any point within one polygon is closer to its centroid than to the other centroids). Then, the ArcGIS Joint Spatial tool was used in order to associate each $200\text{m} \times 200\text{m}$ pixel with the moisture value of the SMAP $1\text{km} \times 1\text{km}$ cell that contains the centroid of the pixel (the option HAVE THEIR CENTER IN is used when performing the Joint Spatial). While covering a large portion of the Swiss territory, the SMAP data does not span over the whole country (the spatial coverage is different depending on the month as

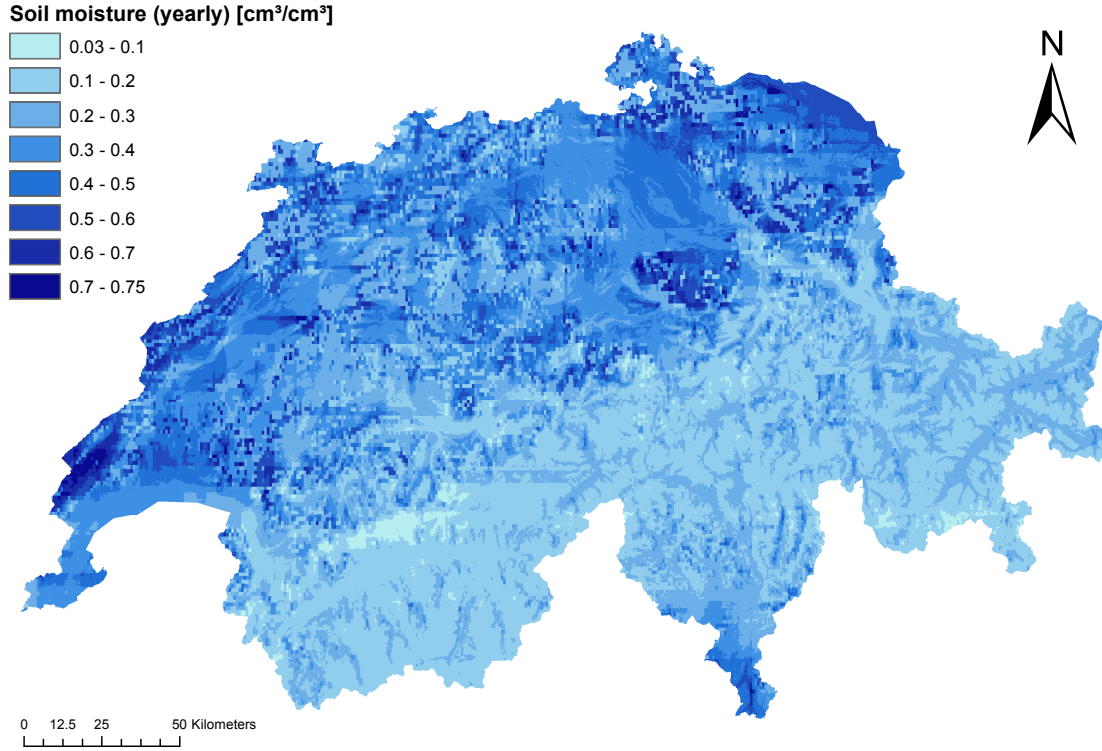


Figure 5.3: Yearly soil moisture map estimated from the SMAP data in cm^3/cm^3 .

well). Therefore, an RF model is trained for each available month using the soil moisture values in pixels as labels, and the previously mentioned space, weather and terrain features. Testing errors, that is, the RMSE and the NRMSE are shown for each model in Table 5.1. The obtained yearly soil moisture (VWC) map of Switzerland is shown in Figure 5.3.

Table 5.1: Testing RMSE (E_R) and NRMSE (E_{NR} , in percentage) for Random Forest models trained for monthly ground soil moisture (Volumetric Water Contents).

Month	Soil Moisture		
	E_R [cm^3/cm^3]	E_{NR} [%]	OOB [-]
Jan.	0.08	19.61	0.82
Feb.	0.08	20.82	0.82
Mar.	0.07	22.75	0.85
Oct.	0.07	26.65	0.87
Nov.	0.07	23.88	0.87
Dec.	0.07	19.78	0.84

Statistics of VWC are computed within each GK500 polygon to form the soil moisture features: VWC minimum, maximum, range, mean, standard deviation and sum. It results in 6 soil moisture features.

Then, a total of 183 features are computed for each GK500 polygon. Note that not all features will be systematically used to estimate each of the three ground thermal variables. Depending on

the variable, an adequate subset of the mentioned features will be used.

5.3 Ground temperature estimation

The first step in the potential study is the estimation of ground temperature maps in Switzerland at different shallow depths in order to assess the shallow thermal gradient. Given that we consider very shallow horizontal ground loop collectors or systems (loops), the seasonal weather variations may have a significant impact on the ground temperature, and yearly temperature values would not bring sufficient information. As indicated above, below the depth of 2-3 m, the soil temperature changes little throughout the year. However, most horizontal ground loop collectors are at the depth of 1-2 m, where there are significant changes in temperature throughout the year. The reason why the collectors are so shallow are partly the lack of thicker soil/sediment cover, the higher costs of deeper trenches, and the empirical results that collectors at this depth have proved suitable for providing economic space heating.

As a result, monthly ground temperature maps had to be computed. The estimation is based on an hourly ground temperature time series data, available from MeteoSwiss (see Table A.1) for multiple locations and at multiple depths, namely 5, 10, 20, 50 and 100 cm. The data is not always available for all the depths at the same locations for the same years. Therefore, the datasets for the five different depths are treated separately. Each data is aggregated monthly through the available years, allowing for twelve typical average monthly values for each location. Intuitively, the seasonal variations should be attenuated with larger depths, as it is shown in Figure 5.4, showing the monthly ground temperature at the five different depths, averaged through all locations available for each depth. To allow for the estimation to be at the resolution of the 200×200 [m²] pixels, the values of the available measurement stations are assigned to their nearest pixels. Each measurement corresponds to the pixel whose centroid is the closest to the location of the considered station. In the case where multiple stations are located within one pixel, the multiple measured values are averaged through the stations in order to provide one ground temperature value for the pixel.

RF models are trained using the pixel ground temperature values as training labels and weather and terrain variables as features: latitude, longitude, altitude, ground aspect and slope, and monthly precipitation, sunshine duration, snow depth, air temperature as defined in section 5.2.2. One RF model is built for each depth and for each month, leading to 60 different models. The models are used to extrapolate the training temperature data and build a ground temperature map for each month and each depth. Testing errors, that is, the RMSE and the NRMSE are shown for each model in Table 5.2. The resulting monthly maps for a depth of 100 cm as well as yearly maps for 5, 10, 20, 50 and 100cm are shown in Figures 5.5 and 5.6. Furthermore, 95% Prediction Intervals (PIs) for ground temperature have been computed at all depths and for all month, both in the test set and for new predicted points. A visualization of the PIs for two months (January and June) are shown in Figure 5.7.

5.4 Thermal conductivity estimation

Due to a lack of existing data, the estimation of the ground thermal conductivity is less straight-forward than that for the ground temperature. It includes multiple steps that aim at taking maximum advantage of the current data available at national level. The steps are as follows: (i) geophysical inversion of

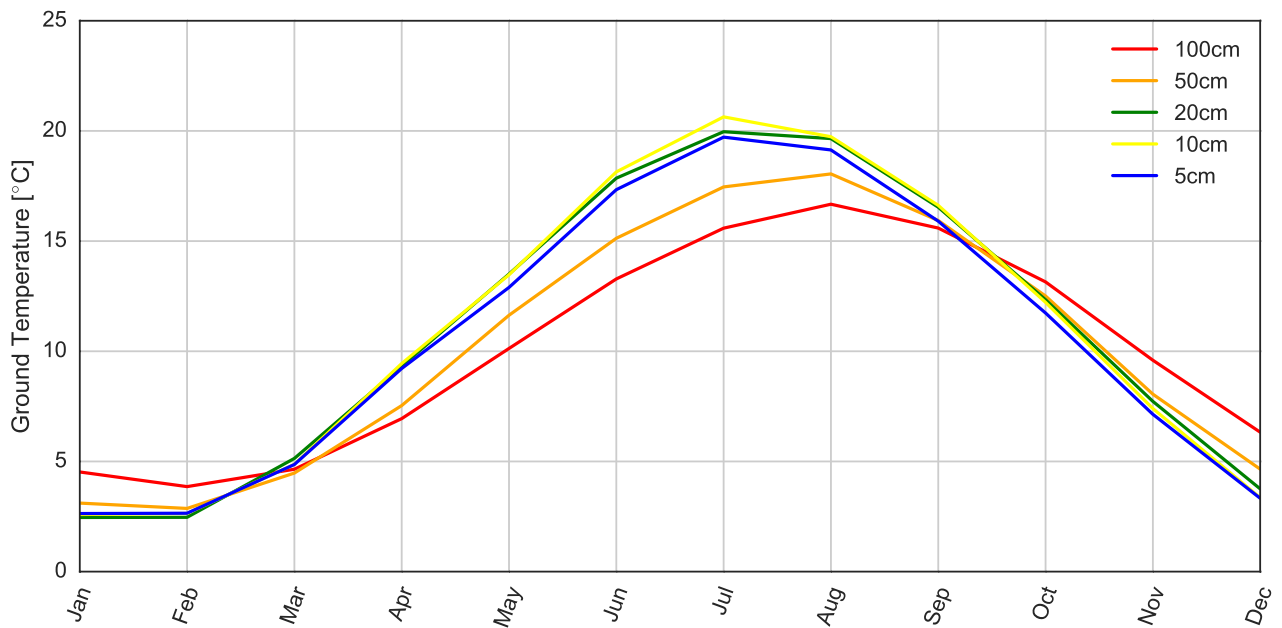


Figure 5.4: Monthly ground temperature in Switzerland. Temperatures are shown at 5, 10, 20, 50 and 100cm, and are averaged through all the measurement stations available in Switzerland.

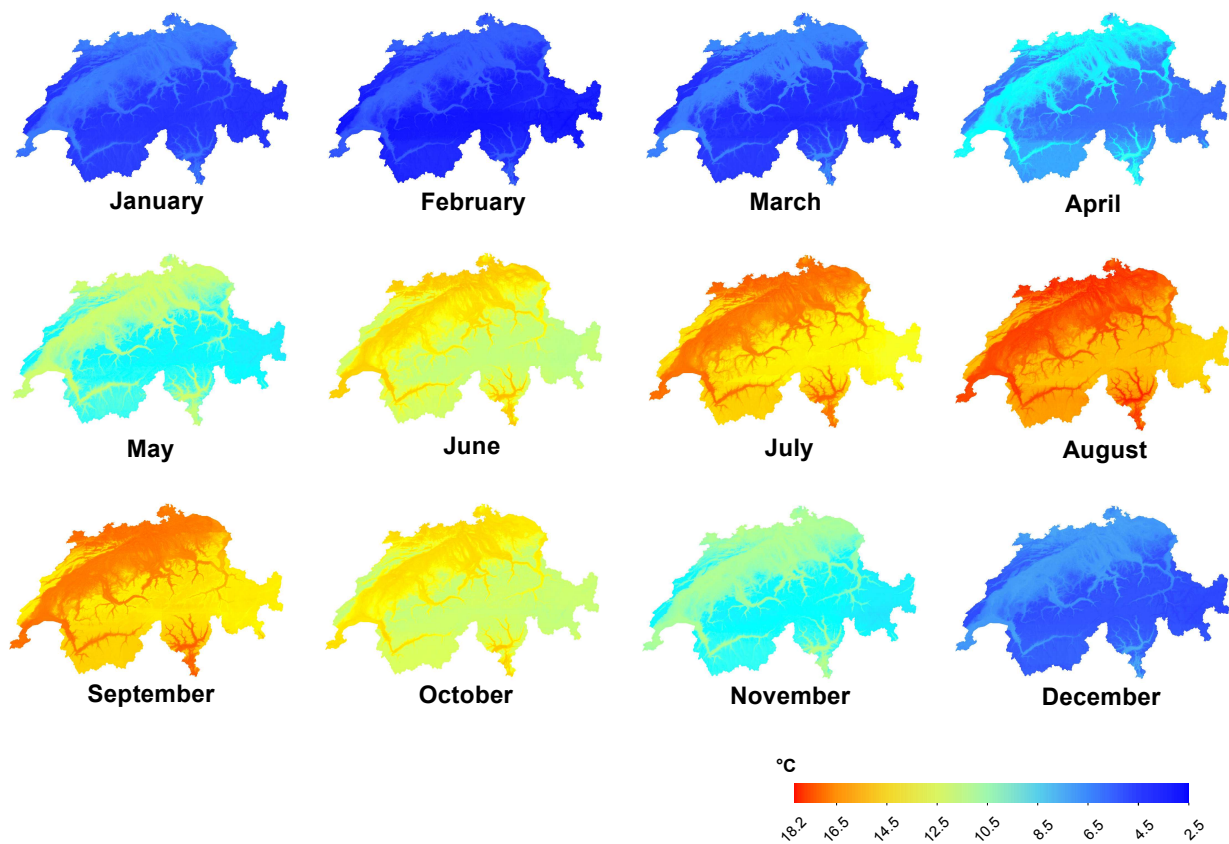


Figure 5.5: Monthly ground temperature maps as estimated for Switzerland for a depth of 100cm.

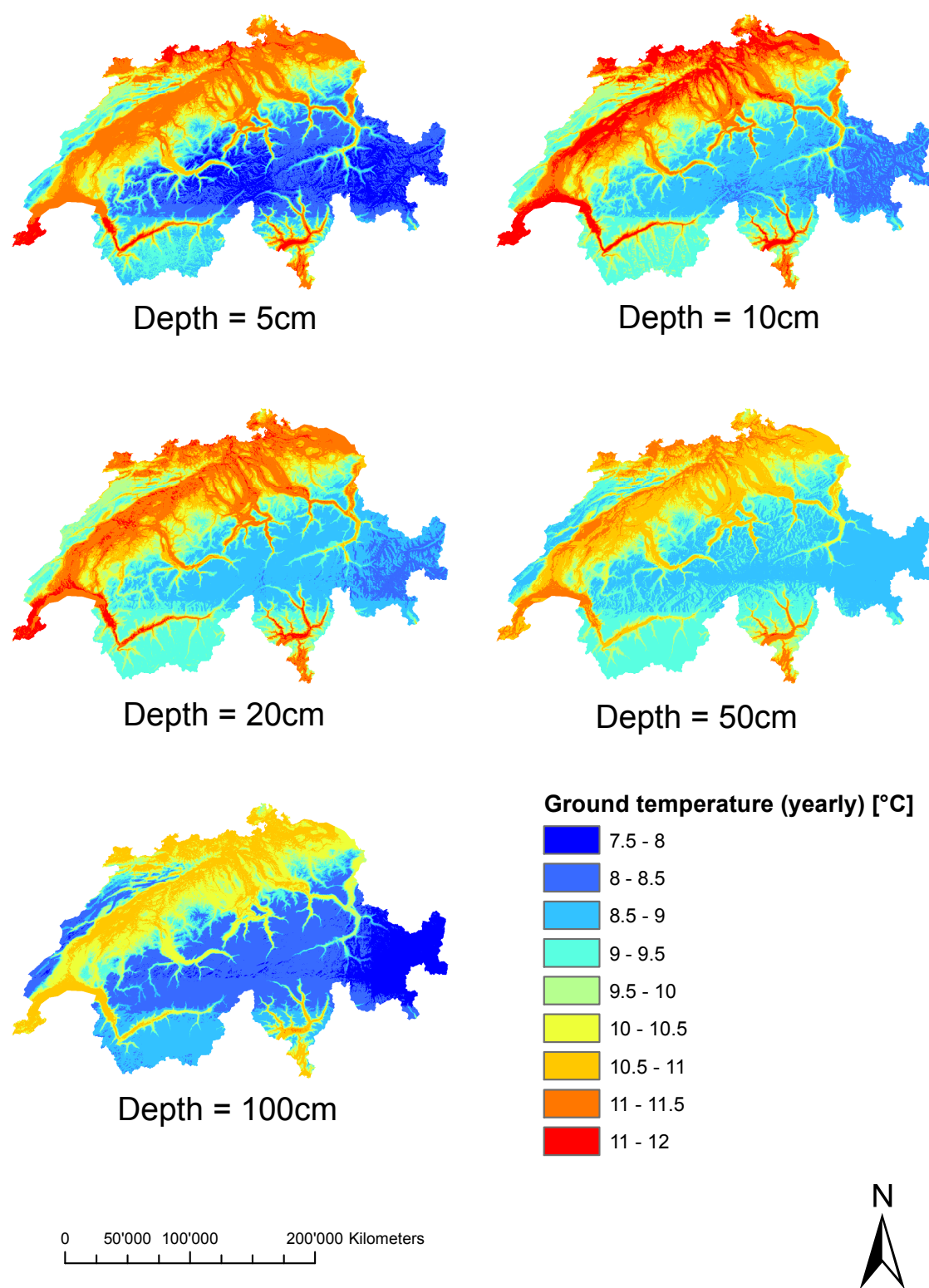


Figure 5.6: Yearly ground temperature maps as estimated for Switzerland at the depths of 5, 10, 20, 50 and 100cm.

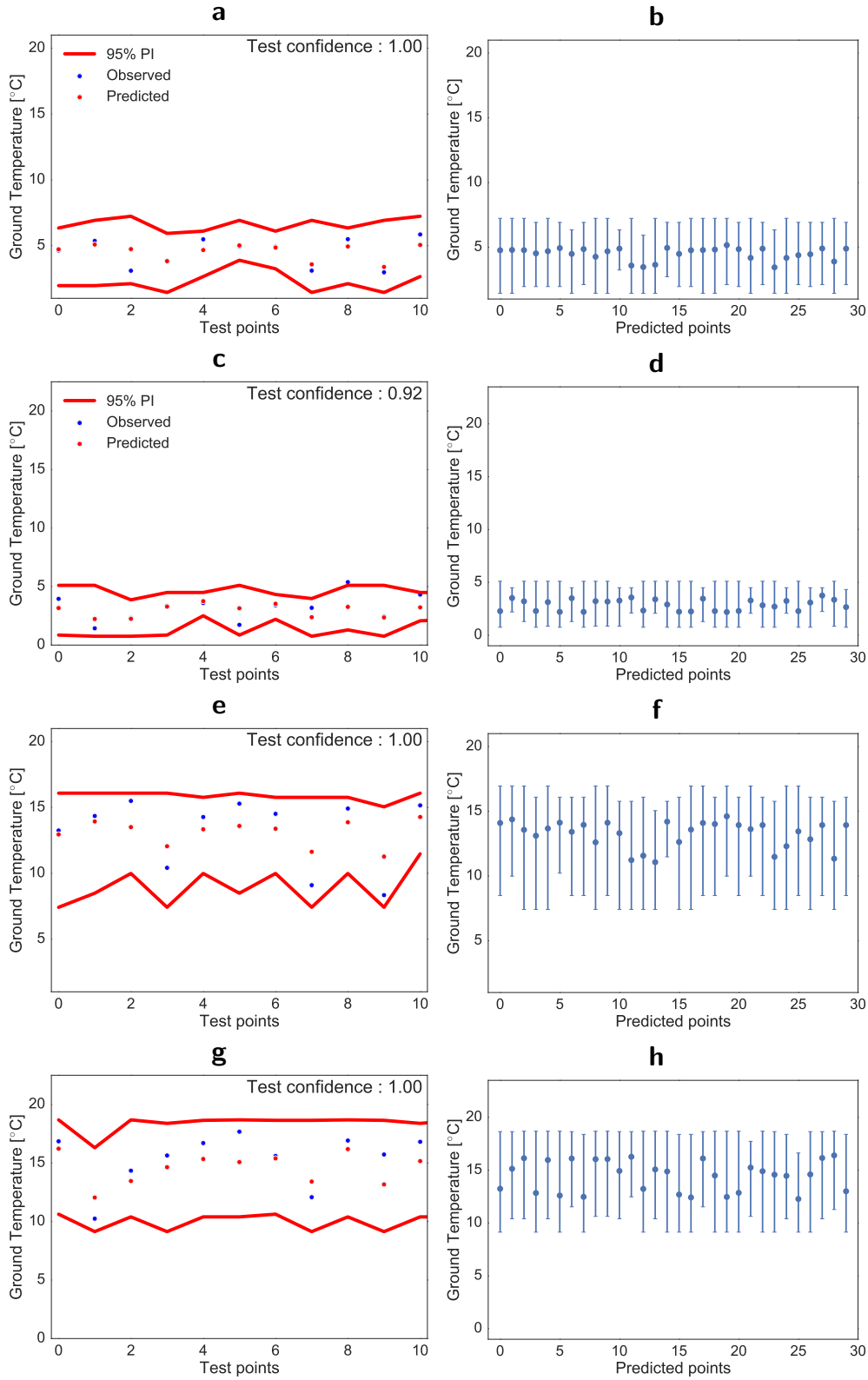


Figure 5.7: Prediction Intervals (with 95% confidence) from Quantile Random Forests for the monthly ground temperature at a depth of 100 and 50cm for an example of 2 months. (a) and (c): PIs in the test set, in January, respectively for 100cm and 50cm; (b) and (d): PIs for 30 random new points, in January, respectively for 100cm and 50cm; (e) and (g): PIs in the test set, in June, respectively for 100cm and 50cm; (f) and (h): PIs for 30 random unknown points, in June, respectively for 100cm and 50cm.

Table 5.2: Testing RMSE (E_R), NRMSE (E_{NR} , in percentage) and OOB score for Random Forest models trained for monthly ground temperature at multiple depths.

Month	5cm			10cm			20cm			50cm			100cm		
	E_R [°C]	E_{NR} [%]	OOB [-]	E_R [°C]	E_{NR} [%]	OOB [-]	E_R [°C]	E_{NR} [%]	OOB [-]	E_R [°C]	E_{NR} [%]	OOB [-]	E_R [°C]	E_{NR} [%]	OOB [-]
Jan.	0.81	25.40	0.17	0.69	20.81	0.46	0.57	19.18	0.43	0.63	23.17	0.25	0.87	17.55	0.26
Feb.	0.57	18.69	0.15	0.65	20.20	0.48	0.69	22.24	0.34	0.66	26.88	0.27	0.86	20.27	0.24
Mar.	0.78	13.74	0.35	0.94	16.24	0.46	1.34	20.44	0.40	1.32	35.38	0.40	1.04	20.58	0.44
Apr.	0.88	08.49	0.59	1.12	10.62	0.46	1.35	12.30	0.46	2.00	32.11	0.44	1.11	15.36	0.52
May.	1.48	10.38	0.52	0.97	06.62	0.34	1.35	08.86	0.34	1.80	17.21	0.33	1.19	11.58	0.47
Jun.	2.34	12.48	0.56	1.22	06.41	0.42	1.22	06.35	0.42	2.02	14.62	0.31	1.34	09.95	0.47
Jul.	2.22	10.55	0.46	1.08	05.06	0.48	1.18	05.59	0.40	2.21	13.77	0.25	1.54	09.74	0.40
Aug.	1.68	08.28	0.53	1.22	05.95	0.48	0.96	04.62	0.46	2.09	12.47	0.29	1.51	08.88	0.37
Sep.	1.10	06.57	0.46	1.02	05.86	0.45	1.05	05.99	0.38	1.90	12.90	0.23	1.44	09.03	0.42
Oct.	0.83	06.61	0.37	0.87	06.66	0.34	0.71	05.43	0.28	1.57	13.65	0.33	1.12	08.25	0.46
Nov.	0.93	11.90	0.24	0.86	10.44	0.32	0.60	07.18	0.31	1.19	16.35	0.30	1.14	11.25	0.38
Dec.	1.10	29.08	0.06	0.78	18.89	0.35	0.60	14.27	0.43	0.80	19.53	0.32	1.05	15.31	0.39

Vertical Electrical Soundings (VES) data, (ii) estimation of electrical resistivity values, (iii) spatial extrapolation of electrical resistivity in Switzerland, (iv) conversion of electrical resistivity into thermal conductivity. VES data is perhaps the most common resistivity data as it involves standard equipment and is a practical non-invasive geophysical study often performed to extract basic ground properties. The methodology presented here can therefore be re-used in other locations, should this sort of data be available. Also note that the conductivity could have been simply assumed from typical rock values. The focus of the study, however, is precisely to attempt to extract more accurate values from real data. The remaining parts of the section aim at explaining the details of the four previously mentioned steps.

5.4.1 Processing and interpretation of Vertical Electrical Soundings data

The estimation of electrical resistivity across Switzerland is based on a Vertical Electrical Soundings (VES) dataset created by the Swiss Geophysical Commission (see Table A.2). The data was built in an effort to gather multiple measurement studies performed over the last few years by multiple Swiss laboratories and universities [210]. The dataset is split into two parts: while a first part of the dataset offers the raw electrical measurements only and require interpretation (inversion), the second part is already interpreted, meaning it was already inverted. As a result, the two parts of the dataset were processed individually. The locations of all the points are shown in Figure 5.8.

The first part of the dataset, which requires interpretation, includes 4144 points. For each point, the 1D inversion set of functions from the pyGIMLi library [81] is used in order to provide an interpretation of the sounding data (see section 3.2.3 in chapter 3 for more details on vertical electrical sounding data). Note that the number of different resistivity layers of the soil n_s at the measurement location is a parameter of the inversion algorithm. Therefore, a simple tuning strategy is performed: for each point, (i) the point is inversed separately with multiple values of n_s from 2 to 10, (ii) the resulting forward model is used to compute the apparent resistivities corresponding to the multiple distances between the electrodes, (iii) the n_s minimizing the RMSE between the original measurements and the forward modeled values is picked. Once n_s is picked, we obtain the resistivity in each different layer of the ground at the measurement point. Figure 5.9 shows an example of the inversion results

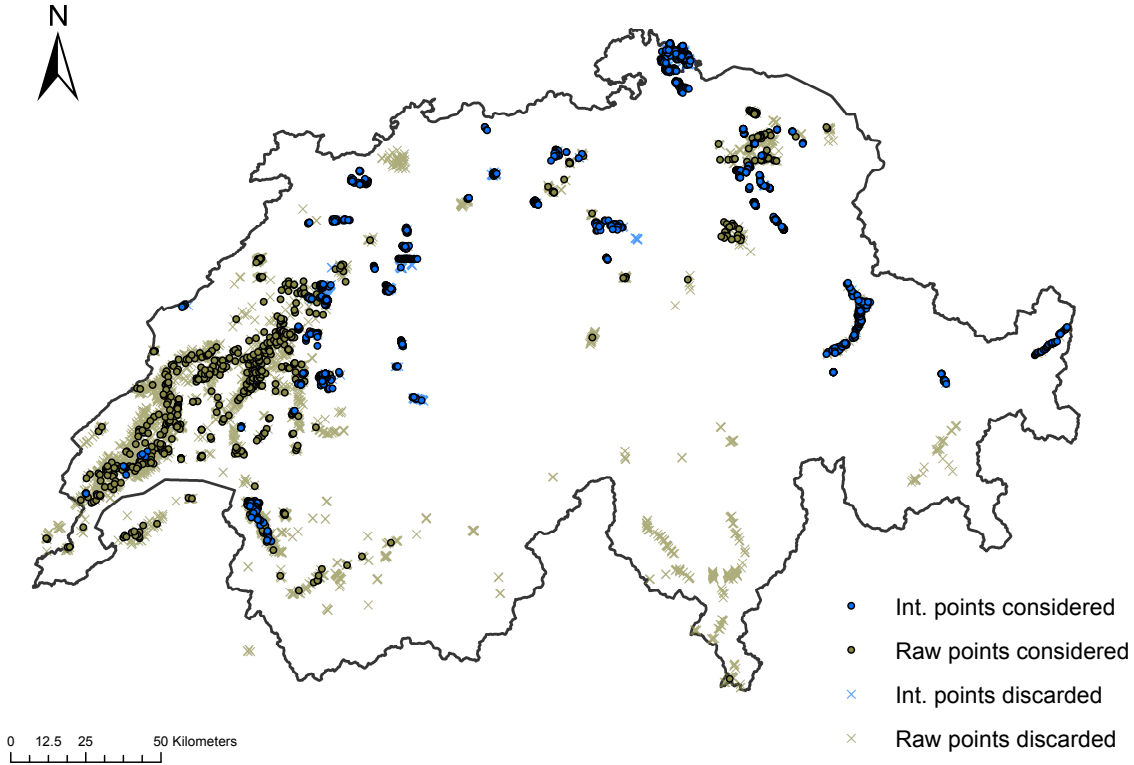


Figure 5.8: Vertical Electrical Sounding data point locations in Switzerland. “Int” stands for “Interpreted”.

for one point. Two additional constraints are considered in order to filter noisy data points: we do not consider a point for which (i) the depths of the layers are non-strictly-increasing, as it has no physical meaning, (ii) the RMSE is greater than 30%. Note that the choice of 30% as a maximum threshold for the RMSE is motivated by a tradeoff between accuracy and number of points considered. While a higher threshold than 30% would signify poor lead to a poor accuracy, a lower one would result in a very low number of points considered (around 5% of the original 4144 points data). It results in 694 interpreted points in the first part of the data.

The second part of the dataset, already interpreted, includes 1915 points. This part of the data specifies the layers’ thickness and apparent resistivity at each measurement location. Note that the method or algorithm used for inversion is unknown, as the studies leading to this data were often performed many years ago and did not specify their strategy. Therefore, after verifying that the values are realistic (between 0 and 100000 $\Omega.m$), the values have to be trusted. Furthermore, information (depth and/or resistivity) is missing in some points, which are excluded from the study. It results in 1521 interpreted points in the second part of the data.

The two parts of the data are merged together once both interpreted. Finally, the two data gather 2215 points across Switzerland.

5.4.2 Estimation and extrapolation of electrical resistivity

The whole VES data is processed to extract resistivity values at shallow depths for all 2215 filtered points. In order to obtain one shallow electrical resistivity value for each point, the resistivities at

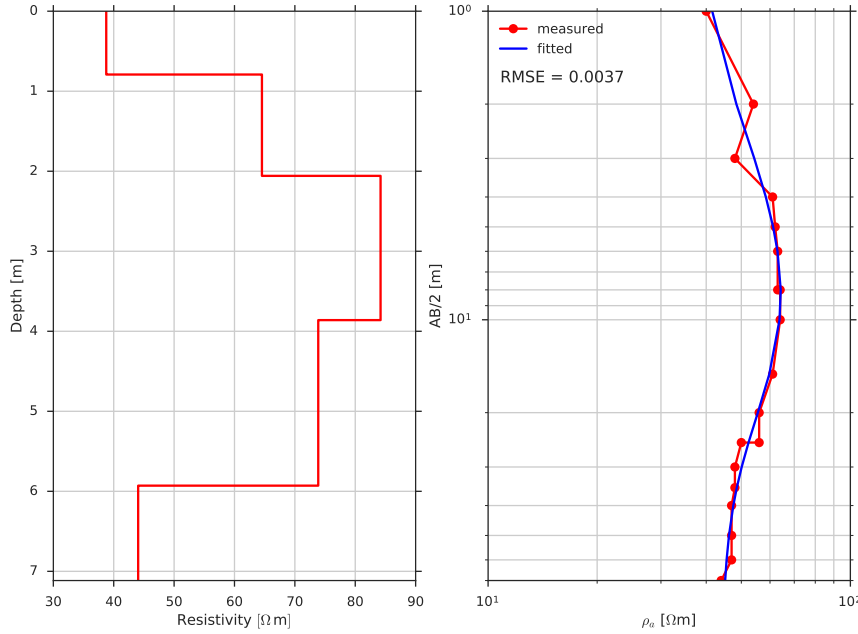


Figure 5.9: Vertical Electrical Sounding data point inversion example. The right graph shows the measured electrical resistivities given by Ohm's law for different distances between electrodes A and B in red and the forward model resulting from the inversion in blue (ρ_a is the apparent resistivity, and the RMSE between the two set of values is specified); the left graph shows the resulting depths and resistivities of the ground layers obtained from the inversion.

different depths are averaged through the first meter and weighted by the width of the corresponding layer. At each measured location, the average shallow resistivity ρ_{moy} is computed as follows:

$$\rho_{\text{moy}} = \frac{1}{h_{n_{s < 1\text{m}}} - h_0} \sum_{i=1}^{n_{s < 1\text{m}}} \rho_i (h_i - h_{i-1}) \quad (5.3)$$

where ρ_i is the interpreted resistivity in layer i , h_i is the width of layer i , and $n_{s < 1\text{m}}$ is the number of layers in the first meter of the ground (the total width of all layers in the first two meters is therefore $h_{n_{s < 1\text{m}}} - h_0$).

The estimated shallow electrical resistivity values are aggregated within the GK500 geology polygons (see Table A.2). Using the Joint Spatial function from ArcGIS to track the GK500 polygons covered by the VES interpreted points, the resistivity values from these points are averaged in each polygon. Note that in the case of a GK500 polygon containing only one interpreted point, this sole point defines the apparent electrical resistivity of the polygon. Eventually, 317 polygons are covered by at least one interpreted VES point and are therefore attached with an estimated shallow resistivity value.

In order to further extrapolate the electrical resistivity over the whole Switzerland, an RF model is trained. The 317 resistivity values are used to build the training labels, and the space, weather and all geological features (defined in section 5.2.2) sampled at the label points locations are used as training features. The label variable, however, is not the resistivity itself but $\ln(5 + \frac{\rho}{C})$ where ρ is the resistivity within the polygon and C is the number of $200 \times 200 \text{ [m}^2\text{]}$ pixels contained in the polygon. This small

modification allows to take advantage of the information of the polygon size while reducing the order of magnitude of the output. The RMSE, NRMSE and OOB score of the RF are respectively 0.25, 13 %, and 0.37, showing a good model performance in the test set. The RF model is finally used in order to estimate the electrical resistivity in each GK500 polygon over the whole Swiss territory. The obtained electrical resistivity map of Switzerland is shown in Figure 5.10. Also, 95% Prediction Intervals (PIs) for electrical resistivity predicted values have been computed, both in the test set and for new predicted points. A visualization of the PIs for 30 test and new points is shown in Figure 5.10. Note that the values plotted on the Y-axis are not the electrical resistivity ρ but the modified output used to train the RF ($\log(5 + \frac{\rho}{C})$ with C the number of pixels in the polygon of interest). Finally, in order to show the distribution of the uncertainty attached to the estimation, PIs have been computed for all GK500 polygons over Switzerland (besides the ones that were not considered from the start because of lack of data, or the ones used for training). As discussed already in chapter 4, the lower and upper width of the PIs can be seen as lower and upper prediction errors (respectively noted $PE_{s,down}$ and $PE_{s,up}$), and the overall prediction error attached to the estimation of the electrical resistivity in each GK500 polygon is the average $(PE_{s,down} + PE_{s,up})/2$. The resulting error map is shown in Figure 5.11.

5.4.3 From electrical resistivity to thermal conductivity

As explained in chapter 3, extracting the thermal resistivity based on existing electrical resistivity values requires information on the local structure/texture of the soil. While multiple models (presented in section 3.2.5) have been studied to perform such a conversion, they cannot be used in the present study because of a lack of a particular variable, namely the saturation degree S_r , over the whole of Switzerland.

We use, instead, the data collected by both studies in order to train a conversion model to predict the thermal resistivity from the electrical resistivity. It is then straightforward to extract the thermal conductivity as the inverse of the thermal resistivity. Both data gather 135 points with experimental values of electrical and thermal resistivity, along with other soil characteristics, for different types of soils with various texture and structure. All 135 points offer, in particular, the dry density γ_d , the gravimetric water content (GWC), the saturation degree S_r and the percentage sum of the sand and gravel fractions F . As saturation values are very challenging to gather at the scale of a country, we rather use the VWC to express the soil water content. The VWC can be obtained from γ_d and the GWC using Eq. 3.5 (3).

The conversion from electrical resistivity to thermal conductivity consists of the following steps:

1. Import the combined data from [90] and [91],
2. Train an RF conversion model, with, as features: the experimental values from the combined data for percentage sum of the sand and gravel fractions F , Volumetric Water Content (VWC) and electrical resistivity; and as outputs: the thermal resistivity values from the combined data. The RMSE, NRMSE, and OOB score of the conversion RF trained in (2) are respectively 0.16, 17.6% and 0.94, which shows very good performance of the conversion model.
3. For each GK500 polygon in Switzerland, extract the electrical resistivity (estimated in section 5.4) and the VWC mean (extracted from NASA SMAP data, as presented in section 5.2.2), and consider the 10 possible values for F (5%, 15%, ..., 95%) corresponding to the center of the 10 possible intervals presented in section 5.2.2 (I_1, \dots, I_{10}).

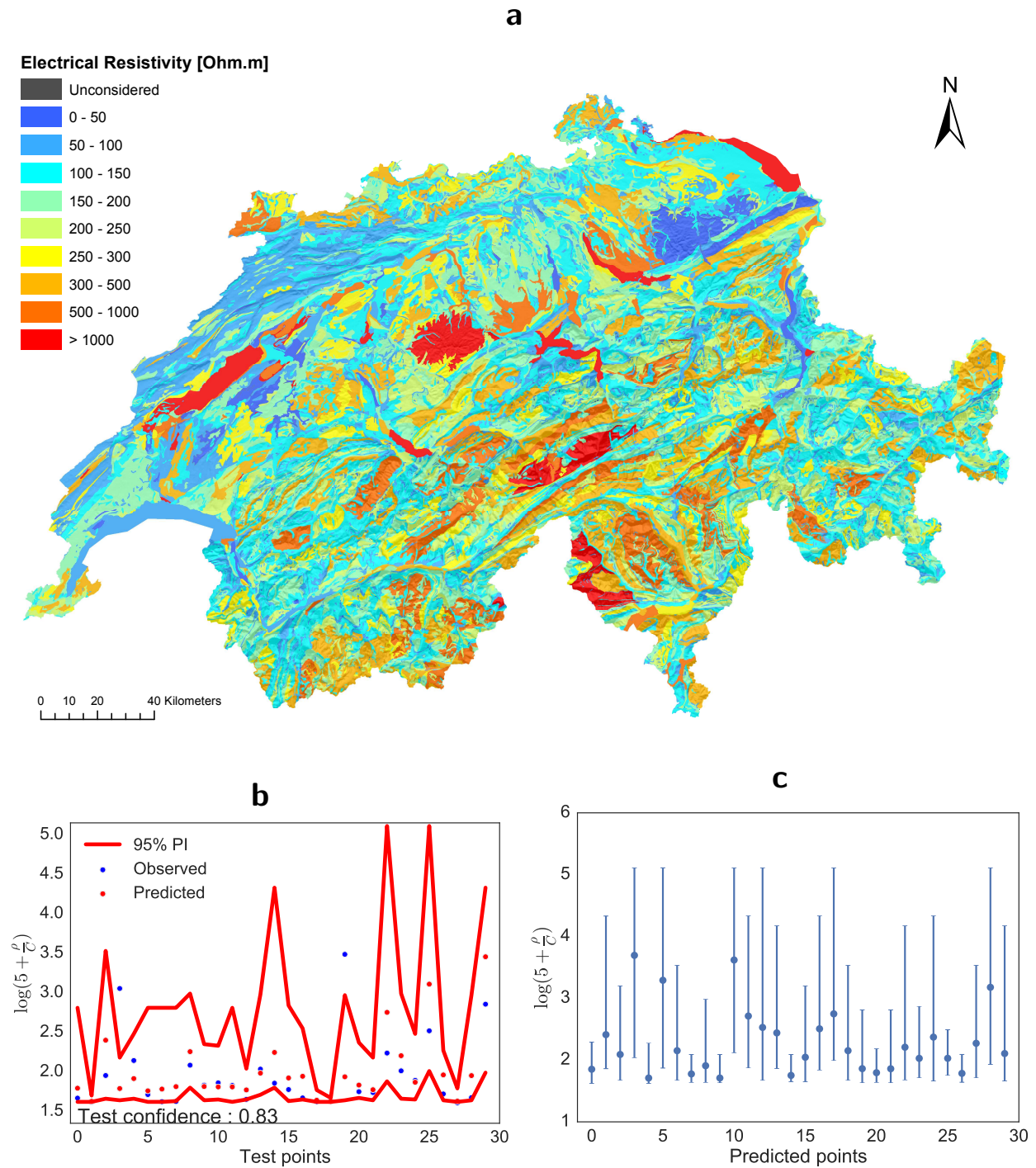


Figure 5.10: Electrical resistivity map as estimated in the study with visualization of PIs (with 95% confidence) both in the test set and for new points. **(a)** Electrical resistivity (ρ) map; **(b)** PIs in the test set for the label variable used ($\log(5 + \frac{\rho}{C})$ and not ρ) while training the RF model; **(c)** PIs for 30 random unknown points for the same label variable.

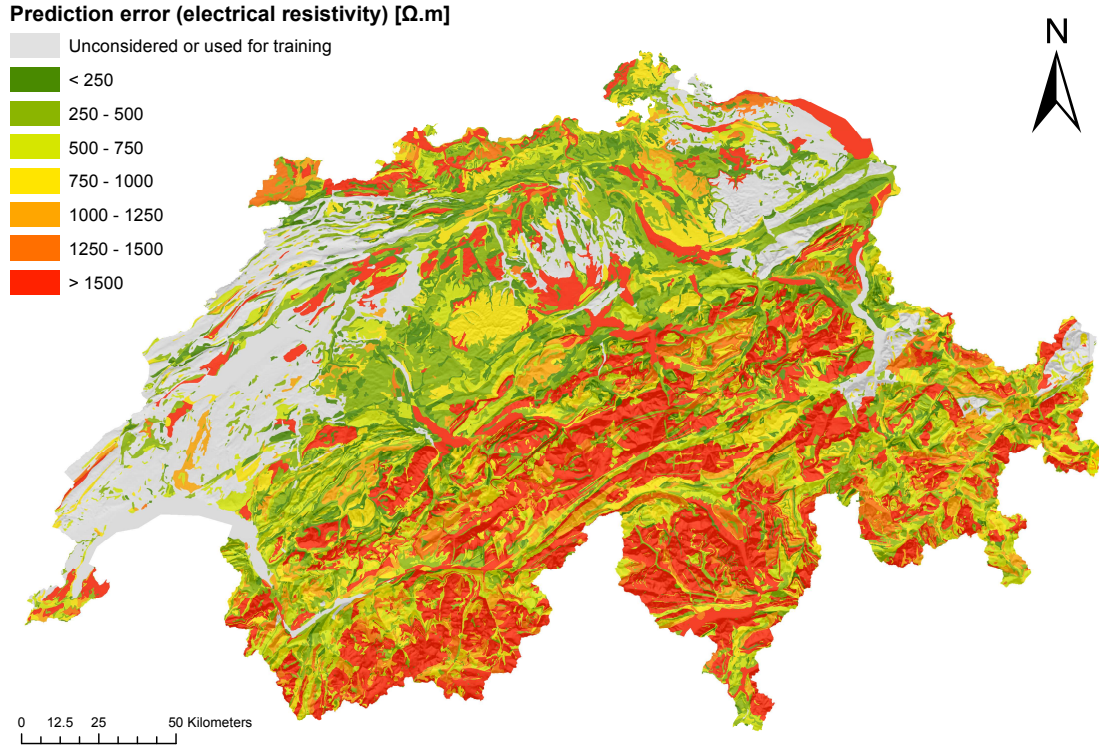


Figure 5.11: Prediction error attached to the electrical resistivity estimation (for each GK500 polygon) obtained from the trained RF model. The error is computed as the average of the down and up ($PE_{s,down}$ and $PE_{s,up}$) width of Prediction Intervals computed with Quantile Regression Forests.

4. For each GK500 polygon in Switzerland, use the trained conversion RF model together with the 10 possible configurations of VWC, electrical resistivity and F in order to estimate the thermal resistivity for all 10 possible F values.
5. Use the full joint PDFs extracted for each GK500 polygon (extracted from the NABODAT data, as presented in section 5.2.2) in order to extract the marginal PDF for F ($\mathbb{P}(F \in I_i)$), use it to weight the 10 possible thermal resistivity values with their respective probabilities and finally obtain the final thermal resistivity value for each GK500 polygon.
6. The thermal conductivity is then computed as the inverse of the thermal resistivity over the whole Switzerland.

Note that the joint PDF for soil texture is not fully used and only the probabilities for F are used; this is caused by the lack of sand, silt and clay content information within the experimental data used for thermal and electrical resistivity values [90, 91]. We computed the full joint PDF nonetheless as it is a valuable piece of information that can be useful for later research. The obtained thermal conductivity map of Switzerland is shown in Figure 5.12. In order to have an estimation of the RF prediction, 95% Prediction Intervals (PIs) were again computed for the conversion from electrical resistivity to thermal resistivity, before applying the F PDFs and the computation of the conductivity, both in the test set and for new predicted points. A visualization of the PIs for 30 test and new points

is shown in Figure 5.12. Note that the values plotted on the Y-axis are not the thermal conductivity λ but the modified output used to train the RF ($\log(\rho_t)$ where ρ_t is the thermal resistivity).

5.5 Thermal diffusivity estimation

The third and final thermal variable to estimate is the shallow thermal diffusivity, for each GK500 polygon in Switzerland. The estimation consists in two main steps: (i) estimation of shallow thermal diffusivity in locations where temperature data is available at various depths, using Fourier modeling of the 1D heat equation, and (ii) extrapolation of the diffusivity to the whole Switzerland.

5.5.1 Fourier modeling for thermal diffusivity estimation

The Fourier modeling strategy presented in section 3.2.4 (chapter 3) to estimate the apparent thermal diffusivity is applied to several locations in Switzerland where shallow ground temperature data is available from MetwoSwiss at 5, 10, 20, 50 and 100cm, as presented in Table A.1.

The Fourier series for the daily ground temperature is estimated at each available station, at each depth and for each available year using the FFT algorithm. The constant term of the series, which is by definition the average yearly temperature ($T_{0,z}$), is first computed. Then, the amplitudes and phases (R_n and ϕ_n) of the frequencies for the multiple harmonics of the Fourier series are computed over a period of one year, meaning $P = 365.24$ days and $\omega = \frac{2\pi}{365.24}$. Although the first three harmonics ($n = 1, 2, 3$) are often enough to reproduce the signal with a good approximation [83], we consider three dominant harmonics, namely the ones with the highest amplitudes. It allows for a better estimation of the signal. The reconstructed signal from Fourier analysis for one station (Bern), at one depth (20 cm), and for one year (2013) is shown in Fig. 5.13, together with the 30 first computed amplitude and phase values.

Following the strategy presented in section 3.2.4, the slope of $\ln(R_n)$ vs. $z\sqrt{n}$ is then estimated, separately for the [5, 10, 20] cm and the [50, 100] cm time series. The estimated slopes in one station (Bern) for the [5, 10, 20] cm time series are shown in Fig. 5.14 for 2013, 2014, 2015 and 2016, where three dominant harmonics are defined by $n = m_1, m_2, m_3$. The linear fit for each of the three dominant harmonics is shown for each year. In Fig. 5.14 the slope does not vary greatly from one harmonic to the other, which validates the uniformity assumption. The mean slope is then computed by averaging all three harmonic slopes, which gives the damping depth and finally the apparent thermal diffusivity (using Eq. 3.11) for each of the 49 locations and each available year.

For each station, the estimated yearly thermal diffusivity estimations are cross-validated with typical values for various common rocks and soils given for two saturation states [211], given the type of rock from the GeoCover500 polygon data (see 5.2.1). The value is validated if it is within the typical minimum and maximum values $\pm 0.5 \times 10^{-6} \text{ m}^2/\text{s}$ of the corresponding rock for each year, otherwise it is discarded. The final diffusivity values for each of the 49 available stations are computed as the average year diffusivity value for that station.

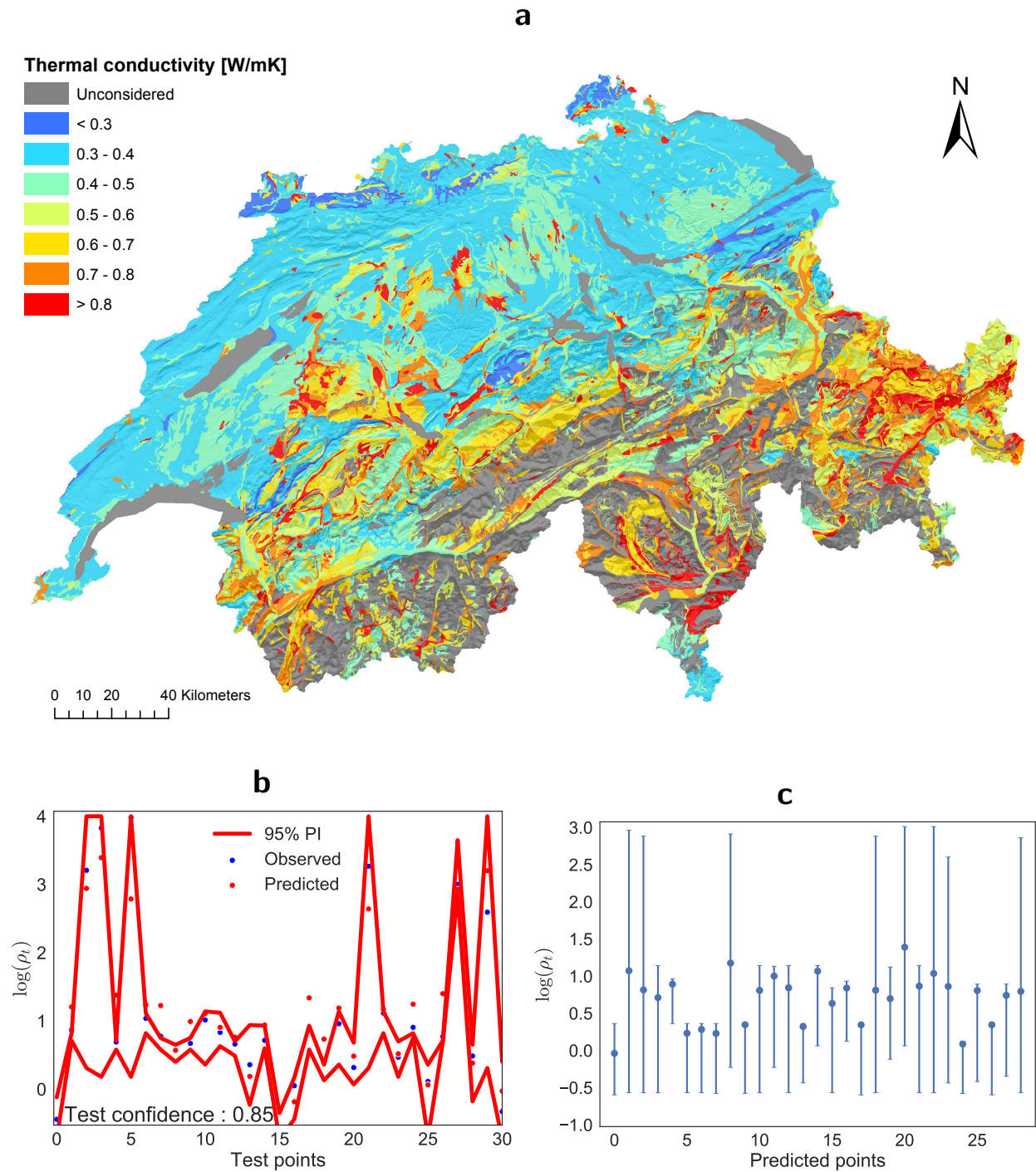


Figure 5.12: Thermal conductivity map as estimated in the study with visualization of PIs (with 95% confidence) both in the test set and for new points. **(a)** Thermal conductivity (λ) map; **(b)** PIs in the test set for the label variable used ($\log(\rho_t)$ where ρ_t is the thermal resistivity) while training the RF model; **(c)** PIs for 30 random unknown points for the same label variable.

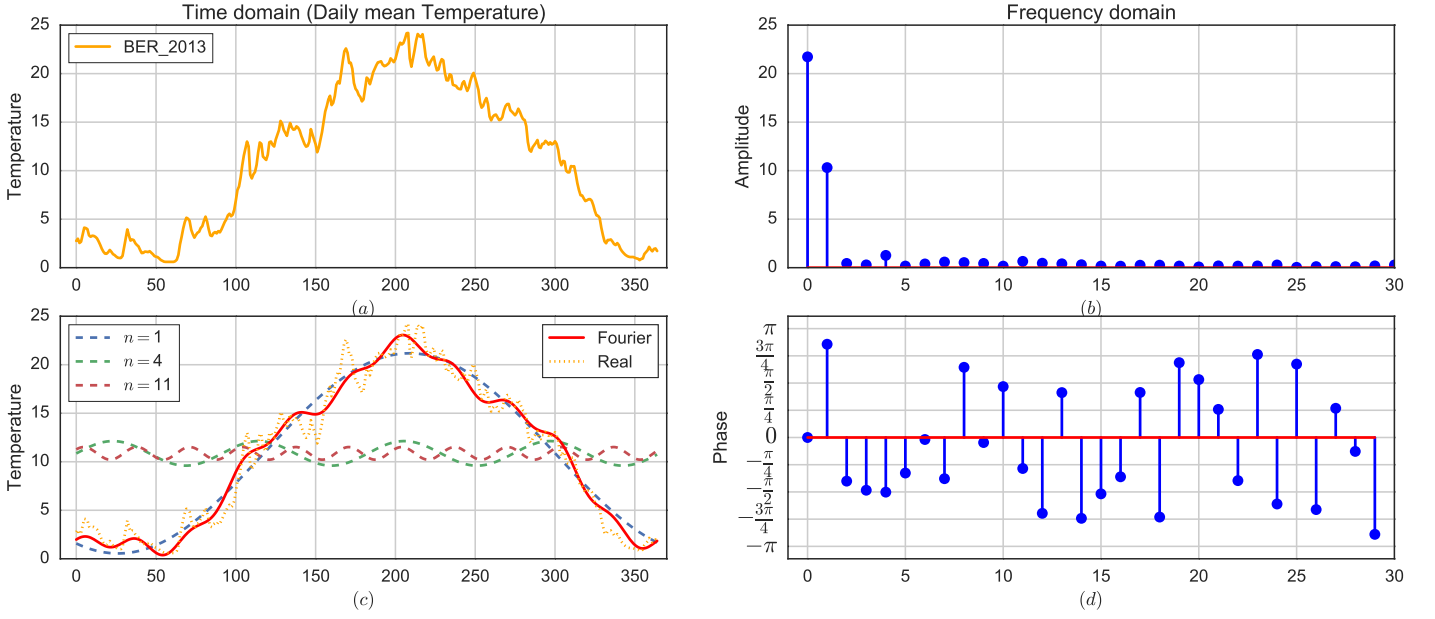


Figure 5.13: Fourier analysis of a ground temperature time series example. The figure shows, for depth of 10cm, in Bern, during the year 2013: (a) daily average temperature time series over the year, (b) and (d) amplitude and phase of the 30 first frequencies, (c) harmonics for the 3 dominant frequencies (here $n = 1$, $n = 4$ and $n = 11$) and resulting Fourier approximation of the signal.

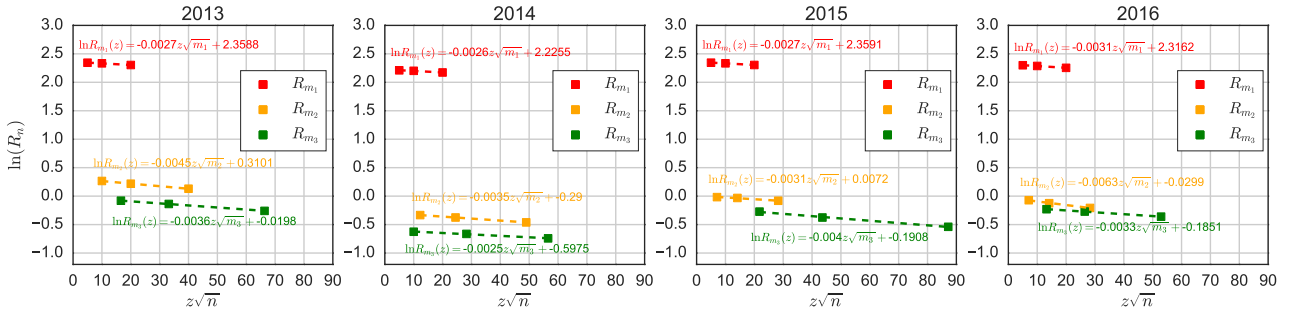


Figure 5.14: Slope fitting for diffusivity estimation example. The linear fit between $\ln(R_n)$ and $z\sqrt{n}$ is shown for the three dominant harmonics ($n = m_1, m_2, m_3$) in Fourier analysis in Bern, in 2013, 2014, 2015 and 2016.

5.5.2 Extrapolation of diffusivity

In order to further extrapolate the estimated thermal diffusivity over the whole Switzerland, an RF model is trained. The training data is represented by the GK500 polygons in which we estimated the diffusivity. In the case of multiple stations within one GK500 polygon, the considered diffusivity in the polygon is the average of the diffusivities at the included stations, otherwise the sole estimated diffusivity value defines the diffusivity over the whole polygon. It results in 47 training (polygon) points.

For each point, the estimated diffusivity is to be used as label for the RF model, and the considered features are the following: the space, weather and GK500 geological features, as well as the soil moisture and the soil texture statistics features (all presented in section 5.2.2). The soil texture statistics, however, are not available for all polygons throughout the country as some of the possible rock types are not represented by a NABODAT soil texture measurement point (as explained in section 5.2.2). As a result,

a second RF model is trained over the 47 training points without the soil texture features. This latter model can later be used to obtain diffusivity estimations in polygons that lack that texture information.

To ease the prediction of the thermal diffusivity, the considered label (output value) during the training process is slightly modified: instead of the diffusivity α , we consider $\ln(\alpha \times C)$, where C is the number of $200 \times 200 \text{ m}^2$ pixels contained in each GK500 polygon. The resistivity estimation makes it possible to take advantage of the information of the polygon size, while reducing the order of magnitude of the output. The testing RMSE, NRMSE and OOB score are respectively $0.69 [10^{-6} \text{ m}^2/\text{s}]$, 13.7% and 0.87, for the RF model considering the soil texture information, and $0.82 [10^{-6} \text{ m}^2/\text{s}]$, 16.4%, and 0.78% for the RF model not considering the soil texture information. While both RF models show good performances in the test set, the model taking soil texture into account has a better accuracy, which aligns with intuition. The apparent thermal diffusivity is then estimated in all GK500 polygons in Switzerland, using the previously trained RF models (taking $e^{(\cdot)/C}$ of the prediction to recover the diffusivity from the predicted modified label). The obtained thermal diffusivity map is shown in Figure 5.15. Finally, 95% Prediction Intervals (PIs) have been computed for the estimated diffusivity values, for both RF models (with and without soil texture information), both in the test set and for new predicted points. A visualization of the PIs for the test points and 30 new points is shown in Figure 5.15. In addition, the prediction error estimated from the PIs computed for all GK500 polygons show the spatial variation of the uncertainty attached to the thermal diffusivity estimation (Figure 5.16).

5.6 Results

5.6.1 Discussion

Three main variables affecting shallow geothermal potential, namely the monthly ground temperature, the ground thermal conductivity and the ground thermal diffusivity, have been estimated at a shallow depth of 1m over the entire Swiss territory. To match the resolution of the wind and PV solar energy related studies, the estimated maps for thermal conductivity and thermal diffusivity are re-aggregated from the GK500 vector polygon resolution into the $200 \times 200 [\text{m}^2]$ raster pixel level in order to match the resolution of the ground temperature maps (and other potential studies performed in the thesis). For both variables, the value attributed to each pixel was computed as follows: (a) if the pixel is fully included within a GK500 polygon, then the value of the pixel is the value estimated for the polygon, (b) if the pixel is located at the boundaries of multiple polygons, then the value of the pixel is the weighted average of the values estimated for the multiple polygons, the respective polygon weights being the percentage of area they cover within the pixel. Note that for pixels in case (b), the polygon area percentages were computed using the Tabulate Intersection tool from the Statistics toolbox within ArcGIS.

Note that in the present case of geothermal heat pumps, it is not straight forward to assess, even crudely, a power or an actual energy generation potential, as it was for wind energy. Indeed, in addition to the suitable areas for heat pump installations and the design and characteristics of heat pumps, an energy calculation requires an estimation of the heating demand and the desired temperature in each building using the pump, as it is the difference of temperature between the cold and hot sources that determines the induced heat flow and ultimately the geothermal potential. These aspects are therefore to be considered for a future technical potential study. It is desirable, however, that each of the maps

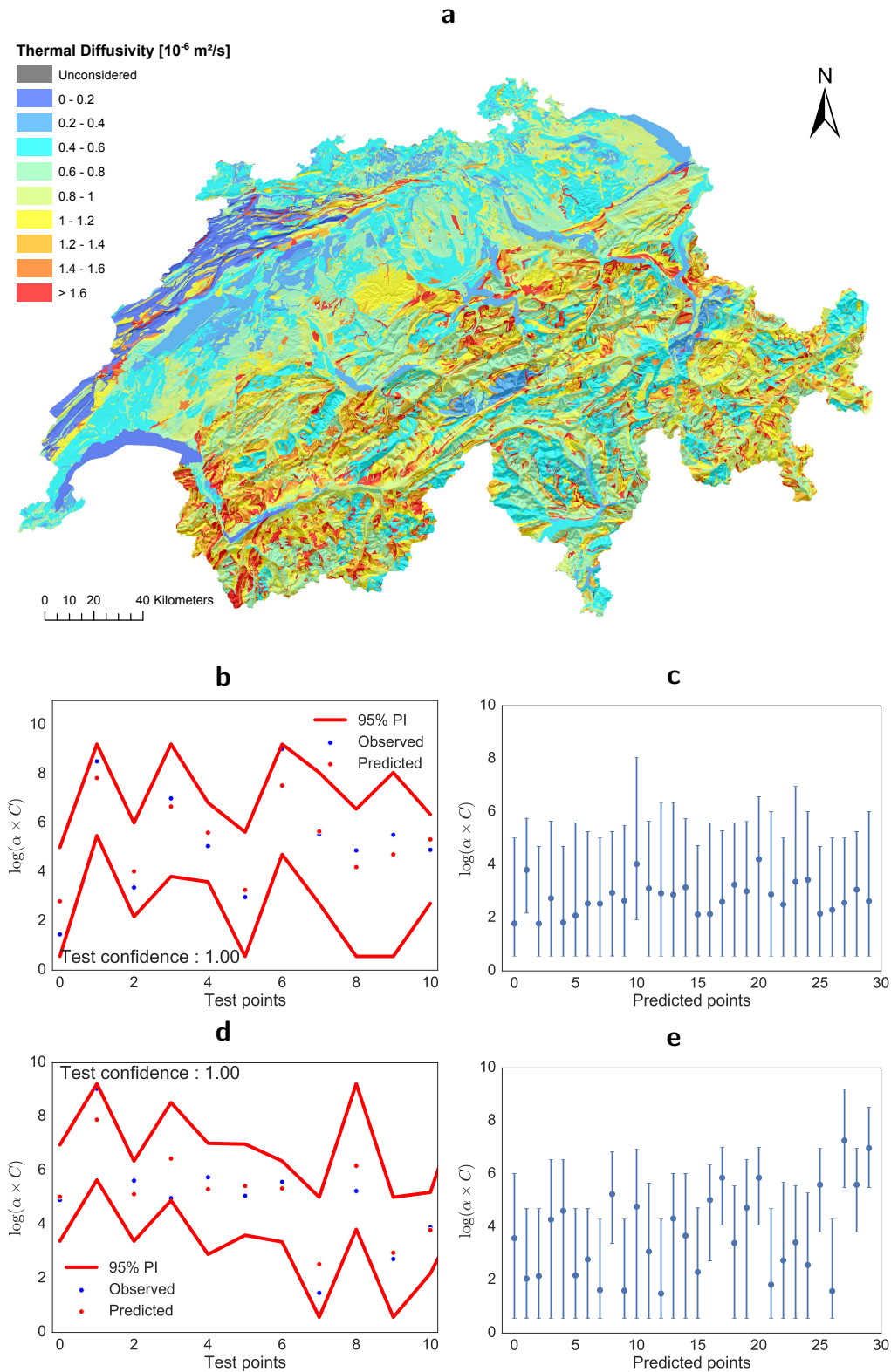


Figure 5.15: Thermal diffusivity map as estimated in the study with visualisation of PIs (with 95% confidence) both in the test set and for new points. **(a)** Thermal diffusivity (α) map; **(b)** PIs in the test set for the label variable used ($\log(\alpha \times C)$) while training the RF model, with soil texture features considered; **(c)** PIs for 30 random unknown points for the same label variable, with soil texture features considered; **(d)** PIs in the test set for the label variable used ($\log(\alpha \times C)$) while training the RF model, without soil texture features; **(e)** PIs for 30 random unknown points for the same label variable, without soil texture features.

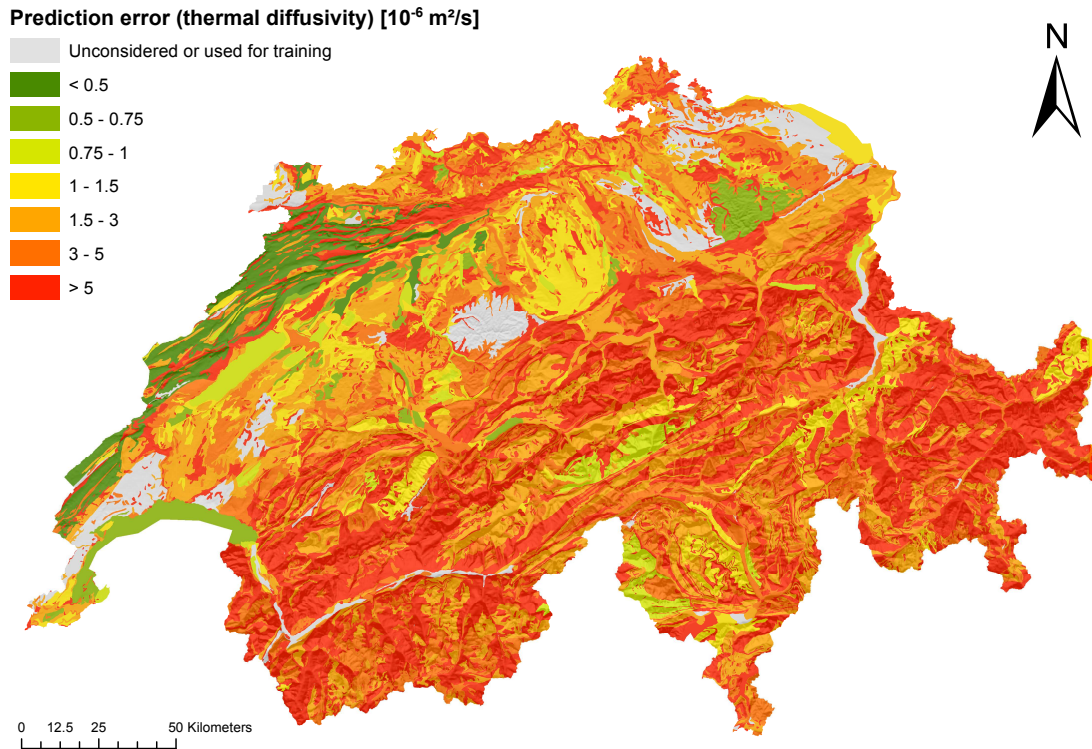


Figure 5.16: Prediction error attached to the thermal diffusivity estimation (for each GK500 polygon) obtained from the trained RF model. The error is computed as the average of the down and up ($PE_{s,down}$ and $PE_{s,up}$) width of Prediction Intervals computed with Quantile Regression Forests.

obtained for the three variables are discussed and validated in some fashion, with the current available information at hand, before proceeding with the further combination of the estimations.

Figure 5.6 presents the obtained yearly ground temperature maps for four different depths (5, 10, 20, 50 and 100 cm). The figure shows an instinctive pattern: the smaller the depth the closer the ground temperature is to air temperature. Conversely, the deeper the depth the less the ground temperature is affected by spatio-temporal variations. In particular, while the ground temperature is always higher in the Swiss Plateau (where the altitude is lower) than in the Alps (where the altitude is higher), one can observe on the yearly maps that the general temperature difference between the two regions is smaller at a depth of 100cm (difference of around 3°) than at a depth 5cm (difference of around 4 to 5°). This latter depth of 100cm is highly relevant for very shallow ground source heat pump installations, the most common ones (horizontal collectors) which are mostly at 1-2 m depth, and hence is the one requiring more attention. In Figure 5.5, one can observe the monthly variations of the ground temperature at 100cm. Although the seasonal variations are significant, they are not so drastic as in the uppermost tens of centimeters, where the temperature very much resembles the air temperature. In particular, the coldest and warmest periods are characterized by less extreme temperatures. The depths below the uppermost tens of centimeters allow the use of ground-source heat pumps in the winter, as 3 to 4°C is sufficient for heating purposes. One can observe on the Prediction Intervals (PIs) on Figure 5.7 that the test confidence is higher than 95% for all four PIs shown in the test set. While the PI on Fig. 5.7c has a confidence of 92%, it is the result of one observed

point being very slightly outside of the PI boundary. Also, even though the test set is very small, the respective impacts of ground depth and seasonality in Switzerland are shown in the PIs extracted for new predicted points: (i) the width of a PI is generally larger at 100cm than 50cm, as the temperature at 100cm is more stable throughout the year and with a slightly broader range of possible values than at 50cm, where a specific month dictates a narrower range of values reflecting the air temperature, (ii) the width of a PI is generally larger in spring/summer than in winter, as the temperature shows often more fluctuations in spring/summer at higher values (between 10°C and 20 °C) than in the winter (between 0°C and 6-7 °C)), where it is naturally cold but rarely gets very cold.

The electrical resistivity map (Fig. 5.10) gives a first intuition of suitable locations for ground source heat pumps installations. Naturally, the most suitable locations are the ones characterized by smaller resistivity values, meaning the blue regions in the map. A significant number of areas seem to be suitable, including a medium to large portion of the Plateau and the Geneva and Vaud cantons. Also, a large amount of small polygons spread all over the country show low resistivity values, notably in the Plateau, in the Valais canton in the south and the eastern cantons (Uri, Glarus, Graubunden). When looking at the PIs in Fig. 5.10, one can observe that the test confidence is, even though acceptable, not very high, and more particularly lower than 95%, which indicates that the PIs for this variable are not always reliable, and must therefore be considered with caution for new predicted points. The PIs in figure 5.10c, however, gives a good indication of the typical uncertainty attached to the prediction of electrical resistivity. In the case of the 30 unobserved random points shown, it is rather high since it is on average around ± 1 (between 2 and 4) for the modified output, corresponding to around an error of $50 \times C \text{ } \Omega.m$, which can be significantly high depending on C . This is partly caused by the extremely high range of possible values for electrical resistivity. In the error map (Figure 5.11), the distribution of uncertainty can be seen across the Swiss territory. While the uncertainty is generally higher in the Alps region, its distribution over Switzerland is rather heterogenous. It depends on the range of the output values, as mentioned previously, but also the density of training data available in different regions (which explains notably the high error in the Alps, where very few polygons are known), and the similarity (with respect to the features) between the unobserved polygons and the training polygons. Note, however, that this resistivity map is a intermediate step to the estimation of the thermal conductivity and that the electrical properties of the ground may differ significantly from the thermal properties. Therefore, the thermal conductivity map must be analyzed to extract valid conclusions.

Fig.5.12 shows the estimated thermal conductivity values in Switzerland. A first natural observation is the range of the estimated values, which seems relatively small. These values, however (ranging from 0.25 to more than 1 W/mK), are perfectly normal for unsaturated clay, silt and sand ([211]), which are the most frequent type of soil in the surface (mostly quaternary) layer of the ground ($< 1m$), the depth considered in the present study. The thermal conductivity seems to be higher in the center-west of the plateau, the southern (Ticino and Valais cantons), and the eastern part of Switzerland (St. Gallen canton), which are therefore suitable locations for very shallow geothermal installations. The northern part of the country, on the contrary, including some of the most dense urban areas of the plateau are not characterized by a high conductivity, notably Zurich, Vaud, Geneva and Neuchatel cantons have also a moderate potential for the very shallow installations. Note, however, that the systems considered in the present study are characterized by a very shallow depth, which are systems that are nowadays more adapted to rural areas, due to the lack of horizontal space in the ground in

urban areas (even though some interesting technologies, such as heat basket, are being developed to avoid that issue). For dense cities, Borehole Heat Exchangers (BHE) are usually preferred as they need less horizontal space and have a very high efficiency.

Ground source heat pumps therefore still remain an excellent option for the densely populated areas of Switzerland, but require a potential study at larger depths, which is not the focus of the present study. Also, note that the lack of measurement data for soil texture results in a significant number of unconsidered (grey) polygons, which are the polygons for which the type of texture was unobserved. Fortunately, the great majority of these unconsidered polygons are either lakes (Neuchatel, Bienne, Leman, Brienz, Zurich etc.), either located in the Alps, where the population is very low. The potential information is then naturally less significant in these areas. When looking at the PIs in Figure 5.12, one can notice that similarly to the first estimation of electrical resistivity, the test confidence is not very high, showing again a difficulty to extract a PI for conductivity/resistivity variables. Furthermore, for the 30 random samples the width of the PI is on average $[-0.5, 1.5]$ in terms of $\log(\rho_t)$, which corresponds to an error of ± 0.13 [W/mK] in terms of thermal conductivity. This uncertainty is significant but acceptable, given the limited information used to perform the conversion from electrical to thermal resistivity. Note, however, that the uncertainty from the electrical resistivity estimation naturally propagates into the conversion step, which is not taken into account by the QRFs when computing the PIs. The present estimation of the uncertainty is therefore not fully accurate and only offers an approximation. That is why the error map was not plotted, to avoid showing partly inaccurate information.

Fig. 5.15 shows the estimated thermal diffusivity in Switzerland. Although the impact of the diffusivity is less important than the conductivity, it is nonetheless of great importance when a potential study is being conducted, as it gives information on how fast the heat is being conducted through the ground and is needed in the modeling of the heat conduction between the ground and tubes of the installation (for further geographical and technical potential estimations). The regions with the highest diffusivity are the Valais Canton in the south and the east cantons: Schwyz, Obwalden, Nidwalden and Glarus. The PIs shown on Fig. 5.15 are notably more reliable than for the thermal conductivity estimation as the test confidence is better than 95% in both feature cases (with soil texture and without soil texture). The uncertainty in the prediction, as shown by the width of the PI for unobserved points, however, is still significant. In the random sample the width of the PI is on average $[1, 5]$ in terms of $\log(\alpha \times C)$, which corresponds to $\pm 45/C \cdot 10^{-6} \text{m}^2/\text{s}$, with C the number of pixels in the polygon. By construction of the label, the uncertainty decreases with an increasing size of the polygon. Also, the uncertainty is slightly lower when the soil texture is not considered (Figures 5.15d and 5.15e), specially for points with high predicted values. It seems that the smaller quantity of information entails a larger confidence for the model, which has fewer possibilities to consider and ultimately finds easier to predict a variable with fewer features used during the training process. It seems that the multiple trees in the forest agree more with each other when the number of features is small, which leads to a lower uncertainty. Note, however, that it does not mean that the accuracy of each prediction is necessarily higher, but that the variance of the tree predictions is on average slightly smaller. In addition, the changes in uncertainty with space can be observed from Figure 5.16. As discussed the resistivity, the uncertainty is again higher within the Alps, which is partly explained by the very small number of training points within this region. The uncertainty

is also higher for small polygons, as previously explained. The general distribution is, like in the case of resistivity, quite heterogeneous, with values of uncertainty ranging from $0.2\text{--}0.5 \cdot 10^{-6} \text{m}^2/\text{s}$ to values of $5 \cdot 10^{-6} \text{m}^2/\text{s}$ and even more in the most uncertain polygons.

The estimated variables were aggregated (averaged) at the canton level in Switzerland in order to have a higher level view of the very shallow geothermal potential. Volumetric Heat Capacity was also included in the aggregation, as it is commonly paired with thermal conductivity in geothermal potential studies. Plots for thermal diffusivity and electrical resistivity, and heat capacity and thermal conductivity can be seen in Figure 5.17. Notably, it can be observed that the cantons with the highest mean thermal conductivity values are Graubunden, Ticino and Valais, followed by Glarus, Uri and Bern.

5.6.2 Preliminary geographical potential estimation

In order to provide a better idea of the actual energy potential for very shallow geothermal systems, it would be desirable to estimate, at least at a preliminary level, their geographical potential (for both heating and cooling) in Switzerland.

To that aim, we use the following steps, in each $200 \times 200 \text{ [m}^2\text{]}$ pixel:

1. The total potential heat Q that one can remove from (for heating) or store (for cooling) in the ground can be approximated by:

$$Q = c_v V \delta T \quad (5.4)$$

where c_v is the volumetric heat capacity estimated in the pixel, V is the available ground volume in the pixel and δT is pixel temperature difference suffered by the ground, here calculated as the difference between the mean underground temperature (at 1m depth) and the mean surface air temperature (during the heating or cooling season). The ground volume is computed considering the ground surface area available in a pixel when discarding the pixel areas covered by building footprints (using ArcGIS). Then $V = 1 \times (\text{available ground surface area})$, considering the ground depth of 1m.

2. We use the Coefficient Of Performance (COP) of a potential very shallow system to compute the final heat transferred to the house (for heating) or heat removed from the house (for cooling).

On the one hand, the definition of the COP is given by:

$$\text{For heating: } \text{COP}_{\text{heat}} = \frac{|Q_h|}{W} = \frac{Q_h}{Q_h - Q_c} \quad (5.5)$$

$$\text{For cooling: } \text{COP}_{\text{cool}} = \frac{|Q_c|}{W} = \frac{Q_c}{Q_h - Q_c} \quad (5.6)$$

where W is the work required by the considered system, Q_h is the heat supplied to the hot reservoir and Q_c is the heat removed from the cold reservoir. When heating, the hot and cold reservoirs are respectively the house and the ground, and when cooling, the roles are inversed.

On the other hand, the COP can be estimated in practice in each pixel based on the Carnot model (offering the best possible COP for a heat pump) for both the heating and cooling season, as follows:

$$\text{For heating: } \text{COP}_{\text{heat}} = \eta \frac{T_h}{T_h - T_c} \quad (5.7)$$

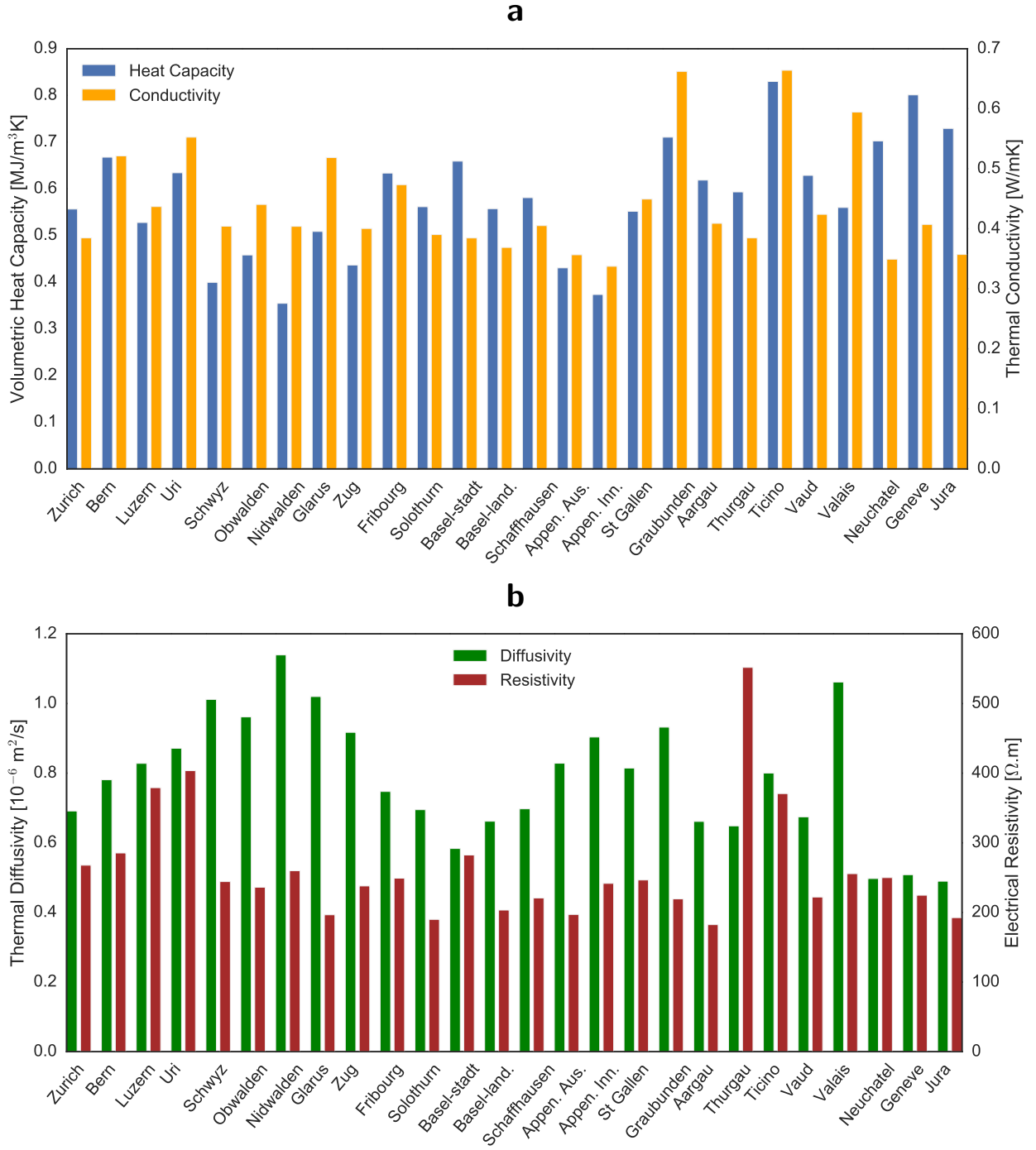


Figure 5.17: Estimated variables aggregated (average) at canton level. **(a)** Volumetric Heat Capacity and Thermal Conductivity aggregated at canton level, **(b)** Thermal Diffusivity and Electrical Resistivity aggregated at canton level.

$$\text{For cooling: } \text{COP}_{\text{cool}} = \eta \frac{T_c}{T_h - T_c} \quad (5.8)$$

where η is the efficiency of the system ($\eta = 1$ corresponds to a Carnot system, assuring the maximum theoretical efficiency) and T_h and T_c are the temperatures of the hot and cold reservoirs

respectively, in Kelvin. For heating, T_c is the mean ground temperature and T_h is the desired incoming temperature to the house for heating, which is here set to 35°C (standard in Europe). For cooling, T_h is the mean ground temperature and T_c is the desired incoming temperature to the house for cooling, which is here set to -5°C (standard in Europe).

The COP can therefore be computed with Eq 5.7 and 5.8 in each pixel, for each month in both cooling and heating seasons (we consider an efficiency of 0.4), and can then be averaged through the months for both seasons. Then, the average season COPs are injected back in Eq. 5.5 and 5.6 and the equation can be inverted to compute:

$$\text{For heating, the heat transferred to the house: } Q_h = Q_c \frac{\text{COP}_{\text{heat}}}{\text{COP}_{\text{heat}} - 1} \quad (5.9)$$

$$\text{For cooling, the heat removed to the house: } Q_c = Q_h \frac{\text{COP}_{\text{cool}}}{\text{COP}_{\text{cool}} + 1} \quad (5.10)$$

where Q_c in Eq. 5.9 is computed using Eq. 5.4 as the average available Q (extractable) during the heating season months, and Q_h in Eq. 5.10 is computed using Eq. 5.4 as the average available Q (storable) during the cooling season months.

3. We assume that the heating season includes 7 months from the beginning of October to the end of April and the cooling season is the 5 remaining months.
4. For simplicity, we assume that we extract/store the total ground potential once per year during the heating/cooling season.

The energy values computed with equations 5.9 and 5.10 finally provide a first estimation of the geographical potential for very shallow geothermal systems in Switzerland, for each pixel, both in the heating and cooling seasons. The obtained maps are shown in Figure 5.18. It results in a total yearly potential of 4.00 TWh and 11.81 TWh respectively for cooling and heating in Switzerland. The latter heating potential corresponds to 17.8% of the Swiss space heating demand in 2017 (of 239.2 PJ, or 66.4 TWh) [212].

5.6.3 Validation with other potential studies

Very few geothermal potential studies have been conducted at very shallow depth, particularly in Switzerland, which makes comparison with other studies difficult. While the ThermoMap created by the European very shallow geothermal project [195] offers values for Switzerland, the resolution of these values is low, and more importantly, the depth does not match as it considers a depth of 10m. In Switzerland, a few local organizations may have values for various thermal variables, but at larger depths. The System d'Information du Territoire a Geneve (SITG), for example, offers heat capacity and thermal conductivity estimations for the Quaternary layer. The thickness of the Quaternary, however, may vary. It is therefore not possible to perform a thorough comparison with the results of the present study.

5.6.4 Limitations

The present chapter suggests new methodologies for the estimation of thermal ground characteristics at a large scale, and valuable information on their range of values in Switzerland. Several limitations, however, are to be mentioned, and possibly improved in further studies. These include: (i) **The lack of labeled training data.** Ideally, the training data should include a few hundreds of points in order to build very reliable models. Particularly in the case of the thermal diffusivity and the conversion from electrical to thermal resistivity, the size of the data was around 50 to 100 points, which is commonly considered the minimum size to perform supervised learning. Consequently, the results are still valuable but would greatly benefit from additional data. Note that the test errors are based on the test set and are therefore only validated of this set. Low test errors unfortunately do not guarantee a good generalization outside of this test set. The bigger the test, the more reliable are the test errors; more data may be added in the future to improve the models. (ii) **The lack of validation data.** As no dataset is available for the shallow ground characteristics of Switzerland over the entire territory, it is impossible to validate the final obtained results in unobserved points. Should another study be available, a comparison between the results would be valuable to provide some validation to the present results. (3) **Uncertainty propagation.** Using a combination of consecutive RF models together with more conventional signal processing methods (e.g. FFT) and numerical models (e.g. iterative inversion schemes) brings confidence, as the conventional methods have been tested and validated through the years, but also additional uncertainty, as the uncertainty of each estimation propagates through the next step, and therefore increases at each step. In that sense, a single step machine learning strategy can be considered to avoid the propagation issue. Labeled data, however, is naturally required for the final variable to estimate. There might be however, an increasing availability for data in the future, as there is currently an effort to digitalize geological and geophysical information, which is still often stored in the form of paper maps or written information. It is particularly the case in Switzerland, where a very significant amount of geo-studies have been performed through the years.

5.7 Summary

This chapter presents a methodology combining GIS data processing, machine learning and traditional modeling strategies in order to estimate the theoretical very shallow (first meter of the ground) geothermal potential (vSGP) in Switzerland. This potential assessment consists of the estimation of three significant thermal variables at shallow depth: (i) ground temperature gradient, (ii) ground thermal conductivity, and (iii) ground thermal diffusivity. The ground thermal heat capacity can then be recovered from the conductivity and the diffusivity. The estimation of the three variables is proposed at the spatial resolution of $(200 \times 200) \text{ [m}^2\text{]}$ pixels, across the Swiss territory, and at a monthly mean temporal resolution for the ground temperature, and a yearly mean temporal resolution for the thermal conductivity and thermal diffusivity. Besides the methodological contributions provided for the estimation of the large-scale vSGP, the present chapter eventually shows that, while traditional shallow geothermal systems (100-200 m deep Borehole Heat Exchangers mainly) have been extensively used in Switzerland, there is also a significant potential for very shallow geothermal energy systems, which can be a viable low-cost solution in adequate locations. There is notably a high potential for such systems in the Valais, Ticino and St. Gallen cantons, where the highest thermal conductivity values were found.

The obtained information on ground thermal characteristics can be of great use for municipalities, stakeholders and private holders who are considering small to large-scale very shallow geothermal installations. With the current development of new efficient and cost effective geothermal systems at shallow depths (including Slinky systems, helicoidal systems and heat baskets), the estimated results could serve as a useful help to identify the optimal locations for geothermal energy and for energy-related decision making in general in Switzerland. Also note that the methodology is mainly based on various sources of data (geological, weather and topographic data) that are currently being digitalized in more and more countries and methods/algorithms that are already implemented in various libraries. Therefore, should similar data be available, the methodology is generalizable to any other location.

A flowchart summarizing the entire methodology proposed in the chapter is shown in Figure 5.19.

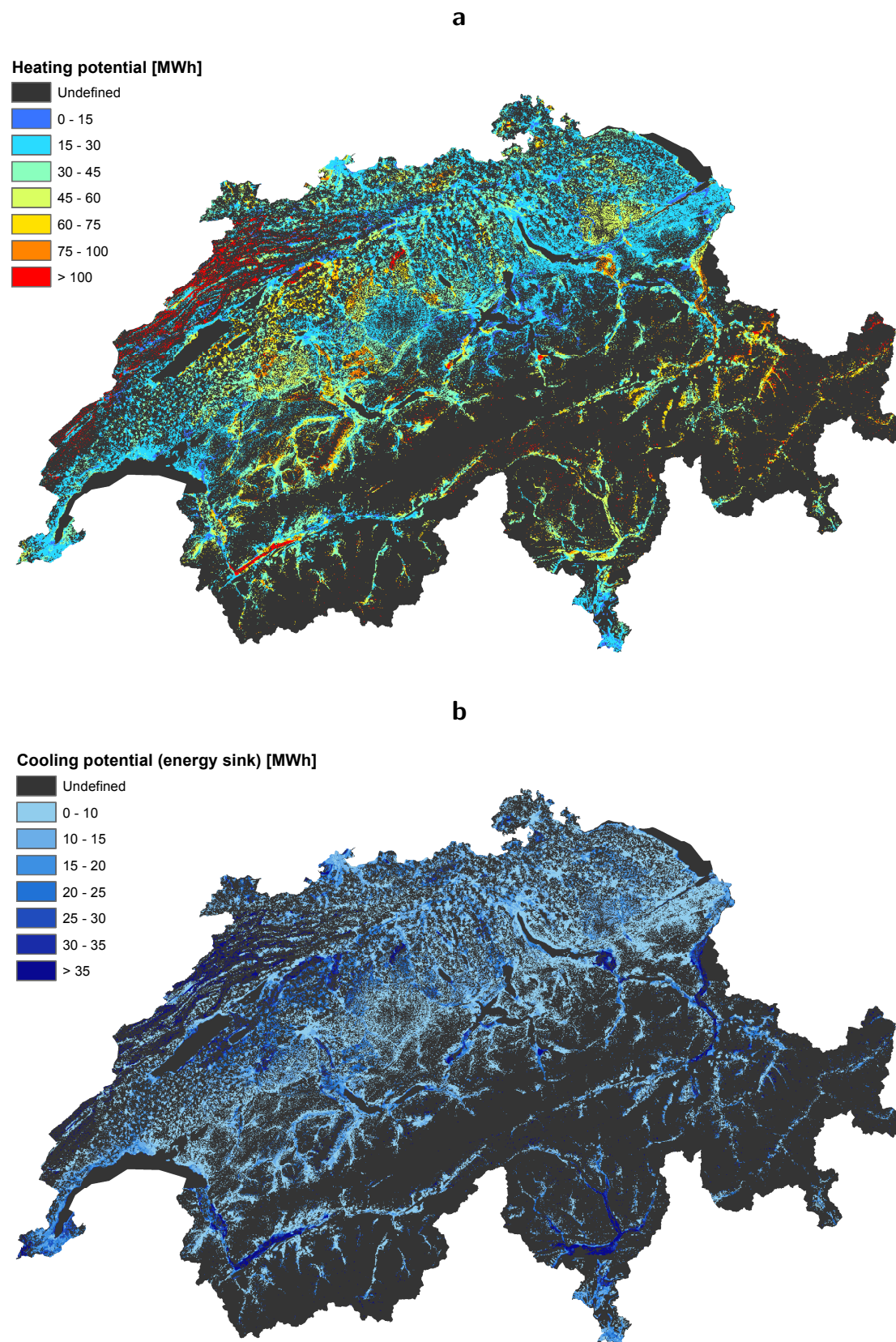


Figure 5.18: Estimated preliminary geographical potential for very shallow geothermal systems. **(a)** Heating potential, during one heating season in a year, **(b)** Cooling potential, during one cooling season in a year.

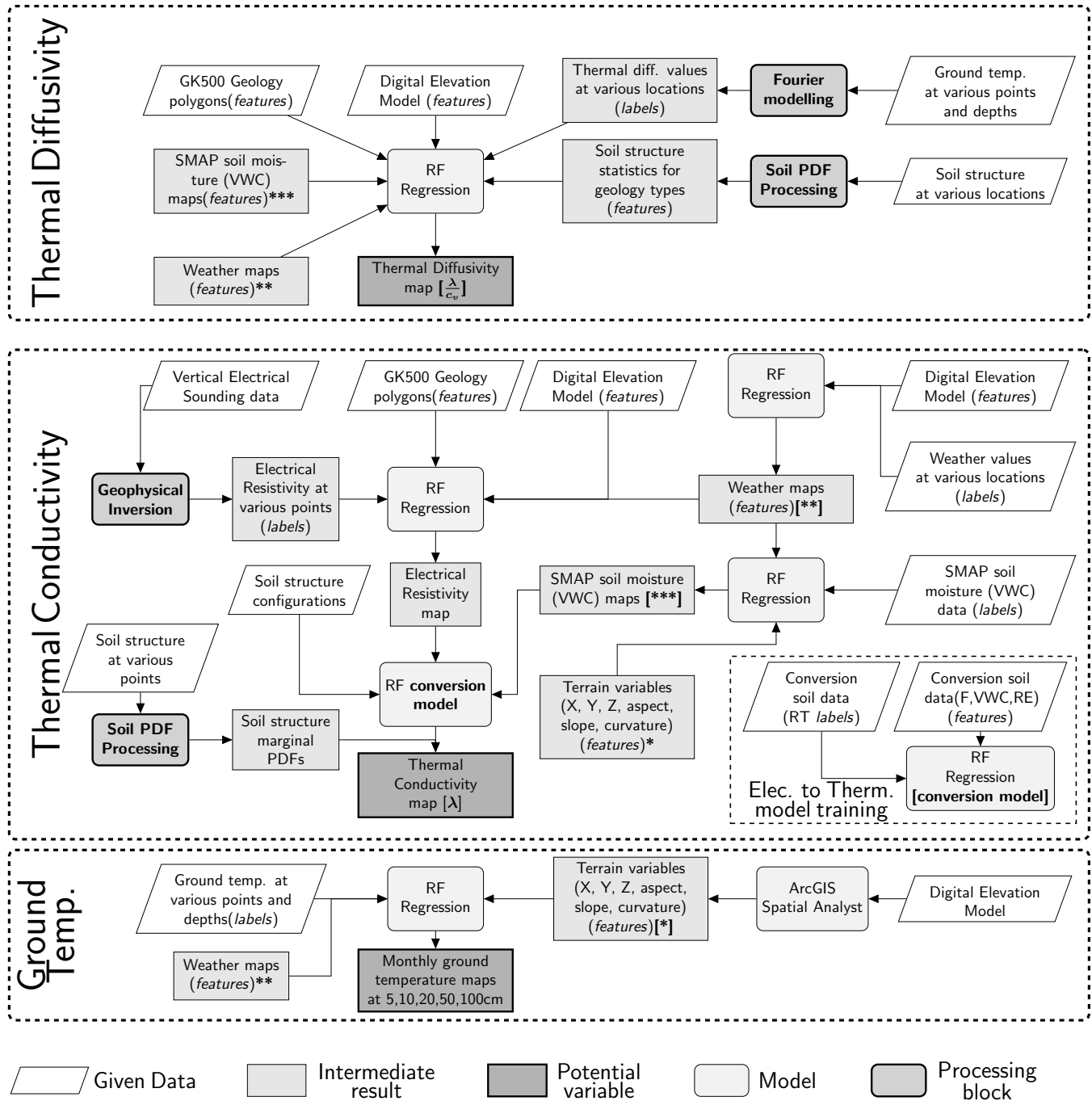


Figure 5.19: Flow chart of the methodology for the theoretical shallow geothermal potential estimation at pixel scale.

6

Solar energy: a technical potential estimation at commune scale

This chapter is based on the article [22]:

Assouline, D., Mohajeri, N., and Scartezzini, J-L. (2017). Quantifying rooftop photovoltaic solar energy potential: a machine learning approach, *Solar Energy* 141 278-296.

and borrows from the article [23]:

Assouline, D., Mohajeri, N., and Scartezzini, J-L. (2018). Large-scale rooftop solar photovoltaic technical potential estimation using Random Forests, *Applied Energy* 217 189-211.

This chapter attempts to derive a general methodology to assess the technical potential for solar energy through the use of PV panels mounted over rooftops, with an application to Switzerland. While the two previous chapters were only extracting the theoretical potential for wind and very shallow geothermal energies, this chapter is aiming at investigating the technical solar potential. Since it is perhaps the most promising energies out of the three and the most deployable at a large scale, it is tackled with more depth than the two former ones. The potential is this time fully determined, accounting for the geographical and technical constraints. The full hierarchical approach (presented in the introduction, section 1.2) is applied so as to estimate not only the physical potential but also the geographical and technical potentials. It requires, however, the estimation of many variables defining the geographical and technical aspects. The general complexity of the study is therefore greatly increased compared to the two previous chapters. Consequently, we chose to aggregate the potential estimations to communes, the smallest administrative divisions in Switzerland, rather than pixels, in order to reduce the computational requirements in a first attempt. Most importantly, such a scale is particularly adequate for policy-making and strategy planning [213] within cities and communes.

As a result of the commune scale, the general amount of handled data points is significantly decreased when compared to the previous studies using 200×200 [m²] pixels. To adapt to this change, we chose to use Support Vector Regression (SVR) as a benchmark algorithm rather than Random Forests (RF). Despite its slower training speed and harder tuning, SVR is indeed known to

provide slightly better accuracy with smaller training sets, particularly when the number of features considered is higher than the size of the data (as explained in chapter 2).

The chapter details the steps allowing the computation of the solar PV potential of building rooftops for each commune in the greater part of the urban areas of Switzerland. Using, similarly to the two previous studies, a combination of machine learning and GIS methods together with available data, the main aim of this chapter is to estimate, at the national scale, all the solar and urban variables of interest leading to the solar rooftop potential everywhere in Switzerland. The variables of interest include: (1) the available area for PV solar installations on rooftops, (2) the slope and direction of rooftops, (3) the global solar horizontal and tilted radiations and (4) the shading factors over rooftops.

The chapter is organized as follows. Section 6.1 offers a literature review on rooftop solar PV potential studies and places the present chapter within its context. Section 6.2 presents the data sources used in the chapter and some of the processing performed to extract significant features. Section 6.3 explains the extraction of the theoretical potential, specifically the estimation of monthly solar horizontal radiation maps in Switzerland. Section 6.4 details the computation of the geographical potential, including the estimation of the available rooftop area for PV panels, the slope and aspects of rooftops, shading factors and global tilted radiation over rooftops. Section 6.5 presents the computation of the technical potential. Section 6.6 provides a discussion on the obtained results. Finally, section 6.7 concludes the chapter and summaries the proposed methodology.

6.1 Related literature

Solar photovoltaic (PV) panels on the existing building rooftops have proven to be an efficient and viable large scale resource of sustainable energy for urban areas [2, 3, 214–217]. In addition, solar panels can have an important role in integration of decentralised renewable energy resources in a neighbourhood [218, 219]. In Switzerland, only 3% electricity generated in 2017 comes from PV [6] but is likely to increase to 13.4% by 2050 [220]. The feasibility of solar PV installations is of importance not only for individual property owners, but also the local governments and municipalities [213].

Depending on the availability of data, regional characteristics, as well as scale of study, several methodologies have been suggested to determine rooftop PV potential [217, 221, 222]. At the scale of Europe as a whole (27 EU members), studies show that there is a large building integrated photovoltaics (BIPV) potential, 840 TWh annually, which equals more than 22% of the expected European electricity demand by 2030 [221, 223, 224]. At national and regional scales, studies show significant values for urban (rooftop) PV applications in many countries. These include the USA [225, 226], Israel [227], Canada [228], and Spain [229], where urban PV deployment could potentially cover 15 to 45% of national electricity consumption. At regional scale, Lopez et al. [230] provide a GIS-based methodology for all the states of USA and their technical rooftop potential. Several studies explore the PV potential for buildings at the city and neighbourhood scale [213, 221, 231–233]. For Hong Kong [233], as an example, the estimated potential of rooftop PV is 5981 GWh which can account for 14.2% of the city's 2011 electricity use. Another example is Seoul in South Korea where deployment of rooftop distributed photovoltaic systems can cover 30% of the city's annual electricity consumption.

Several studies propose a hierarchical approach for estimating the rooftop potential of PV solar electricity on a regional or national scale [2–4, 229, 234]. While at the neighbourhood and city scale,

the use of 3D building data is frequently suggested [235], at the regional and national scale, due to lack of data, significant effort is usually dedicated to estimate the existing roof area and the available roof area for PV installation. For example, Izquierdo et al. [3] use a sampling method to compute the available rooftop area for PV installations in different municipalities in Spain. By contrast, Wiginton et al. [2] use GIS-based Feature Analyst tool to compute the available rooftop area and PV solar potential. Then a sampling technique with additional variables is used to explore the relation between population density and the PV solar potential. The International Energy Agency [215] uses statistical information to estimate the building area (roofs and facades) and to obtain the potential for solar energy. Several other studies use aerial images [236] and ArcGIS together with LiDAR (Light Detection and Ranging) data to determine roof geometries and associated roof areas for the PV solar potential [237–243]. Studies based on advanced cartographic information and high-resolution images derived from remote sensing technologies such as LiDAR is expected to produce high accuracy results for the corresponding potential [221, 234]. Such studies, require, however, very large precise remote sensing data to be able to extract reliable information over a whole region. More recently, GIS and LiDAR data was combined with a simple statistical model to extract the available area for PV where LiDAR data is not available and estimate the rooftop PV potential all over the US [244]. In addition to the available roof area for PV solar power plants, another important variable to estimate is the shading from neighbouring buildings and trees over rooftops. There are hardly any studies developing a precise methodology to estimate the shading impacts at a national scale. Most studies use simple coefficients and apply the shading coefficient to the final potential [3, 235]. Detailed estimation of shading over rooftops has been suggested by [245, 246], however, their study focus on the neighbourhood scale and city scale.

As In recent years, Machine Learning (ML) algorithms (e.g. support vector machines, artificial neural networks) have been widely used for forecasting solar radiation on the horizontal and tilted surfaces [8, 9, 198–201, 247–252]. However, using machine learning algorithms so as to estimate urban characteristics for solar prediction on building roofs, including the available roof area and roof geometries, has been very rarely explored. In 2014, Joshi et al. [205] used a ML method for image processing in order to detect and classify building rooftops and estimate the corresponding solar PV potential. More recently, Mohajeri et al. [253] used Support Vector Machines (a machine learning algorithm) together with GIS building data in order to classify rooftops in the city of Geneva Switzerland and estimate the impact of rooftop geometries in relation with their access to solar energy.

This chapter is motivated by the lack of large scale (regional and national) solar potential methodologies. Furthermore, while most studies use constant average coefficients to estimate the available area for PV, the geometrical characteristics of the rooftops and the shading impacts over rooftops, we provide a methodology which train Machine Learning models based on real data and use them to assess these three variables at unknown locations. In particular, we here use Support Vector Regression (presented in chapter 2, section 2.2.5) to train the multiple required models. It ultimately results in the estimation, for the first time, of the technical rooftop PV potential in all communes of Switzerland.

6.2 Data

6.2.1 Data sources

All data sources used within this chapter are presented in Appendix A and signified by a ✓ symbol for Chapter 6 within tables A.1 and A.2. They include weather monitored and solar radiation data, digital elevation models, land cover data as well as building vector data. Note that the solar radiation data (global, diffuse, and direct) have different quality depending on: (i) the computation methods (e.g. classical interpolation or geo-statistical methods as well as algorithm methods) and (ii) the source of data including ground monitoring stations, satellite images, and combination of both types of data. The general consistency of the different data has been verified in overlapping areas. In particular, the location of the buildings in the building clusters data (VECTOR25) has been verified with respect to the Geneva canton data (SITG).

6.2.2 Data processing

As in the earlier chapters, the first step in training the ML models is naturally to extract representative features. While the physical potential does not require very large feature processing, the geographical potential does, particularly concerning roof characteristics in urban areas. These urban features need to be available and extractable from databases that cover all the Swiss urban areas. The urban areas are defined according to the CORINE Land Cover data for Switzerland. The data is freely available from the Swiss Federal Institute for Forest, Snow and Landscape Research [254]. The CORINE Land Cover data has a vector polygon format which defines different land uses across Switzerland. From the CORINE Land Cover Switzerland, we only select the polygons indicating continuous and discontinuous urban areas. The following features have been used in both SVR training and testing processes in order to estimate roof characteristics for unavailable building data in urban areas in Switzerland:

- Average of building ground floor area and total building ground floor area (A_f and $A_{f,sum}$) for each commune. We use the VEC25 shapefile data from Swisstopo, that is, vector polygons of building ground floor. Using GIS tools, we calculate the average ground floor area (A_f) and total building ground floor area ($A_{f,sum}$) for each commune.
- Building density (D_b). Building density is measured by site coverage which is the total ground floor area divided by the total urban area in each commune.
- Population density (D_p). The commune population divided by the total urban area in each commune.
- Building typology statistics (GEOSTAT) (See Table A.1). Building typology includes type of residential buildings, year of construction, number of floors, main space heating source, and main water heating source. The data is freely available from the Swiss Federal Statistical Office [255]. Type of buildings consists of two classes namely, individual houses and multi-family houses. Year of construction consists of 10 classes (before year 1919, 1946 to 1966, 1961 to 1970, 1971 to 1980, 1981 to 1990, 1991 to 2000, 2001 to 2005, 2006 to 2010, 2011 to 2015). Number of floors includes 9 classes (1, 2, 3, 4, 5, 6, 7 to 9, 10 to 14, 15 floors and above). Main space heating source includes

10 classes (fuel, wood, heat pumps, electricity, gas, district, heating, coal, solar, others, and no heat). Main water heating source includes 10 classes (fuel, wood, heat pumps, electricity, gas, district, heating, coal, solar, others, and no heat). Thus, the total possible number of features for building typology is 41. Since this data is only available for pixel size of 100 by 100m all over Switzerland, we calculate the percentage of buildings for each feature in each pixel size and then estimate the average values for each commune. Adding A_f , $A_{f, \text{sum}}$, D_b and D_p to building typology features (41 features) constitutes the final input data of the 45 features for SVR training and testing.

While the total number of input data for SVR is 1901 communes, the labelled data consists of 42 labels (for 42 commune roof characteristics are available) for testing and training. The rest of communes (1859) have unknown roof characteristics and is remained to be predicted. As mentioned, the total number of features is 45. The entire data need to be scaled. To do so, features and labelled data are normalized by subtracting their mean value and divided by the standard deviation. Although 45 features is not a very large number for machine learning practitioners, it is still a reasonable number and could benefit from feature selection and/or Principal Component Analysis [256, 257]. We reduce slightly the dimensionality of the data to 39 by performing Principal Component Analysis, that is, analyzing the spectrum of the covariance matrix of the training data on 45 features.

6.3 Theoretical potential estimation

The total amount of energy received from the sun by the urban areas of Switzerland, independently of urban characteristics, is presented as monthly and yearly raster maps for diffuse horizontal (G_D), global horizontal radiation (G_h), and extra-terrestrial horizontal radiation (G_{oh}). The monthly and yearly solar estimation is based on the existing satellite solar data [258] and weather data from MeteoSwiss [259] for specific location as well as a Digital Elevation Model [260] for estimating latitude, longitude, and altitude. The direct horizontal radiation (G_B) is estimated based on the difference between global horizontal and diffuse horizontal radiation. To estimate G_D , G_h and G_{oh} for the whole of Switzerland, the 200×200 [m²] pixel grid presented in chapter 4 (section 4.2.2) as a the base for predictions. To predict the data for unknown pixels, we use support vector regression (SVR) for spatial extrapolation. Given the large number of data points, a binary data format called Hierarchical Data Format 5 (HDF5) is used to process the data in python. H5py, an interface for HDF5 in python, is used in order to manipulate the data.

6.3.1 Estimation of horizontal solar radiation

Monthly G_D , G_h and G_{oh} are estimated using SVR monthly models trained with the following input features: (i) sunshine duration, temperature, precipitation and cloud cover, for which monthly maps are extracted using the same strategy as the one presented in section 4.2.2 (chapter 4), but using SVR instead of RF (it is notably interesting to compare results obtained from both methods, which are ultimately rather similar in this case) and (ii) space variables: latitude, longitude, altitude. The labeled values is based on satellite data for G_D , G_h and G_{oh} which are extracted for 100 locations throughout Switzerland from the SoDa database [258]. The satellite solar radiation data is available in a 15min resolution in

Table 6.1: Testing RMSE (E_R) and NRMSE (E_{NR} , in percentage) from SVR models trained for weather and solar variables.

Month	Temp.		Cloud		Prec.		Suns.		G_h		G_D		G_{oh}	
	E_R [°C]	E_{NR} [%]	E_R [%]	E_{NR} [%]	E_R [mm]	E_{NR} [%]	E_R [hours]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]
Jan.	1.85	0.68	12.82	21.81	21.71	27.58	10.57	13.09	71.04	5.22	29.28	4.01	2.87	0.09
Feb.	1.30	0.48	8.05	8.05	19.64	26.97	12.66	12.46	100.79	4.64	53.04	5.19	2.89	0.06
Mar.	0.73	0.26	9.97	9.97	23.40	26.48	10.67	7.65	118.01	3.31	79.75	5.37	2.41	0.03
Apr.	0.61	0.22	7.39	7.39	19.92	21.17	14.94	9.78	193.05	1.93	165.80	8.93	1.69	0.02
May.	0.62	0.22	6.67	6.67	17.00	13.45	15.64	9.07	224.31	4.24	141.41	6.25	0.65	0.01
Jun.	0.48	0.17	9.06	9.09	17.54	12.99	21.92	11.71	336.88	5.90	149.52	6.39	0.23	0.01
Jul.	0.44	0.15	5.71	5.71	21.23	15.21	16.22	7.60	322.23	5.68	160.95	7.31	0.42	0.01
Aug.	0.48	0.17	4.23	4.25	20.24	14.23	18.53	9.48	208.44	4.36	117.15	6.26	1.23	0.01
Sep.	0.46	0.16	4.59	4.16	17.11	14.59	12.61	8.06	155.25	4.01	90.45	5.92	2.47	0.03
Oct.	0.50	0.18	5.49	4.05	15.44	15.89	16.82	13.85	124.76	4.79	61.18	5.44	2.62	0.05
Nov.	0.96	0.35	7.21	8.26	19.00	19.61	9.92	12.63	95.68	6.36	35.50	4.46	2.92	0.08
Dec.	1.47	0.54	9.52	10.86	23.10	25.02	10.03	15.21	61.39	5.75	38.71	4.48	2.89	0.11

irradiance form (in W/m²); the corresponding solar energy is then aggregated hourly and summed on a daily (24-hour) basis, to obtain daily irradiation values (that we simply call radiation), (in kWh/m²). Aggregating the radiation values at a daily basis allows to work with energy instead of power, and the resulting is data more smooth and suitable for the learning procedure. The input features for training (weather and space variables) are extracted for the same 100 locations as the solar radiation data. SVR is applied for the three solar radiation data to estimate monthly mean daily values for G_D , G_h and G_{oh} .

Testing errors are shown in Table 6.1 and maps are visualized in Fig. 6.1. We would like to assess the impact of weather features in some fashion, as provided by the Variable Importance measure in the Random Forests algorithm. SVR, however, does not have an embedded feature importance measure. To provide an idea of this impact, we simply compare the monthly solar models using SVR including the weather parameters (sunshine duration, temperature, precipitation, cloud cover) with models where the weather parameters were omitted (latitude, longitude, and altitude were considered in both models). In particular, we compare the RMSE and RRMSE values for the two models to assess the usefulness of the weather parameters as extra model information. The results show that including weather parameters offers a better performance for G_h estimation for all months, whereas for G_D the performance varies depending on the month. The yearly performance for G_D is, however, very similar for both models. For G_{oh} the performance is very good and similar in both models, partly because G_{oh} depends mostly on latitude and longitude and the weather parameters do not have much effect, as it can be observed in Table 6.2, where the performance of the models were compared while considering the weather features or not.

6.4 Geographical potential estimation

To predict monthly solar energy over building roofs suitable for PV solar installations in Switzerland, several variables need to be estimated. These include the available roof area for PV panels (A_R), shading factors from neighbouring buildings and trees on the building roofs (S_{Sh} and S_{hill}), as well as the monthly global tilted solar radiation on non-horizontal surfaces (G_t). To estimate G_t , data on

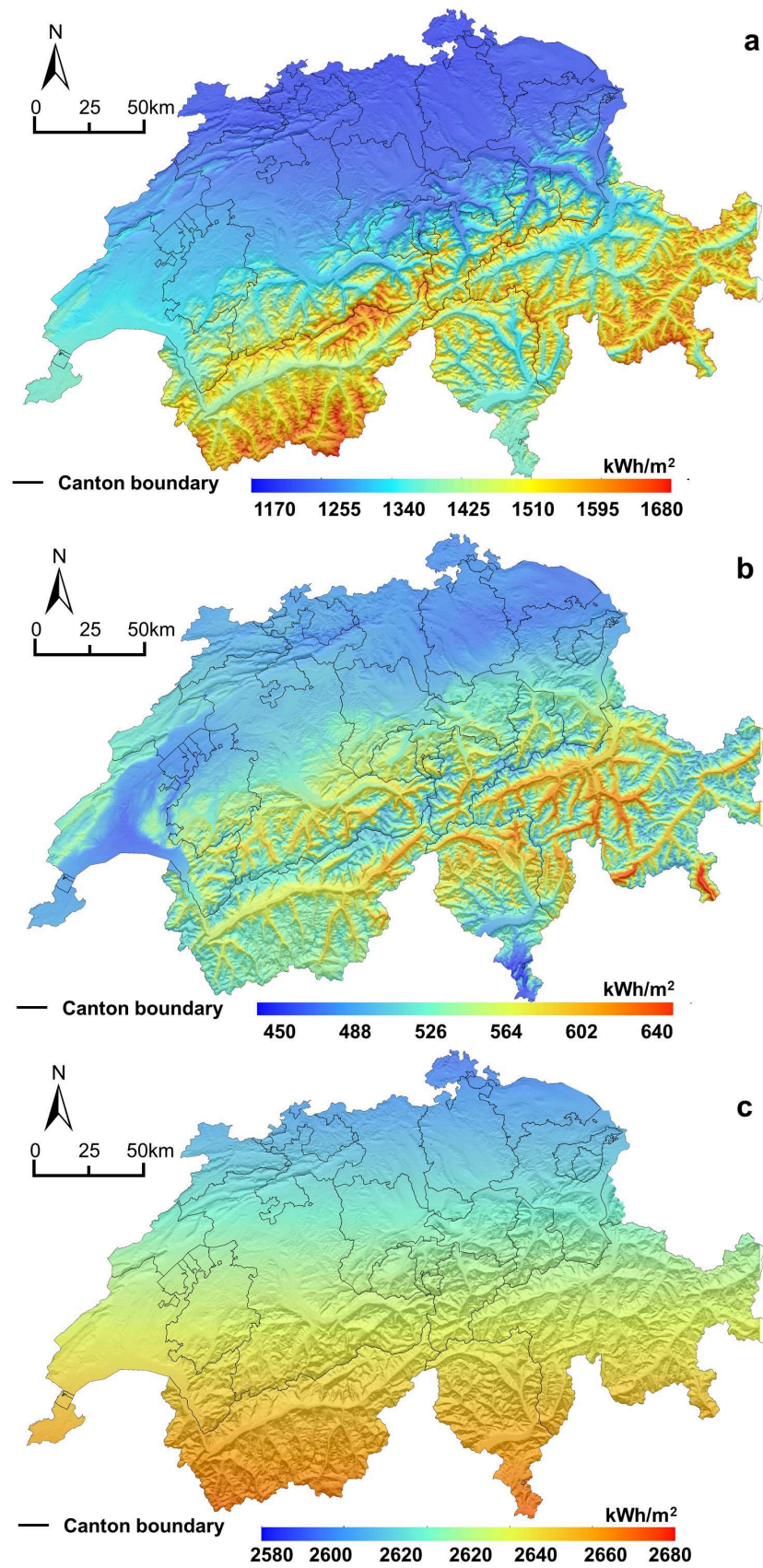


Figure 6.1: Prediction of horizontal solar radiation maps using SVR. **(a)** Yearly mean global horizontal radiation, kWh/m², **(b)** Yearly mean diffuse horizontal radiation, kWh/m², **(c)** Yearly mean extraterrestrial, kWh/m².

Table 6.2: Testing errors (RMSE and NRMSE) for G_h , G_D and G_{oh} , while no weather variables are considered (no w.) and weather variables are considered (w.).

Month	G_h , no w.		G_D , no w.		G_{oh} , no w.		G_h , w.		G_D , w.		G_{oh} , w.	
	E_R [kWh/m ²]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]
Jan.	76.81	5.64	33.43	4.57	2.89	0.09	71.04	5.22	29.28	4.01	2.87	0.09
Feb.	117.42	5.4	64.54	6.31	2.87	0.06	100.79	4.64	53.04	5.19	2.89	0.06
Mar.	145.59	4.08	88.04	6.16	2.42	0.04	118.01	3.31	79.75	5.37	2.41	0.03
Apr.	177.7	3.68	156.21	6.26	1.7	0.02	93.05	1.93	165.8	8.93	1.69	0.02
May	248.22	4.69	133.77	5.47	0.65	0.01	224.31	4.24	141.41	6.25	0.65	0.01
Jun.	373.27	6.54	126.42	5.4	0.22	0.01	336.88	5.9	149.52	6.39	0.23	0.01
Jul.	324.34	5.71	152.83	6.94	0.39	0.01	322.23	5.68	160.95	7.31	0.42	0.01
Aug.	242.4	5.07	125.9	6.73	1.21	0.01	208.44	4.36	117.15	6.26	1.23	0.01
Sep.	166.52	4.3	86.8	5.68	2.37	0.03	155.25	4.01	90.45	5.92	2.47	0.03
Oct.	96.16	3.69	57.24	5.09	2.61	0.05	124.76	4.79	61.18	5.44	2.62	0.05
Nov.	86.56	5.75	32.59	4.09	2.89	0.08	95.68	6.36	35.5	4.46	2.92	0.08
Dec.	68.82	6.45	25.28	3.95	2.85	0.11	61.39	5.75	28.71	4.48	2.89	0.11
Mean	176.98	5.08	90.26	5.56	1.92	0.04	159.32	4.68	92.73	5.83	1.94	0.04

several roof characteristics including roof slope and roof azimuth, in addition to latitude, longitude, and altitude, are needed. Precise roof characteristics data is freely available only for 42 communes in Switzerland located in the canton of Geneva (<http://ge.ch/sitg/>). The strategy is therefore to train machine learning models (using SVR) over these 42 communes, and use the trained models to estimate the required variables in the rest of the considered communes in Switzerland (1859 number of remaining communes). Note that the Geneva canton must therefore be somehow representative of the rest of the communes in Switzerland, in order for the training process to be meaningful. Even though the availability of more data (from other cantons) would have been preferable, we believe that the representativity is indeed partly there: Geneva canton includes rural and urban communes, with different levels of building densities and architecture styles. There may be, however, legal restrictions regarding buildings which are specific to Geneva canton. As a result, it still constitutes a limitation that can be avoided in the future if more data become available. The following sections will present the details of the estimation of the roof characteristics mentioned above as well as prediction of monthly and yearly solar energy on tilted roofs suitable for PV solar installations.

6.4.1 Available rooftop area estimation

The total available roof area and the average available roof area for PV installation (A_R) for each commune are estimated based on the detailed roof surface data including roof superstructures (e.g. chimney, dormer, staircase). This data is freely available in vector format for only 42 communes in Switzerland [261]. We calculate the ratio of average available roof area for PV (A_R) to the average ground floor area (A_f) for each commune. To estimate A_R , as shown in Fig. 6.2, we first remove the superstructures from roof surfaces using Erase tool from ArcGIS. Second, we erase a band of 1m around the sides of the roof surfaces using Buffer tool in ArcGIS. This is based on regulations for roof-mounted solar PV installations which must be at least 30 cm (0.3 m) away from the external edge of the roof. Third, we remove surfaces with an area smaller than 28 m² from the roofs. This last constraint has two purposes. It allows to: (i) Add a reasonable geometric constraint. This value

Table 6.3: Testing errors (RMSE and RRMSE) in SVR training for slope distribution (β) and for the available area ratio (C_R), that is, the ratio of the average available area for PV to the average ground floor area.

Error	C_R	$\beta_{[0,10]}$	$\beta_{[10,20]}$	$\beta_{[20,30]}$	$\beta_{[30,40]}$	$\beta_{[40,50]}$	$\beta_{[50,60]}$	$\beta_{[60,70]}$
E_R (no unit for C_R , in $^\circ$ for β)	0.042	6.2	4.42	6.56	9.16	3.42	1.4	1.16
E_{NR} (in %)	5.59	39.56	50.58	26.69	24.97	32.19	56.3	96.33

of 28 m² makes it possible to discard small portions of roofs with intricate shapes, unsuitable for PV panel installation. (This geometrical constraint will be improved in next chapter 7) (ii) Add an economic constraint. Depending on the size of PV system, the number of PV panels on the roof, the cost of installation and the return on investment the required roof area for PV installation may be different [262]. We consider here a typical size of PV system of 4.7 kW as a minimum profitable system, which approximately corresponds to 28 m² roof area. After we obtain A_R , the ratio (C_R) for 42 commune in Geneva is calculated using the following 6.1:

$$C_R^j = \frac{A_R^j}{A_f^j} \quad (6.1)$$

where j indexes the commune, A_R is the average available area for PV and A_f is the average ground floor area. We use C_R as label and the features presented in section 6.2.2 for the training and testing process together with SVR to predict the ratio for the rest of the communes in Switzerland. We use the C_R ratio because we would like to profit from the extra information, that is, the ground floor area, which is available for all communes in Switzerland. The obtained maps for the available area ratio and the total available roof area for each commune are shown in Fig. 6.3. Testing errors for C_R are shown in Table 6.3 (together with errors for the slope variable estimated further in the chapter).

6.4.2 Shading factors estimation

Different methods for incorporating losses due to shading in estimating solar rooftop PV potential have been suggested [245, 246, 263]. The shading considered in the present chapter include S_{Sh} and S_{hill} which are estimated in raster format and shown in Fig. 6.4. S_{Sh} is defined as the ratio of fully shaded cells to the total rooftop cells (shaded and unshaded). S_{hill} is the average value of partially shaded or non-fully shaded cells. Both S_{Sh} and S_{hill} are computed based on monthly mean daily average of Hillshade. The two shading factors are computed for the 42 communes in Geneva, using the Hillshade function from ArcGIS. Hillshade maps are based on DOM (Digital Orthophoto Map) which is from Swisstopo (<http://www.swisstopo.admin.ch/>) and has a 2 m by 2 m resolution, and include building and vegetation. The value for Hillshade in ArcGIS ranges from 0, which means that the surface is entirely in the shadow, to 255, which means that the surface is entirely enlightened. The following steps, shown in Fig. 6.4, have been used to compute S_{Sh} and S_{hill} :

- To decrease the computational time, we combine the individual building polygons with their detailed roof geometry (Fig. 6.4a) into continuous polygons (with the same outer boundaries as the original polygons) using Dissolve tool in ArcGIS (Fig. 6.4b). Then we extract a reverse vector map where buildings assign as void and the surroundings as filled polygons (Fig. 6.4c).

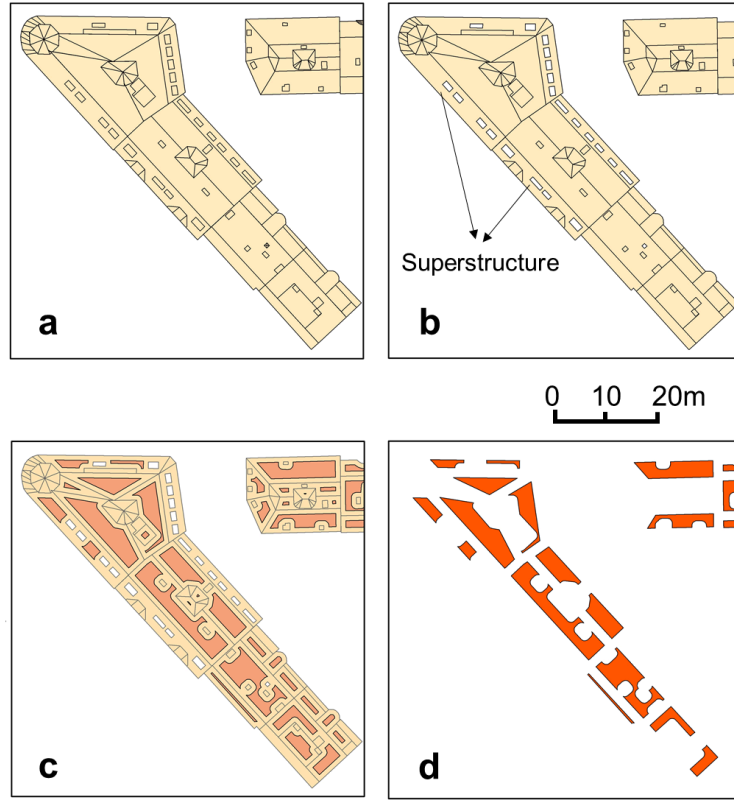


Figure 6.2: Schematic presentation of different steps to estimate available roof area using ArcGIS. **(a)** building polygons with detailed roof geometries including superstructure (e.g. chimney, dormers, staircase), **(b)** removing the superstructures from roof surfaces, **(c)** creating 1 m² buffer around each remaining roof surfaces, **(d)** the final available roof area for PV installation after removing the areas less than 28 m².

- To obtain fully shaded areas and non-fully shaded areas over rooftops (S_{Sh} and S_{hill}), we use raster clipping (clip Hillshade raster map over building layer). Raster clipping process is computationally extremely time-consuming due to the large number of buildings. To make the clipping process faster, the following steps are needed, marked in red broken-line in Figs. 6.4c to 6.4f: (i) As mentioned before, we assign building polygons as voids and the surroundings as filled vector polygons (Fig. 6.4c). This allows us to obtain “reverse” building polygons for each commune. (ii) We then extract “negative” DOM raster map (Fig. 6.4e) by clipping the DOM map (Fig. 6.4d) over the “reverse” building polygons (Fig. 6.4c). We obtain a negative DOM raster map showing the buildings as voids and the rests are building surroundings (iii) we create a boolean raster map using the IsNull function from the Raster Calculator in ArcGIS (Fig. 6.4f). In IsNull raster map (Fig. 6.4f), Null cells (void cells) assign to value 1 indicating there are buildings, whereas not Null cells assign to value 0 indicating there are no buildings, (Fig. 6.4f). (iv) To compute hourly Hillshade raster maps (from 8am to 18pm; Fig. 6.4g) for each month using the representative day of the month considering the altitude and azimuth of the sun [109], we use DOM (Digital Orthophoto Map) data with a 2m by 2m resolution. (v) To obtain the clipped Hillshade map for buildings (Fig. 6.4h), we clip Hillshade map (Fig. 6.4g) over IsNull raster map (Fig. 6.4f). For each Hillshade map we set to Null cells (void cell) the corresponding cells with 0 value in the IsNull raster. To compute binary shading map (Fig. 6.4i), we use the Raster Calculator from

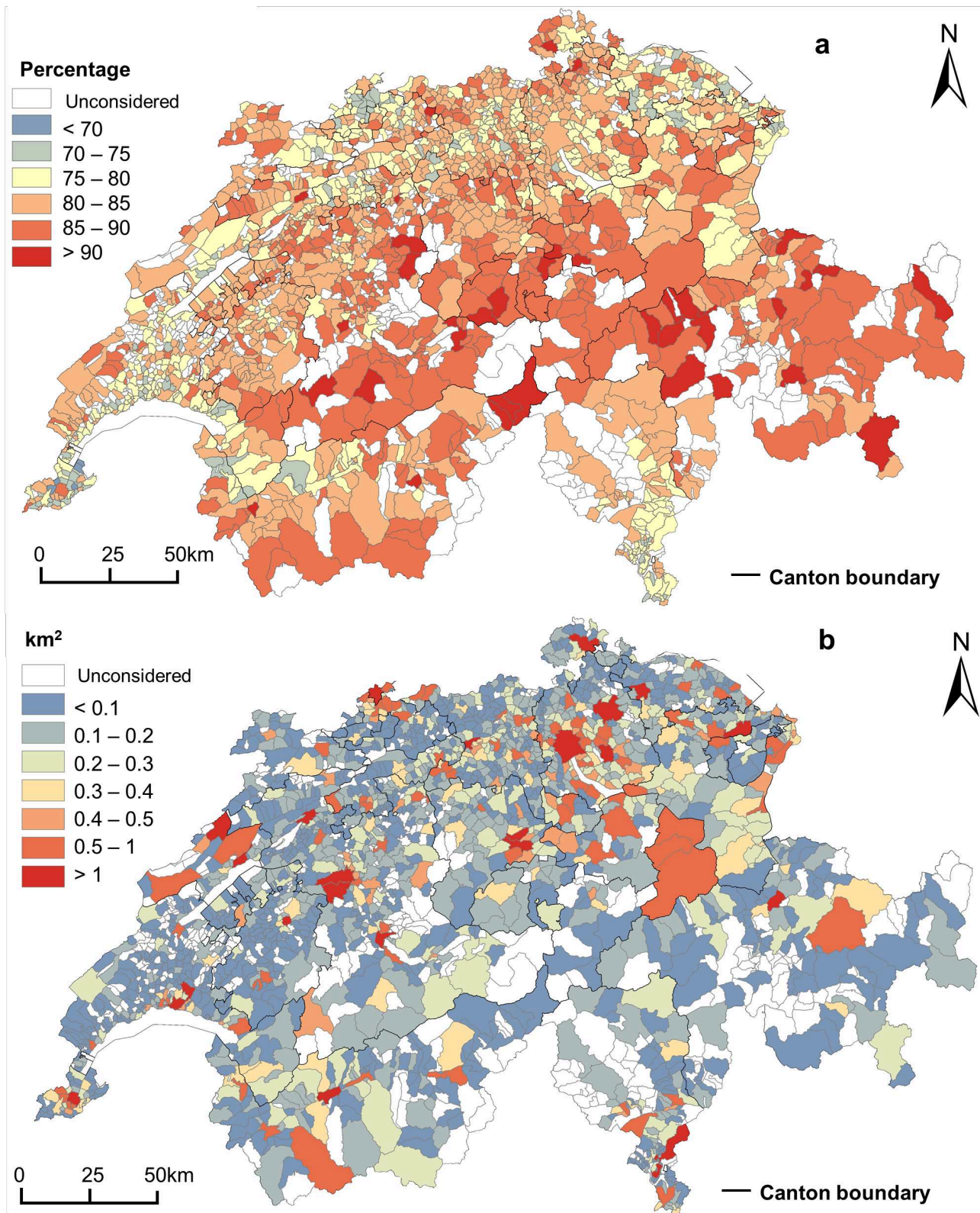


Figure 6.3: Final available area maps in Switzerland. **(a)** the ratio C_R of average available roof area for PV solar (A_R) to the average ground floor area (A_f) for each commune, percentage **(b)** total available roof area in each commune, expressed in km^2 .

Spatial Analyst toolbox in ArcGIS. In the binary raster map (Fig. 6.4h) 0 indicates cells that are non-fully shaded (shown in yellow) and 1 indicates cells that are fully shaded (shown in blue).

- To compute S_{Sh} for each commune j and for each month, considering only fully shaded cells with value 1 in the binary raster (corresponding to Hillshade value = 0), S_{Sh} is given by:

$$S_{Sh}^j = \frac{N_{sh,8\text{ am}}^j + N_{sh,9\text{ am}}^j + \dots + N_{sh,6\text{ pm}}^j}{11N_{cells}^j} \quad (6.2)$$

where $N_{sh,8\text{ am}}^j$ is the number of fully shaded cells in rooftops in commune j at 8am (during the representative day of the considered month), and N_{cells}^j is the total number of rooftop cells (shaded and unshaded) in commune j . Repeating that process for each month, S_{Sh} is given as a monthly mean daily average for each commune and for each month.

- To compute S_{hill} for each commune j and for each month, we only consider non-fully shaded cells, that is, partially shaded, cells with value 0 (corresponding to Hillshade value > 0). We extract the raster map of partially shaded cells using the SetNull function in ArcGIS. Then, we compute the average daily Hillshade values (10 hours for each representative day) for partially shaded cells in each commune. Repeating that process for each month, S_{hill} is finally given as a monthly mean daily average for each commune and for each month.

Note that while S_{hill} is estimated to measure the impact of shading over the direct solar radiation, S_{Sh} is computed as an additional constraint to discard (in the potential estimation) fractions of the rooftop area, and by extension PV modules, which are in shade. Note that, in practice, however, if PV modules are connected in series within a panel, and some of the modules are in the shade, the whole generation of the panel will drop to zero. Similarly, cells can be installed in series within a module, which entails an even larger effect of shading over the modules and therefore the panel. Even though a series arrangement for PV panels is frequently used (because it is easier and cheaper to install), it is very challenging to account for the large-scale shading impact in this case as it requires precise modelling of individual PV cells, at a high time resolution ([264]). Within this study, we therefore consider parallel arrangement of modules within an installed array, which partly prevents the latter effects from happening.

Once the monthly mean daily average values of S_{Sh} and S_{hill} for 42 communes are computed, SVR models are trained for both S_{Sh} and S_{hill} for each month. We use the same features as presented in Section 6.2.2. S_{Sh} and S_{hill} are used as labels to perform training and testing over 42 communes in Geneva. We then apply the regressor to the rest of the communes in Switzerland in order to predict S_{Sh} and S_{hill} . The process is shown in Fig. 6.4 and testing errors for shading factors are shown in Table 6.4.

6.4.3 Global solar tilted radiation estimation

The global tilted solar radiation is estimated combining the tilted direct, diffuse and reflected solar radiations, as explained in chapter 3. We consider S_{hill} , estimated in section 3.2.3, for the direct radiation incident on an tilted surface (G_{Bt}) and the sky view factor (SVF) for the diffuse radiation on an tilted surface (G_{Dt}). A constant SVF for urban areas in Switzerland is assumed as 0.9 over roofs [246, 265]. The modified G_t equation is given by:

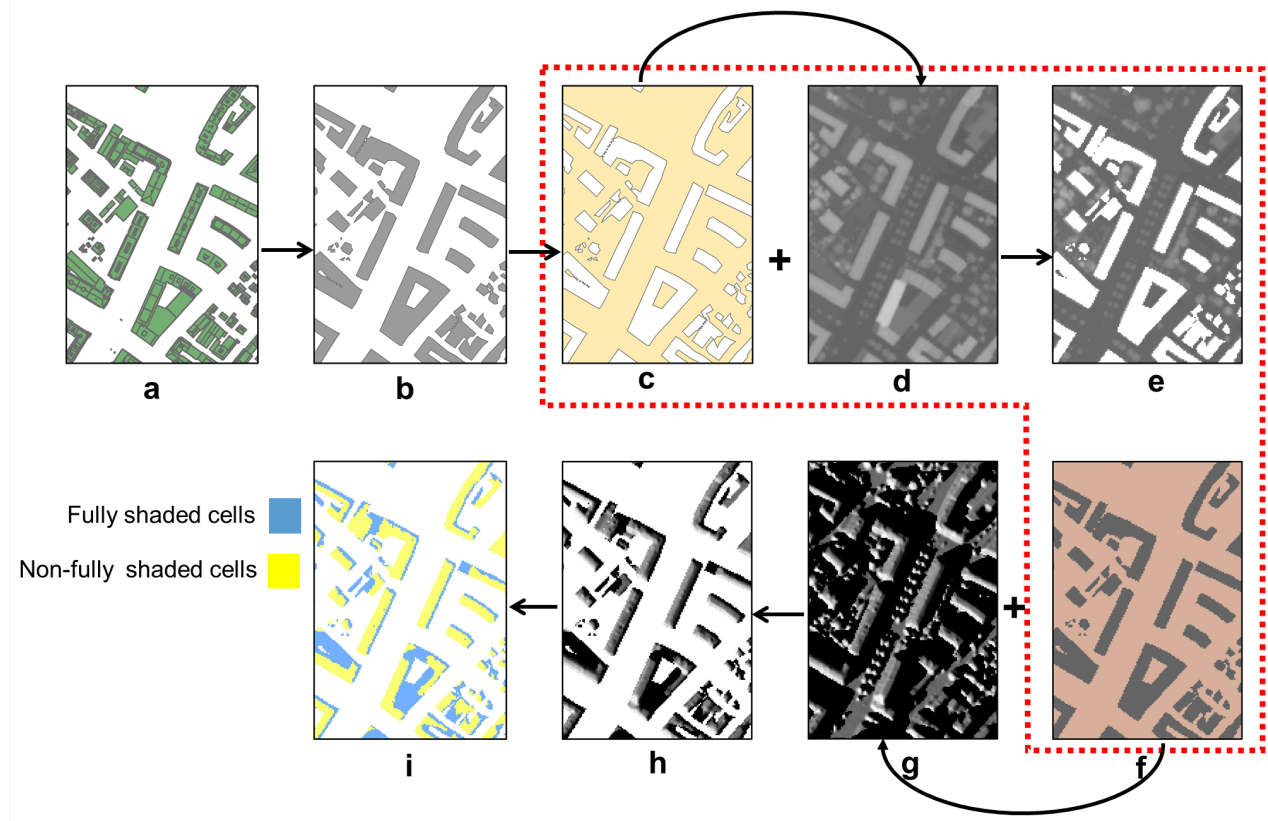


Figure 6.4: Schematic presentation of calculating shading factors (S_{Sh} and S_{hill}). **(a)** buildings with detailed roof geometry, **(b)** dissolve detailed roof geometry into a continuous polygon, keeping the same outer boundaries as the original polygons, **(c)** reverse vector map of buildings as void and the surroundings as filled polygons, **(d)** DOM (Digital Orthophoto Map) map with a 2 m by 2 m resolution, **(e)** “negative” DOM raster map extracted by clipping the DOM map (d) over the “reverse” building polygons (c), **(f)** a boolean raster map or IsNull raster map, Null cells (void cells) assign to value 1 indicating there are buildings, whereas not Null cells assign to value 0 indicating there are no buildings, **(g)** Hillshade map for buildings and their landscape surroundings in urban areas, **(h)** clipped Hillshade map for buildings extracted by clip Hillshade map (g) over IsNull raster map (f), **(i)** binary raster map showing cells that are non-fully shaded in yellow and cells that are fully shaded in blue.

$$G_t = G_{Bt} + G_{Dt} + G_{Rt} = \frac{S_{hill}}{255} (G_h - G_D) R_b + 0.9 G_D R_d + G_h R_r \quad (6.3)$$

We normalize the S_{hill} values by 255, that is the maximum Hillshade value, in order to have a ratio between 0 and 1. The monthly solar radiation for G_h , G_D and G_{oh} was estimated in section 6.3.1. We extract the monthly solar radiation values for urban area i in each commune using the Zonal Statistics tool from Spatial Analyst. To estimate R_b , R_d and R_r using equations presented in chapter 3, several variables including roof slope and roof azimuth are needed.

For 42 communes, in which the detailed roof shape data exist [261], we estimate the frequency distribution of roof slopes and the frequency distribution of roof azimuths using building polygons in ArcGIS. The following steps are proposed so as to estimate roof slope distribution for buildings in 42 communes and predict their distribution for the rest of the communes in Switzerland:

- (i) Using the existing slope data for each roof surfaces, we only consider slope values for roof surfaces that have suitable area for PV installation,

Table 6.4: Testing errors (RMSE and NRMSE) for shading variables using SVR.

Month	S_{Sh}		S_{hill}	
	E_R	E_{NR}	E_R	E_{NR}
	[-]	[%]	[-]	[%]
Jan.	0.045	8.1	0.028	7.3
Feb.	0.048	11.1	0.019	4.41
Mar.	0.048	16.94	0.015	4.99
Apr.	0.028	16.58	0.005	0.98
May	0.02	19.91	0.011	2.78
Jun.	0.018	19.61	0.014	2.18
Jul.	0.02	19.68	0.012	1.94
Aug.	0.26	18.41	0.008	1.31
Sep.	0.039	16.95	0.009	1.88
Oct.	0.049	12.96	0.019	4.4
Nov.	0.049	9.16	0.024	6.02
Dec.	0.049	7.82	0.026	6.89

- (ii) For simplification, in order to have one slope value for each building, we calculate the average slope value of different roof surfaces of each building,
- (iii) We assign the calculated average slope value to the PV panels to be installed on the roofs.
- (iv) We plot the frequency distribution of roof slope for the buildings in each commune ranging from 0° to 70° with 10° degree bin width,
- (v) We consider the frequency distribution of roof slopes for the 42 communes as labels in training and testing process,
- (vi) Using SVR we predict the frequency distribution of roof slopes for other communes in Switzerland and consider only the central value of each bin for simplification. Testing errors for roof slopes are given in Table 6.3.

The roof azimuth distribution for buildings in each commune is estimated directly from the ground floor polygon shapefiles in ArcGIS. The following steps are used to estimate the roof azimuths:

- (i) Convert building polygons to polylines and do the segmentation of polylines to create building sides,
- (ii) Measure the length of building sides and assume that the azimuth of the main roof surface is perpendicular to the longest side of building,
- (iii) Calculate the azimuth of the longest side of each building polygon in ArcGIS and assign it for the roof,
- (iv) To simplify, when calculating the roof azimuths assume the roof shape to be symmetric.

- (v) Compute the frequency distribution of the roof azimuths in each commune in Switzerland ranging from 0° to 180° with 20° bin width,
- (vi) Compute the frequency distribution of roof slopes for other commune in Switzerland, and consider the central value of each bin for PV panel installation.

The slope and azimuth frequency distributions have been estimated for each commune. For simplification, building roof shapes are considered to be symmetrical, with two rooftop sides characterized by the same slope angle and with opposed azimuth values (e.g. 90° and -90°). The probability of each slope value (the central value of each bin, that is, 7 values for the 7 bins) to be within a bin l is p_l . Similarly, the probability of each azimuth value (the central value of each bin, that is, 9 values for the 9 bins) to belong to a bin m for one side and $-m$ for the other side is q_m . Considering Klein-Andersen solar model, we only take into account roof azimuths within $\pm 90^\circ$ of due south, where south is 0° , north is -180° , east is -90° , and west is 90° [109, 110]. To estimate the possible configuration of slope and azimuth (β, γ) together, we use joint probability. We assume that roof slope and roof azimuth values are statistically independent, thus the joint probability is the product of their marginal probabilities and given by:

$$\mathbb{P}(E_{\beta \in l}, E_{\gamma \in m}) = \mathbb{P}(E_{\beta \in l}) \mathbb{P}(E_{\gamma \in m}) \quad (6.4)$$

where $E_{\beta \in l}$ refers to the event in which the slope β belongs to bin l . $E_{\gamma \in m}$ refers to the event in which the azimuth γ belongs to bin m . $\mathbb{P}(E_{\beta \in l})$ and $\mathbb{P}(E_{\gamma \in m})$ are the probabilities of the two respective events $E_{\beta \in l}$ and $E_{\gamma \in m}$, which respectively equal to p_l and q_m as defined above. Eq. 6.4 has been used to calculate 63 configurations of slope and azimuth for each urban area i . In Eq. 6.3, R_b and R_d are calculated for each slope and azimuth configuration. Finally the global tilted radiation G_t is computed for each commune for the 63 different configurations and for each month of the year. The annual results for one specific slope and azimuth configuration (e.g. slope= 35° and azimuth= 10°) is illustrated in Fig. 6.5.

6.4.4 Final geographical potential estimation

The total geographical potential (P_j^{geo}) is finally computed for each commune j and is given by:

$$P_j^{\text{geo}} = \sum_{i=1}^{N_j} b_i \left(1 - S_{\text{sh}}^j\right) \sum_{l,m} p_l q_m \frac{A_R^j}{2} G_{t,i}(\beta_l, \gamma_m) \quad , \gamma_m \in [-90^\circ, 90^\circ] \quad (6.5)$$

where b_i is the number of buildings in urban area j , N_j is the number of urban areas within commune j , S_{sh}^j is the ratio of fully shaded cells to the total rooftop cells in commune, β_l and γ_m are the slope and azimuth value considered for bins l and m , p_l and q_m are probabilities for a roof to be characterized by a slope in bin l and an azimuth in bin m , A_R^j is the average available roof area for PV installation in commune j , $G_{t,i}$ is the global tilted solar radiation for urban area i . When considering S_{sh}^j in Eq. 6.5, we assume that no significant PV electricity is generated by the PV cells that are fully shaded. The solar potential for the available roof area is then calculated for a typical year and for each month of the year considering only the roofs facing to the south. We assume that the PV panels have the same slope as the roofs. However, for the PV panels over the flat roofs we consider a minimum slope angle of 10° . The slopes are taken into account in Eq. 6.5 as one of the

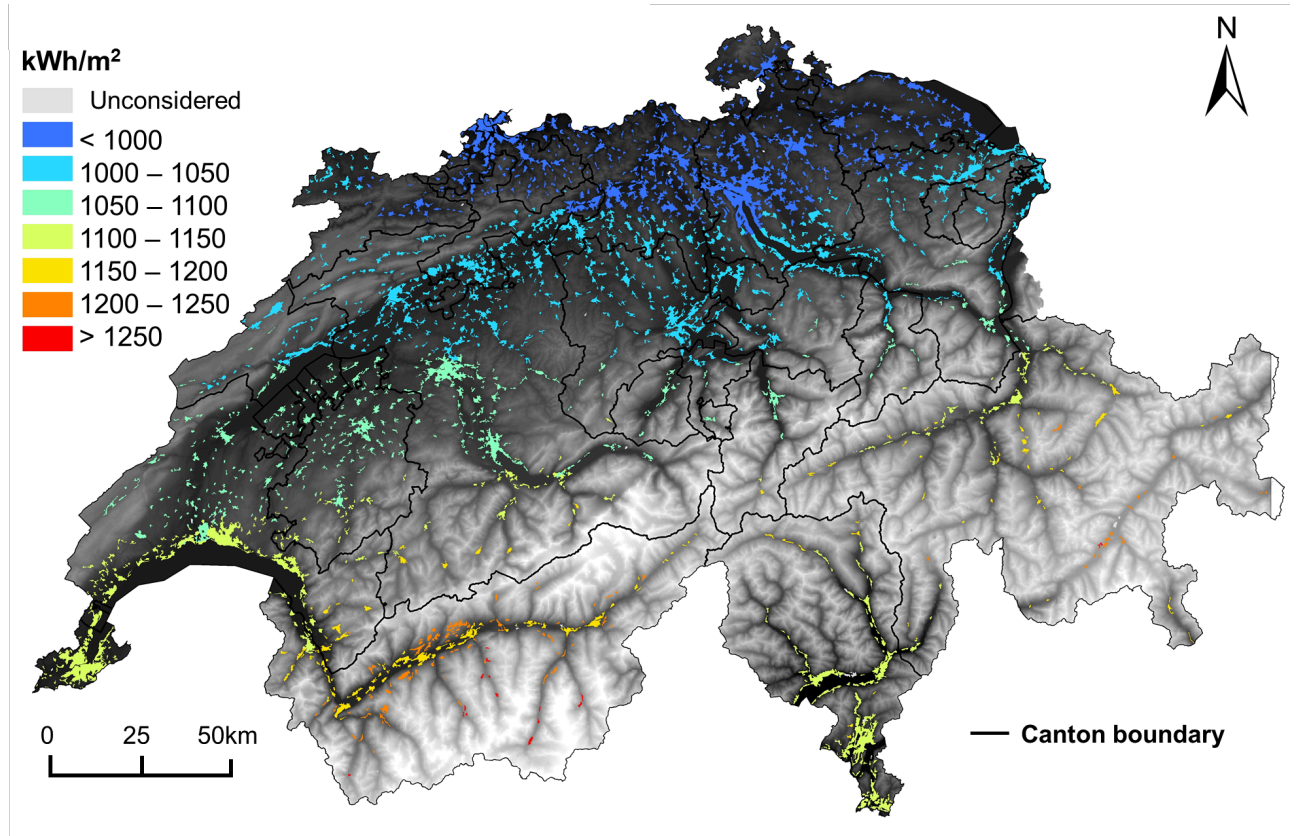


Figure 6.5: Yearly solar radiation for one specific slope and azimuth configuration (slope=35° and azimuth=10°) out of 63 possible configurations.

63 configurations of slope and azimuth. Note that in order to consider a slope for PV panels on flat roofs, for β_l we assign 10 for the central value of bin $[0^\circ-10^\circ]$ instead of 5° .

6.5 Technical potential estimation

To determine the final technical potential, corresponding to the potential electricity production (in GWh/year and kWh/month) from the installed PV panels on useful rooftops, we consider the following equation:

$$P_j^{\text{tech}} = PR \times \eta \times P_j^{\text{geo}} \quad (6.6)$$

where P_j^{tech} is the technical potential, that is the PV electricity production, in each commune j , in GWh per year or kWh per month, P_j^{geo} is the geographical potential, η is the average PV panel efficiency (%) and PR is the performance ratio (in %). Note that in practice varying additional factors can have an effect on the technical potential, including converter losses, dirt, etc; we consider these latter factors to be ideal (i.e equal to 1) in this study. While different PV technology exists on the market, crystalline silicon (c-Si) technology dominates the worldwide PV production [134]. Since 2004, there has been great improvement in the average efficiency of commercial silicon panels, reaching 16% in 2013. These panels are usually guaranteed for a life time of 25 years at a minimal 80% performance ratio [134]. Assuming an annual increase in cell efficiency of the mainstream PV technology, namely by

0.3% per year for the crystalline silicon wafers, we estimate the expected panel efficiency for the year 2016. In Eq. 6.6, the expected panel efficiency η for 2016 is considered as 17%. However, by 2050 the panel efficiency is expected to reach an efficiency of 27.2%. The performance ratio (PR) is defined as the difference between performance under “standard test conditions” (1000 W/m², air mass 1.5 spectrum, panel temperature 25°) and the actual output of the system [224]. The difference is primarily because of losses due to the deviation from standard test conditions, losses due to panel mismatch and dirt, as well as cable and inverter losses. In this analysis, we use 80% performance ratio, because it is expected that PR becomes gradually better due to reliability improvements of the system and remote monitoring [224]. Also note that the choice of PR=80% and η =17% is rather conservative since it is also notably used by the Sonnendach project [266], which provide estimates for roof PV solar potential in a part of the country, as discussed further in the discussion section 6.6.2. The monthly technical potential P_j^{tech} , that is, the PV electricity production for each commune j in Switzerland are shown in Fig. 6.6.

6.6 Results

6.6.1 Discussion

The results presented here relate to the urban areas in 1901 communes in Switzerland. These include the total photovoltaic solar energy potential (yearly and monthly), the total photovoltaic solar energy potential per capita (yearly and monthly), and the available roof area for PV installations. Results presented as a map in Fig. 6.7a and as a histogram in Fig. 6.7b. Fig. 6.7a shows that the annual PV solar electricity production in Switzerland for the urban areas in the 1901 communes reaches 17.86 TWh (assuming 80% performance ratio and 17% efficiency). The total domestic electricity consumption in 2015 (including losses occurring in transmission and distribution) in Switzerland was 62.6 TWh (<http://www.bfe.admin.ch/>). The estimated PV solar electricity production from buildings in urban areas can provide 28% of annual Switzerland electricity consumption based on a relatively conservative approach. The monthly variations of PV solar electricity production and Swiss electricity consumption in 2015 are shown in Fig. 6.8.

While in the majority of communes the PV solar electricity production is less than 15 GWh/year, in a significant minority (about 15% of communes) the production exceeds 15 GWh/year (Fig. 6.7b). More specifically, these latter communes can produce 53% of the total Swiss electricity use in 2015. The last category of the PV electricity production in Fig. 6.7a shows the values greater than 18 GWh/year. The values vary between different communes in Switzerland, as indicated in the plot in Fig. 6.7b. More specifically, Zurich (574 GWh/year), Basel (291 GWh/year), and Bern (255 GWh/year) are the top three highest communes for the potential PV electricity production in Switzerland. The map (Fig. 6.7a) also shows that some northern and central cities have relatively high PV electricity production despite the low solar potential Fig. 6.1 and 6.5. This is primarily due to the fact that large cities have larger number of buildings and, thus, large availability of roof areas for PV installation Fig. 6.3a.

We also calculate the PV solar electricity production per capita for each commune, dividing the total electricity production by the commune population, as MWh per capita. The annual map (Fig. 6.9a) shows that the lowest PV electricity production per capita is in the communes in the central part of Switzerland and in the most densely populated region of Switzerland (Swiss Plateau). By contrast, the highest PV solar electricity production per capita is in the communes located at the Jura

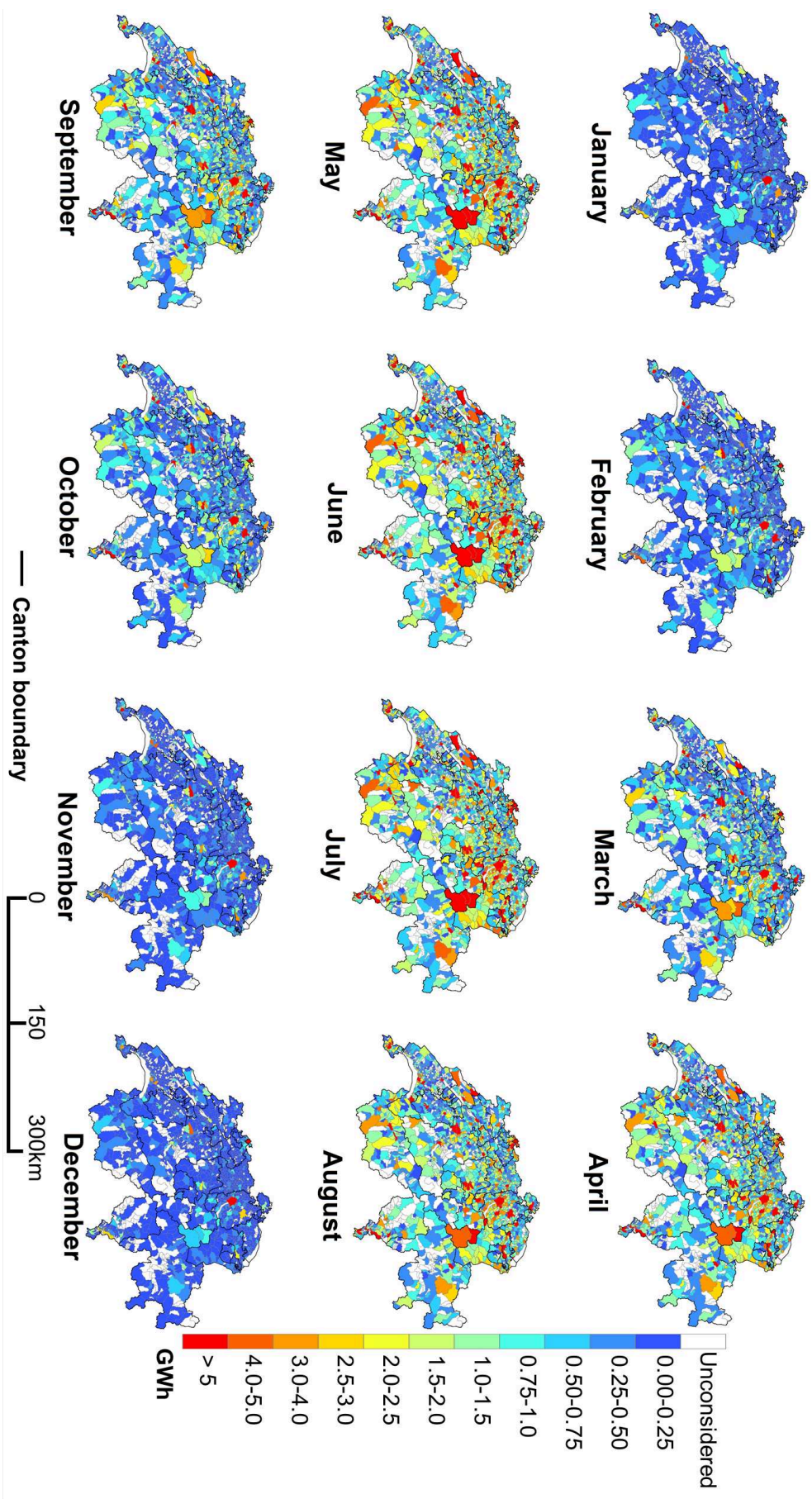


Figure 6.6: Technical potential of rooftop PV solar electricity production for each commune in Switzerland (in GWh/month)

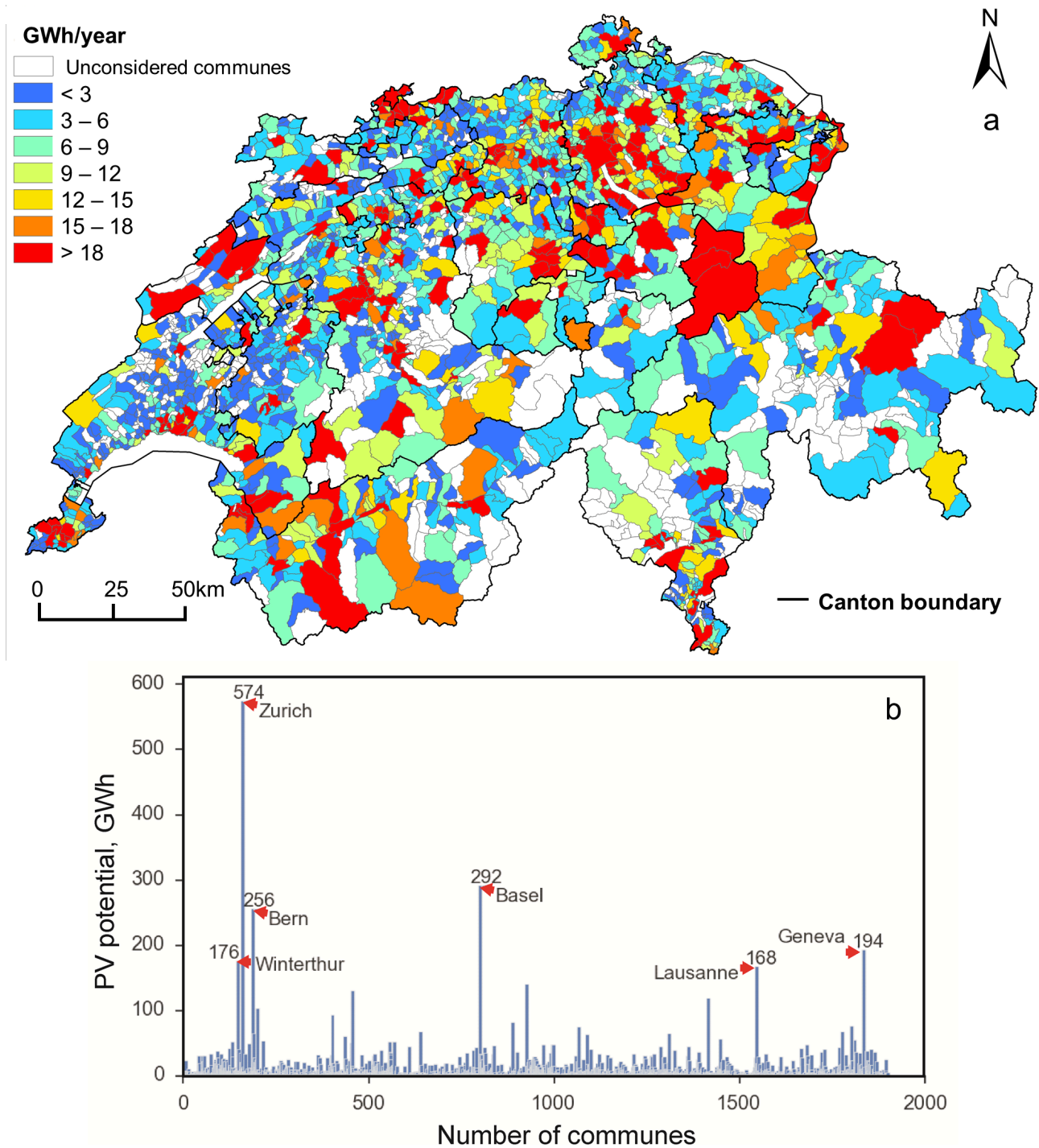


Figure 6.7: (a) Technical potential of rooftop PV electricity production (GWh/year) for each communes in Switzerland; (b) Histogram showing the distribution of rooftop PV electricity production (GWh/year) among 1901 communes in Switzerland. The maximum values belong to large cities (e.g. Zurich, Bern, Basel).

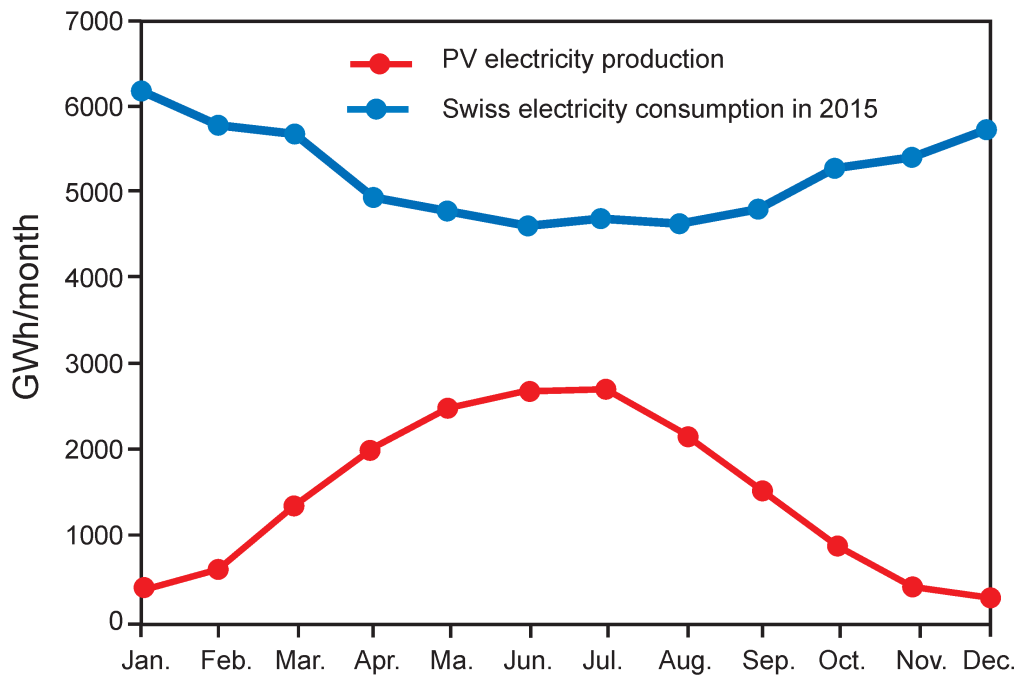


Figure 6.8: A comparison between monthly PV electricity production (GWh/month) for 1901 communes in Switzerland and Swiss electricity consumption in 2015, in GWh/month.

Mountains and the Swiss Alps. The variation of PV electricity production per capita for different communes is shown in Fig. 6.9b. Also, the PV solar electricity production per capita throughout the year (for each month) in Switzerland is shown in Fig. 6.10. More specifically, the spring and summer months (April to September) have the highest potential, as expected; in particular, June and July are the highest. The central part of Switzerland, the Swiss Plateau, is clearly with a lower potential than the Swiss Alps area in the southeast and somewhat lower than the highlands (the Jura Mountains) in the northwest. Nevertheless, the PV solar electricity production for the Swiss Plateau is, by European standard, reasonably high [134, 267, 268].

A comparison of the total potential of PV solar electricity production (TWh/year) with the PV solar electricity production per capita (MWh/year) for 26 cantons in Switzerland is shown in Fig. 6.11a. Zurich, Bern, Aargau, Vaud and St. Gallen are the top fifth cantons in Switzerland as regards the total PV solar electricity production (TWh/year). However, the electricity production per capita (MWh/year) is much more uniform than the total PV potential indicating that the latter is primarily positively related to the population. The total PV solar electricity production (TWh/year) is also naturally positively related to total available roof area (Fig. 6.11b).

The total ground floor area for buildings in urban areas in Switzerland is 407 km²; the estimate is based on the existing data for building stocks in Switzerland [260]. As explained in Section 6.2.2, we consider only buildings in the urban areas using the data from CORINE land cover [254]. The total available area for rooftop PV panels in the urban areas of Switzerland is 328 km². The analysis shows that 81% of the total ground floor area can be used as equivalent to the available roof area for the PV installation. The available roof area per capita is estimated as the total cumulative available roof area divided by the Swiss population in 2015; thus, the total available roof area per capita

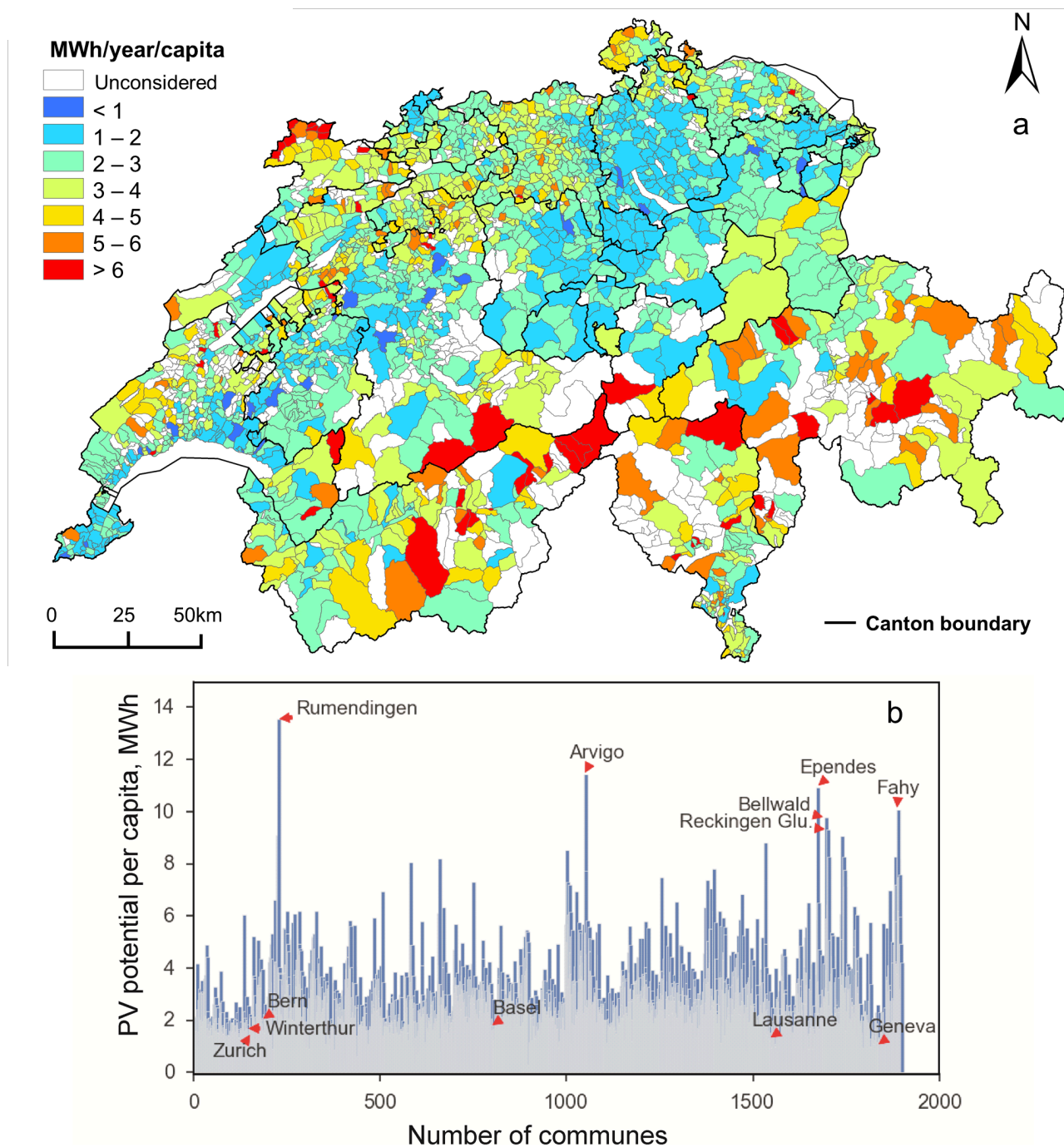


Figure 6.9: (a) Technical potential of rooftop PV electricity production normalized by the population (MWh/year/capita); (b) Histogram showing the distribution of rooftop PV electricity production (MWh/year/capita) among 1901 communes in Switzerland.

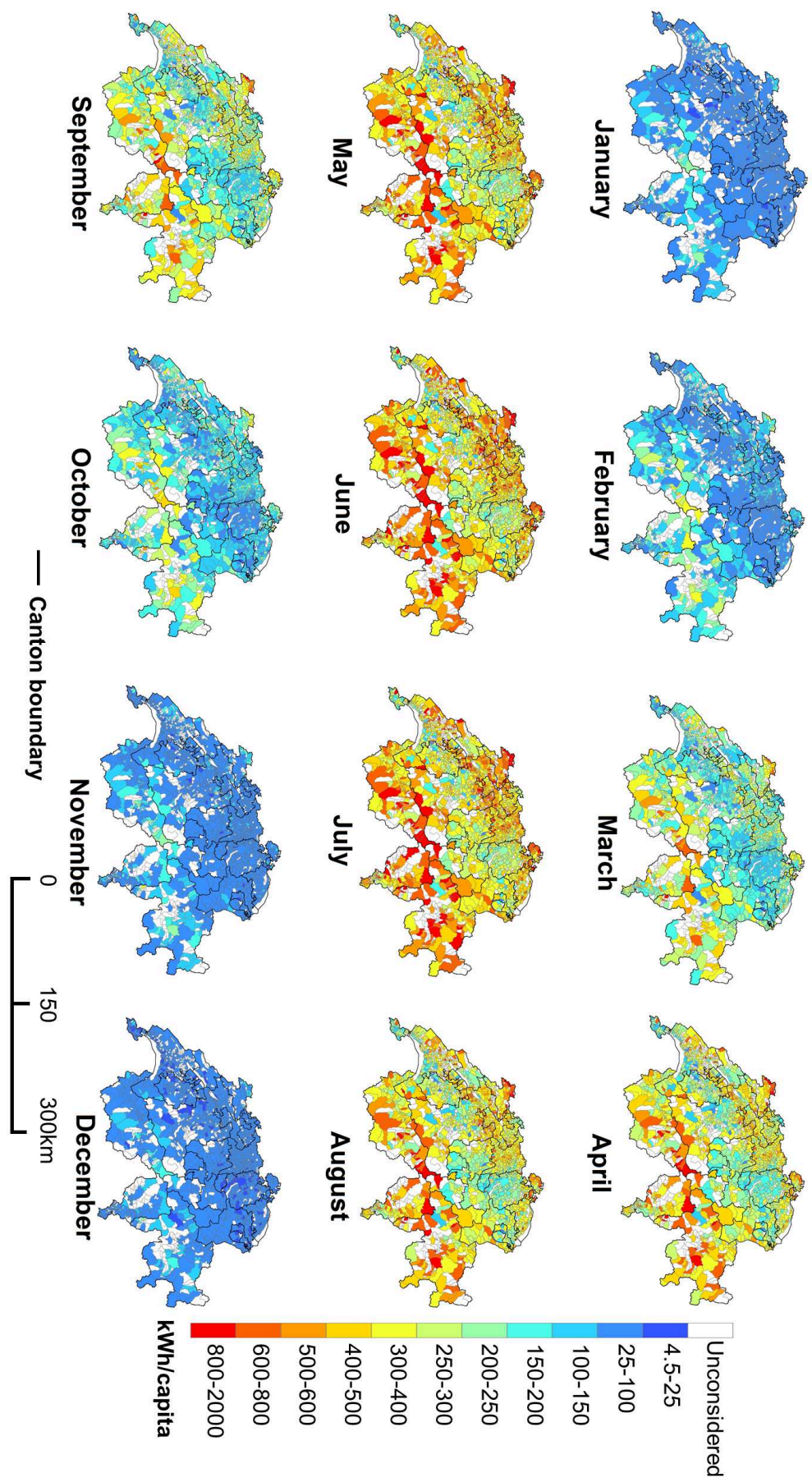


Figure 6.10: Technical potential of rooftop PV electricity production normalized by the population for each commune in Switzerland, in kWh/month/capita.

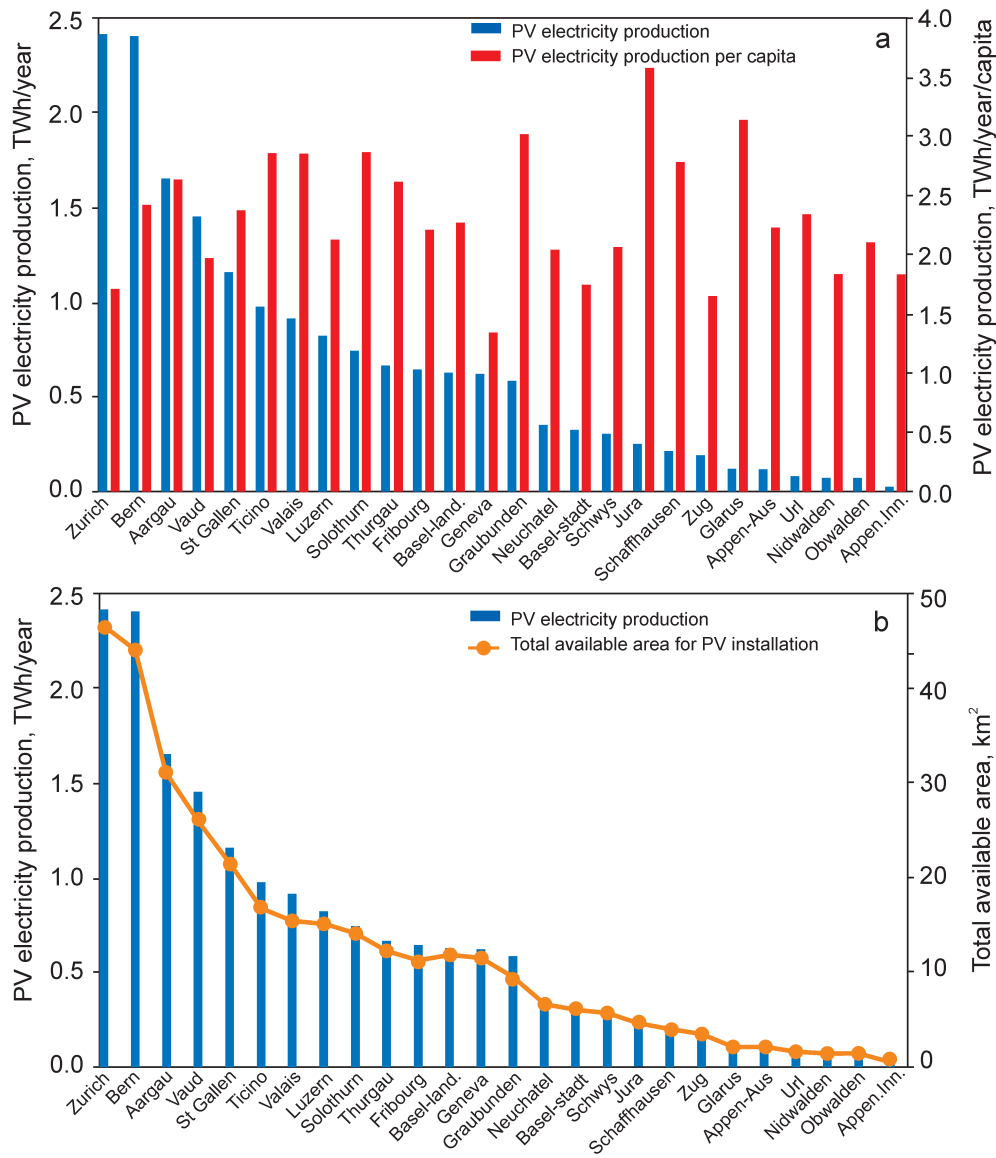


Figure 6.11: (a) Cumulative rooftop PV solar electricity production for each canton in TWh/year and cumulative rooftop PV solar electricity production per capita for each canton in MWh/year/capita, in Switzerland; (b) cumulative rooftop PV solar electricity production for each canton, TWh/year and total available roof area, km², for each canton in Switzerland.

is 41 m²/capita. Note, also, that the estimated available roof area for PV panels decreases when the shading factor and the solar availability of the rooftops are taken into account. These factors are considered in the Eq. 6.5 and explained in Section 6.4.4.

6.6.2 Validation with other potential studies

The results obtained in this chapter were first compared with those provided by [215]. The differences between the two studies, notably concerning the total available roof area, total available roof area per capita, and PV solar electricity production are as follows: In reference [215], the total available roof area and the total available roof area per capita is estimated as 138.22 km² and 18 m² /capita,

respectively; our estimation, however, yields 328 km^2 and $41 \text{ mm}^2 / \text{capita}$, respectively. The annual PV solar electricity production is obtained by [215] is 15.044 TWh/year ; our estimate, however, is 17.86 TWh/year . The reasons for these differences are as follows: our estimate is based on buildings in the urban areas whereas, where reference [215] takes into account all the buildings (urban and rural areas) in Switzerland. In [215], the final value of available roof area takes into account (i) the architectural suitability includes corrections for superstructure on the roofs (e.g. elevators, chimney), (ii) the shading effects, and (iii) the solar availability on the roofs depending on the roof slope, roof orientation and the location. By contrast, in the present chapter we estimate the three above steps separately and the final value of the available roof area (section 6.4.1) is based on architectural suitability and some technical considerations regarding the PV panel installation on the roofs. This is the reason why the available roof area provided here is much higher than that by [215]. The available roof area for PV installation decreases when shading factor and the solar availability of the roofs are taken into account. Another difference regarding the annual PV solar electricity production relates to the efficiency of the PV panels. In [215], the efficiency of PV panels is assumed to be 10%. Given the improvements of the PV panels during the time, we estimate the efficiency of 17% for the year 2016.

We also compare the PV solar electricity production of the canton of Geneva, estimated in this study as 622 TWh/year , with the value estimated by SITG (as presented in the data section 6.2.1) [261] of 662 TWh/year . The main difference is related to the methods of calculation for solar radiation, shading factors, available roof areas for PV installation, and roof characteristics including slope and azimuth. In addition, while they consider all the existing buildings in the canton of Geneva [261], we focus on the buildings in urban areas (70% of all the existing buildings in the canton). There are also slight differences in the efficiency and performance ratio of the PV panels. They assume efficiency of 16% and 79% performance ratio for the PV panels, whereas, for the year 2016, we assume the efficiency of 17% and 80% performance ratio.

In 2016, the PV solar electricity production has been estimated for several thousand buildings in Switzerland by the Sonnendach.ch project [266]. This project is expected to be completed by 2018, whose aim is to estimate the PV solar electricity production for all the roofs of buildings in Switzerland. To be able to compare our results, we aggregate the data extracted by the project within the urban areas for several communes in Switzerland [269]. The main differences between the present results and those of the Sonnendach.ch project are, first, related to the estimation of the total available roof area for PV installation. In the present study we remove all the roof superstructures (e.g. chimney, dormer, staircase) as well as surfaces with an area smaller than 28 m^2 from roofs (details in section 6.4.1). By contrast, in the Sonnendach.ch project all the roof areas including dormers or annexes with a minimum dimension of 8m side length have been considered [269]. The second difference is related to the roof azimuth for solar estimation. In the present study, only roof azimuths within ± 90 of due south are taken into account. By contrast, in the Sonnendach.ch project for solar rooftop estimation roof surfaces in all directions have been considered [269]. By harmonizing the above assumptions between the two studies, the values of PV solar electricity production obtained in Sonnendach.ch project agree well with those reported in this work.

6.7 Summary

This chapter proposes a methodology using combination of Support Vector Regression and Geographic Information Systems to estimate the technical solar rooftop PV potential, that is, the potential solar PV solar electricity production, focusing on the urban areas at the commune level (the smallest administrative division) in Switzerland. For the first time, Machine Learning is used in order to compute multiple variables of interest leading to the technical solar potential in Switzerland, including the global horizontal radiation, the available rooftop area for PV, the slope and aspect of rooftops and shading factors over rooftops. This study shows that Switzerland has a large PV roof top potential, providing a considerable share of future Switzerland grid. Therefore, the design of the future electric grid, the capacity of the existing grid to receive additional power and the contribution of rural and urban areas for rooftop PV deployment in the Swiss grid should be of primary concern for future study. Our results show that, on average, 81% of the total ground floor area of each building can be used as the available roof area for the PV installation. Also, considering the total available roof area, e.g. 328 km², and all rooftops directed within $\pm 90^\circ$ of due south, the annual PV solar electricity production for the urban areas in 1901 communes in Switzerland is estimated to be 17.86 TWh. This amount corresponds to 28% of the Swiss annual electricity demand in 2016.

A flowchart summarizing the entire methodology proposed in the chapter is shown in Figure 6.12.

Note that the limitations of the chapter have not been highlighted. These will be discussed in the following Chapter 7, which will further extend the solar rooftop PV potential study to improve its accuracy. The previously used (200×200) [m²] pixel grid will be considered, and several aspects of the methodology will be improved, notably using additional data sources, as well as updated and more detailed processing steps for the multiple estimations, in particular for the available rooftop area for PV panels and the geometrical characteristics of rooftops across Switzerland. Also, a measure of uncertainty will be provided through the computation of Prediction Intervals obtained from the use Random Forests use to estimate the variables of interest.

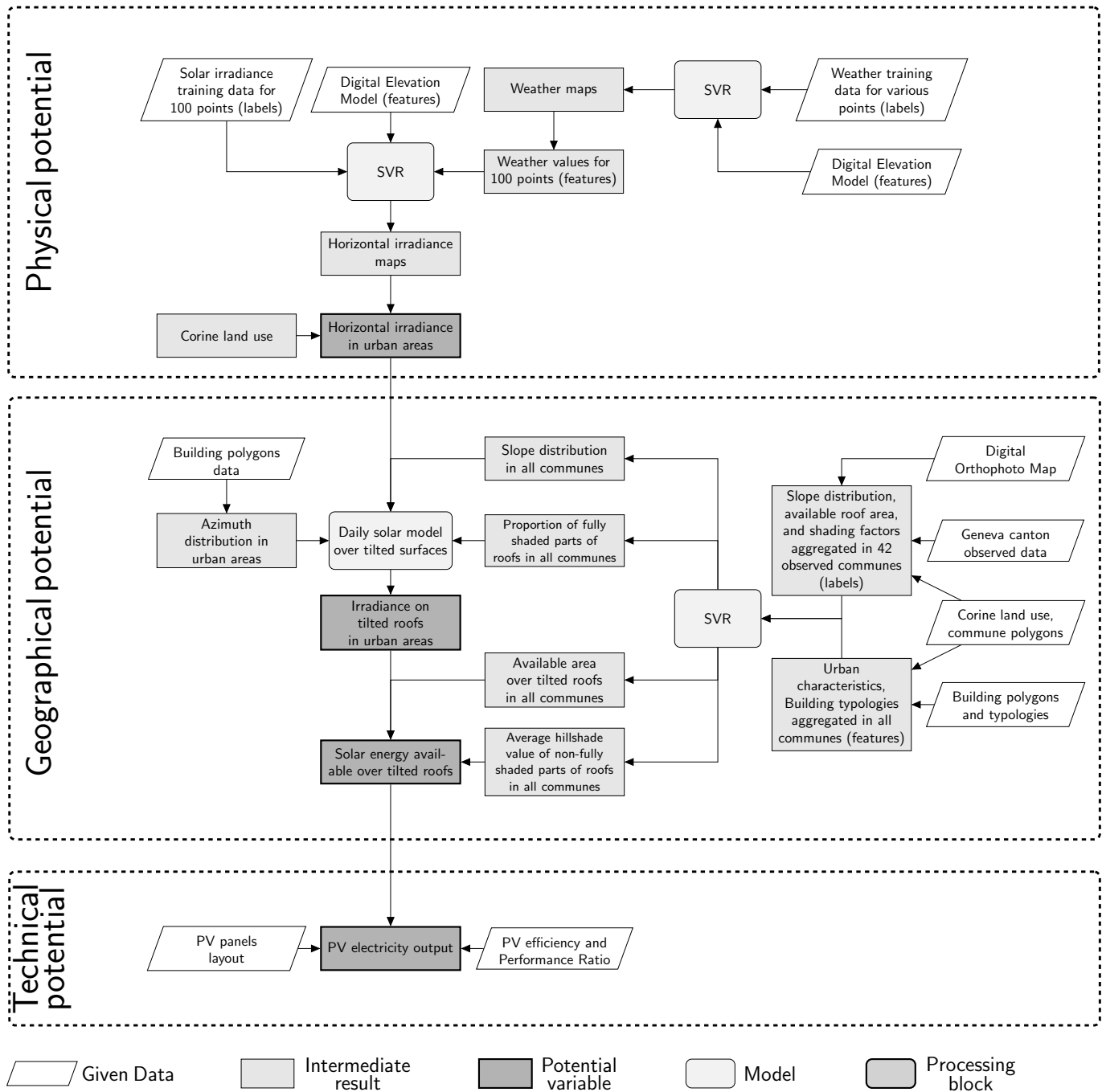


Figure 6.12: Flow chart of the methodology for the solar PV potential estimation at commune scale.

7

Solar energy: an improved potential estimation at pixel scale

This chapter is based on the article [23]:

Assouline, D., Mohajeri, N., and Scartezzini, J-L. (2018). Large-scale rooftop solar photovoltaic technical potential estimation using Random Forests, *Applied Energy* 217 189-211.

It also borrows from the article [270]:

Assouline, D., Mohajeri, N., and Scartezzini, J. L. (2017, October). Building rooftop classification using random forests for large-scale PV deployment. In *Earth Resources and Environmental Remote Sensing/GIS Applications VIII* (Vol. 10428, p. 1042806). International Society for Optics and Photonics.

and the book chapter [21]:

Assouline, D., Mohajeri, N., and Scartezzini, J. L. (2018). Estimation of Large-Scale Solar Rooftop PV solar potential for Smart Grid Integration: A Methodological Review. In *Sustainable Interdependent Networks* (pp. 173-219). Springer, Cham.

This chapter aims at further investigating and ultimately improving the solar rooftop potential study initiated in the previous chapter. While the latter chapter brought valuable contributions in terms of PV solar potential estimation at a large scale, there are a few aspects that could benefit from a revised viewpoint. As such, the first aspect to rethink is the scale and the resolution of the study. While estimating PV solar potential at a city/commune scale is very useful as a support system for policy-making and strategy planning [213], it also has two significant disadvantages including: (i) it is not adapted for private households, smaller cities or neighborhood to estimate the return on investment of solar PV panels, (ii) it makes it hard to integrate decentralized PV solar electricity production with energy storage and electrical grid so as to manage the energy demand of neighbourhoods or communities, and (iii) the official boundaries of communes are subject to changes each year, which means the potential would theoretically also have to be slightly updated accordingly each year. To

resolve the above issues, the resolution must be changed to a finer one. We therefore consider the grid size of (200×200) [m²] pixels spanning over the whole Swiss territory, as presented in chapter 4 (section 4.2.2). As this higher resolution increases significantly the computational requirements, the machine learning methodology needs to be adapted as well. In particular, SVM have trouble handling very large data and are quite challenging to tune, given their three hyperparameters. As a consequence, Random Forests (RF) are considered as the benchmark algorithm, given its multiple computational advantages over SVMs (as presented in section 2).

The higher resolution considered through the use of pixels automatically implies that a higher precision should be achieved in the multiple steps of the potential estimation. As a result, besides the scale of the study, this chapter attempts to improve some of the strategies that lacked in precision and fill some of the most important gaps left by the previous study. The main improvements, detailed in the rest of the chapter, are the following:

- **Spatial resolution and coverage.** In addition to the newly considered pixel scale, a larger number of buildings is taken into account across the country, since additional data sources are considered and because the Corine Land Cover database (defining the urban areas in the previous chapter) is not considered anymore. All buildings (available in vector format) are therefore considered, regardless of the land use. While the use of the Corine Swiss database made sense at a commune scale to extract density factors within urban areas and not across entire communes, it is not significant at a pixel scale. Also, multiple additional data sources became available since the first study, offering ground truth information in previously unknown locations.
- **Rooftop geometrical characteristics estimation.** The previous study aimed at estimating the aspect (direction) of the roofs using building footprints, as well as their slope using an additional ML model trained on Geneva canton data. Here, we propose a more precise methodology taking advantage of a (2×2) [m²] LiDAR data (DOM - a high resolution elevation data) available over the whole Swiss territory to obtain the geometrical characteristics of the rooftops over the whole country. More specifically, raster processing over the DOM is used to extract significant features in order to perform a multi-layer building rooftop classification over the entire country to extract the roof type (shape), slope and aspect of the surfaces of all Swiss rooftops.
- **Available area for PV installation over rooftops.** The estimation of the available rooftop area ratio is perhaps the most relevant in the potential estimation, simply because it is directly proportional to the final estimation. In order to improve its computation, its estimation in the Geneva canton (to build labels/examples for the ML model) was completely redesigned to account for the geometrical characteristics of PV solar panels, by simulating the presence of PV modules over the detailed rooftops of the canton, using GIS vector processing. The computation of the total available area over each rooftop is performed based on the estimated available area ratio and the classified shape of the rooftop.
- **Uncertainty assessment.** An important gap in the previous solar study is the lack of uncertainty measures attached to the multiple estimation. We here propose a simple way to compute the uncertainty attached to a prediction made by a Random Forest model through the computation

of prediction intervals, based on Quantile Regression Forests [48] and the Random Forests implementation from the Scikit-Learn library [25].

The chapter is organized as follows. Section 7.1 positions the present chapter within the context of the literature. Section 7.2 presents the data sources used in the chapter and some of the processing steps performed to extract significant features. Section 7.3 presents the re-computation of the theoretical potential in Switzerland, based on a new Digital Elevation Model and Random Forests. Section 7.4 details the re-computation of the geographical potential within the (200×200) [m²] pixels, including the new estimation of the available rooftop area ratio for PV panels, the re-computation of the shading impact in pixels, the roof shape, slope and aspect classification and the estimation of the total available area over the rooftops of Switzerland. Section 7.5 presents the re-computation of the technical potential. Section 7.6 provides a discussion on the obtained results, and a comparison with the results obtained in the previous chapter. Finally, section 7.7 concludes the chapter and summaries the proposed methodology.

7.1 Literature Context

As the literature review concerning the solar rooftop PV solar potential studies was already presented in the previous chapter, we will here position this chapter in the context of the previously discussed literature, highlighting its novelties in the domain.

As outlined in the literature review of the previous chapter, most studies aiming at estimating the available area for PV solar panels over rooftops apply simple strategies using coefficients over the footprint building area. Many studies use predefined coefficients accounting for multiple constraints (minimum solar irradiance threshold, obstructing superstructures, shading impacts, historical considerations) [227, 271], often adapted for various slope values and building types [224–226, 230, 233, 235, 272, 273]. Other strategies are based on sampling methods. The idea of sampling methods is to compute a variable of interest solely for a given sample of points, or locations, and use an adequate strategy to extrapolate it to the whole data set, or whole region. It allows to provide a more reliable estimate of the variable than an assumed constant coefficient, while keeping the computational requirements of the method reasonably low. A few studies used this strategy to provide an estimate of the available PV area and ultimately the PV solar potential at a large scale [3, 229, 234, 274]. A more precise strategy used in the literature consists in using GIS processing and high resolution LiDAR data to try to extract the available area directly from precise rooftop polygons and elevation data. Notable studies using GIS for area estimation include [231, 232, 241, 275]. Such a strategy is very precise, but requires very large high resolution data to allow a large scale estimation, and has therefore been used mainly for relatively small scale case studies.

Another important aspect related to the estimation of the available roof area is the geometrical aspect estimation (shape, slope, aspect). Many studies use LiDAR data to determine slope and aspect of all rooftops individually [276, 277] or to develop a digital surface model to estimate area, geometrical aspects and shading impact to ultimately extract the PV solar potential [245, 278, 279]. A few studies used LiDAR to classify roof shapes across a region, using thresholds to define multiple slope and aspect classes, and a catalog of common roof shapes for the shape classes [242, 280].

This chapter proposes, for the first time to the best of our knowledge, to combine the accuracy of GIS and LiDAR together with the predictive power of machine learning in order to estimate the geometrical characteristics of the roofs and their available area for PV installation for a whole country. As explained previously, the sole issue of the GIS and LiDAR approaches is the need for a very large high resolution data and the high computational requirements as a consequence of this large data. The suggested methodology is to take advantage of the available data or a reasonable portion of it (large enough to gather valuable information but small enough to be handled easily) to extract a precise training data through GIS processing (for available area examples) and LiDAR raster processing (for the roof classification), before predicting the desired variables at the remaining unknown locations with the help of machine learning. Also, studies very rarely provide uncertainty values attached to the multiple rooftop PV related variables. The use of Random Forests allows us to extract such values through the computation of prediction intervals.

7.2 Data

7.2.1 Data sources

All data sources used within this chapter are presented in Appendix A and signified by a ✓ for Chapter 7 within tables A.1 and A.2. They naturally include some of the data already used in the previous chapter, but also additional digital elevation models, an additional building vector data for rooftops available for a part of the country (Sonnendach project) and an additional building information data (RegBL) for the whole country. The general consistency of the newly used data has been verified in common areas. In particular, the location of the buildings in the building clusters data (VECTOR25) has been verified with respect to the building data from the Sonnendach project with GIS processing. The respective vector shapefiles were overlapped and the centroids of each dataset were matched.

Given the spatial resolution of the available data and their different levels of precision, we divide the Switzerland territory in three different zones. The data in these three zones treated differently throughout the study. These three zones are: (1) the Geneva canton (GEN), (2) the zone spanned by a 3D building data for around 800 communes around Zurich, available from the Sonnendach project [260, 266, 281] (SON) and (3) the remaining Swiss territory, once the two previous zones have been considered (OOSG). The data available from the Sonnendach project (see Table A.2) is a vector polygon data set offering an estimation of the rooftop PV electricity production for buildings included in the fraction of Switzerland shown in Fig. 7.1 referred to as SON [266]. It also provides information on the rooftop characteristics of these buildings. Fig 7.1 shows a visualization of the three mentioned zones.

7.2.2 Data processing

The main updated building data used for Switzerland is the RegBL (Registre des batiments et logements) dataset, available in a GIS vector point format. Each point represents the location of one building and contains related information. Since it is the most comprehensive building data at national level in Switzerland, it was pre-processed separately to extract useful features.

The features available from the RegBL data exist for each building and are of two types: (i) real-valued features and (ii) categorical (class-based) features. The real-values features include number of

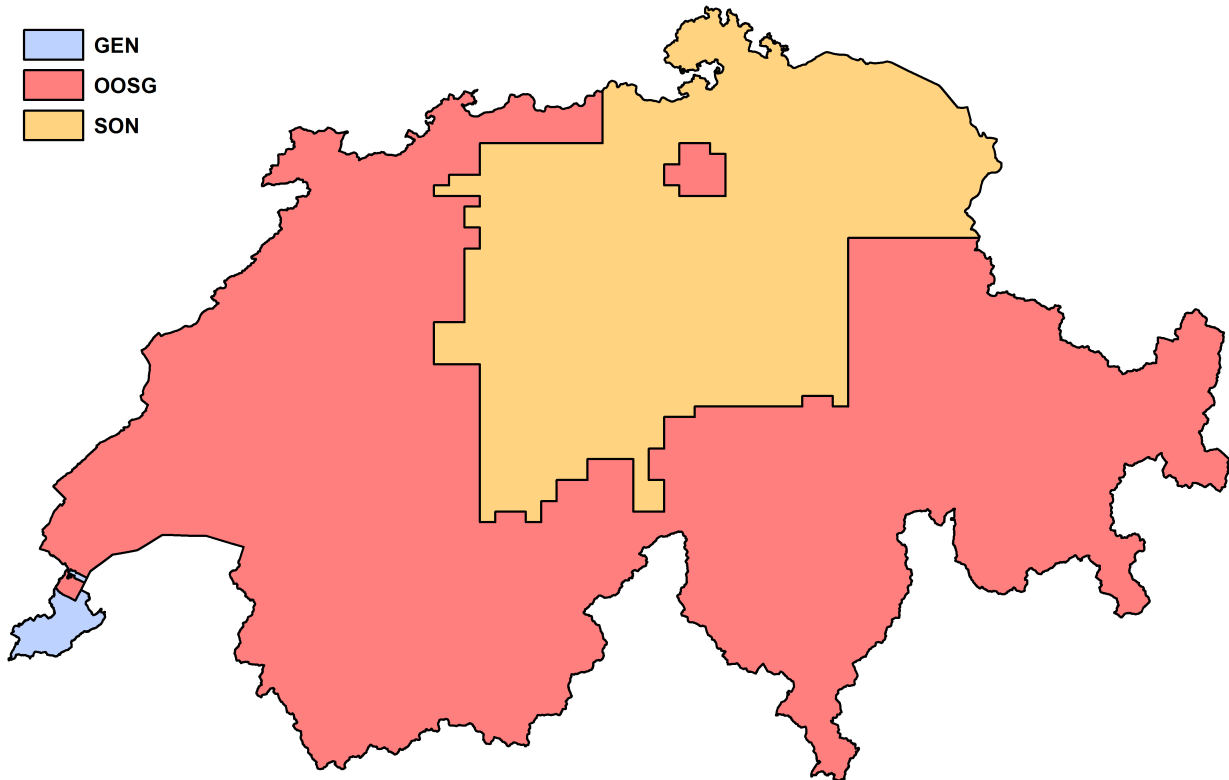


Figure 7.1: An schematic map of Switzerland showing the location of three zones for different data availability and different data processing (see section 7.2.1). GEN: Geneva canton zone, SON: zone spanned by the “Sonnendach” data, OOSG: the “Out Of SON and GEN” zone, meaning the remaining territory in Switzerland.

floors, number of flats, and footprint area. The categorical features include information on energy usage and building typologies. More specifically the categorical features include the following information:

- **Heating type.** The classes include no heating, wood stove, central heating for one flat, central heating for one building, central heating for multiple buildings, long-distance heating, other heating type.
- **Main heating source.** The classes include no heating source, oil, coal, gas, electric, wood, heat pump, solar cells, district heating, other source.
- **Main water heating source.** The classes include no heating source, oil, coal, gas, electric, wood, heat pump, solar cells, district heating, other source.
- **Construction period.** The classes include before 1919, 1919-1945, 1946-1960, 1961-1970, 1971-1980, 1981-1985, 1986-1990, 1991-1995, 1996-2000, 2001-2005, 2006-2010, 2011-2015.
- **Residential class.** The classes include single family, individual, multi-family, mainly residential, partially residential, not residential.

- **Building typology.** The classes include single flat building, two flats building, three flats building, community flat, hotel, other residential building, offices building, commercial building, transport and communication, garage, industrial, storage, recreational building, museum/library, academic/teaching/research building, health building, sports building, agriculture religious building, historical building, non-classified building.

We aggregate the above features within the pixels using averages for real-values features and class probabilities for categorical features. As regards real-value features, for each pixel we obtain the following averages: (i) number of building floors, (ii) number of flats and (iii) ground floor area. As regards categorical features, we obtain the probability for each class within each categories. Four other features are added. These include the building density calculated as (i) number of buildings per pixel and (ii) number of building clusters per pixel. Also added site coverage calculated as (c) the ratio of the total ground floor area of individual buildings to the total pixel area, (d) the ratio of the total ground floor area of building clusters to the total pixel area.

Therefore, a total number of 73 urban features extracted from the building information data (RegBL) and the building clusters data (VECTOR25), to be used for training, test and apply RFs for the estimation of available roof area and shading factors in each pixel. Due to the intrinsic properties of RFs (see section 2.3 in chapter 2), no further feature selection is needed even though there is a relatively large number of features. Besides, it should be acknowledged that the RegBL data suffer from missing features in a significant amount of buildings, specially for the ground floor area. The building data from the Sonnendach project was used to fill the ground floor area gaps for all the buildings in the SON zone 7.1.

7.3 Theoretical potential estimation

Following the strategy adopted for the commune level study in previous chapter, we first estimate monthly maps for global horizontal radiation (G_h), diffuse horizontal radiation (G_D), and extraterrestrial horizontal radiation (G_{oh}), yet using Random Forests instead of SVR in order to obtain prediction intervals for the three solar variables. The SoDa satellite data (see Table A.1) still provides examples of horizontal radiation for 100 locations in Switzerland in order to build the output training data for the RF model. The 200×200 [m²] pixel grid presented in section 4.2.2 serves as the base of for the location of unknown data points, and the DHM25 data provides spatial features (latitude, longitude, and altitude) for the RF model. The SoDa data values, provided hourly, are aggregated as monthly mean values, leading to average daily radiations in each month. For each pixel, if the data is available from the SoDa data, the value of the data point is still allocated to the closest pixel identified by its centroid. To predict the data for the rest of the pixels, we use Random Forests (RFs) for spatial interpolation.

The monthly maps for monthly temperature, precipitation, sunshine duration and cloud cover estimated in 4.2.2 (chapter 4) using RF models are first considered. Other RF models are then trained in order to compute monthly G_h , G_D and G_{oh} , using the mentioned meteorological variables as well as latitude, longitude and altitude as features for the training process. The labels are the monthly satellite data for G_h , G_D and G_{oh} extracted at 100 different locations across Switzerland from the SoDa database [258]. The trained RFs models are used in order to estimate monthly G_h , G_D and G_{oh} in unavailable locations and build monthly maps.

Table 7.1: Testing RMSE (E_R) and NRMSE (E_{NR} , in percentage) for Random Forest models trained for weather and solar variables.

Month	Temp.		Cloud		Prec.		Suns.		G_h		G_D		G_{oh}	
	E_R [°C]	E_{NR} [%]	E_R [%]	E_{NR} [%]	E_R [mm]	E_{NR} [%]	E_R [hours]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]	E_R [kWh/m ²]	E_{NR} [%]
Jan.	1.63	0.60	9.49	16.68	21.58	27.42	12.95	16.29	84.19	6.18	32.35	4.43	12.81	0.41
Feb.	1.40	0.52	8.21	14.67	20.06	27.55	9.18	9.07	104.06	4.79	64.70	6.33	11.46	0.25
Mar.	0.88	0.32	9.17	15.13	21.46	24.28	11.22	8.05	158.04	4.43	106.72	7.47	8.88	0.13
Apr.	0.82	0.29	8.37	12.87	19.59	20.82	13.32	8.61	175.35	3.63	154.87	8.34	5.41	0.06
May.	0.78	0.27	6.61	10.03	18.88	14.94	15.41	8.78	252.28	4.77	173.77	7.68	2.43	0.02
Jun.	1.03	0.36	8.41	13.24	19.50	14.44	16.41	8.66	325.49	5.71	171.79	7.34	0.57	0.00
Jul.	0.99	0.34	6.84	11.82	22.65	16.23	19.09	8.83	251.21	4.43	144.34	6.55	1.41	0.01
Aug.	0.84	0.29	5.42	9.26	20.71	14.56	15.58	7.87	209.43	4.38	134.63	7.20	3.75	0.04
Sep.	0.73	0.26	4.24	7.09	19.08	16.27	12.99	8.24	172.42	4.45	108.01	7.07	7.99	0.10
Oct.	0.71	0.25	5.08	8.51	16.86	17.36	12.72	10.43	113.48	4.35	66.97	5.95	10.53	0.20
Nov.	1.16	0.42	6.34	10.36	20.00	20.65	12.01	15.41	86.53	5.75	38.32	4.81	9.97	0.28
Dec.	1.65	0.61	9.13	15.02	21.20	22.96	11.18	17.30	73.73	6.91	32.19	5.03	11.94	0.44

The testing errors for monthly solar components and weather variables are shown in Table 7.1. The NRMSE for weather variables show a better performance in summer. This is partly explained by the fact that cloud cover, precipitation and sunshine duration show on average values that are smaller and more stable in the summer than in the winter. It can also be noted that the accuracy of G_D is on average smaller than G_h , which can be partly related to the stochastic behavior of G_D . G_{oh} shows an excellent prediction accuracy, which is expected given its more deterministic behaviour, as it is not subject to clouds and other atmosphere related issues. Yearly averaged maps for G_h and G_D can be seen in Fig. 7.2.

We also compute 95% Prediction Intervals (PIs) for G_h , G_D and G_{oh} . Examples are showed for G_h and G_D in January and March in Fig. 7.3. We do not show the PIs for G_{oh} as its patterns are highly predictable and not subject to the weather fluctuations. One can observe that the test confidence is lower than 95% in March, which shows a certain difficulty to derive reliable PIs in this period of the year. It can be partly related to the stochastic behavior of the weather in mid-season (March). Also, the test confidence is clearly lower for G_D than for G_h , which confirms that G_D is more arduous to predict than G_h , as depicted from the testing errors in Table 7.1. Finally, in order to assess the uncertainty attached to the estimations for G_h and G_D , average monthly prediction errors are computed for a random sample of 10000 unobserved pixels across Switzerland and shown in Table 7.2. As discussed in previous chapters 4 and 5, the lower and upper prediction errors are computed as the lower and upper width of the PIs. The latter prediction errors confirm the patterns observable from the test errors, in particular the lower general accuracy achieved during summer.

7.4 Geographical potential estimation

In order to extract the portion of the physical potential available over tilted rooftops, that is the geographical potential, several modifications have been carried-out to update the methodology adopted in the commune scale study. The two principal changes are the following:

Table 7.2: Prediction Errors related for G_h and G_D . Monthly Prediction Errors, computed using Quantile Regression Forests, averaged over a random sample of 10000 unobserved pixels. $PE_{s,down}$ is the average bottom error above the mean predicted value, $PE_{s,up}$ is the upper error above the mean predicted value, PE_s is the average of $PE_{s,down}$ and $PE_{s,up}$.

Month	G_h			G_D		
	$PE_{s,down}$ [kWh/m ²]	$PE_{s,up}$ [kWh/m ²]	PE_s [kWh/m ²]	$PE_{s,down}$ [kWh/m ²]	$PE_{s,up}$ [kWh/m ²]	PE_s [kWh/m ²]
Jan.	181.48	229.08	205.28	53.29	61.25	57.27
Feb.	256.27	337.26	296.76	112.65	114.00	113.32
Mar.	322.50	390.10	356.30	168.29	181.97	175.13
Apr.	253.27	350.21	301.74	199.89	274.64	237.26
May	441.65	471.88	456.76	252.37	321.87	287.12
Jun.	621.65	548.08	584.86	220.70	318.11	269.41
Jul.	588.16	532.59	560.37	179.97	251.10	215.54
Aug.	460.20	413.65	436.92	152.54	221.64	187.09
Sep.	286.63	413.99	350.31	107.01	163.11	135.06
Oct.	263.40	399.48	331.44	72.23	100.61	86.42
Nov.	165.84	203.32	184.58	56.54	79.20	67.87
Dec.	133.02	163.48	148.25	46.39	62.75	54.57

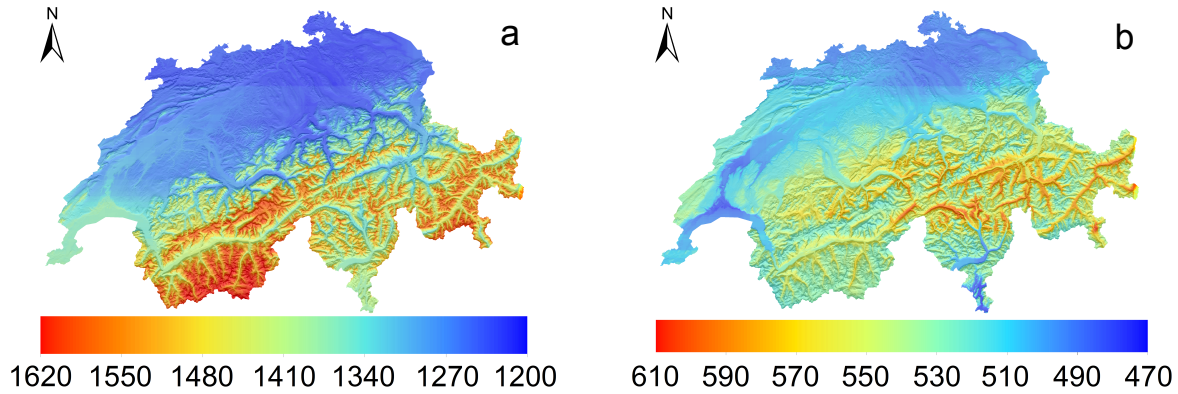


Figure 7.2: Prediction of horizontal solar radiation maps using RFs. **(a)** Yearly mean global horizontal radiation (G_h), in kWh/m², **(b)** Yearly mean diffuse horizontal radiation (G_D), in kWh/m².

- The estimation of the available area has been completely redesigned to provide a better accuracy, and its treatment is different for each zone of Switzerland (GEN, OOSG, SON as seen in Fig. 7.1)
- A roof shape, aspect and slope classification using raster data is suggested to provide a better estimation of the rooftops' geometry and replaces the previous slope and aspect computation.

The general steps of the geographical potential estimation is as follows: (i) estimate the available area for PV installation over the roofs, (ii) estimate the shadowing effects from surrounding obstacles,

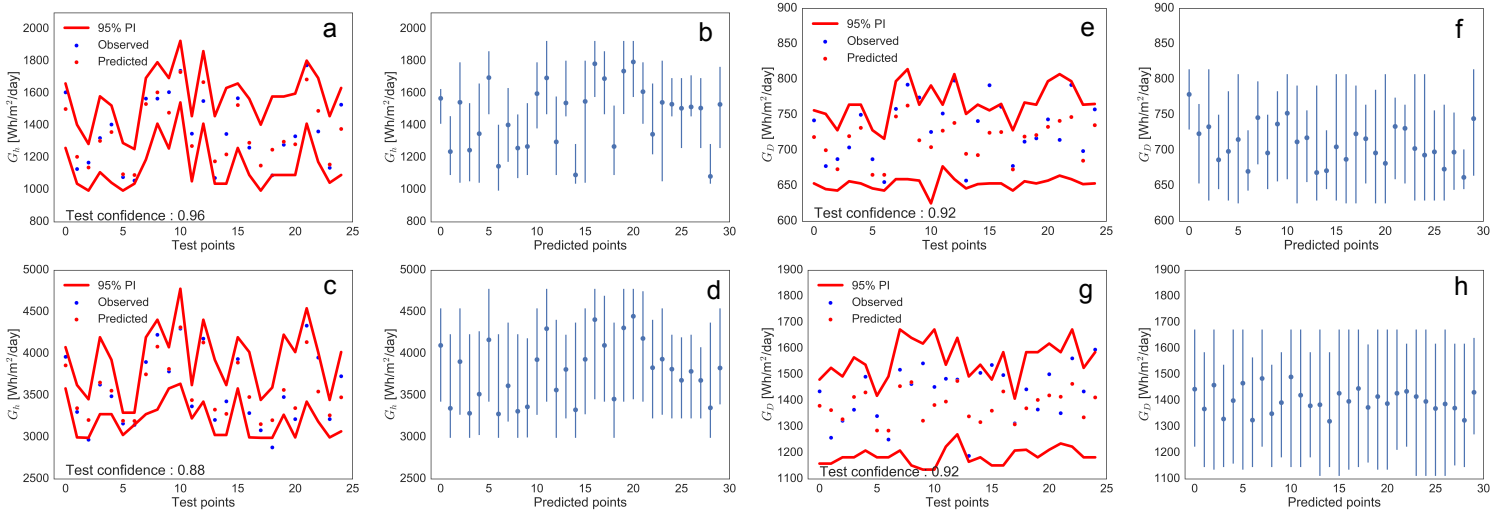


Figure 7.3: Prediction Intervals (PIs) from Quantile Random Forests for global horizontal and diffuse horizontal radiations (G_h and G_D). **(a)** and **(c)**: PIs for G_h in the test set, respectively in January and March, **(b)** and **(d)**: PIs for G_h for 30 random unobserved points, respectively in January and March, **(e)** and **(g)**: PIs for G_D in the test set, respectively in January and March, **(f)** and **(h)**: PIs for G_D for 30 random unobserved points, respectively in January and March. The test confidence (percentage of observed points within the interval) is given for PIs in the test set.

(iii) estimate the roofs geometry characteristics, including the tilt angle, the aspect, and the type of shape of the roofs, and finally (iv) estimate the global radiation over tilted rooftops (combining horizontal radiations, roof geometric and shading effects). The variables are all assessed at the aggregated level of a (200×200) $[m^2]$ pixel grid. In the following sections we explain in details the estimation process for each of the mentioned variables.

7.4.1 Available rooftop area estimation: an updated methodology

The available area estimation is based on a detailed rooftop data already used in the commune study (chapter 6) which includes superstructures (chimneys, HVAC systems etc.) for around 3200 pixels in the canton of Geneva in Switzerland (the SITG dataset, see Table A.2). This data will provide examples to train an RF model to estimate the available area for PV panels over rooftops. The general strategy leading to the computation of the latter area variable is as follows: (i) extract the available area for PV panels over each rooftop in the canton of Geneva using the previously mentioned data and GIS processing, (ii) choose and compute an appropriate label (output value) at pixel level based on the available area for PV installation and the zone of interest in Switzerland, (iii) train an RF model with the previously computed labels and urban features (as presented in section 7.2.1), (iv) use the trained RF model to estimate the labels and finally the available area for PV installation in the remaining pixels in Switzerland.

The steps performed to extract the available area for PV solar systems over the roofs in Geneva canton are the following:

1. Remove the superstructures from the roof with the ArcGIS Erase tool.

2. Form a buffer of 40 cm around the roof, with the ArcGIS Buffer tool, and erase it from the roof polygon. This corresponds to a typical security measure for roof-mounting PV modules, allowing a space between the sides of the panels and the edges of the roof. (The space of 1m considered in previous chapter is unnecessarily large compared to recent security restrictions adopted for most PV panel installations.)
3. Create a rectangle grid of $(1 \times 1,6)$ $[\text{m}^2]$ rectangles originating from (x_o, y_o) the left corner of the longest wall of the roof polygon, using 150 rows and $\lfloor l \rfloor$ columns, where l is the integer part of the length of the longest wall. It geometrically simulates the location of PV modules over the roof surface, as $(1 \times 1,6)$ $[\text{m}^2]$ is the average area of a typical monocrystalline PV module. The number of 150 rows is chosen large enough to guarantee that all surfaces will be covered entirely. When creating the modules, we account for the slope of the roof by using the projected height of the panel $1.6 \cos(\beta)$, where β is the tilt angle of the roof. The directions of the PV grid are defined by the vectors originating in (x_o, y_o) and passing through the following points:

$$\begin{cases} x = x_o + M \cos(\theta) \\ y = y_o - M \sin(\theta) \end{cases} \quad (7.1)$$

where θ is the azimuth of the longest wall (the ArcGIS convention sets the azimuth at 0 at due north and computes it clockwise), and M some large number.

4. Compute the number of rectangles N_{PV} that fits within the boundaries of the roof polygon. It expresses the number of PV modules that, in practice, can be installed over the roof. We use the Join Spatial tool to count the number of fitting modules.
5. Compute the area available for PV installation over the roof as $N_{PV} \times 1.6$ $[\text{m}^2]$ (the area of each module is 1.6 $[\text{m}^2]$). We consider 8 $[\text{m}^2]$ (5 modules) as a minimum available area for PV installation, as it represents the standard minimum installation in practice.

Illustrations for Steps 1, 2 and 3 are shown in Fig. 7.4. Since steps 1 and 2 use simple ArcGIS tools, they are performed for all detailed roof surfaces at the same time. Steps 3, 4 and 5 include multiple computations using various tools and thus were performed through an external python code (connected to ArcGIS with the ArcPy library) which loops over the surfaces. Note that this process could be done with the ModelBuilder interface for ArcGIS, but would suffer from a slower processing time, which is a significant issue in the present case of a very large number of surfaces.

The computed available area is further processed to extract appropriate labels for training in each pixel of the Geneva canton. In order to take advantage of the higher level of details of the building data available in the SON zone (see Fig. 7.1) the corresponding area is treated and predicted differently in SON and OOSG/GEN zones (see Fig. 7.1), which results in two different RF models. The differences in treatment are two-fold: (i) the considered available area labels (training outputs) used to train the RF model are different (hence a different predicted output) and (ii) the total rooftop available area for PV, needed for the final geographical potential, is computed differently in each zone.

In the OOSG zone (see Fig. 7.1), the labels are defined using the average available rooftop area for PV installation over building clusters (available from the VECTOR25 data, see Table A.2). For each pixel j in Geneva canton, the label is computed as follows: (i) we compute the total available area for

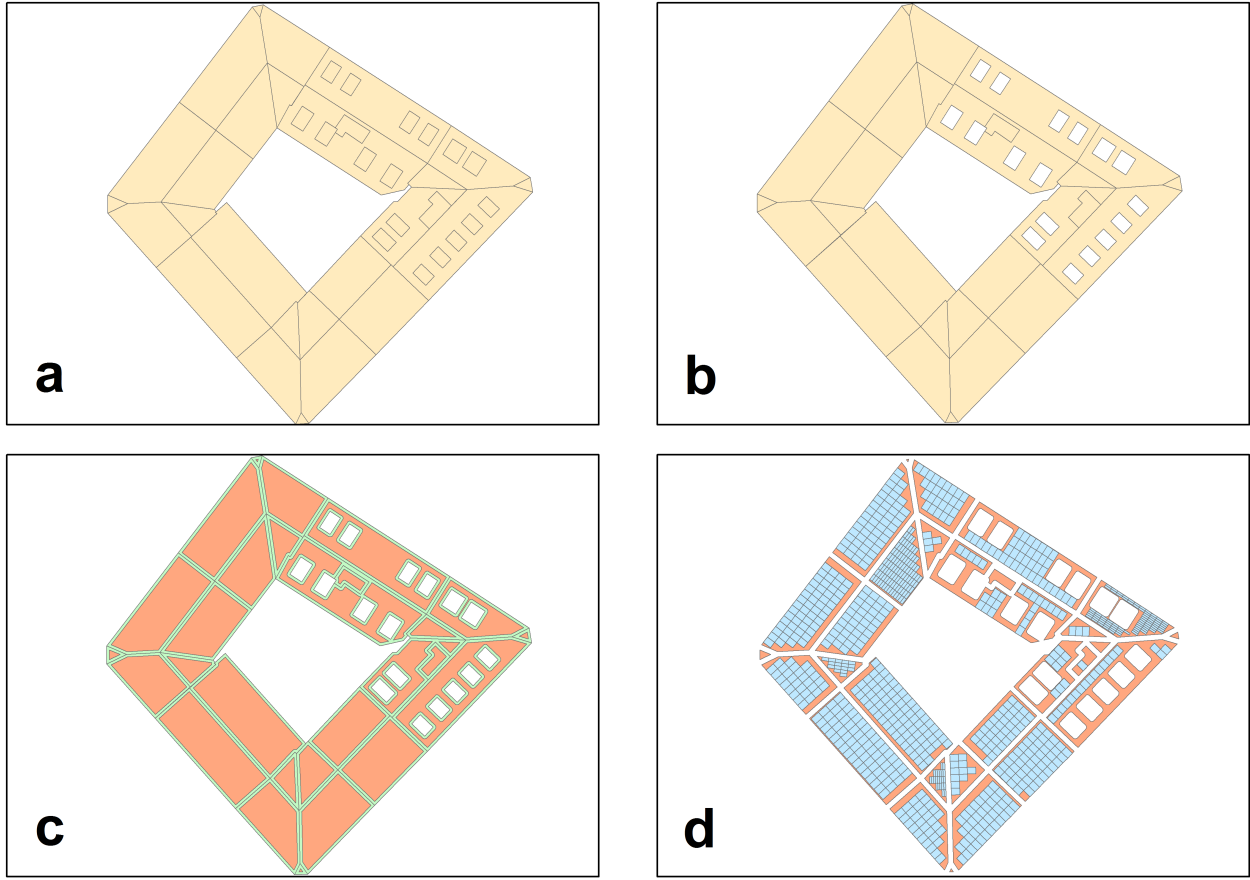


Figure 7.4: Available Area labelling process. It is divided in steps **(a)** Detailed building polygons, containing all roof characteristics, including all superstructures, **(b)** the superstructures are removed from the roofs, **(c)** a buffer of 40 cm is created around the roof surfaces, **(d)** PV panels are virtually installed over the roofs where it is possible, as described in step 3 in section 7.4.1.

PV installation A_R^c over each building cluster c , by adding the available areas from all detailed roofs within the boundaries of the cluster, (ii) we compute the ratio of available area to the ground floor area $C_R^c = A_R^c / A_f^c$ for each cluster, (iii) we finally average the ratios to obtain the mean available area ratio $C_R^{c,j}$ in the current pixel j , which is the considered label in OOSG. Note that we use this ratio instead of the average available area to take advantage of the clusters' ground floor area information. A RF is trained using all urban features presented in section 7.2.2, together with the $C_R^{c,j}$ labels, in the pixels of Geneva Canton: the testing errors are shown in Table 7.3. The trained RF is used to predict $C_R^{c,j}$ in all pixels in the OOSG zone. In each new pixel j , the average available area on a polygon roof is given by $A_R^{c,j} = C_R^{c,j} A_f^{c,j}$ where $A_f^{c,j}$ is the average building cluster ground floor area in the pixel. The total available rooftop area for PV installation in pixel j is then computed as $A_{R,tot}^{c,j} = b_j A_R^{c,j}$, where b_j is the number of building clusters in the pixel.

In the SON zone (see Fig. 7.1), the labels are defined using the average available area for PV panels over the rooftop surfaces. This choice is motivated by the precise building data available in the SON zone, offering information on the geometry of the rooftops including the slope and aspect of each surface composing the rooftops. For each pixel j in Geneva canton, the label is computed as follows: (i) we consider the tilted area A_t^s of each roof surface s , meaning the total area of the roof

Table 7.3: Testing RMSE (E_R) and NRMSE (E_{NR} , in percentage) for Random Forest models trained for available area ratios (C_R^c , C_R^s). We provide also the Out Of Bag (OOB) score to highlight the difficulty to estimate C_R^c : although the E_R and E_{NR} errors for C_R^c are very high, the OOB score is larger than 0, which means our model brings improvement over a simple average estimate. PE_s is the average prediction error estimated on a random sample of 10000 unknown pixels.

Error	C_R^c	C_R^s
E_R (no unit)	0.97	0.11
E_{NR} (in %)	98	27.19
OOB (no unit)	0.17	0.12
PE_s (no unit)	1.01	0.22

surface considering its tilt (slope), instead of the projected area (ground floor area) (ii) we compute the ratio of available area for PV installation to the tilted area $C_R^s = A_R^s/A_t^s$ for each roof surface s and (iii) we compute the average available area ratio $C_R^{s,j}$ for a roof surface in the current pixel j , which is the considered label in SON. A RF model is trained using all urban features presented in section 7.2.2, together with the $C_R^{s,j}$ labels. The testing errors are showed in Table 7.3, along with the ones in OOSG area. The trained RF is used to predict $C_R^{s,j}$ in all pixels in SON zone. For each new roof surface s in SON zone, in each pixel j , the available area for PV installation over the surface is given by $A_R^s = C_R^s A_t^s \approx C_R^{s,j} A_t^s$, where A_t^s is the tilted area of the surface. Note that as the ratio C_R^s is not known for each surface s in SON zone, we approximate it with the average ratio $C_R^{s,j}$ predicted in the pixel j containing the surface. The total available area in a pixel can then be recovered by adding all available areas from the roof surfaces. A map showing the total rooftop available area for PV throughout the whole Switzerland is shown in Fig. 7.5.

We compute prediction intervals for the estimated available area ratios in both OOSG and SON zones (C_R^c and C_R^s), which are shown in Fig. 7.5. While the test confidence is satisfactory in both cases, the large width of the intervals for a few points concerning the estimation of C_R^c shows the difficulty in predicting the available area over the building clusters in the OOSG zone. Average prediction errors are also computed over a random sample of pixels and shown in Table 7.3 and further show the challenge in estimating the C_R^c , but the relative simplicity in estimating C_R^s , given its reasonable prediction uncertainty.

7.4.2 Shading factors estimation

The shading losses are added to the geographical potential based on (2×2) [m²] LiDAR elevation data considering buildings and trees called a Digital Orthophoto Map (DOM), available for the entire Switzerland. The methodology applied to compute the shading effects over buildings is in principle the one we presented in the commune study (chapter 6, section 6.4.2), with two updates: (i) the shading variables are aggregated in each (200×200) [m²] pixel in Switzerland instead of being aggregated in each commune and (ii) RFs are used instead of SVR, and different features are used to train the model.

Based on the (2×2) [m²] DOM, the hillshade function from ArcGIS is used to compute the two shading variables S_{Sh} and S_{hill} used in the commune study (presented in chapter 6, section 6.4.2) over buildings of the Geneva canton, similarly to what is presented in the commune study. After having computed the hourly hillshade values, for each month, using the DOM and the sun

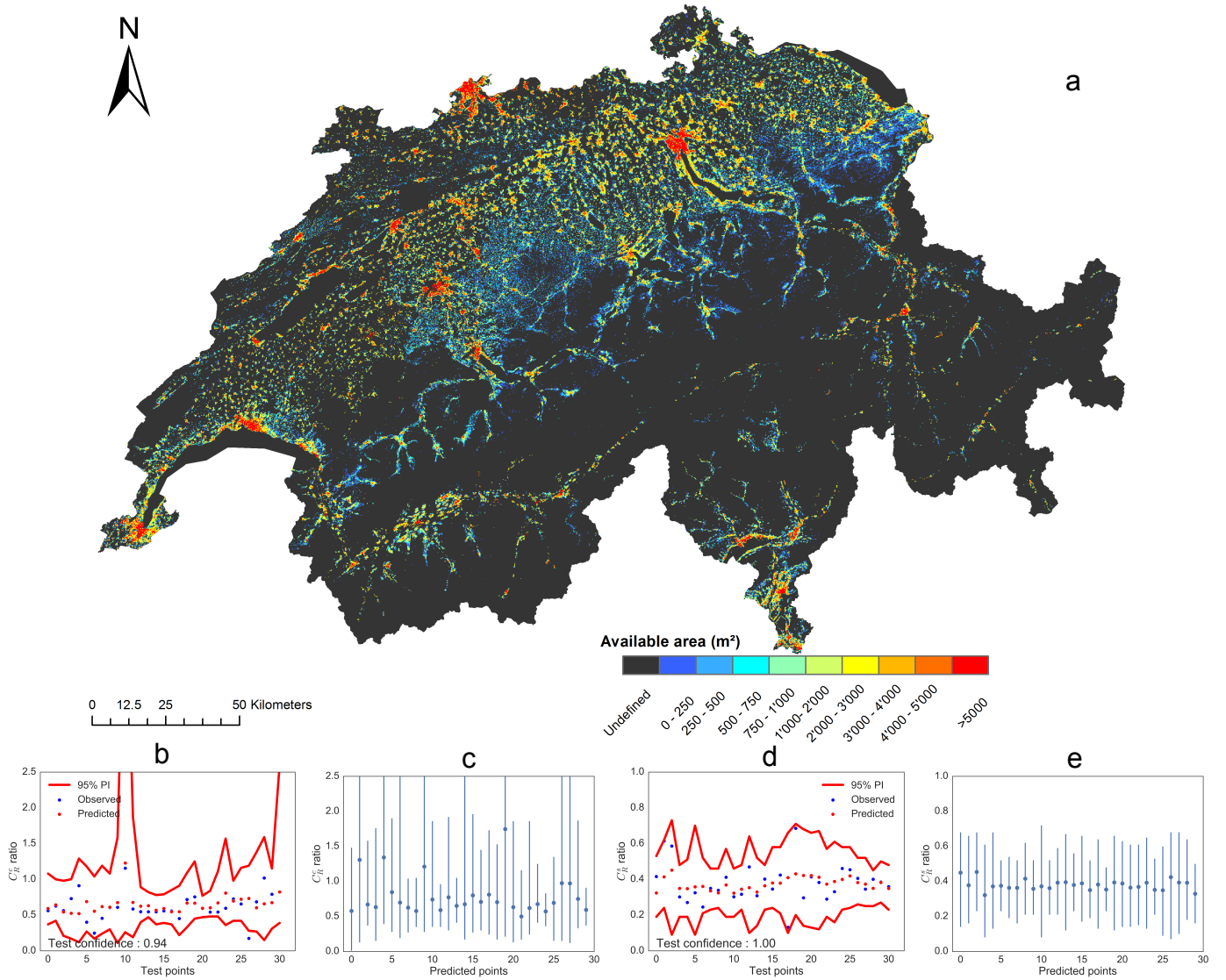


Figure 7.5: (a) Available roof area map. It provides the total available area over rooftops for PV installation in each considered pixel, combining the results obtained in both SON and OOSG regions. **(b)** and **(d)**: prediction intervals from Quantile Random Forests for the available area ratios, respectively C_R^C and C_R^S , in the test set, **(c)** and **(e)**: prediction intervals from Quantile Random Forests for the available area ratios, respectively C_R^C and C_R^S , for 30 random unobserved points. The test confidence (percentage of observed points within the interval) is given for intervals in the test set.

altitude and azimuth values corresponding to the representative day of the month [109] from 8am to 6pm (6pm is included), we clip hillshade raster cells over the building clusters and separate fully shaded and partially shaded cells, using the processing steps detailed in the commune study. The two variables are then aggregated in each pixel as follows:

- S_{Sh} in each pixel j , and for each month, is given by:

$$S_{Sh}^j = \frac{N_{sh,8\text{ am}}^j + N_{sh,9\text{ am}}^j + \dots + N_{sh,6\text{ pm}}^j}{11N_{cells}^j} \quad (7.2)$$

Table 7.4: Left side of the table: Testing RMSE (E_R) and NRMSE (E_{NR} , in percentage) for Random Forest models trained for shading variables (S_{hill} , S_{Sh}); Right side of the table: Monthly Prediction Errors, computed using Quantile Regression Forests, averaged over a sample of 10000 unobserved pixels. PE_s is the average of the lower and upper error (half width of PIs)

Month	Test errors				Prediction errors	
	S_{hill}		S_{Sh}		S_{hill}	S_{Sh}
	E_R	E_{NR}	E_R	E_{NR}	PE_s	
	[-]	[%]	[-]	[%]	[-]	[-]
Jan.	24.62	23.01	0.14	23.71	42.36	0.24
Feb.	20.28	17.40	0.15	32.97	35.41	0.25
Mar.	15.58	11.97	0.14	44.82	26.96	0.21
Apr.	10.95	7.47	0.11	56.53	21.09	0.17
May.	9.49	5.97	0.09	66.95	18.44	0.15
Jun.	9.57	5.84	0.08	71.22	17.98	0.12
Jul.	9.48	5.88	0.09	68.53	17.82	0.13
Aug.	9.96	6.56	0.10	59.98	19.87	0.16
Sep.	13.44	9.82	0.13	50.34	24.44	0.20
Oct.	18.79	15.54	0.15	37.77	34.76	0.22
Nov.	23.16	21.18	0.15	26.84	40.46	0.24
Dec.	26.28	25.34	0.15	22.97	42.71	0.24

where $N_{sh,8\text{ am}}^j$ is the number of fully shaded cells over the rooftops in pixel j at 8am, and N_{cells}^j is the total number of rooftop cells in pixel j .

- S_{hill}^j is computed for each month by calculating the average hillshade value of the non-fully shaded DOM cells over all rooftops in pixel j through the 11 considered hours in the representative day of the month.

RF models for both S_{Sh} and S_{hill} are trained using Geneva canton computed values as labels and the urban features presented in 7.2.2 as features. These models are then applied to the remaining part of Switzerland to estimate, for each month, the two shading factors in each pixel populated by buildings. The resulting testing errors are shown in Table 7.4.

Prediction intervals (with 95% confidence) are computed for both S_{Sh} and S_{hill} , and shown for January and March in Fig. 7.6. One can observe from the latter intervals that the variability of the shading variables remain similar across the points (pixels) for a particular month, as they show flat envelopes; also, the estimators seem to have a low variability in their prediction. This is partly due to the relatively few changes in the training inputs and outputs, making it troublesome for the RF models to differentiate between two new points with similar input values. Table 7.4 also show average monthly prediction errors computed over a random sample of 100000 pixels. Similarly to the prediction of the solar horizontal components, prediction errors confirm a pattern already observable from the test errors: in the case of shading factors, the uncertainty is higher during wintertime, simply because the average shading impact is higher during that period. Values are therefore generally higher, but may still vary unexpectedly, resulting in higher errors than summertime.

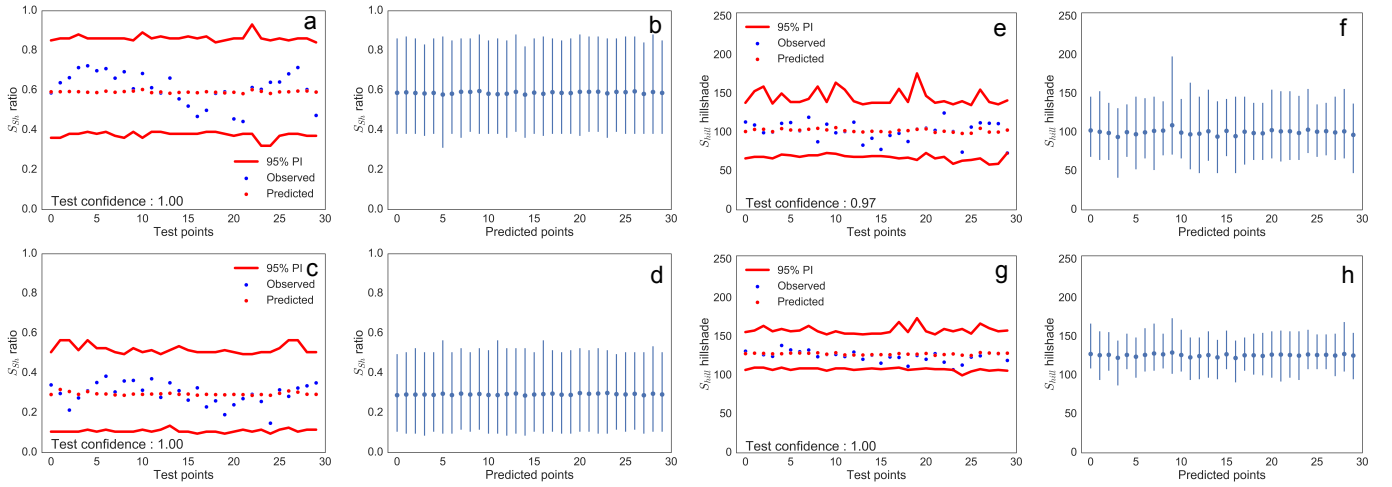


Figure 7.6: Prediction Intervals (PIs) from Quantile Random Forests for the two estimated shading variables (S_{sh} and S_{hill}). **(a)** and **(c)**: PIs for S_{sh} in the test set, respectively in January and March, **(b)** and **(d)**: PIs for S_{sh} for 30 random unobserved points, respectively in January and March, **(e)** and **(g)**: PIs for S_{hill} in the test set, respectively in January and March, **(f)** and **(h)**: PIs for S_{hill} for 30 random unobserved points, respectively in January and March. The test confidence (percentage of observed points within the interval) is given for PIs in the test set.

7.4.3 Rooftop geometrical properties estimation (in OOSG and GEN zones)

In order to finalize the geographical potential estimation, geometrical characteristics of the rooftops, including the roof type, slope and aspect of the surfaces forming the roof, are estimated for all building clusters in Switzerland, based on (2×2) [m²] LiDAR data. As the direct estimation of the geometrical characteristics of rooftops' surfaces would result in poor accuracy, the regression problem is transformed into a classification problem. As a result of this transformation, we do not estimate the exact value of each geometrical characteristic, but rather the class (among multiple predefined classes) to which it belongs. Random Forests are used in order to classify the rooftops of the building clusters into 9 aspect classes, 5 slope classes, and 6 roof type classes. The probabilities for each type, slope and aspect class are computed in each pixel, and will serve as weights in the final computation of the geographical potential, as explained later in the text. Note that this estimation is used in OOSG and GEN zones, and not in SON zone (see Fig. 7.1), where precise slope and aspect data are already available. As the three estimations include multiple steps, we explain their various details within further following subsections.

Geometrical statistics extraction

ModelBuilder (an ArcGIS tool) and Python codes were used to automate the process of extraction of useful statistics from the DOM elevation data, which will serve as features for the machine learning models leading to the classification tasks. We first split the entire DOM into medium-sized parts (about 25 parts for the entire remaining Switzerland territory once the communes covered by Sonnendach data are discarded), to allow for reasonable processing time on each of them. Then, we built a model using Model Builder which extract aspect and slope data over the building clusters (VEC25 data). These data is used as features for the classifications. The convention for aspect values is

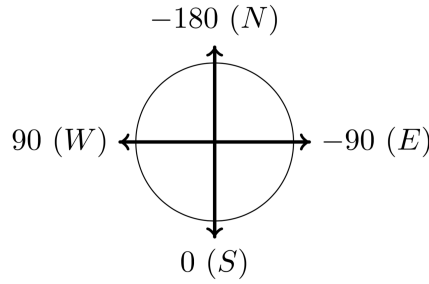


Figure 7.7: Convention used for aspect calculation.

shown in Figure 7.7. A python script was written to run the model steps autonomously outside of ArcGIS to speed up the computational time.

The built geometrical ArcGIS model, performs, for each of the 25 parts of Switzerland, the following tasks:

- (i) Upsample it to a resolution of $(0,5 \times 0,5)$ $[m^2]$ to gain in precision;
- (ii) Compute aspect and slope raster from the DOM raster using the Spatial Analyst toolbox;
- (iii) Perform a Re-classification of raster values (from Spatial Analyst toolbox) by bins:
 - For slope, with 9 bins: $[0^\circ, 10^\circ]$, $[10^\circ, 20^\circ]$, $[20^\circ, 30^\circ]$, $[30^\circ, 40^\circ]$, $[40^\circ, 50^\circ]$, $[50^\circ, 60^\circ]$, $[60^\circ, 70^\circ]$, $[70^\circ, 80^\circ]$, $[80^\circ, 90^\circ]$.
 - For aspect, with 2 different bins configuration: (a) 5 bins, including flat, $[-135^\circ, 135^\circ]$ (North), $[-135^\circ, -45^\circ]$ (East), $[-45^\circ, 45^\circ]$ (South), $[45^\circ, 135^\circ]$ (West); (b) 19 more precise 20° bins, including flat, $[-170^\circ, 170^\circ]$, $[-170^\circ, -150^\circ]$, $[-150^\circ, -130^\circ]$, $[-130^\circ, -110^\circ]$, ..., $[130^\circ, 150^\circ]$, $[150^\circ, 170^\circ]$. These two different set of bins are used separately for two tasks. The first is used to build features for roof classification, as it expresses the main changes in aspect across the roof and to avoid dilution of the feature information. The second more complete configuration is used to build labels for the aspect estimation.
- (iv) Compute frequencies of raster cells for each slope and aspect bin over each building cluster to obtain the frequencies of cells with an aspect in each of the 5 aspect bins, the frequencies of cells with an aspect in each of the 19 aspect bins, and the frequencies of cells with a slope in each of the 9 slope bins;
- (v) Compute statistics for each of the three histograms frequency data extracted in (iv) (mean bin, mode bin etc.);
- (vi) Export these frequencies and statistics in csv format, ready to be used as features for further classifications. Illustrations of reclassified slope and aspect rasters can be seen in Figure 7.8 for two different roof types.

The extracted raster statistics will be used to form various features that will be used for the roof type classification first, and then for the aspect and slope classification, which is more straightforward than the roof type one.

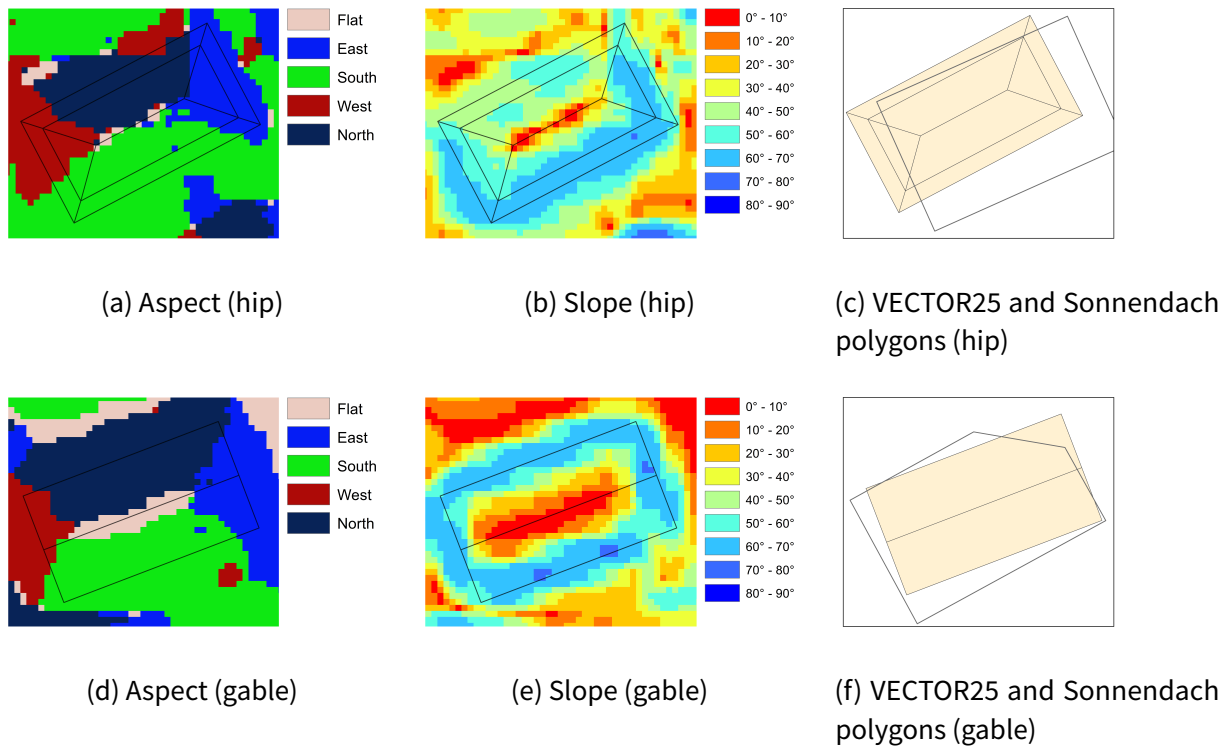


Figure 7.8: Aspect and Slope reclassified ($0,5 \times 0,5$) [m^2] rasters, along with the building polygons from VECTOR25 (grey thick line) and Sonnendach data (black thin line) for a building with a hipped (a,b,c) and a gabled roof (d,e,f). One can observe the significant delay of position between the two polygons. Also, the different aspect and slope patterns between the two types are clearly shown, specially regarding the amount of roof raster cells showing a flat surface, significantly larger in the gable case.

Roof type classification

The first step in the roof type classification is to choose the different classes that cover all possible types of roof. There are many possible roof types considered in the literature [242, 282]. For our purpose, we accounted for the differences both in roof shape and the general footprint geometry of the building. For example, one building can have a pyramidal roof, but with a rectangular or an L-shaped footprint, which will result in a very different aspect distribution. Roof shapes include mainly flat, gable, hipped, pyramidal, shed, mansard, and gambrel. Footprint geometries of gatherings of buildings can considerably vary. However usual forms include: rectangular, L-shaped, T-shaped, U-shaped, O-shaped, Triangle-shaped. Some of these shapes and geometries can be difficult to differentiate from one another due to their similarities and the relative lack of precision imposed by the large scale of our study. Thus, some choices were made to decrease the complexity of the task, thus increasing the performance of our classifier. It was decided to gather some of them in the same class (by similarity) in order to reduce the total number of classes. The classes are as follows: Gable and Shed, Hip and Pyramidal, L and T-shaped, O and U-shaped, and Complex. The complex class includes the Triangle-shaped buildings, and all roofs that do not fit in existing classes. Gambrel and Mansard were discarded because of their complex structure and the lack of examples found in the training process, as discussed later in the study.

Since the footprint geometry and roof shape do not depend on the same features, we decided to perform two layers of classification, separating three main polygon classes in the first layer, and

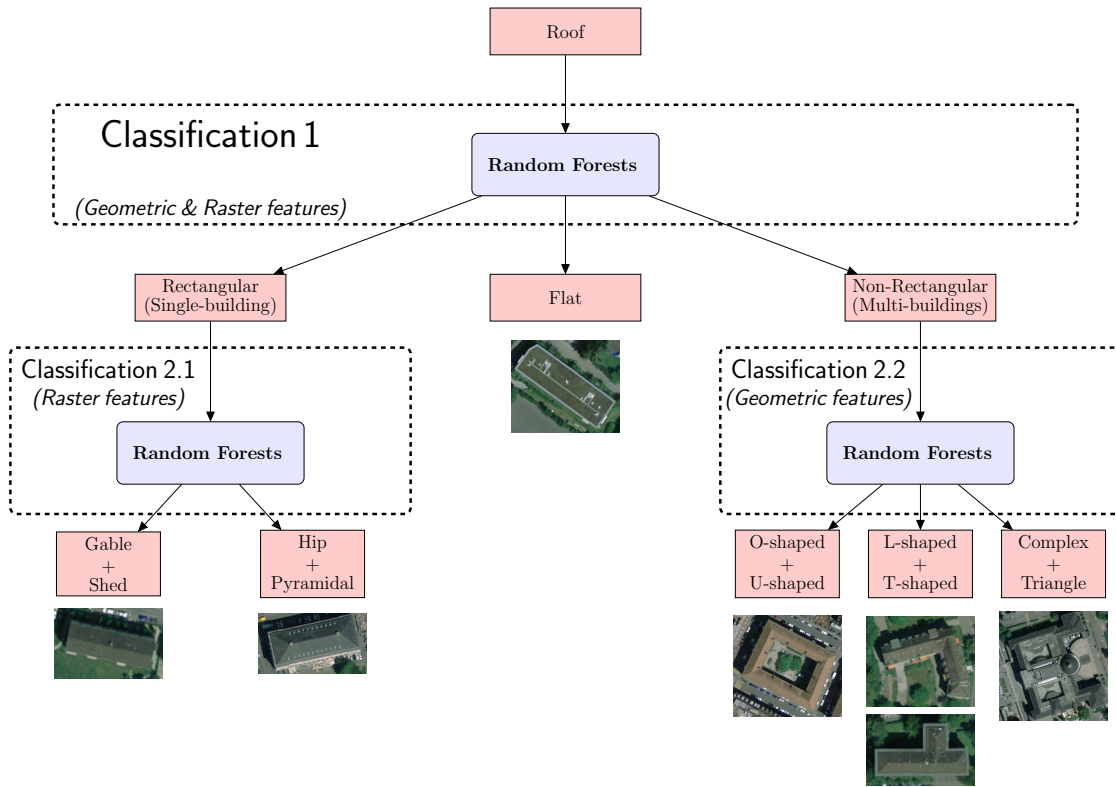


Figure 7.9: Roof classification scheme.

treating with shapes and geometry independently in the second layer. More Specifically, we perform: (i) Classification 1 to differentiate between Flat, Rectangle, and Non-Rectangle polygons and (ii) Classification 2.1 on Rectangle polygons, to differentiate between Gable, and Hip; Classification 2.2 on Non-Rectangle polygons, to differentiate between O-shaped, L-T-shaped, and Complex. The whole classification process is illustrated in Figure 7.9. Note that in performing this classification, we did not differentiate between the different roof shapes when classifying the non-rectangular polygons. We show further in the study that the simple binary classification between Gable and Hip for rectangle polygons is a very hard task at a large scale, resulting in a quite poor accuracy. As it would lead to an even poorer accuracy for multi-buildings polygons, we focus on the geometry of the footprint and consider they are gable shaped.

The chosen roof classes are then considered to build the labeled set of examples. The labeled set was obtained by manually detecting different classes of buildings from examples using high resolution satellite images and the Sonnendach data. The VECTOR25 building clusters polygons were layered over satellite images from Swisstopo (Swissimages 25cm) so that roof classes could be attributed to each polygon by visual observation. A total number of 1252 building clusters were manually labelled, which approximatively corresponds to 1% of the total number of building clusters all over Switzerland. We considered 3 regions containing both rural/suburban parts and dense urban parts, including contemporary and old city center buildings: Baden region, Luzern region, and Winthertur region. A number of 268, 556, 232, 32, 88, and 76 building clusters were labeled respectively for Flat, Gable, Hip, O-shaped, L-shaped, and Complex classes. The training

set was composed by 75% of the labels and the remaining 25% of the labels was used for the test set, leading to the accuracy computation for the classifiers.

Going in pairs with the labels, the features used for each polygon in the classification tasks are of two types: (i) the geometric features will serve as simple features to differentiate between different geometric footprint shapes and (ii) the raster based features will be used to differentiate between different roof shapes (Flat, Gable, Hip) and will be extracted from the slope and aspect raster data.

The geometric features characterize the shape of the building footprints. They must be simple enough to be computable directly from the polygons, and aim at differentiating the different geometric footprint shapes. A natural feature is the number of vertices. Yet, it is clearly not sufficient to characterize the footprint shapes. To add information about the compactness of the polygon, the iso-perimetric quotient (isoQ) of the polygon is used as an extra geometric feature. This coefficient is defined as the ratio of the polygon area and the area of a circle with the same perimeter. A straight-forward calculation leads to the isoQ expression:

$$\text{isoQ} = \frac{4\pi A}{P^2} \quad (7.3)$$

where A and P are respectively the area and the perimeter of the polygon.

The raster features characterize the roof shape based on the elevation data. Since a roof shape is intuitively described by the arrangement of the roof different directions and tilts, the raster features used for training will be combination of various slope and aspect statistics extracted in the DSM raster processing. These raster features include:

- **Statistics from the 5 bins aspect raster data:** mean, standard deviation, variety, majority, minority and median.
- **Statistics from the 9 bins slope raster data:** mean, standard deviation, variety, majority, minority and median.
- **Frequencies and percentages from the 5 bins aspect raster data:** number of cells with aspect in each aspect bin, and proportion of cells with aspect in each aspect bin.
- **Frequencies and percentages from the 9 bins slope raster data:** number of cells with slope in each slope bin, and proportion of cells with slope in each slope bin.
- **Ratios of flat cells frequencies and other directions frequencies:** East/Flat ratio, South/Flat ratio, West/Flat ratio and North/Flat ratio.
- **Ratios of flat cells frequencies and other slope bins frequencies:** $[10^\circ, 20^\circ] / \text{Flat}$ ratio, $[20^\circ, 30^\circ] / \text{Flat}$ ratio, etc.
- **Ratios of slope bins with one another:** $[10^\circ, 20^\circ] / [20^\circ, 30^\circ]$ ratio, $[10^\circ, 20^\circ] / [30^\circ, 40^\circ]$ ratio, etc.
- **Boolean variables to identify symmetry in roofs:** EWsym, NSsym, BothSym respectively indicates an east-west symmetry, north-south symmetry, and a symmetry in both directions. They are simply computed: if the number of east cells is equal to the number of west cells, plus or minus 100, EWsym = 1, otherwise EWsym = 0. The computation is similar for NSsym. BothSym is given by EWsym \times NSsym.

Table 7.5: Roof shape estimation confusion matrix for Classification 1.

OOB = 0.85	Flat	Rect	Non-Rect	Acc.
Flat	45	13	9	70%
Rect	5	185	7	94%
Non-Rect	4	4	41	84%

Table 7.6: Roof shape estimation confusion matrix for Classification 2.1.

OOB = 0.72	Gable	Hip	Acc.
Gable	118	21	85%
Hipped	40	18	31%

Table 7.7: Roof shape estimation confusion matrix for Classification 2.2.

OOB = 0.65	O-sh.	L-Sh.	Complex	Acc.
O-sh.	8	0	0	100%
L-Sh.	3	16	3	73%
Complex	2	0	17	90%

The first classification (Classification 1) uses both geometric and raster based features to differentiate between flat roofs, non-flat rectangular polygons and non-flat non-rectangular polygons. The features of Classification 1 include: number of vertices, isoQ, percentages from 5 bins aspect data, and ratios of flat cells $[20^\circ, 30^\circ] / \text{Flat}$, $[30^\circ, 40^\circ] / \text{Flat}$, $[40^\circ, 50^\circ] / \text{Flat}$ slope bins, for a total of 8 features. A choice of $B = 500$ trees is found to be sufficient to obtain optimal results, and m is chosen by 6-fold cross validation. The same number of trees and strategy for m tuning is used in the other roof classifications. The performance of the trained RF classifier is summarized in Table 7.5, in the form of a confusion matrix. This matrix exposes, for each class (each row), the number of polygons well classified, and the number of polygons wrongly classified in other classes. For example, the first row of the matrix shows that, out of the $45 + 13 + 9 = 67$ flat roofs considered in the validation set, 45 were well classified as flat roofs, 13 were wrongly classified as rectangular non-flat polygons, and 9 were wrongly classified as non-flat non-rectangular polygons. The last column gives the accuracy of the classifier for each class, meaning the percentage of well classified polygons. The Out-Of-Bag (OOB) score is also provided in the table.

The second classification (Classification 2.1) uses purely raster based features to differentiate between gable and hipped roofs. The features of Classification 2.1 include aspect and slope statistics, frequencies and percentages respectively for aspect and slope, and 13 different slope ratios. The performance of the trained RF classifier is summarized in Table 7.6.

Finally, the third classification (Classification 2.2) uses purely geometric features to differentiate between O-U-shaped, L-T-shaped and complex buildings. The features are simply the number of vertices and the isoQ. The performance of the trained RF classifier is summarized in Table 7.7. It is straight forward to obtain the final classification accuracy for each roof type, by multiplying the accuracies in each classification layer: $AE_{\text{final}} = AE_{\text{class1}} \times AE_{\text{class2}}$. It can be observed in Table 7.8. After the classifiers are built with the labeled data, they are used on the remaining polygons to determine their shape.

The results of the classification vary greatly depending on the class. While the model identifies Gabled roofs quite well, it is very poor in classifying hipped roofs. This is mainly caused by the relatively low resolution of the LiDAR data and the lack of precision of the VECTOR25 polygons in shape and more particularly in the location. Thus, this prevents the model from detecting changes in aspect values

Table 7.8: Final accuracy of the overall classifier to detect each roof class.

Flat	Gable	Hip	O-shaped	L-Shaped	Complex	Mean
70%	80%	30%	84%	61%	76%	67%

in small areas, which is the key to detect hipped roofs, characterized by the two lateral small “hips” (Figure 7.8). Besides the quality of the data at hand, the model performance is heavily depending on the size of the training data and the number of labels for each class. In case of the hipped roofs, a higher number of labels is desirable, and will be used in the future, to distinguish them from gabled roofs.

Aspect and slope estimation

The aspect and slope angle of the roofs are of course very significant when it comes to estimating the solar energy available over the roofs. As we are particularly interested in this information over the VECTOR25 building clusters in OOSG and GEN zones, aggregated values of aspect and slope are desirable for each building. More specifically, we aim at one aspect value and one slope value for each building cluster polygon. Consequently, the modes (most frequent value) of the aspect and slope distributions were considered. These two quantities are real numbers, and naturally call for a regression estimation. Nevertheless, they revealed themselves to be quite delicate to predict, solely based on our raster features. As a consequence, we decided to relax the problem into a classification task by creating bins that act as classes, for both aspect and slope. The resulting classifications for each polygon consist of: (i) classification of the main aspect, meaning the center of the most frequent aspect bin represented, and (ii) classification of the main slope, meaning the center of the most frequent slope bin represented. In order to capture the different aspects of the various roof sides in each building, we consider as a prior information the predicted roof types from the previous section 7.4.3. We use the symmetry of each roof type to virtually distribute the different roof aspects from the main aspect estimation. Note that this symmetry allows us to gather aspect bins that are in the same direction, (meaning delayed by 180° , as for example $[-50^\circ, -30^\circ] \cup [130^\circ, 150^\circ]$), which divides by two the number of aspect classes. Random Forests were used for both slope and aspect classifications. A table summarizing how the roof aspects are distributed from the main aspect for each roof type is shown in Table 7.11.

Classes were created as 20° bins for aspect estimation, and 10° bins for slope estimation. More specifically, the following bins were used:

- 5 bins for slope: $[10^\circ, 20^\circ]$, $[20^\circ, 30^\circ]$, $[30^\circ, 40^\circ]$, $[40^\circ, 50^\circ]$, $[50^\circ, 60^\circ]$, corresponding respectively to classes C_{s1} , C_{s2} , C_{s3} , C_{s4} , C_{s5} . Slope values beyond 60° are very rare and thus not considered.
- 9 bins for aspect:
 1. $[-10^\circ, 10^\circ] \cup [-170^\circ, 170^\circ]$
 2. $[-170^\circ, -150^\circ] \cup [10^\circ, 30^\circ]$
 3. $[-150^\circ, -130^\circ] \cup [30^\circ, 50^\circ]$
 4. $[-130^\circ, -110^\circ] \cup [50^\circ, 70^\circ]$
 5. $[-110^\circ, -90^\circ] \cup [70^\circ, 90^\circ]$

Table 7.9: Aspect estimation confusion matrix.

OOB = 0.63	C_{a1}	C_{a2}	C_{a3}	C_{a4}	C_{a5}	C_{a6}	C_{a7}	C_{a8}	C_{a9}	Acc.
C_{a1}	128	19	1	1	18	5	1	0	12	69%
C_{a2}	14	267	22	3	4	35	37	1	1	70%
C_{a3}	1	21	186	11	2	0	24	18	3	70%
C_{a4}	1	2	5	127	7	0	5	18	14	71%
C_{a5}	17	6	4	5	124	6	1	0	12	70%
C_{a6}	28	37	2	1	11	83	7	1	2	48%
C_{a7}	1	37	31	1	0	6	174	5	0	68%
C_{a8}	2	2	15	20	4	0	5	81	11	58%
C_{a9}	10	4	1	17	21	0	2	1	97	63%

6. $[-90^\circ, -70^\circ] \cup [90^\circ, 110^\circ]$
7. $[-70^\circ, -50^\circ] \cup [110^\circ, 130^\circ]$
8. $[-50^\circ, -30^\circ] \cup [130^\circ, 150^\circ]$
9. $[-30^\circ, -10^\circ] \cup [150^\circ, 170^\circ]$

corresponding respectively to classes C_{a1} , C_{a2} , C_{a3} , C_{a4} , C_{a5} , C_{a6} , C_{a7} , C_{a8} , C_{a9} .

The labeled set was extracted from the Sonnendach data in SON zone, containing aspect and slope values for each surface of all building rooftops in the covered area. The main aspect and slope were computed by extracting the most frequent aspect and slope value classes across the surfaces of each polygon, thus forming the label for each polygon. The entire Sonnendach data was considered, gathering 11449 polygons. The training and test set were built respectively with 75% and 25% of the labeled set. In both aspect and slope classifications, we use the respective frequencies and ratios to serve as features. For the aspect classification, the reclassified aspect values from the 20° bins were used to form the features of the input data. More specifically, the features include 9 aspect percentages and 20 ratios of aspect frequencies. For the slope classification, similarly, the reclassified slope values from the 10° bins were used to form the features of the input data. More specifically, the features include 7 slope percentages and 12 ratios of slope frequencies.

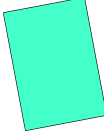
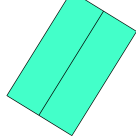
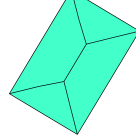
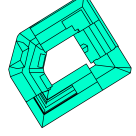
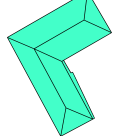
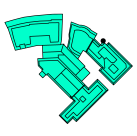
The strategy used for parameter tuning is similar than in section 7.4.3. A number of 500 trees is used and m is chosen by 6-fold cross validation in both classifications. A summary of the classifiers' performances is given in the form of the accuracy matrices described in Tables 7.9 and 7.10. The two classifiers can now be used on the unlabeled polygons to determine their main aspect and slope, and the distribution of aspect and slope for the remaining sides of each roof is assumed to be as shown in Table 7.11.

As in the roof type classification step, the performance of the model changes significantly depending on the class of aspect or slope. While the aspect estimation offer a reasonable accuracy of around 67% for almost all aspect classes, it seems still very challenging to estimate slope at a large scale without a very high resolution data. Indeed, if the most frequent slope class in Switzerland ($[20^\circ, 30^\circ]$) is relatively well identified, the other classes show a very low accuracy. Note that the use of Random Forests classification is not the first natural idea for aspect and slope estimation. One can simply

Table 7.10: Slope estimation confusion matrix.

OOB = 0.50	[10,20]	[20,30]	[30,40]	[40,50]	[50,60]	Acc.
[10,20]	83	126	24	11	0	34%
[20,30]	55	414	123	24	0	67%
[30,40]	32	257	304	86	0	45%
[40,50]	30	113	177	115	5	26%
[50,60]	3	18	20	21	1	2%

Table 7.11: Roof characteristics considered for each roof type. β_l and γ_m are the center value of respectively the main slope (tilt) and the main aspect (direction) class predicted for the roof of interest.

Roof Type	Flat	Gable	Hip	O-shaped	L-Shaped	Complex
						
Num of sides	1	2	4	8	6	8
Num of directions	1	2	4	8	4	8
Roof aspects	γ_m	γ_m $\gamma_m + 180$	γ_m $\gamma_m + 180$ $\gamma_m + 90$ $\gamma_m - 90$	γ_m $\gamma_m + 180$ $\gamma_m + 90$ $\gamma_m - 90$	γ_m $\gamma_m + 180$ $\gamma_m + 90$ $\gamma_m - 90$	γ_m $\gamma_m + 180$ $\gamma_m + 90$ $\gamma_m - 90$
Roof sides Slope	10°	β_l	β_l	β_l	β_l	β_l

compute the number of pixels in each aspect and slope bin, and assume that the bin with the highest frequency of cells is the main one. The center of the bin is then the estimated aspect or slope value. Unfortunately, the lack of precision of the raster data resulted in poor results while using this simpler approach. Random forests offered significantly higher performance.

Application of the classifiers in OOSG and GEN zones

We use the roof shape classifier to estimate the roof type of all building clusters in OOSG and GEN zones. We further compute roof type probabilities in each pixel with a natural frequentist approach, meaning that we consider that the probability for a roof type in a pixel is the proportion of roofs of this type in the pixel. More formally, the probability for a building cluster to be characterized by a roof type c_n (meaning in the n^{th} roof type class, with $n = 1, \dots, 6$), given that it is located in pixel j is given by r_n^j , expressed as:

$$\mathbb{P}(\text{building cluster roof type} = c_n \mid \text{pixel } j) = \frac{\#(bc \in c_n)}{\#(bc \in j)} = r_n^j \quad (7.4)$$

where $\mathbb{P}(x \mid y)$ is the conditional probability of x given y , $\#(bc \in c_n)$ is the number of building clusters that have been estimated to have a roof type of class c_n in pixel j and $\#(bc \in j)$ is the total number of building clusters in pixel j .

We apply the two aspect and slope classifiers to the building clusters in the OOSG and GEN zones in order to extract the main slope and main aspect classes for each cluster. The slope and aspect values assigned to each class is the center value of the bin (e.g. 25° for slope bin $[20^\circ, 30^\circ]$). We also compute the aspect and slope probabilities in each pixel j with a frequentist approach. Therefore, the probabilities for a building cluster to be characterized by a main slope of β_l (meaning in the l^{th} slope class, with $l = 1, \dots, 5$) and main aspect of γ_m (meaning in the m^{th} slope class, with $m = 1, \dots, 9$), given that it is located in pixel j , are respectively given by p_l^j and q_m^j , expressed as:

$$\mathbb{P}(\text{building cluster main slope} = \beta_l \mid \text{pixel } j) = \frac{\#(bc \in \beta_l)}{\#(bc \in j)} = p_l^j \quad (7.5)$$

$$\mathbb{P}(\text{building cluster main aspect} = \gamma_m \mid \text{pixel } j) = \frac{\#(bc \in \gamma_m)}{\#(bc \in j)} = q_m^j \quad (7.6)$$

where $\mathbb{P}(x \mid y)$ is the conditional probability of x given y , $\#(bc \in \gamma_m)$ is the number of building clusters that have been estimated to have a main aspect value of γ_m in pixel j , $\#(bc \in \beta_l)$ is the number of building clusters that have been estimated to have a main slope value of β_l in pixel j and $\#(bc \in j)$ is the total number of building clusters in pixel j .

7.4.4 Global tilted radiation estimation

The global tilted radiation over rooftops is again computed using the tilted radiation model defined in chapter 3, in each pixel and with added shading impacts. These impacts consist in S_{hill} , which allows to take into account the fraction of light reaching the rooftops in the computation of the direct radiation (G_{Bt}), and the Sky View Factor (SVF), which accounts for the shading impact over the diffuse radiation (G_{Dt}). We still assume a constant SVF value of 0.9 over rooftops in Switzerland [265], as in the commune study, and the modified tilted radiation is

$$G_t = \frac{S_{\text{hill}}}{255} (G_h - G_D) R_b + 0.9 G_D R_d + G_h R_r \quad (7.7)$$

where G_h and G_D are horizontal global and diffuse radiation, and R_b , R_d , and R_r are respectively the direct, diffuse, and reflected coefficients as defined in section 3. G_h and G_D are estimated for each pixel as explained in section 7.3. We normalize S_{hill} by 255 to obtain a fraction of light, defined between 0 and 1 (hillshade values vary from 0, when the cell is fully enlightened, to 255, when the cell is fully in shade). Since R_b , R_d , and R_r are functions of the slope and azimuth, which are computed differently in zone SON from OOSG and GEN zones (see Fig. 7.1), the computation of the tilted radiation is also treated differently depending on the zone.

In SON zone, the global tilted radiation G_t is determined over each roof surface (a building rooftop contains multiple surfaces) in all pixels. It is allowed by the knowledge of the slope β_s and azimuth γ_s of each roof surface s in the zone, along with the estimated horizontal solar radiation. We use Eq. 7.7 to compute $G_t(\beta_s, \gamma_s)$ by the way of Reindl and Klein models for the diffuse and direct coefficients described in section 3.

In the OOSG and GEN zones, the global tilted radiation G_t cannot be computed for specific roof surfaces and is rather computed over the virtual surfaces of each possible roof configuration (roof type = c_n , main slope = β_l , main aspect = γ_m) for a building cluster, as considered in section 7.4.3, with 6 possible roof types, 9 possible aspect values, and 5 possible slope values; it results in

270 possible roof configurations. In each of the 270 configurations the tilted radiation is computed for an aspect of γ_m , $\gamma_m + 90$, $\gamma_m - 90$, and $\gamma_m + 180$, to capture the four main directions (aspects) of the surfaces forming each rooftop. The probabilities to belong to a roof type c_n and to be characterized by a main slope β_l and main aspect γ_m , are calculated respectively as r_n^j , p_l^j , and q_m^j , in each pixel j as explained in 7.4.3. We consider slope, aspect and roof type to be independent variables, which leads to a joint probability of $r_n^j p_l^j q_m^j$ for each configuration in pixel j . The statistical independence between the three variables is validated through the computation of conditional and marginal probabilities, as shown in Appendix C, section C.1.

The probabilities will be used to associate a weight to each configuration using the corresponding tilted radiation $G_t(\beta_l, \gamma_m)$ in each pixel. The specific slope and aspect values of the surfaces for each configuration as well as the repartition on the available area over these surfaces are two crucial issues which will be explained in the following section.

7.4.5 Final geographical potential

The geographical potential is eventually estimated in each (200×200) $[m^2]$ pixel in Switzerland. As mentioned before, the geographical potential is computed differently in zone SON from the two other zones namely, OOSG and GEN (see Fig. 7.1). In all zones we only consider roof surfaces characterized by an aspect within the $[-90^\circ, 90^\circ]$ domain, in order to select roof surfaces oriented “towards the south”, for which the potential is maximum in central Europe. This selection is performed to add an economical constraint to the PV solar potential.

In SON zone, the geographical potential in a pixel j is obtained from the sum of the individual potentials of each roof surface in the pixel. The available rooftop area for PV installation over roof each surface s is computed by multiplying the known tilted area A_t^s of the surface (total area of the roof surface considering its tilt) by the estimated average availability coefficient $C_R^{s,j}$ (see section 7.4.1). The shading factor S_{Sh}^j (see section 7.4.2) is used to discard the entirely shaded portions of the roof surfaces (we assume that fully shaded cells do not produce electricity output from PV panels). Finally, the final geographical potential in each pixel j in SON zone is given by:

$$p_{j,geo}^{Son} = (1 - S_{Sh}^j) \sum_{\text{surfaces } s} C_R^{s,j} A_t^s G_t(\beta_s, \gamma_s) \text{ with } \gamma_s \in [-90^\circ, 90^\circ] \quad (7.8)$$

where $G_t(\beta_s, \gamma_s)$ is the global tilted radiation over roof surface s characterized by a slope and an aspect of respectively β_s and γ_s . Note that $\gamma_s \in [-90^\circ, 90^\circ]$, which means that we do not consider roof surfaces pitched towards the north direction (with an aspect value within $[-90^\circ, -180^\circ]$ or $[90^\circ, 180^\circ]$). Other variables in the equation are defined in the above paragraph; note that we use the tilt of the roof surfaces in the equation, meaning that we consider that the PV panels are directly installed on the pitch of the roofs.

In OOSG and Geneva zones, the geographical potential is determined for a typical (average) building cluster rooftop in each pixel j and multiplied by the number of building clusters to obtain the total potential in the pixel. To compute the potential for a typical building cluster, we use the probabilities and global tilted radiation values defined in section 7.4.4, along with a spreading function F_{c_n} which is specific to each roof type c_n . The spreading function is defined using the assumed symmetry of the roof type in order to distribute the aspects (directions) and the available area for

PV installation over the different sides of the roof. The slope is kept identical in each side of the roof. F_{c_n} adds the (available area $\times G_t$) products corresponding to the different sides of the roof. For example, an L-shaped roof uses the following spreading function:

$$F_{c_n=L} [A_R^{c,j}, G_t(\beta_l, \gamma_m)] = \frac{A_R^{c,j}}{4} G_t(\beta_l, \gamma_m) + \frac{A_R^{c,j}}{4} G_t(\beta_l, \gamma_m - 90) \\ + \frac{A_R^{c,j}}{4} G_t(\beta_l, \gamma_m + 90) + \frac{A_R^{c,j}}{4} G_t(\beta_l, \gamma_m + 180) \quad (7.9)$$

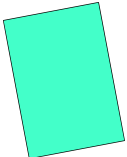
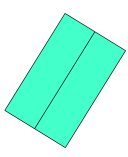
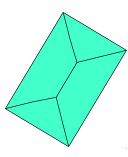
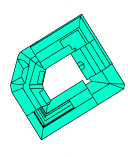
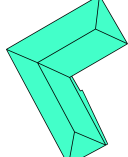
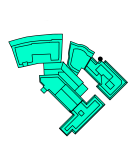
where $A_R^{c,j}$ is the average available area for PV installation over the rooftop of a building cluster in pixel j , $G_t(\beta, \gamma)$ is the global tilted radiation over a surface characterized by a slope of β and an aspect of γ , β_l and γ_m are the main slope and aspect values corresponding respectively to the l^{th} and m^{th} classes of slope and aspect defined in section 7.4.3. We consider that the available area is equally spread over the different directions spanned by the roof. This assumption was motivated by the following points: (i) the location of obstacles is random over the roofs, and therefore it cannot be known without a very precise roof surface data and (ii) the total tilted area of the surfaces (considering the pitch of the roof), however, is approximately equally distributed over the surfaces of a single roof. Point (ii) was validated with the SON data, for which we found an average NRMSE of 25% between the areas of the surfaces and the mean surface area within each building in SON zone. A summary of the considered characteristics of the spreading function assigned for each roof type is given in Table 7.12, extending the previous Table 7.11.

Several assumptions were made concerning the distribution of the available area depending on the roof type. In order to model O-shaped, L-shaped and Complex types the number of sides is simplified and the area is split equally across all sides. Note that for the Complex type the choice of 8 sides having 4 different directions is simplistic and was made in order to attempt to capture all the main directions possibly covered by the shape. For the Flat type we consider panels oriented parallel to the walls of the building of interest (rather than oriented south) in order to maximize the area of panels. We also consider a tilt of 10° for the panels (rather than panels laid flat on the roof) to optimize the electricity output while avoiding self-shading issues. The Hip type, since rather frequent, was treated differently to model the change of available area for PV installation between the two main surfaces and the “hips”. As the hips are usually small and not suitable for PV installation, we intend to remove it from the available area estimated. As a result, we want to compute the area virtually allocated to the hips in our estimation of the rooftop available area for PV, in order to remove it. Let us denote the total hip area h and the available hip area (available area for PV installation on the hips) h_R . The estimation is achieved by setting h and h_R according to the following cases, using the available area for PV over a building cluster A_R^c previously estimated:

$$\begin{cases} h = 20 \text{ m}^2 \text{ and } h_R = 0.8h & , \text{ if } A_R^c \geq 4 \times 20 \text{ m}^2 \\ h = 7 \text{ m}^2 \text{ and } h_R = 0.8h & , \text{ if } 4 \times 7 \text{ m}^2 \leq A_R^c \leq 4 \times 20 \text{ m}^2 \\ h_R = 0 & , \text{ if } A_R^c \leq 4 \times 7 \text{ m}^2 \end{cases} \quad (7.10)$$

The values in these three cases result from the following information: (i) a sample of about 700 hips polygons was extracted from the SON zone, and it showed a mode h of 20 m^2 , and a minimum h of 7 m^2 , which are used as thresholds in the previous computation (ii) a pyramidal roof shape, with an area of $4h$, represents the limiting case of Hip roofs (where the hips are identical to the

Table 7.12: Roof characteristics considered in the spreading function F_{c_n} , depending on the roof type c_n . A_R^c is the average roof available area for PV installation. Note that the available area is used for all sides, besides the special case of Hip roofs, where the area is different in the two main sides and the two “hips” (the two small triangles). See previous Table 7.11 for more details on the considered roof characteristics.

Roof Type	Flat	Gable	Hip	O-shaped	L-Shaped	Complex
						
Roof sides av. area	A_R^c	$A_R^c/2$	2 main sides: $(A_R^c - 2h_R)/2$ 2 hips: 0	$A_R^c/4$	$A_R^c/4$	$A_R^c/4$

two other sides), and offers the smallest area for the same hips area, (iii) we computed in Geneva canton the average ratio between the total hip area and the available area which amounted to 0.8, therefore, we use 0.8 as an approximation of the ratio of the two. The second shading factor S_{Sh}^j is used to discard only the fully shaded fractions of the rooftops. The final geographical potential in each pixel j in OOSG and GEN zones is given by:

$$P_{j,geo}^{OOS} = b_j \left(1 - S_{Sh}^j\right) \sum_{l,m,n} r_n^j p_l^j q_m^j F_{c_n} \left[A_R^{c_j}, G_t(\beta_l, \gamma_m)\right] \quad (7.11)$$

where F_{c_n} is the spreading function defined earlier in the section, r_n^j, p_l^j and q_m^j are the probabilities defined in section 7.4.3 and b_j is the number of building clusters in pixel j . The aspect values considered for each side of the roof classes are presented in Table 7.11 ($\gamma_m + x$, where $x \in [-90, 0, 90, 180]$). The terms of F_{c_n} for which the aspect is out of range of the $[-90^\circ, 90^\circ]$ domain are set to 0, meaning that we do not consider roof portions pitched towards the north direction, as we did in SON zone. Note that the sum in Eq. 7.11 is the sum of 270 potential terms corresponding to the 270 possible roof configurations, weighted by their respective probability (naturally, $\sum_{l,m,n} r_n^j p_l^j q_m^j = 1$ for any j). Finally, as in SON zones, the PV panels are considered to be installed at the pitch of the roofs.

7.5 Technical potential estimation

Similarly to the commune scale study, the technical potential in each pixel j is given by the following equation:

$$P_{j,tech} = PR \times \eta \times P_{j,geo} \quad (7.12)$$

where $P_{j,tech}$ is the technical potential in pixel j , PR is the Performance Ratio (in %), and η is the PV panel efficiency (in %). The assumptions concerning the two coefficients remain identical from the previous commune study: we consider still typical values, with $PR = 80\%$ [134], and $\eta = 17\%$. The technical potential is then obtained for all populated pixels, and for each month, as a daily average in the month (in Wh/day). The cumulated potential in each month can be easily derived by multiplying the daily average by the number of days in the month to obtain a monthly potential (in Wh/month or kWh/month). Figure 7.11 shows the twelve monthly technical

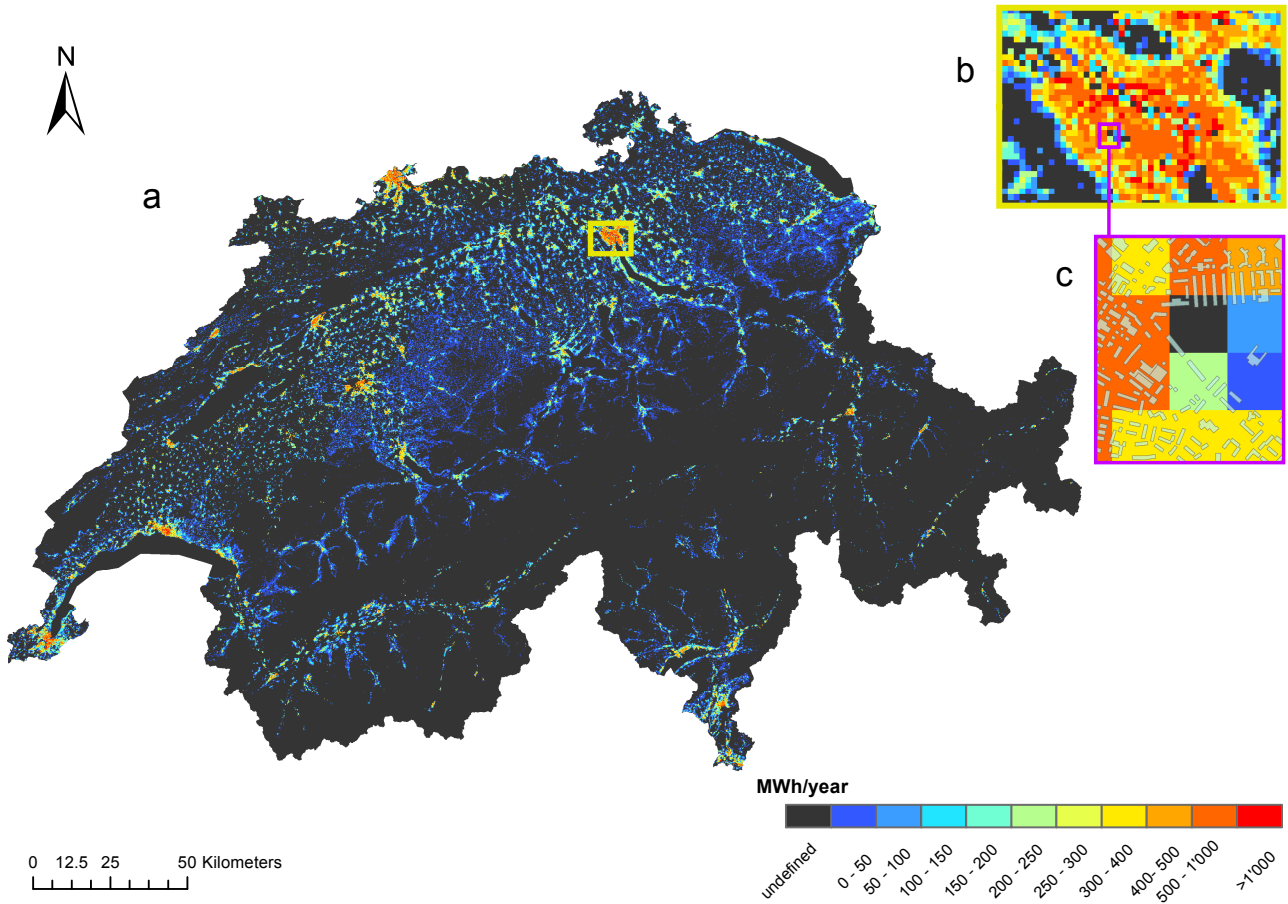


Figure 7.10: (a) Yearly technical potential for rooftop PV electricity production, in MWh/year; (b) Zoom in the Zurich urban area (the zoom location is signified by the yellow window within the map); (c) Further zoom within the Zurich urban area at the pixel level.

potential maps, in MWh/month. Summing the month potentials results in the yearly technical potential, presented in Figure 7.10, in MWh/year.

7.6 Results

7.6.1 Discussion

The technical rooftop PV solar solar potential corresponding to the total electricity production from PV installation on the rooftops, has been estimated monthly and yearly in Switzerland for 159015 pixels of size (200×200) [m²]. The total estimated PV electricity production from building rooftops in Switzerland is 16.29 TWh/year using an efficiency and a performance ratio of respectively 17% and 80%. Considering that the domestic electricity demand in Switzerland in 2017 (freely available data from the SwissGrid website [283]) was estimated to be 64.4 TWh/year, the rooftop PV solar electricity could potentially provide 25.3% of the yearly demand. The areas offering the biggest potential are the dense urban areas, as they offer the highest number of buildings, and thus a generally high total available area for PV installation. Specifically, the potential is concentrated in the Swiss

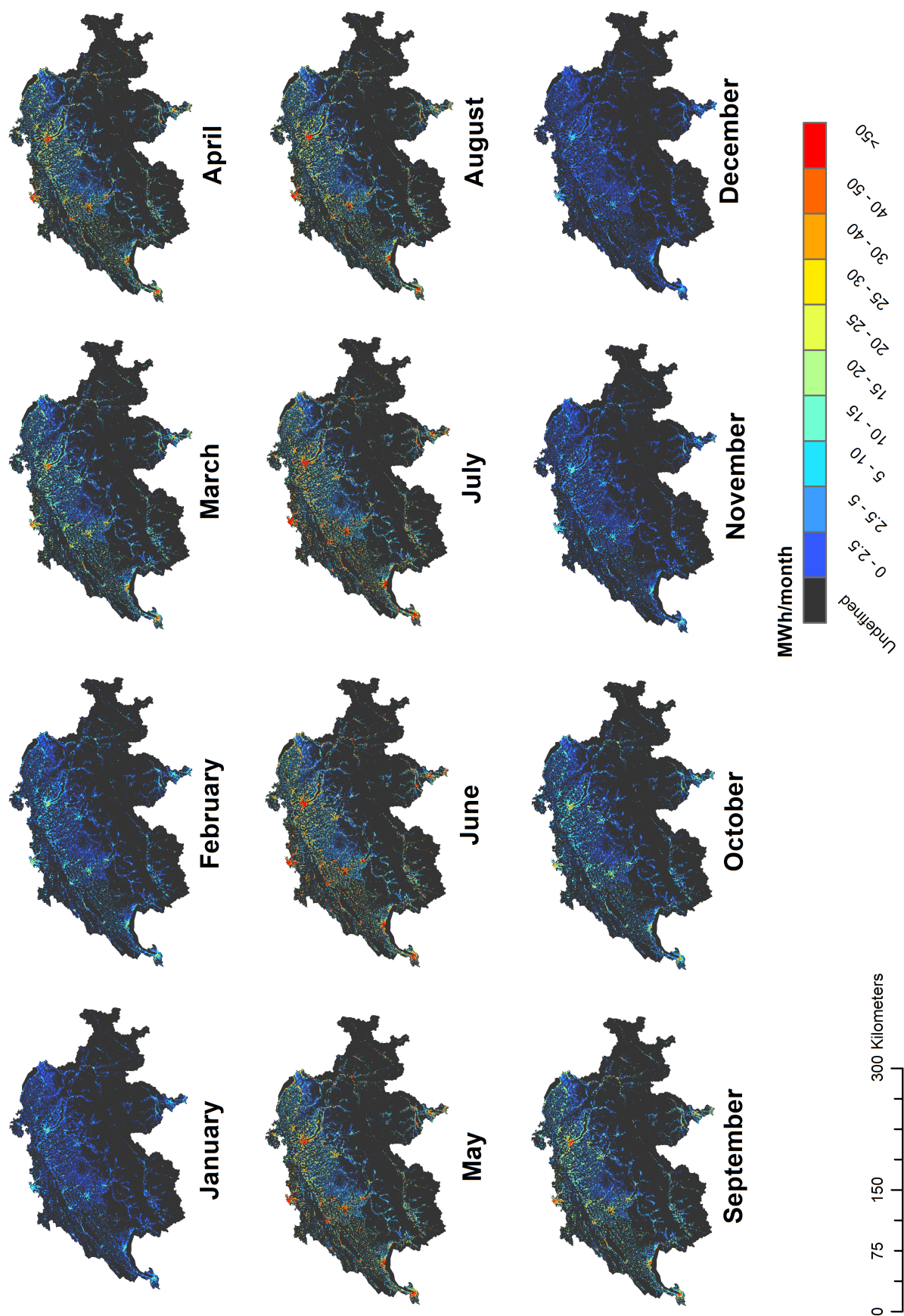


Figure 7.11: Monthly technical potential for rooftop PV solar electricity production, in MWh/month.

plateau, around the Geneva lake, and in the Ticino area. The Alps and the small cities in between large agglomerations unsurprisingly suffer from a lower solar potential.

The electricity demand profiles of the different Swiss cantons are compared to their total estimated rooftop PV solar potential in order to assess the capacity for rooftop PV to meet the national demand. The SwissGrid dataset [283], offering load values at a resolution of 15 minutes, is aggregated in order to extract the monthly electricity demand of each demand in 2017. Figure 7.12 shows demand profiles along with the estimated PV solar potential for all Swiss cantons of Switzerland. Note that some cantons are grouped within the demand dataset, and are therefore aggregated accordingly for their PV potential. It can be seen that, as one can predict, the gap between the electricity demand and the PV solar potential reduces significantly during summertime, and is filled in some cantons (e.g. in Neuchatel - NE). The comparison between the estimated PV solar potential and the electricity demand can also be seen at two aggregated levels in Figure 7.13: (i) at the national level, monthly, and (ii) at the cantonal level, yearly. Note that storage of the surplus PV electricity (e.g. through batteries) is a sound solution to reduce the demand-potential gap without automatically requiring electricity from the grid. It requires, however, hourly values for the electricity demand and the PV solar potential, which allow for the sizing of the batteries [284, 285].

The rooftop PV solar potential per capita is also computed in each pixel by normalizing the potential by the population in each pixel, using the STATPOP Swiss population data from 2015 [286]: the yearly potential per capita map is shown in Fig. 7.14. An interesting feature of such a map is that it neutralizes the overpowering effect of the number of buildings, and rather focuses on the actual potential for PV electricity production to be a viable solution for the inhabitants in the pixel. When comparing both potential and potential per capita maps the general patterns are clearly reversed, as the dense area in the Swiss Plateau has low PV electricity production due to their high population. The biggest potential per capita can be observed in the Swiss alps and the small cities located in the Plateau. One can notice, however, that some larger cities well exposed to solar radiation manage to maintain a high potential per capita, for instance in the east side of the country.

As discussed previously, the available area for PV solar installations over rooftops plays a crucial role. The total available area is estimated to be 252 km² over all considered buildings in the pixels; this estimate results in an average available area of 31.5 m² per capita. The total ground floor area in Switzerland extracted from more than 2 million buildings is 269 km². It follows that, in average, in Switzerland 94% of the ground floor area can be used to install PV panels over the rooftop of a building. Unsurprisingly, the ground floor area appears to have a very high impact on the estimation of the available area, as shown in the variable importance [41] graph in Fig. 7.15. The two most important factors, according to the variable importance figure, are the site coverage (DS_area(build_area_per_km2)) and the mean building ground floor area (footprint_mean) in the pixels. It is however known that feature importance can have bias if real-value and categorical features are mixed[287], which is the case in the present study. Therefore, although the mentioned results concerning the variable importance seem to align with intuition, they must be considered with caution. It should also be mentioned that available roof area for PV installation is estimated independently from the shading impacts; the 94% ratio therefore does not reflect the impacts of shadings. We have applied the shadings separately as indicated in the Equation 29. Therefore, the final available roof area for PV installation: (i) does not include the shadings from neighboring buildings and trees on rooftops

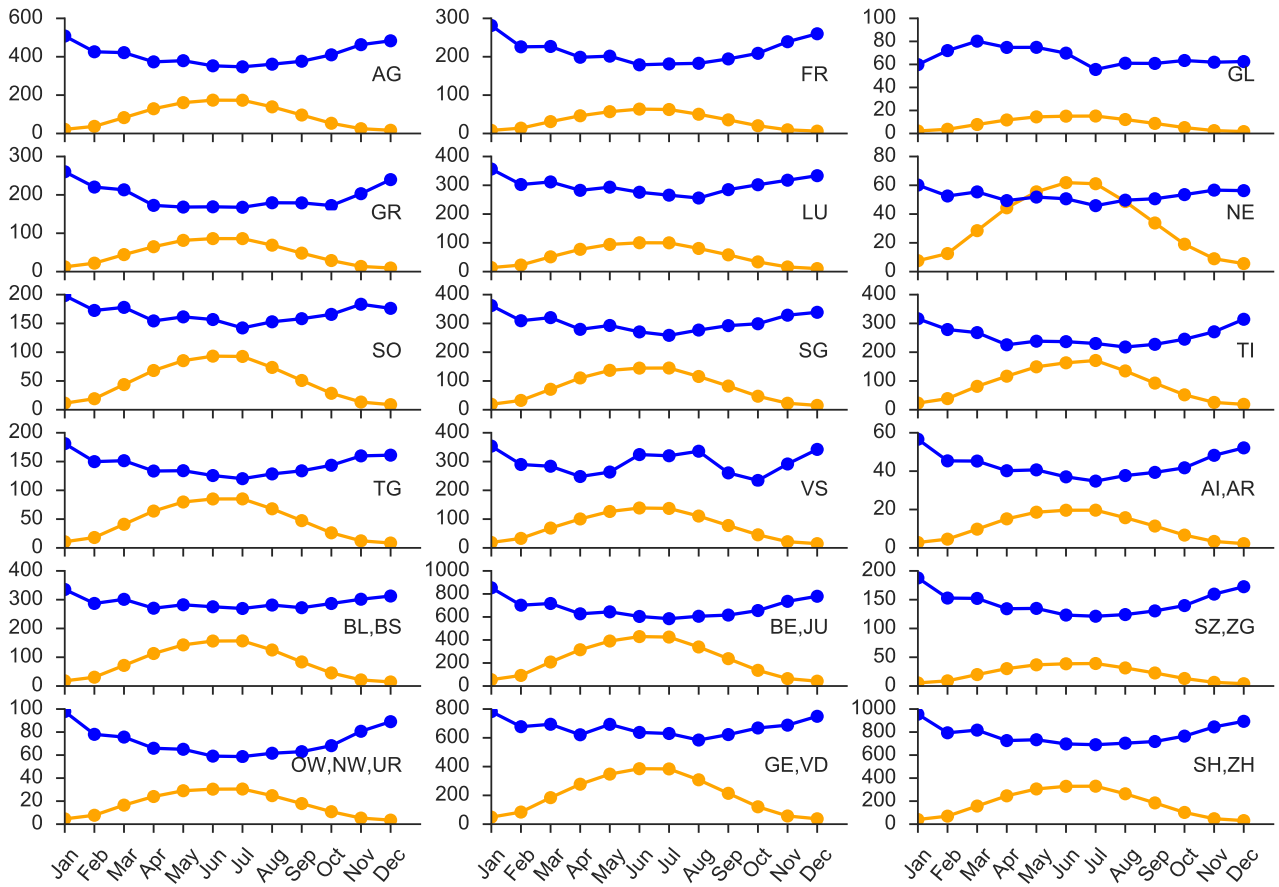


Figure 7.12: Monthly demand profile (for 2017), in blue, and rooftop PV solar potential for the Swiss cantons, in yellow, in GWh/month. The abbreviations used for the canton names are the following: AR:Appenzell Ausserrhoden, AI:Appenzell Innerrhoden, BL:Basel-Landschaft, BS:Basel-Stadt, BE:Bern, FR:Fribourg, GE:Geneve, GL:Glarus, GR:Graubunden, JU:Jura, LU:Luzern, NE:Neuchatel, NW:Nidwalden, OW:Obwalden, SH:Schaaffhausen, SZ:Schwyz, SO:Solothurn, SG:St. Gallen, TG:Thurgau, TI:Ticino, UR:Uri, VS:Valais, VD:Vaud, ZU:Zug, ZH:Zurich.

and (ii) include only the roof surfaces oriented towards the south direction. In the final potential estimation (Eq. 7.8 and 7.11), however, we separately consider these two parameters.

7.6.2 Comparison with the first estimation (commune level)

This study brings significant improvements over a previously presented commune level analysis. As such, the respective results of the two studies would benefit from a comparison, and a discussion on the possible reasons for the significant differences between the two studies. The total ground floor area is estimated at 407 km² in the previous estimation, whereas in the current study is 269 km². This significant difference is mainly explained by the use of the building information data (RegBL) in the present study. This data provides a more precise and ultimately smaller estimation of the ground floor area than the building clusters' polygons (used in the previous study), which englobe multiple buildings. The total available rooftop area for PV installation in Switzerland is estimated at 328 km² in the previous estimation, while in the present study is 252 km². This smaller available area value is in parts explained by the following points: (i) as the ground floor area and the available area are clearly

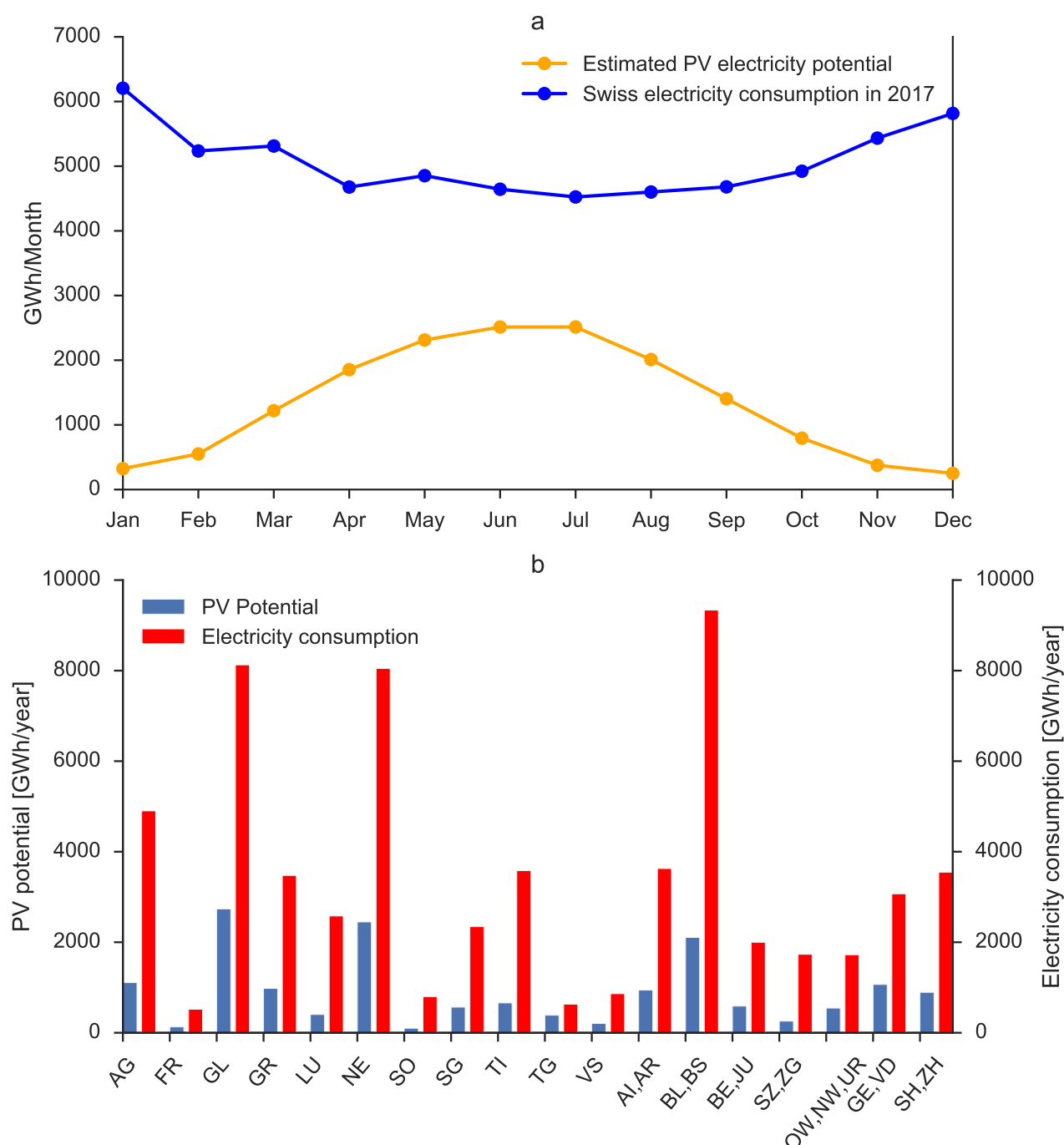


Figure 7.13: Comparison between the electricity consumption (for 2017) and the estimated rooftop PV solar potential at two aggregated levels in Switzerland: (a) at the national level, monthly, and (b) at the cantonal level, yearly. The abbreviations used for the canton names are the following: AR:Appenzell Ausserrhoden, AI:Appenzell Innerrhoden, BL:Basel-Landschaft, BS:Basel-Stadt, BE:Bern, FR:Fribourg, GE:Geneve, GL:Glarus, GR:Graubunden, JU:Jura, LU:Luzern, NE:Neuchatel, NW:Nidwalden, OW:Obwalden, SH:Schaaffhausen, SZ:Schwyz, SO:Solothurn, SG:St. Gallen, TG:Thurgau, TI:Ticino, UR:Uri, VS:Valais, VD:Vaud, ZU:Zug, ZH:Zurich.

positively correlated, the smaller estimation of the first entails a smaller value of the latter and (ii) the update on the area labelling in Geneva (considering the rectangular constraints and the standard

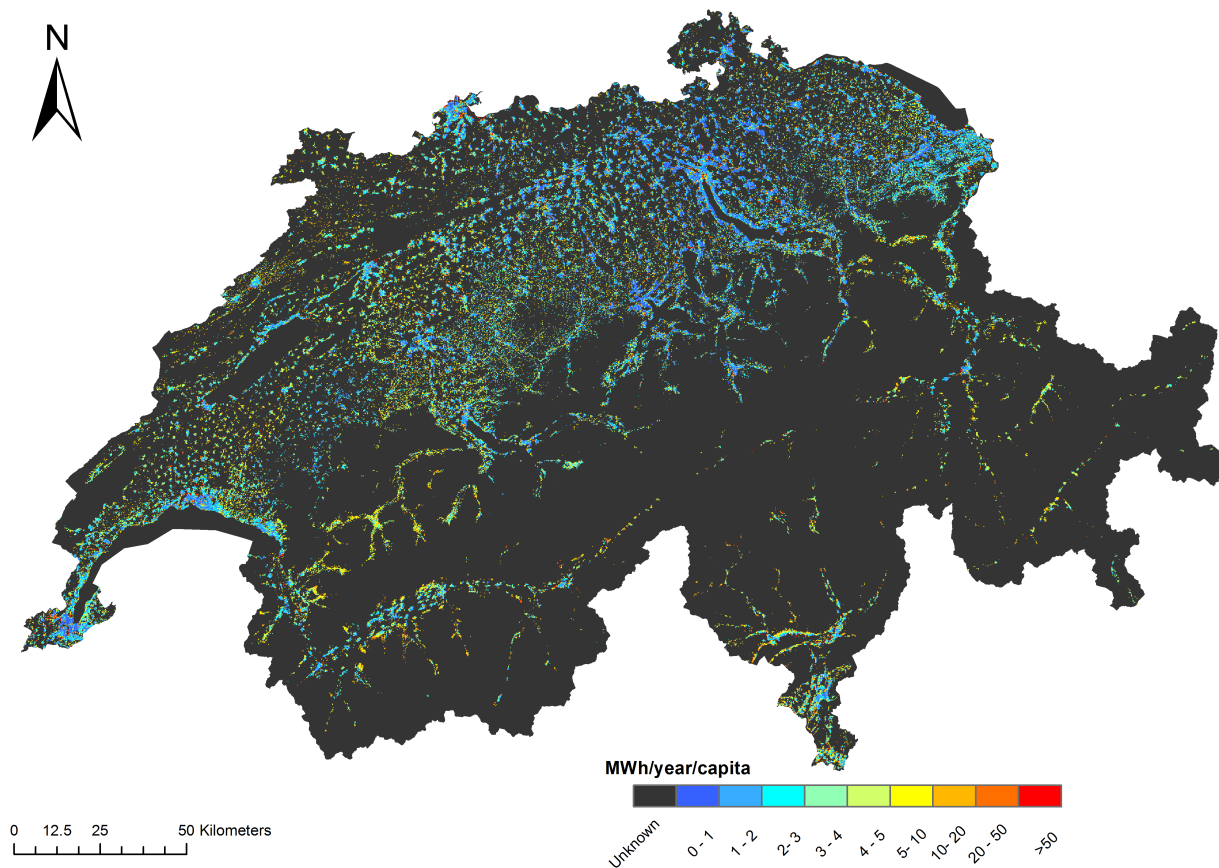


Figure 7.14: Yearly technical potential normalized by the population in each pixel for rooftop PV electricity production, in MWh/year/capita.

dimensions of PV panels) makes the estimation closer to reality and ultimately significantly smaller. It follows that the total PV electricity production estimated for Switzerland in the current study is 16.29 TWh/year, whereas in the previous study was 17.86 TWh/year. The smaller technical potential determined by the present study is in parts explained by the significant smaller estimated value of the available area for PV installation. The difference in potential seems, however, small compared to the difference in available area estimation. This is mainly explained by the fact that a much larger number of building clusters is considered in the present study (1'825'678 throughout the whole Switzerland) while only those contained in urban areas (defined by the Corine land cover database) were considered in the previous study (1'264'050). This higher number of buildings counter-balances the decrease in available area, and explains the decrease of approximately 7% from the potential estimated in the previous study.

7.6.3 Validation with other potential studies

The total annual PV electricity production in Switzerland can be compared with the study carried-out by the International Energy Agency in 2002. There are noticeable differences between the present study and the IEA study [215], including: (i) the total available rooftop area for PV, estimated as 138.22km² by IEA, and 252km² in the present study and (ii) the total PV electricity potential, estimated as 15.044 TWh/year by IEA, and 16.29 TWh/year by the present study. It should be first noted that the aim of

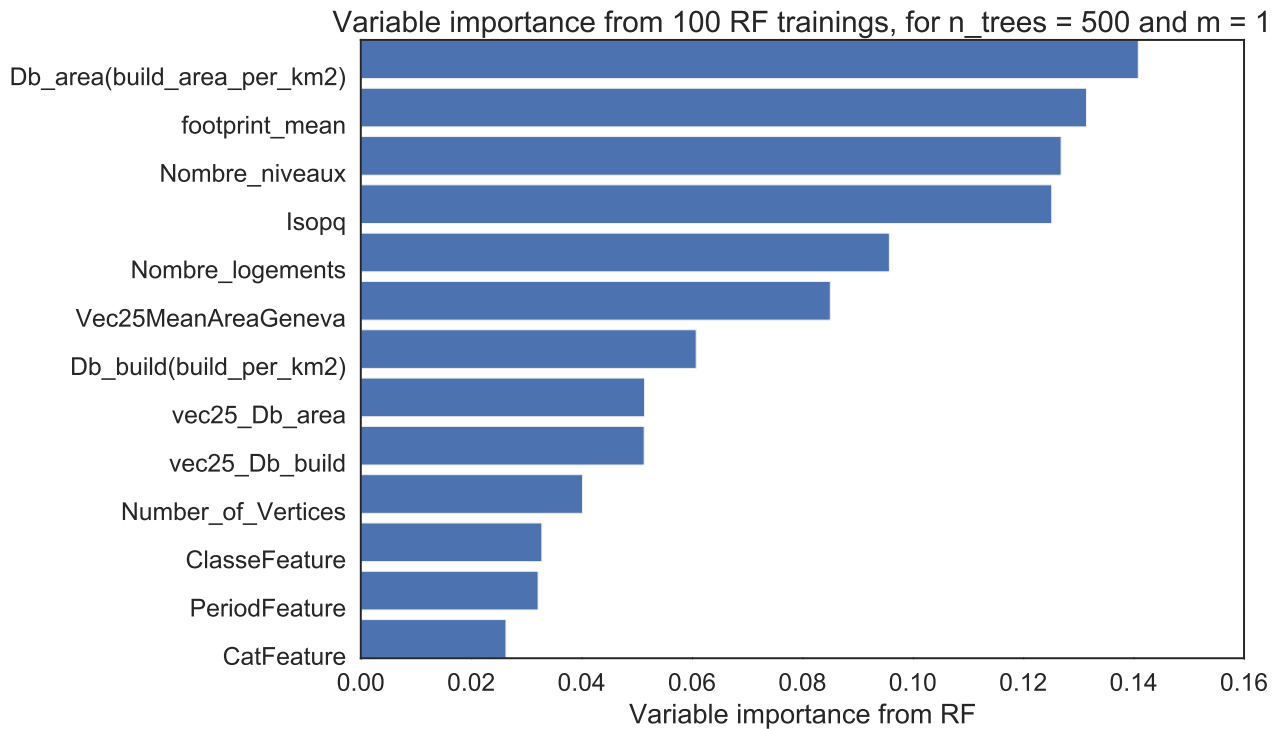


Figure 7.15: Variable Importance (VI). The graph shows the importance of each feature during the training of a Random Forest for the estimation of the available area in OOSG/GEN zones (describe in section 7.4.1). Here the categorical features are aggregated to reduce the original number of features for visualization purposes (e.g. instead of considering 12 construction period features, we aggregate them to one feature expressing the most frequent period in a pixel). The features in the x axis are as follows: CatFeature: most frequent residential class, PeriodFeature: most frequent construction period, ClasseFeature: most frequent building typology, Number_of_Vertices: average number of vertices of building cluster polygons, vec25_Db_build: number of building clusters, vec25_Db_area: ratio of the total ground floor area of building clusters to the total pixel area, Db_build(build_per_km2): number of buildings, Vec25MeanAreaGeneva: average ground floor area of building clusters, Nombre_logements: average number of building flats, Isopq: average isoperimeter quotient of building cluster polygons, Nombre_niveaux: average number of building floors, footprint_mean: ground floor area, Db_area(build_area_per_km2): ratio of the total ground floor area of individual buildings to the total pixel area.

the IEA study was to extract a global approximated estimate of multiple European countries. Therefore, the study uses rule of thumbs and does not claim to provide very precise estimates. It is however useful to compare these estimations with the present study ones, to validate the order of magnitudes of the different estimations. One might observe that while the final electricity potential is relatively matching between the two studies, the estimated area suitable for PV is significantly different. This is mainly explained by the available area definition the IEA study considers. In particular, in the IEA study, the available area includes two factors: (i) construction elements and shading (60 % on average) and (ii) solar suitability, to select surfaces with a sound solar yield (55 % on average). The present study, in contrast, considers that the available area include solely the construction constraints, and estimates the shading impact and the good solar yield factor separately. Note that the so-called good solar yield factor is equivalent with considering the south oriented roof surfaces only [288]. Furthermore, the tilted solar radiation in the IEA study is estimated using an average country specific value and a panel

efficiency of 10%, which adds to the explanation for the smaller final potential estimation.

The estimated PV electricity production in this study is also compared with the results from SITG for the canton of Geneva [261]. The comparison is shown in Figure 7.16. One can notice from the comparison that the present study shows a lower potential value for seven communes and a very similar value for the rest of them. This is partly related to the aggregation required to extract commune potential values using the pixel potential values. In particular, a significant amount of pixels lying at the boundaries of the communes were discarded in the computation of the potential of the communes, in order to avoid considering their potential twice. It resulted in an underestimation of the potential of certain communes showing relatively high values in SITG estimation. Note also that the strategy used by SITG differs significantly from the one used in the present study, which further explains the differences in the potential estimation. In particular, SITG study estimates the available area for PV installation using the following main criteria: (i) it selects areas with a reasonable solar yield, using a threshold of 1000 kWh/m², (ii) it removes surfaces that have a total area smaller than 5m², and (iii) it forms a buffer of 1m around the roof boundaries, and erase the selected roof portions. As mentioned previously, the corresponding solar yield factor is equivalent to considering the roof surfaces oriented towards the southern direction (with an aspect within $\pm 90^\circ$), as it is done in the present study. Criteria (ii) and (iii), however, are significantly less constraining than the criteria used in the present study (as presented in section 7.4.1), and most probably lead to a generally higher available area estimation than in the present study. As a result, the underestimation found in the present study with respect to the SITG study is further explained.

The rooftop PV solar potential has been estimated in 2016 for several thousand buildings in Switzerland based on 3D building data in the Sonnendach project [260, 266, 281]. As described in sections 7.4.1 and 7.4.4, slope and aspect information were extracted from the Sonnendach project in order to compute the PV electricity production in SON zone (see Fig. 7.1). The present results, however, do not use Sonnendach project final PV solar potential values and hence the two estimations can be compared. To do so, we aggregate the data published in the Sonnendach project within the pixels considered during the study. The differences in the final results between the present study and the Sonnendach project are mainly related to three key points. The first point is the estimation of the available rooftop area for PV installation. In the present study we remove all superstructures (HVAC systems, chimneys, etc.) and take the geometrical properties of PV solar panels into account in order to extract the actual roof areas suitable for PV (details in section 7.4.1). The Sonnendach project, however, considers all roof surfaces with a minimum side length of 8 m [269] and therefore considers the total pitched area of the roofs as being potentially available for PV. The second point is the treatment of the rooftop aspects during the study. The present study considers only with an aspect (direction) within $\pm 90^\circ$ from due south. By contrast, the Sonnendach project considers all roofs, regardless of their aspect [269]. Finally, the third point is the estimation of the shading impact over the rooftop PV solar potential. The present study estimates two factors to account for the shading impact both over the solar radiation and the possible areas for PV installations (respectively S_{hill} and S_{sh} , see details in section 7.4.2). The Sonnendach project, on the other hand, only considers the shading impact over the radiation [269]. As a result, the Sonnendach project data is re-scaled so that it fits the present study assumptions. The general NRMSE computed between the two studies is 26%. We also provide the detailed comparison

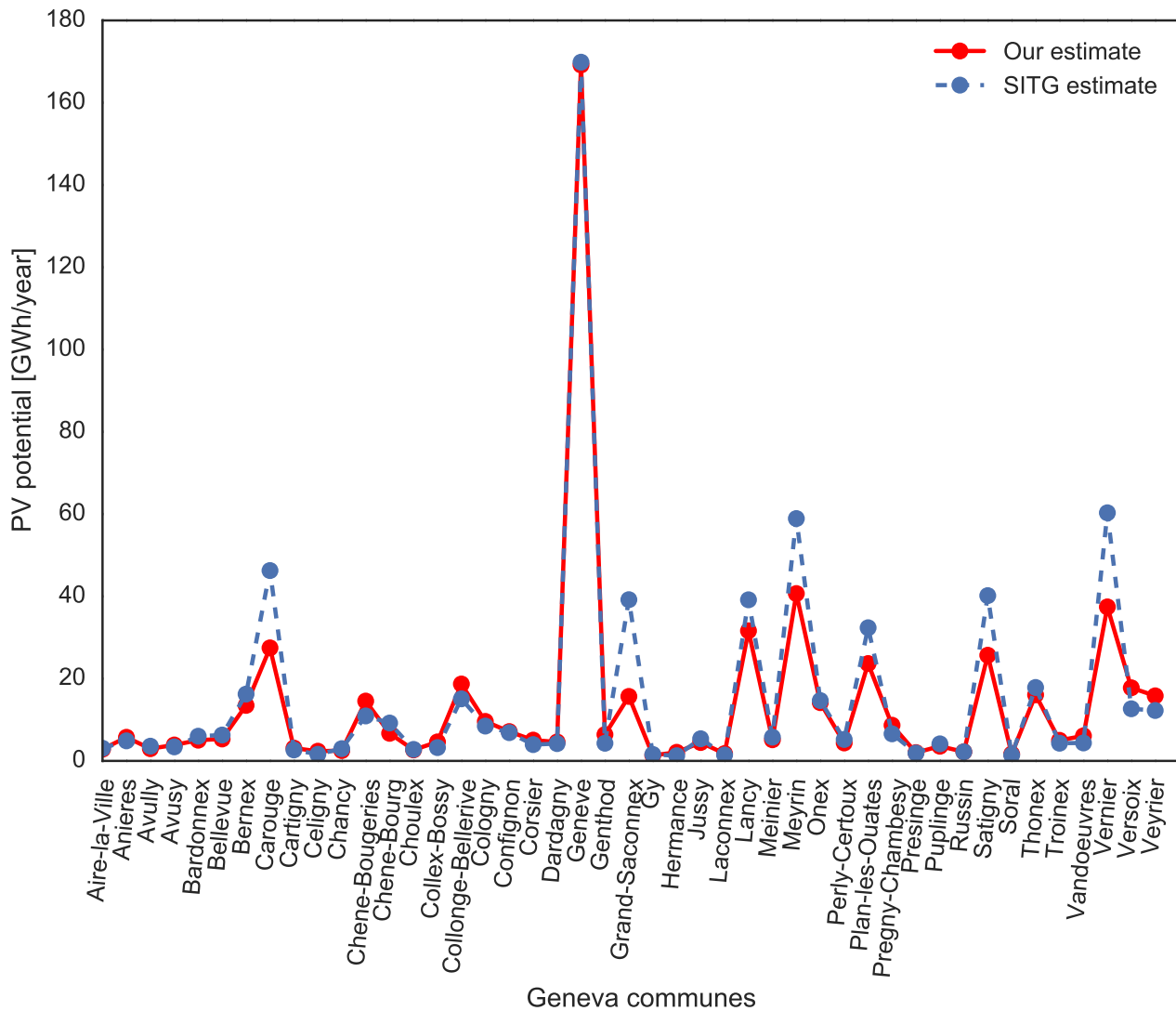


Figure 7.16: Comparison between the yearly rooftop PV solar potential estimates from SITG and the current study, for all communes in the Geneva canton in GWh/year.

20 pixels randomly chosen in the SON zone in order to show local differences between the potential by the present study and Sonnendach study. This comparison is shown in Table 7.13.

7.6.4 Limitations

The present chapters brings significant methodological improvements in terms of large scale estimation of the PV solar potential over rooftops and a literature contribution on the rooftop PV solar potential of Switzerland. It should also be mentioned that the complexity brought by the machine learning in our study did not increase significantly the total required computational time, while improving the general estimation accuracy of the estimation. In particular, the training time for a Random Forest model, using a training data of around 3200 points and 73 features (the largest training data in our study), with 500 trees, and including a 6-fold CV to tune the m parameter, showed to be on average 3.14 minutes.

Table 7.13: Comparison of the estimated PV solar potential in 20 randomly chosen pixels between Sonnendach project potential study and the present study. Note that the Relative error is the absolute error between the two estimations, in MWh/year, while the Absolute error is the ratio of the Relative error to the sonnendach estimation value, in %

Pixel ID	sonnendach study [MWh/year]	Present study [MWh/year]	Relative error [MWh/year]	Absolute error [%]
110735	22.034	17.338	4.697	11.029
1413673	18.048	22.282	4.234	9.942
588095	48.202	40.021	8.181	19.210
330453	65.677	54.733	10.944	25.698
561206	116.113	141.486	25.373	59.580
418828	45.387	38.655	6.732	15.807
588067	42.355	39.576	2.779	6.526
611207	154.940	138.102	16.839	39.541
711382	44.700	38.016	6.683	15.695
1005625	33.460	27.645	5.815	13.655
532421	3.745	2.814	0.931	2.185
1213638	4.017	3.065	0.952	2.235
553384	41.852	51.288	9.436	22.158
641880	22.580	17.720	4.859	11.410
901685	11.095	11.501	0.407	0.955
386227	85.383	67.581	17.801	41.801
566755	14.528	13.754	0.774	1.818
722803	13.139	9.363	3.775	8.866
684581	20.366	19.812	0.554	1.300
266811	44.095	33.672	10.423	24.475

Several limitations, however, remain challenging and should be consider in future studies. They include: (ii) **Discrepancies in PV solar potential at pixel boundaries.** The grid resolution induces some issues at the boundaries of pixels, in case some buildings overlap multiple pixels. Various choices were made in order to minimize that issue (e.g. the total available area for PV installation in pixels of the Geneva canton was computed as the sum of the simulated modules within the pixel, instead of the total building available area). The buildings overlapping multiple pixels have their potential divided between the multiple pixels. In addition, some pixels might have their potential slightly overestimated or underestimated as a result of overlapping modules between pixels and very large building clusters overlapping multiple pixels. This issue could be avoided by considering natural clusters of buildings instead of predefined pixels. This would require an efficient method to cluster buildings according to multiple variables at the national scale. It defines, however, an interesting topic for future work. (iii) **Uncertainty propagation.** Prediction intervals were computed for each estimated variable and provide a measure of the uncertainty of their estimation. As already mentioned in the previous chapter, however, the multiple consecutive steps inputting the estimations from machine learning into physical models, however, induce a propagation of the errors through these analytical models. While the increased accessibility of data might allow for a more direct estimation of the potential, therefore reducing the propagation of uncertainty, the present study does suffer from such a propagation. A thorough study using multivariate error analysis and combining the prediction intervals for all estimated variables would be needed to produce final prediction intervals for the PV solar potential value of each pixel [289–292] (separately in SON and OOSG regions, as the approach is

different in the two regions). The errors induced by the multiple GIS steps shall also be considered [292]. It represents an extensive study that could be the focus of future work.

7.7 Summary

This chapter further explores the potential for PV panels over rooftops in Switzerland, initiated by the previous chapter. While the previous study was valuable for decision making at communal level, it was desirable to provide a study adapted to single buildings and neighborhoods, or small communities. As such, a new resolution of (200×200) [m²] pixels is suggested, and several improvements are achieved to match this higher resolution. These ameliorations translate into methodological improvements, in particular regarding the large scale estimation of (i) geometrical rooftop characteristics, by extracting classes for roof shapes, slopes and aspects based on LiDAR raster analysis and Random Forest prediction, (ii) available rooftop area for PV installation, by simulating the presence of PV panels over rooftops based on GIS vector processing and Random Forest prediction, and (iii) uncertainty assessment for PV solar potential variables, by computing prediction intervals based Quantile Regression Forests. These improvements were supported by the use of additional data sources, improving the general accuracy and the spatial coverage of the estimation. Note that the procedure proposed for the different estimations can be applied to other countries. It would only require datasets which are nowadays available for most of them, e.g. a general low quality GIS building data, some building information data, a more precise GIS roof building data for at least a small region of the country, some weather data, and a Digital Elevation Model.

A flowchart summarizing the entire new methodology proposed in the chapter is shown in Figure 7.17.

The chapter eventually shows a more realistic estimation of the Swiss potential for rooftop PV installations, with a total of 16.29 TWh/year, which corresponds to 25.3% of the electricity demand in 2017. The extracted PV solar potential estimation for each (200×200) [m²] pixel can be ultimately useful for decision making, regarding two main aspects: (i) the identification of optimal locations for cost effective PV installations, leading to an increased integration of decentralized solar energy through PV panels over the building rooftops and (ii) energy system design and optimization for managing the electricity demand in neighborhoods and small communities.

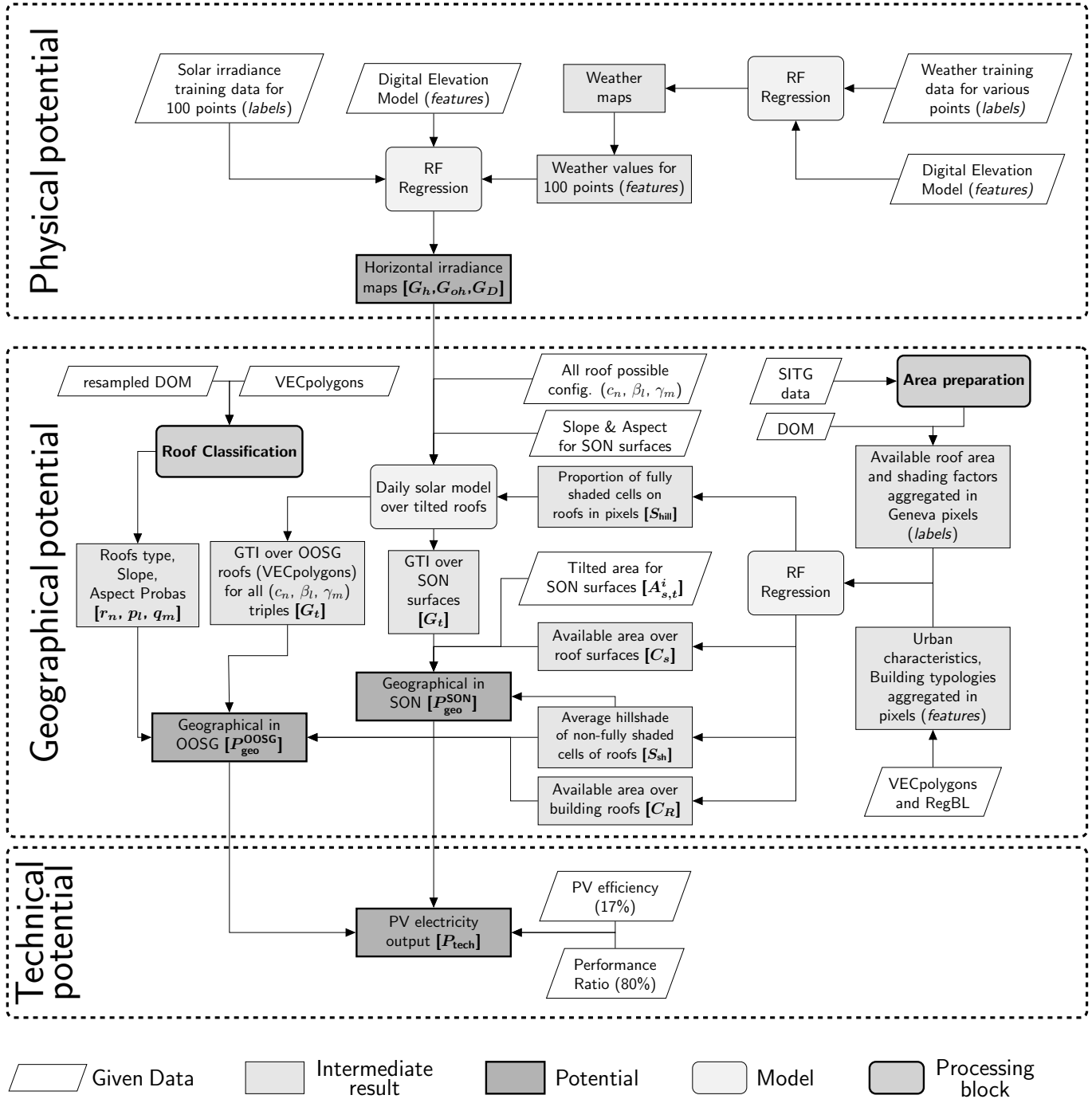


Figure 7.17: Flow chart of the methodology for the PV solar potential estimation improved at pixel scale.

8

Conclusion

In this concluding chapter, Section 8.1 presents the main contributions of this thesis. Section 8.2 examines practical implications related to the current implementation of the estimated potential for all considered renewable forms of energy. Section 8.2 closes the present thesis by discussing the possible areas of future research.

8.1 Main findings

In this thesis, we propose a general strategy relying on available data and traditional models together with Machine Learning methods to assess the large-scale energy potential for three very popular forms of renewable energy: wind energy (through the use of horizontal axis wind turbines), very shallow geothermal energy (through the use of ground connected heat pumps) and solar energy (through the use of PV panels over rooftops). Each energy-specific study includes the estimation of multiple variables of interest related to the potential, variables which often have very different behavior (meteorological variables, topographic variables, soil-related variables, building-related variables, etc.) and depend on different parameters. The use of Machine Learning notably allows, given adequate features, for the prediction of the latter variables at unknown locations, along with the uncertainty attached to the predictions. In each case, the methodology is applied to the Swiss territory; in particular, this means that it is based on data available in Switzerland. Similar data is however more and more accessible in a significant number of regions and countries, allowing the considered strategies to be widely generalizable.

We first investigate the theoretical potential for wind energy (**chapter 4**), at a 200×200 [m²] pixel resolution (partly adopted to match the typical building energy demand resolution imposed by privacy issues). In this case, the potential is primarily dictated by the estimation of the monthly wind speed over the country. The main difficulty in assessing the wind speed lies within its change of behavior with different obstacles configuration, which translates into a different treatment for rural and urban areas. For rural areas, wind speed measurements (at a height of 10m) are used together with features of interest (meteorological, topographic, and roughness-related variables) to train Random Forests models in order to spatially extrapolate the monthly wind speed in all rural parts of Switzerland. Reasonable average test error and prediction error of respectively 0.7 m/s and ± 1.15 m/s are achieved. The wind speed is then vertically extrapolated at 100m with using a classical log-law (installation height for

large rural horizontal wind turbines). For urban areas, however, very few wind measurements are performed. We therefore use a deterministic approach to derive wind speed values above buildings (for small turbines installation) at the periphery of urban areas, using urban boundary layer assumptions to allow for the re-use of rural area estimations, and empirical expressions for urban characteristics. The obtained values are validated with available measurements in urban areas. Finally, a theoretical power computation leads to a promising potential installation of, for one turbine, on average 80 kW (up to 1600 kW) in rural areas and 15 W (up to 1.1 kW) in urban areas. Considering an average yearly operation time of 1600 hours [6], it corresponds to an average potential electricity generation of 128 MWh per year per turbine in rural areas, and 24 kWh per year per small turbine in urban areas. Following a crude estimation of the number of wind turbines installable in both rural and urban areas, we obtain a total yearly geographical potential approximation for wind, amounting to 1.17 TWh in rural areas, and 4.2 GWh in urban periphery areas.

As a second study, the theoretical potential for very shallow geothermal potential (vSGP - in the first two meters of the ground) is examined (**chapter 5**), at the same 200×200 [m²] pixel resolution. This time, the theoretical assessment is more convoluted since it involves the estimation of multiple variables reflecting the thermal behavior of the ground: (i) monthly ground temperature gradient, (ii) ground thermal conductivity, and (iii) ground thermal diffusivity. For each of the three variables, we rely on ML models (Random Forests) which are separately trained using a large number of meteorological and soil/geology-related features. In addition, the use of several traditional modeling strategies was required to build some of the training output set using related data, notably for the diffusivity and conductivity, which are rarely measured in practice. The test errors for the ground temperature are of 1 to 2 °C, which is very acceptable. On the other hand, the estimation for diffusivity and conductivity, due to more arduous process, is more challenging. While the normalized test errors for these two variables are around 15%, showing a sound accuracy, the choices of modified output labels entail a rather large prediction error (as seen from maps 5.16 and 5.11). Nevertheless, the study shows a significant potential for very shallow geothermal energy systems, notably in the Valais, Ticino and St. Gallen cantons which show large thermal conductivity values. Following a preliminary estimation of the heat content during the heating season and possible storage during the cooling season, as well as the average COP in both seasons, we also provide a first estimation of the yearly geographical potential for very shallow geothermal energy systems. The results show a total yearly potential of 4.00 TWh for cooling, and 11.81 TWh for heating. The latter heating potential notably corresponds to 17.8% of the annual space heating demand in Switzerland in 2017.

Chapter 6 considers the potential for electricity generated by solar PV panels mounted over rooftops. Given the popularity and the large deployment capabilities of the technology, the full technical potential is tackled; this brings significant complexity to the study compared to the two previous chapters. The resolution chosen for this first technical study, however, is the communes, the smallest administrative division in Switzerland. While being very useful for stakeholders and global assessment, this resolution allows for a smaller computational complexity. Support Vector Machines are used to allow the spatial estimation of the multiple variables required when dealing with the discussed potential: global horizontal radiation over rooftops, available rooftop area for PV installation, slope and aspect distribution of rooftops and shading factors over rooftops. Several sources of meteorological and

building data together with elevation models are used as features for these multiple estimations. The global tilted radiation over rooftops is then computed using traditional solar models, by combining the estimated variables as inputs. The achieved Normalized RMSE for the monthly horizontal radiation components are very satisfactory, with an average of around 6%. Similarly, the accuracy of the shading estimation is very acceptable, offering NRMSE values under 20%. The estimation of slope distribution, however, is more problematic and offers high test errors. This is notably explained by the intuitive difficulty to predict the slope solely based of non-geometrical building characteristics; it is therefore improved in the next chapter. Finally, the prediction of the ratio of available area to ground floor area show a very reasonable normalized RMSE test error of 5%. Results show that the rooftop PV potential is significant in Switzerland. Indeed, considering all urban rooftops directed within $\pm 90^\circ$ deviation from due south, the potential annual PV solar electricity production is estimated at 17.9 TWh, which corresponds to 28% of the Swiss annual electricity demand in 2017.

The rooftop PV potential discussed in the earlier chapter is further explored and improved within **chapter 7**. In a desire to assess the potential for single buildings and neighborhoods, the study is re-designed at the 200×200 [m²] pixel resolution. Multiple improvements are implemented within the methodology to answer the resulting higher computational needs, and generally increase the accuracy offered by the previous study. The improvements include methodological ameliorations, notably regarding the large-scale estimation of (i) the available rooftop area for PV installation, using ArcGIS to simulate the presence of PV panels over rooftops and Random Forests prediction for extrapolation, and (ii) the geometrical rooftop characteristics, by classifying the shapes, slopes and directions of rooftops over the country with on LiDAR raster analysis and Random Forest classification. The use of RF allows for prediction uncertainty assessment, as performed in chapters 4 and 5. Note that the latter improvements relied on the use of additional data sources available since the previous study. While the classification of geometrical characteristics offer reasonable results, it should be noted, however, that the estimation of available area, based on general building characteristics, remains challenging. Even though the examples offered by the ArcGIS methodology are of good quality, the estimation itself showed high error rate (in the test set and for new predictions) when the slope of roofs were not known (in OOSG zone). It is therefore reasonable to conclude that an accurate estimation of the area available over rooftops require very precise data such as detailed 3D models or high resolution LiDAR data to make it possible to precisely assess the tilt of roofs and detect obstacles (e.g. chimneys, windows, etc.) with confidence. The final potential values, however, were compared and globally validated with existing studies in Switzerland. Results showed a more realistic and eventually lower potential than the first solar study, with 16.3 TWh per year, corresponding to 25.3 % of the yearly electricity demand in 2017. In particular, it is worth noting that the installed PV capacity of 1683 GWh in 2017 (as given by SFOE [6], presented in the chapter introduction, section 1.2.1) corresponds to only 10.3 % of the estimated potential generation.

8.2 Practical implementation

The present thesis provides methodologies to estimate, at the scale of a country, the potential of solar, wind, and geothermal energies, using some of the most widely used and available systems (PV solar panels, ground connected heat pumps, horizontal axis wind turbines). It also highlights the significant potential contributions these energies could bring as renewable resources in Switzerland. In particular, it is clear from the comparison with the installed power in 2017 that the potential is not fully utilized, particularly for solar PV and wind energy.

Unfortunately, while the estimations provided in this thesis are rather realistic (e.g. by considering only roof directed towards ± 90 degrees from due south), they are nonetheless most probably slightly optimistic regarding the actual implementation of such renewable energy systems. This is not the case because the potential was overestimated, but rather because some practical factors were not taken into account, as they would require a whole study to be considered thoroughly. These factors mainly include additional constraints which define the *economic* and *market* potentials (presented in the introduction, section 1.2), and further reduce the potential. These constraints include notably cost issues, implementation policies, social acceptance and legal considerations. In practice, these factors are difficult to estimate because they are directly related to the behavior of decisioners and consumers. While the acknowledgment of environmental issues is fortunately being spread out and the general behavior is slowly becoming aware of these issues, these economic and social limitations are still significant for most countries and must be taken into account when drawing feasible energy scenarios for the future. In Switzerland, however, the current federal law is very much in favour and supporting the development of sustainable technologies. Regarding PV solar and geothermal heat pumps in particular, the current subsidies are behind the citizen's wishes, which limits the effects of additional constraints not considered in the present thesis, and comforts the idea that a large potential is still available in Switzerland for renewables (8.1).

Note that studies examining economic and market constraints exist in literature, in particular for solar PV, which remains the most popular technology available. Notably, social acceptability [293], policies [294] and economic potential analysis [295, 296] have been studied. Such studies are, however, rarely embedded with a thorough technical study. Ideally, all social and technical aspects should be combined to provide a fully realistic potential estimation strategy, especially for countries suffering from many economic and market limitations; this is a worthy area of research to be considered in the future.

8.3 Future outlook

This thesis tackles an issue with practical implications, and adopts a multi-disciplinary approach to solve it by taking advantage of the current availability of data and utilizing techniques from a variety of domains. There are therefore a vast array of possible future directions of research. They can be classified in two main categories: (i) directions aiming at improving the present thesis work, (ii) directions aiming at tackling related subjects which constitute a natural extension of the thesis.

Under the first “improvement” class of potential future routes lie the following:

- *Uncertainty propagation.* The uncertainty propagation resulting from the use of consecutive estimation steps is an important aspect to tackle in order to improve the general methodology. Note that the uncertainty attached to GIS operations should also be considered in the process. First, it requires an assessment strategy. It is not trivial, yet achievable using the prediction errors computed (with Quantile Random Forests), and either (i) a thorough analytical computation, based on the used models and a priori assumptions on the distributions of the considered variables, either (ii) a computational approach such as Monte Carlo methods. Once assessed, one can attempt to reduced the induced uncertainty by simplifying or decreasing the number of estimation steps. The accuracy lost by the simplification can be balanced out by the decrease of uncertainty propagated through the method. This is an interesting trade-off to investigate, which, however, may require more output data, allowing for a more “direct” and less hierarchical estimation.
- *Take advantage of new data / design new strategies based on this data.* The availability of data is currently growing exponentially. In fact, the availability data at the beginning of this doctoral study, for Switzerland in particular, was significantly lower than it is now, as this conclusion is being written. It is thus naturally that the present data-driven work would benefit from the use of all new sources of data, in terms of accuracy (e.g. by using larger training datasets and additional features for ML models) and methodology, by redesigning new approaches for the estimations tackled in the thesis. For example, the estimation of the available area for PV installation could greatly be improved with newly available building rooftop data (<https://shop.swisstopo.admin.ch/en/products/landscape/build3D2>); and the geothermal potential could surely benefit from a soon available precise geological cover data (https://shop.swisstopo.admin.ch/en/products/maps/geology/GC_VECTOR). As this type of data is getting more and more widely available, the resulting methodology would likely be still generalizable to other locations than the Swiss territory. Also, the present results could be validated with some of this new data.

Under the second “related” class of potential future routes lie the following:

- *Tackle higher levels of potential.* As discussed earlier in the conclusion, a natural extension of the present thesis would be to further study the levels of potential which were not considered (for time constraints, or lack of knowledge in a domain which is not the focus of the thesis). This would include notably a geographical and technical potential study for wind and geothermal energies, as well as an economic and implementation potential methodology for solar energy.
- *Confront the estimated potential with precise demand values.* Another natural step is the comparison of the presently estimated potential values with current or forecasted demand loads, in order to clearly assess the potential contribution of each renewable energy and shape efficient solutions to respond to the daily, monthly or yearly energy needs. Although a canton aggregated comparison with demand was performed in the thesis for the PV solar potential, it is a short global analysis which does not bring significant information to draw conclusions at urban or even commune scale. Also, it should be reminded that the 200×200 [m²] pixel scale was partly adopted to match the typical resolution offered by demand data, which cannot be too high due to privacy issues. This is therefore a worthy topic to pursue in future research.

- *Optimize the hybrid combination of multiple energies.* Following the previous point, a last topic of interest is the optimization of the renewable energy use. Renewable energies are, in practice, used together in a hybrid system in order to meet the demand in an optimal way. The estimated potential values can therefore be used as inputs for an optimization to be performed spatially based on demand loads estimated at multiple time resolutions. This could notably help the design of future efficient energy scenarios. Also, the potential for other sustainable sources of energy, such as solar thermal and biomass, which can be assessed following the same methodology presented in this thesis, therefore adding to the renewable energy mix.

Appendices



Data presentation

The proposed strategies to extract the potential values for the considered renewable energies use several datasets offering information on various domains which define the specificities of Switzerland. These datasets are processed in multiple ways to extract information, gather examples, train models and eventually provide estimations for all the required variables in the computation of the potential. As a result, they are at the core of this thesis, and their availability ultimately drives the complexity and the multiple steps of the suggested methodologies. This appendix presents all the data sources used in the thesis, with their reference along with their type, characteristics and some illustrations. When available, a link is provided along with the reference to access the data.

A.1 Time series data

Meteorological times series. Monthly averaged measurements for multiple meteorological variables, including precipitation (P), air temperature (T), sunshine duration (SD), and cloud cover (CC), air pressure (AP), and snow depth (SND) (cumulated fresh daily sum over the month), over different periods of time. P is in mm of rain [mm], T is in degree Celsius [$^{\circ}\text{C}$], SD is in hours, CC is in percentage of the sky, AP is in hPa and SND is in cm. The locations of the stations for each variable are shown in Figure A.1 and A.2. Details on their characteristics are given in Table A.1.

Wind speed. Time series for monthly average wind speed (scalar values) from multiple measurement stations, at a height of 10m, in m/s. The locations of the stations for each variable are shown in Figure A.2. Details are given in Table A.1.

Solar radiation [SoDa]. Time series for solar horizontal global (G_h), direct (G_D) and extra-terrestrial solar irradiance (G_{oh}) (in W/m^2), derived from satellite images, and originally at a time resolution 15 min, from the SoDa database. The locations of the stations for each variable are shown in Figure A.1. Details concerning the data characteristics are given in Table A.1.

Ground temperature. Hourly ground temperature measurement data available for 47 stations in total across Switzerland, in $^{\circ}\text{C}$, at multiple depths: 5cm, 10cm, 20cm, 50cm, 100cm. The measurements were performed at various times, from 2000 to 2018. Locations of stations are given in Figure A.2. Details concerning the data characteristics are given in Table A.1.

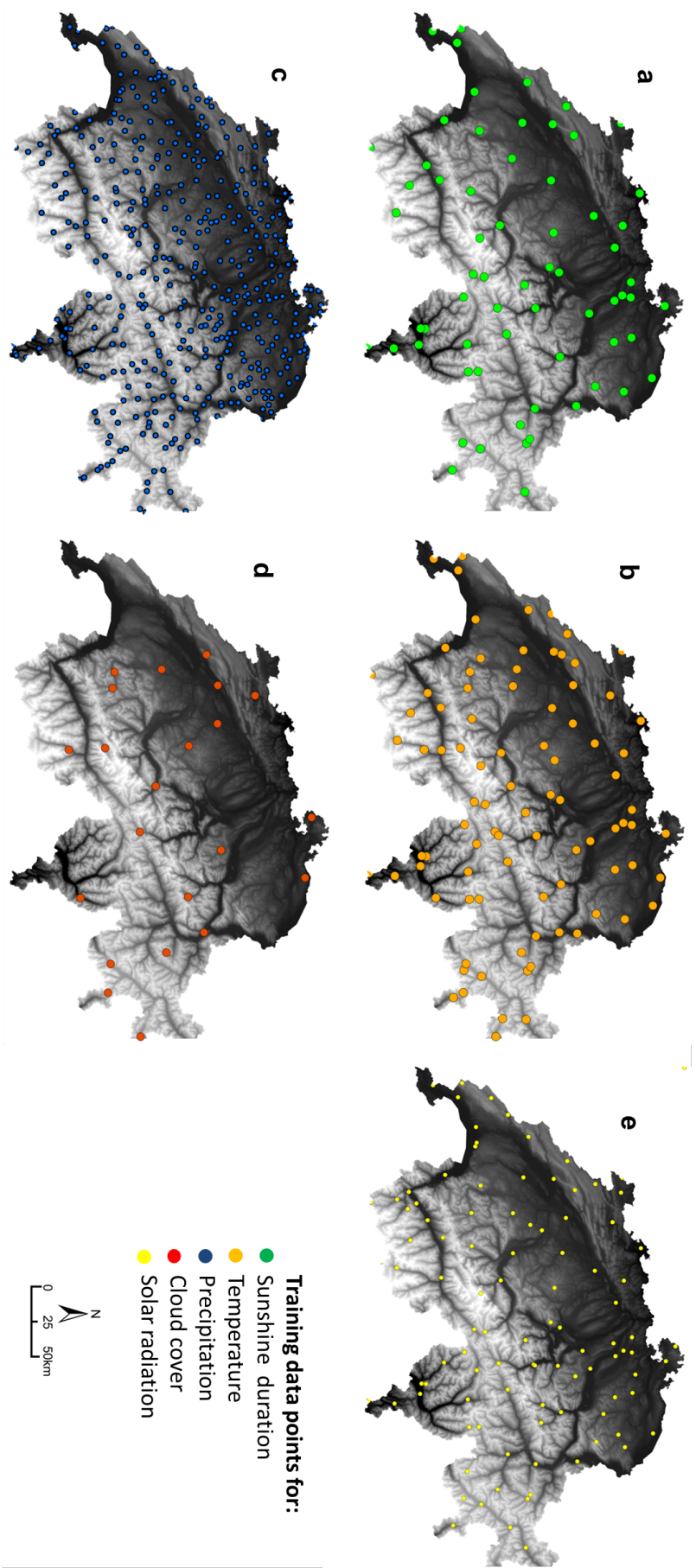


Figure A.1: Locations of measurement stations for weather data, used for training the weather models. (a) solar radiation, (b) sunshine duration, (c) precipitation, (d) cloud cover, (e) temperature.

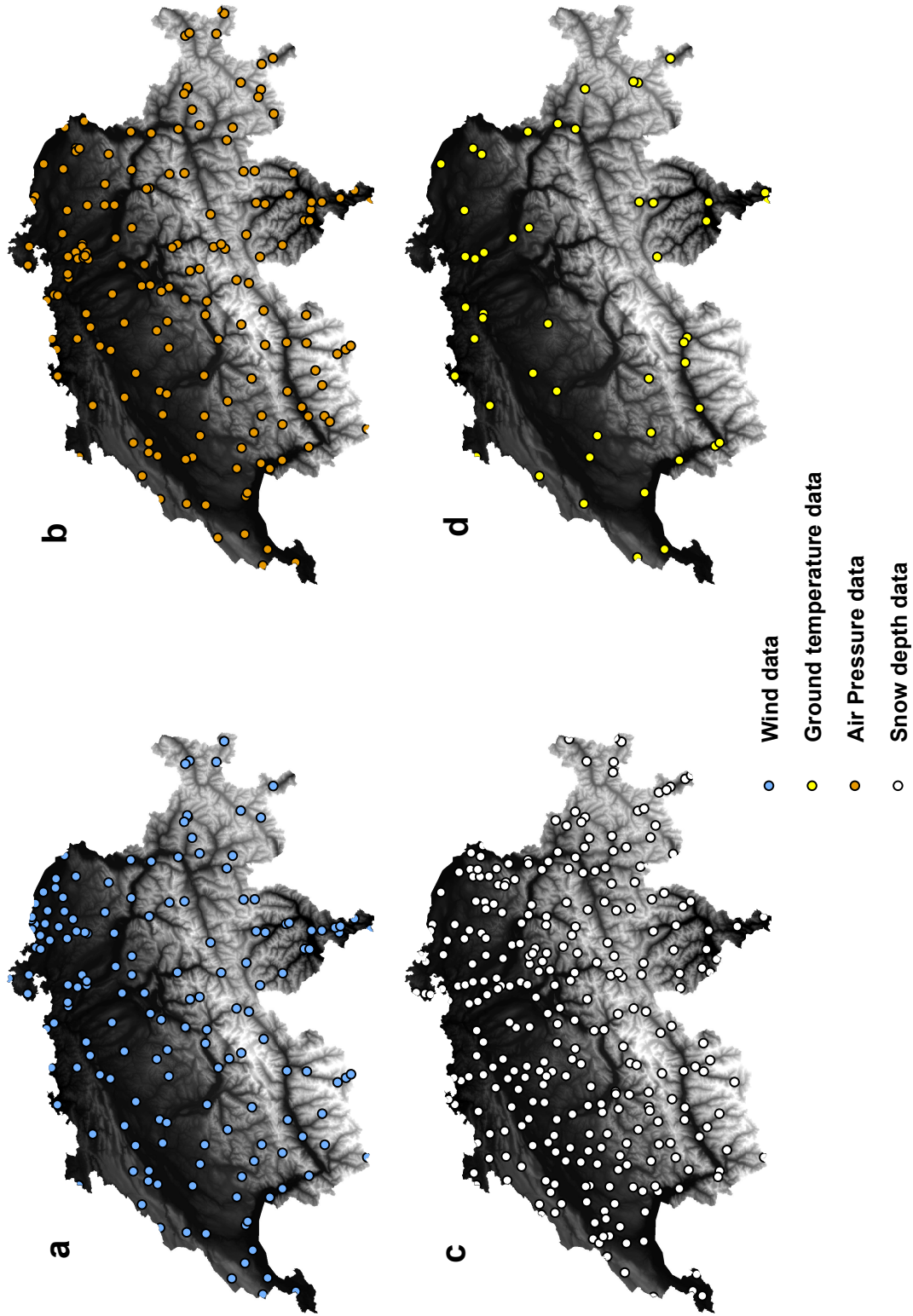


Figure A.2: Locations of additional measurement stations for data, used for training various models. (a) Wind, (b) Air pressure, (c) Snow cover, (d) Ground temperature.

Soil moisture. SMAP (Soil Moisture Active Passive) satellite data from NASA offering surface soil moisture (approximately 0-5 cm) in cm^3/cm^3 derived from brightness temperatures and sigma nought measurements. The values are available on a fixed 3 km and 1 km EASE-Grid 2.0. The data was available for various months and various years, covering a different area for each month and year available. Details concerning the data characteristics are given in Table A.1.

A.2 Raster data

Digital Elevation Models. Raster data offering altitude values at different spatial resolutions, and accounting or not for various obstacles. Digital Elevation Models used in the present thesis include

- Digital Elevation Model (DEM): Basic altitude model of Switzerland, accounting only for the terrain.
- Digital Height Model (DHM): More precise altitude model accounting only for the terrain.
- Digital Terrain Model (DTM): High resolution altitude model accounting only for the terrain, with the best accuracy.
- Digital OrthophotoMap (DOM, also referred to as the Digital Surface Model or DSM): High resolution altitude model accounting for all obstacles, including trees, buildings, etc.

Illustrations concerning elevation models are given in Figures A.3 and A.4 Details concerning the data characteristics are given in Table A.1.

Building typology statistics. Building-related data offering information on various typologies, including period of construction, energies used, etc. The description is given in section 6.2.2, chapter 7. Details concerning the data characteristics are given in Table A.1.

A.3 Vector polygon data

Swiss administrative boundaries (swissBOUNDARIES^{3D}). Vector polygon data offering all administrative boundaries of Switzerland (national, cantonal, district and municipal/communal boundaries).

CORINE land cover. Land cover GIS polygon data defining the use of land in Switzerland. It is based on the European CORINE database, and was re-computed at a high resolution over Switzerland. The resulting polygons are illustrated in Figure A.5. Details concerning the data characteristics are given in Table A.2.

Building vector data. Several GIS building vector data were used in the thesis, regarding the footprint, rooftops or facades of buildings across Switzerland, in polygon or polyline format. They include:

- **Building clusters [VEC25].** Footprint polygons for building clusters (gatherings of buildings).

Table A.1: Time series and raster datasets used in the thesis, along with their characteristics and reference.

Data	Data type	Region	Inputs	Period	Time resolution	Spatial resolution	Error	Source	Chap. 4 [Wind]	Chap. 5 [Geoth.]	Chap. 6 [Solar 1]	Chap. 7 [Solar 2]
Sunshine duration	Time series	Switzerland	Meteo stations	1981-2010	Monthly means	66 stations	Negligible	MeteoSwiss [259, 297]	✓	✓	✓	✓
Precipitation	Time series	Switzerland	Meteo stations	1981-2010	Monthly means	417 stations	4-40%	MeteoSwiss [259, 297]	✓	✓	✓	✓
Temperature	Time series	Switzerland	Meteo stations	1981-2010	Monthly means	91 stations	Negligible	MeteoSwiss [259, 297]	✓	✓	✓	✓
Cloud Cover	Time series	Switzerland	Human observations	1981-2010	Monthly means	23 stations	NA	MeteoSwiss [259, 297]	✓	✓	✓	✓
Air Pressure	Time series	Switzerland	Meteo stations	2000-2017	Daily means	257 stations	NA	MeteoSwiss [259, 297]	✓			
Snow depth	Time series	Switzerland	Meteo stations + Human observations	>2000	Daily sum	291 stations	Negligible	MeteoSwiss [259, 297]		✓		
Ground temperature (various depths)	Time series	Switzerland	Measurement stations	>2000	Hourly	47 stations	$\pm 0.1-0.3^{\circ}\text{C}$	MeteoSwiss [259, 297]		✓		
Soil Moisture (various depths)	Time series	World	Satellite images	>01/04/2015	30 sec.	$3\text{km} \times 3\text{km}$	$\pm 0.05\text{m}^3/\text{m}^3$	NASA SMAP [298]		✓		
Wind speed	Time series	Switzerland	Meteo stations	>2000	Monthly means	197 stations	NA	MeteoSwiss [259, 297]	✓			
Solar radiation (HelioClim3)	Time series	World	Satellite images (Meteosat 8, 9)	>2004	15 min	3 km	7.5%	SoDa [258, 299]			✓	✓
Digital Elevation Model (DEM)	Raster	Switzerland	Elevation maps + aerial images	NA	NA	$250\text{m} \times 250\text{m}$ upsampled to $200\text{m} \times 200\text{m}$	$\pm 2\text{m}$	Swisstopo [260, 300]			✓	
Digital Height Model (DHM)	Raster	Switzerland	Elevation maps + aerial images	NA	NA	$25\text{m} \times 25\text{m}$	$\pm 2\text{m}$	Swisstopo [260, 300]	✓	✓		✓
Digital Surface Model (DOM)	Raster	Switzerland	LiDAR points	NA	NA	$2\text{m} \times 2\text{m}$	$\pm 0.5\text{m}$	Swisstopo [260, 300]	✓		✓	✓
Digital Terrain Model (DTM)	Raster	Switzerland	LiDAR points	NA	NA	$2\text{m} \times 2\text{m}$	$\pm 0.5\text{m}$	Swisstopo [260, 300]	✓			
Building typology data (GEOSTAT)	Raster	Switzerland	Building/housing census data	2014	NA	$100\text{m} \times 100\text{m}$	NA	FSO [255]			✓	

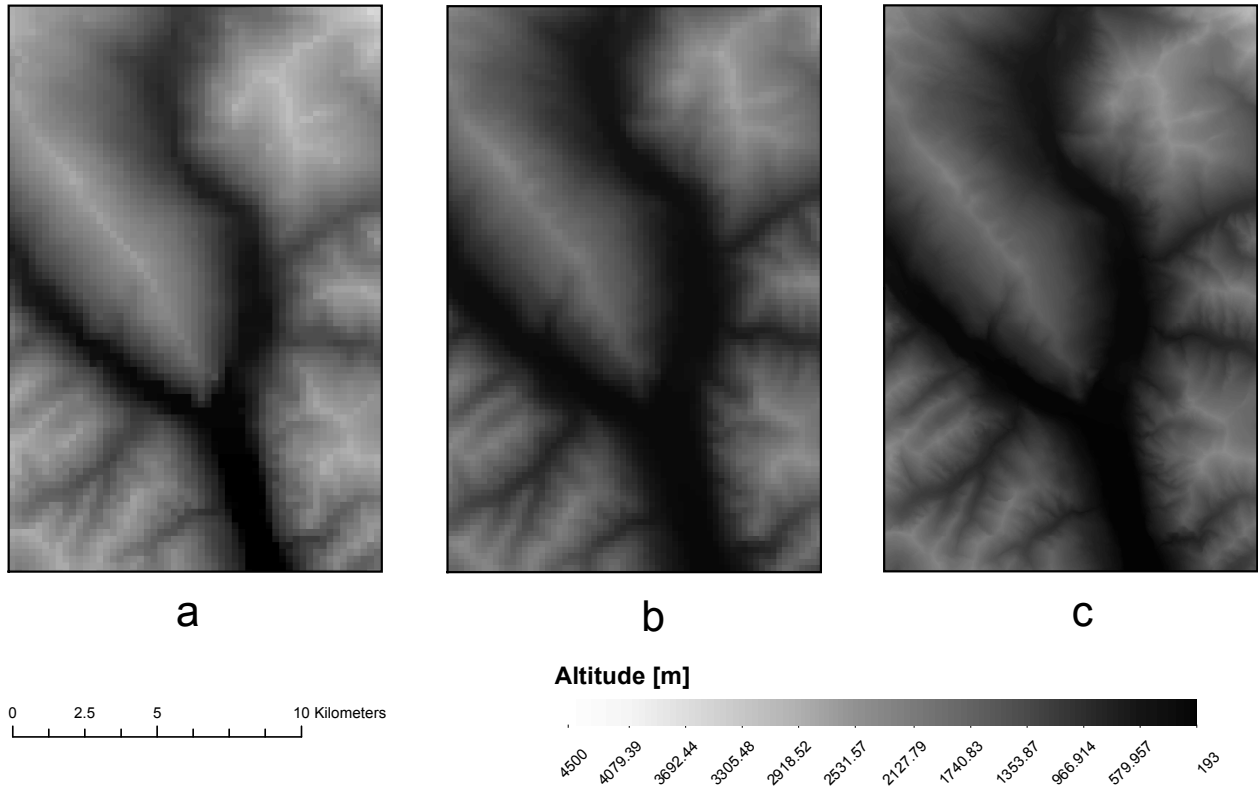


Figure A.3: Illustration for some of the elevation models available for Switzerland. **(a)** Digital Elevation Model (DEM), at a resolution of 250×250 [m²], **(b)** Digital Height Model (DHM), downsampled to a resolution of 200×200 [m²], **(c)** Digital Height Model (DHM), at a resolution of 25×25 [m²],

- **Building rooftops for SON zone [swissBUILDINGS3D/Sonnendach].** Polygons defining rooftops in the zone covered by the Sonnendach project (SON zone, as defined in chapter 7). They notably include the slope and aspect values of all rooftops in this zone.
- **Building footprint [TLM3D].** Footprint polygons for buildings.
- **Building rooftops for Geneva [SITG rooftops].** High-resolution polygons for building rooftops in the canton of Geneva (GEN zone, as defined in chapter 7).
- **Building superstructures for Geneva [SITG superstructures].** High-resolution polygons for the superstructures (chimneys, HVAC systems, etc.) present over rooftops in the canton of Geneva (GEN zone, as defined in chapter 7).
- **Building facades [swissBUILDINGS3D/Sonnendach].** Polyline data for the walls of buildings in Switzerland, covering a certain zone (as covered by the Sonnendach project at the time of the wind energy study, defined in chapter 4). It notably includes the facade area of each building wall in this zone.

Illustrations concerning building vector data are given in Figures A.6 and A.7

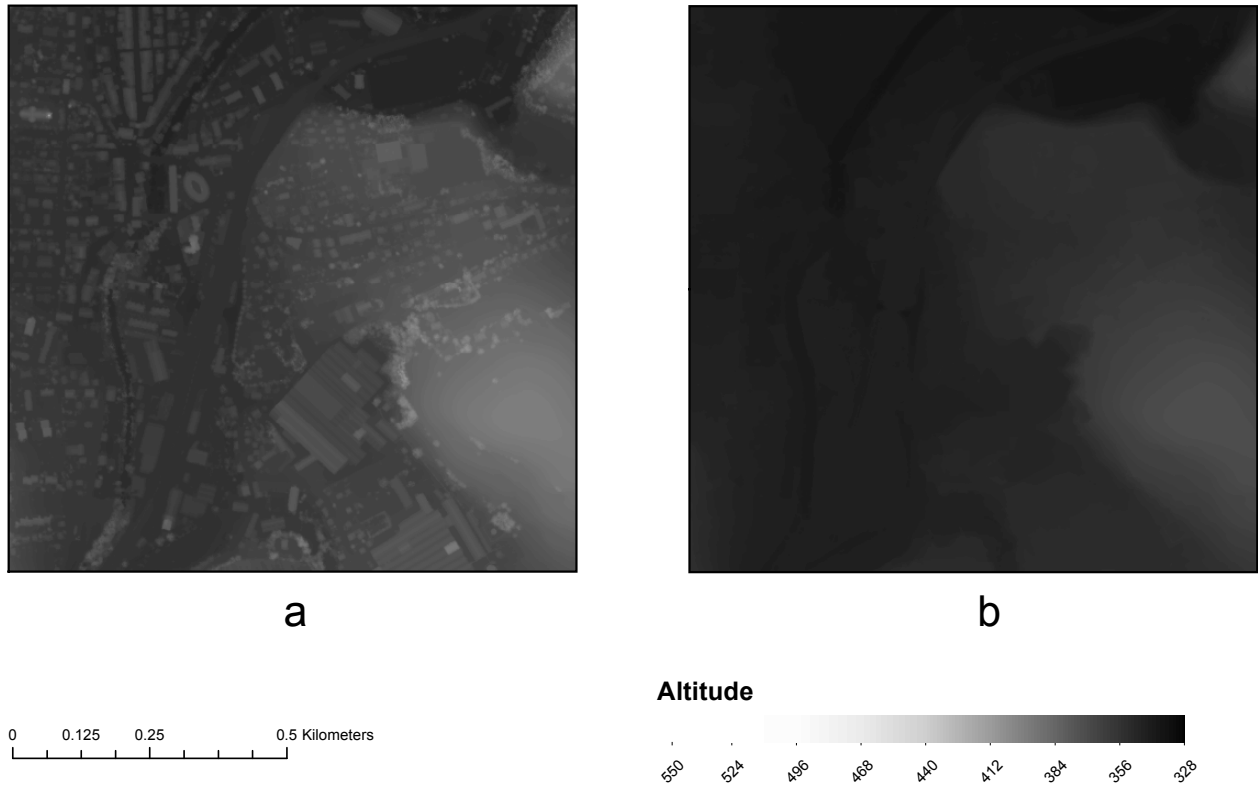


Figure A.4: Illustration for the most precise elevation models available for Switzerland. **(a)** Digital Orthophoto Map (DOM), at a resolution of 2×2 [m²], **(b)** Digital Terrain Model (DEM), at a resolution of 2×2 [m²], for the same area than **(a)**.

Geology cover [GK500]. Polygons offering information regarding geological formations the uppermost layer of the ground in Switzerland, surface geological formations. Various categories are covered by the data (as presented in 5.2.2) The GK500 polygons are shown in Figure 5.1 (section 5.2.1). Details concerning the data characteristics are given in Table A.2.

A.4 Vector point data

Building information [RegBL]. Point data offering building typology statistics information in Switzerland, for each building separately. Details on the content of the data are given in section 7.2.2 (chapter 7), and details concerning the data characteristics are given in Table A.2.

Soil texture [NABODAT]. Measurement data for soil texture information of the ground, notably the fraction of sand, silt and clay in soils. Details on the definition of soil texture are given in section 3.2.2, details on the content of the data are given in section 5.2.2 (chapter 5), and details concerning the data characteristics are given in Table A.2.

Vertical Electrical Soundings [VES]. Vertical Electrical Sounding data, gathered from many studies in Switzerland through the years. Locations of VES points are given in Figure 5.8 in

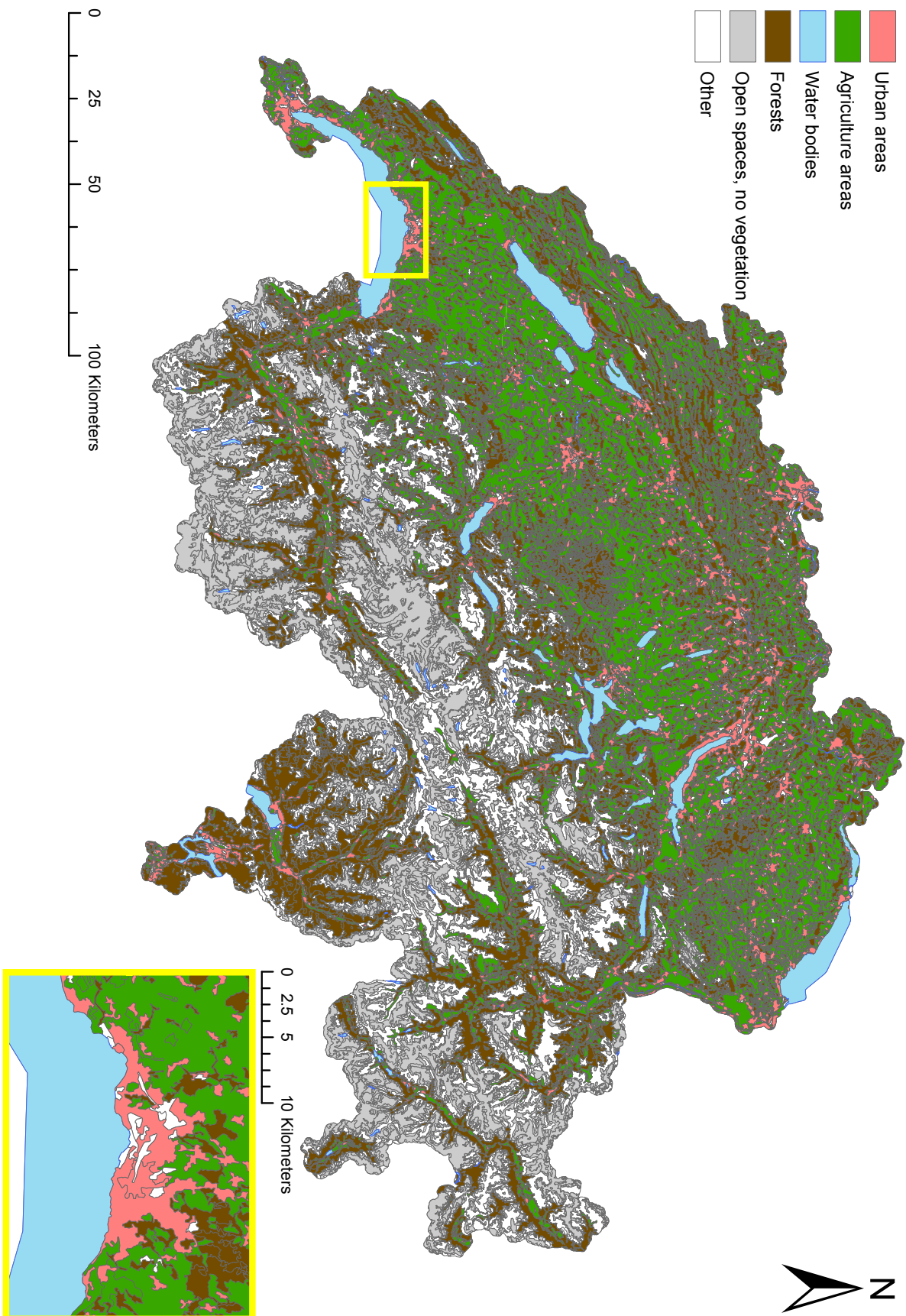


Figure A.5: Corine polygons defining the land use in Switzerland. Land use categories are simplified for illustration purposes.

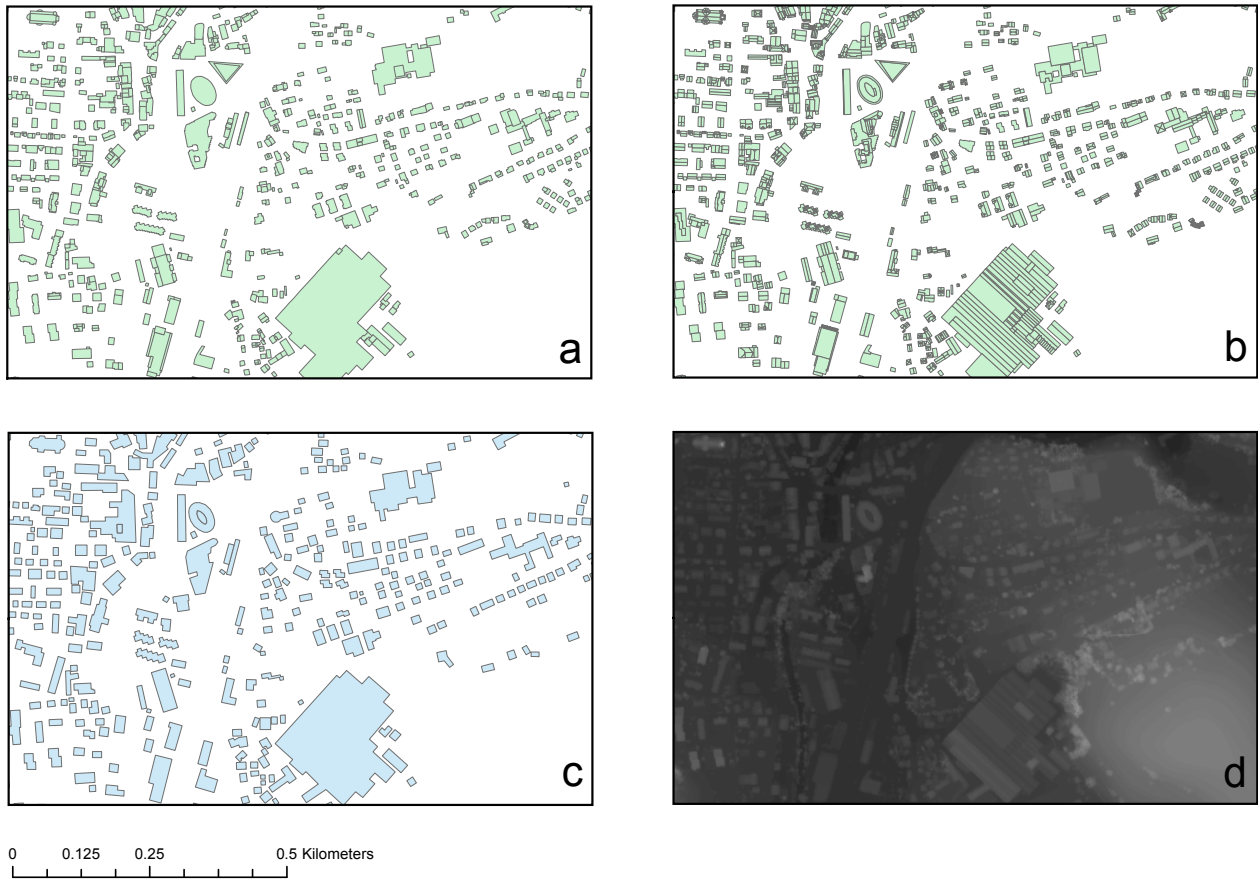


Figure A.6: Illustration for some of the vector polygon building data available in Switzerland. (a) Building footprint data (TLM3D), (b) Building rooftop data (swissBUILDINGS3D/Sonnendach), (c) Building clusters (VEC25), (d) Digital Orthophoto Map (DOM).

section 5.4.1. Details on the definition of VES data are given in section 3.2.3, details on the content of the VES Swiss data are given in 5.4.1 and details on the characteristics of this data are given in Table A.2.

Table A.2: Vector polygons, points and other datasets used in the thesis, along with their characteristics and reference. (FSO: Federal Statistical Office, SITG: Systeme d'Information du Territoire Genevois.)

Data	Data type	Region	Inputs	Last updated	Error	Source	Chap. 4 [Wind]	Chap. 5 [Geoth.]	Chap. 6 [Solar 1]	Chap. 7 [Solar 2]
Swiss administrative boundaries (swissBOUNDARIES3D)	Vector polygons	Switzerland	Cadastral Survey + Federal Statistical Office information	2015	±3-8m	Swisstopo [260]			✓	
CORINE land cover data (Switzerland)	Vector polygons	Switzerland	aerial images + various vector and raster layers	2012	±3-8m	WSL [254]	✓		✓	
Building clusters data (VECTOR25)	Vector polygons	Switzerland	Pixel map +aerial photos	2008	±3-8m	Swisstopo [260, 300]			✓	✓
Building roof data (SITG)	Vector polygons	Geneva canton	Numerical 3D buildings	2011	±30cm	SITG [261]			✓	✓
Roof super-structures (SITG)	Vector polygons	Geneva canton	Numerical 3D buildings	2011	±30cm	SITG [261]			✓	✓
Building data (swissBUILDINGS3D Sonnedach)	Vector polygons	800 communes around Zurich	LiDAR + 3D buildings	2015	±30-50cm	Swisstopo [260, 266, 281]				✓
Building Facades (swissBUILDINGS3D, Sonnedach)	Vector polylines	Parts of Switzerland	LiDAR + 3D buildings	2018	±30-50cm	Swisstopo [260, 266, 281]	✓			
Building footprint data (TLM3D)	Vector polygons	Switzerland	LiDAR + 3D buildings	2018	±30-50cm	Swisstopo [260, 281]	✓			
Geology cover polygons (GK500)	Vector polygons	Switzerland	Geological Atlas of Switz. + other maps	2014	±0.02m	Swisstopo [260, 300]		✓		
Building information data (RegBL)	Vector points	Switzerland	Building/housing census data	2015	NA	FSO [255, 301]				✓
Soil texture (NABODAT)	Vector points	Switzerland	Measurements (Probes)	2017	NA	FOAG [302]		✓		
Vertical Electrical Soundings	Vector points	Switzerland	Measurements	2008	Coords: ±10-30m	SGPK [210]		✓		
Electrical/Thermal resistivity data	Tabled experimental values	India	Experimental Measurements ("electrical res. box")	2004/2010	±4%	various studies [90, 91]		✓		



Figure A.7: Illustration for some of the building vector polygon data available for the canton of Geneva. **(a)** Building rooftops data for Geneva canton (SITG rooftops), **(b)** Building superstructures for Geneva canton (SITG superstructures).

B

Classes in GeoCover data (GK500)

Table B.1: Hydrogeology [HYDRO] categories in GK500 dataset. N_{GK500} is the number of polygons for each category and Area is the total area spanned by all polygons of the category.

Description	N_{GK500}	Area [km ²]
No information	1	69.8
Areas without productive aquifer reservoirs	2930	10816.72
Areas without productive reservoirs	250	508.32
Surface water	250	1377.6
Glacier, Neve	377	866.28
Aquifer tanks in karstifiable coherent rocks	1796	6144.08
Low productive aquifers	3218	7758.64
Low productive aquifers in unquantifiable, cracked and porous coherent rocks	2945	10608.32
Productive aquifers partly out of valley bottoms	1112	2061.04
Highly productive aquifer reservoirs of valley bottoms	441	1243.24

Table B.2: Geological period [PERIOD] categories in GK500 dataset. N_{GK500} is the number of polygons for each category and Area is the total area spanned by all polygons of the category.

Description	N_{GK500}	Area [km ²]
No information	1229	4976.52
Carboniferous	220	1040.04
Cretaceous	1264	4331.08
Devonian	86	762.6
Jurassic	1772	6826.72
Lower Paleozoic	126	642.8
Permian	342	770.64
Permian-Cretaceous	100	100.84
Quaternary	5541	13733.56
Tertiary	1712	6815.68
Triassic	928	1453.56

Table B.3: Aquifer productivity [PRODUCTIV] categories in GK500 dataset. N_{GK500} is the number of polygons for each category and Area is the total area spanned by all polygons of the category.

Description	N_{GK500}	Area [km ²]
Barely exploitable, usually in fine sands	250	508.32
Surface water	250	1377.6
Glacier, Neve	377	866.28
Not locally or barely exploitable	2930	10816.72
Not very productive	2461	9197.0
Not very productive, in the moraines	2263	6507.36
Productive, variable or low productivity	1796	6144.08
Variable productivity	484	1411.32
Variable productivity in loamy gravels	956	1321.08
Usable saturated area for a depth of 10 to 20 m	284	771.68
Usable saturated area for a depth of 2 to 10 m	1112	2061.04
Usable saturated area for a depth of more than 20 m	157	471.56

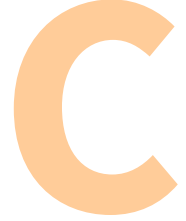
Table B.4: Rock formation types [TYPE ROCHE] categories in GK500 dataset. N_{GK500} is the number of polygons for each category and Area is the total area spanned by all polygons of the category.

Description	N_{GK500}	Area [km ²]
Watercourses, lakes	250	1377.6
Glaciers, snowfields	377	866.28
Magmatic rocks	323	1530.64
Unconsolidated rocks	5023	11630.76
Metamorphic rocks	2078	7511.0
Sedimentary rocks	5269	18537.76

Table B.5: Rock/soil types [LITH PET] categories in GK500 dataset. N_{GK500} is the number of polygons for each category and Area is the total area spanned by all polygons of the category. $N_{\text{soil samples}}$ is the number of NABODAT soil texture points available in each category.

Description	N_{GK500}	$N_{\text{soil samples}}$	Area [km ²]
Amphibolites with diorite passages or a hornblende gneiss	226	5	287.48
Slates with intercalations of dolomites, corneules, gypsum, limestone and sandstone	6	0	27.0
Phyllite slates with intercalations of sandstone and pudding stones	82	0	309.64
Ferruginous Argillites	49	7	15.12
Breccia or limestone pudding stones	7	8	22.96
Breccia and pudding stones	4	0	30.04
Sandy limestones with siliceous limestones with marly shale levels	224	8	761.64
Silicious limestones	163	4	840.8
Limestones, sometimes marbled	112	4	103.68
Conglomerates and breccia rich in sericite	38	0	105.68
Rivers, lakes	250	0	1377.6
Diorites and gabbros	23	1	64.52
Dolomies and Corneules	175	0	205.44
Dolomies with gypsum levels	47	4	42.44
Landslide and scree deposits	783	89	903.68
Glaciers	377	0	866.28
Gneiss and micaschists rich in biotite and muscovite	350	33	1525.88
Gneiss rich in biotite or muscovite, sometimes chloritic, sometimes with calc-silicates	7	0	124.96
rocks or quartzitic horizons (hornfelse)			
Gneiss rich in feldspar	203	13	992.32
Gneiss rich in feldspar, schistous with sericite, epidote and chlorite formation	88	3	443.12
Gneiss with two micas or biotite rich in feldspar and varied structures	98	0	679.4
Gneiss with two micas or biotite rich in feldspar, platelets	5	2	268.28
Gneiss with two micas or biotite, rich in feldspar, mainly homogeneous	28	0	252.56
Gneiss with two micas or biotite, with white, often green (phengite)	22	0	137.8
Gneiss with Sericite-chloritic schist	291	2	1278.0
Granite with passages of quartzic diorite or quartzic syenite	111	3	823.84
Granite with sericite, epidote and chlorite	98	6	526.52
Gravel and sand	754	740	1500.08
Gravel and sand, sometimes clayey or silty	738	636	1597.96
Hard and compact sandstone with marly shale and limestone phyllites	189	43	1096.04
Sandstone and marls with low to moderately consolidated pudding stone levels	207	70	945.68
Quartz sandstone with sandy slates	15	0	7.2
Glauconite siliceous sandstone and echinoderm debris	52	0	126.2

Description	N_{GK500}	$N_{\text{soil samples}}$	Area [km ²]
Sandstone mainly calcareous and porous, with marl levels	362	396	1622.92
Sandstone with marl levels	84	20	482.8
Dolomitic marble	23	0	19.04
Marly and shale clay with limestone bench, dolomite and sandstone	162	145	226.32
Marls with breach levels with hard and sandy shells	1	0	34.96
Well-consolidated sandstone slabs	51	0	166.76
Marbles with sandstone levels with shell-rich breccia	12	23	35.84
Marls with sandstone levels, conglomerates or poorly consolidated pebbles	376	698	1146.72
Quartz Phyllites	1	0	0.48
Phyllites with limestone micaschists	206	0	765.84
Phyllites with limestone micaschists with limestone levels and marbled dolomites	17	0	36.68
Phyllites with limestone micaschists with greenstone levels	1	0	1.04
Quartz porphyry	59	0	73.36
Porphyrites and porphyry tuffs	15	0	19.6
Poudingues à brèches avec arkoses et grès	80	0	174.04
Pudding stones with breccia with arkoses and sandstone	66	22	1406.8
Peridotites and olivine rocks	18	0	8.48
Quartzites	70	0	112.12
Radiolarites	11	0	3.72
Dolomitic rock, sometimes with limestone levels	480	19	817.04
Limestone rocks in general, often with marly intercalations	1175	560	4636.2
Limestone rocks with dolomitic levels	48	64	184.8
Limestone rocks with important levels of marl, shale and marly limestone	628	159	1370.96
Limestone rocks, often marly	89	0	480.68
Volcanic and pyroclastic rocks	5	0	9.64
Sand, gravel, pebbles and blocks	454	45	457.28
Marly shale with calcareous phyllites and interbedded sandstone	382	79	1226.52
Marly shale with limestone phyllites with tuffitic sandstone levels	42	0	90.48
Green shale with passages of eruptive rocks or eclogites	173	5	252.76
Serpentinities	78	13	84.68
Clayey silts, with clay with sand levels	264	499	686.24
Sand silts with gravel and blocks	1962	1614	6331.44
Silts with silty sands, often clayey, mainly calcareous	68	170	154.08
Syenite	12	0	13.16
Amphibolite and gneiss mixing zone	23	0	30.72



Extra calculus for chapter 7

C.1 Statistical independence of rooftop slope, aspect, and roof type.

Two discrete random variables U and V are said to be statistical independent (we note $U \perp V$) if $\mathbb{P}(U = u | V = v) = \mathbb{P}(U = u)$ for all $u \in S_U$ and $v \in S_V$, where S_U and S_V are the support of U and V , gathering all possible realizations (values) of the two random variables.

We want to verify the statistical independence between the three following variables: main slope (most frequent tilt), main aspect (most frequent direction), and roof type for building rooftops in OOSG zone in Switzerland. Let us note S , D , and T the random variables respectively corresponding to the three mentioned variables. All buildings in OOSG zone are considered, and the marginal and conditional probabilities are computed for the three couples of random variables. The average conditional probabilities are also computed in order to be compared with the marginal probabilities. The results are shown in Tables C.1, C.2, and C.3. Note that in order to simplify the notations, the possible values for the three variables have been replaced by simple class values (e.g. $S = 15^\circ$ becomes $S = 1$, $S = 25^\circ$ becomes $S = 2$, etc.). Also, note that the three events $S = 0$, $D = 0$ and $T = 0$ all correspond to a flat roof. Therefore, as a roof cannot be flat and characterized by a non-zero slope at the same time, or flat and characterized by a particular direction at the same time, all the conditional probabilities involving one of the three mentioned events are null, except $S = 0 | D = 0$, $S = 0 | T = 0$ and $T = 0 | D = 0$ (whose probabilities equal 1), as it can be seen from the first column and first row in the three Tables C.1, C.2 and C.3. It can be observed that the marginal probabilities are relatively close to the average conditional probabilities for each possible value of S or T .

Therefore, it can be inferred from the 3 tables respectively that:

Table C.1: Marginal and conditional probabilities for S (main slope variable) and $S | D$ (D being the main aspect variable). **Mean Cond.** is the average conditional probability $\mathbb{P}(S = \beta | D = \gamma)$ for each β value.

\mathbb{P}	$D = 0$	$D = 1$	$D = 2$	$D = 3$	$D = 4$	$D = 5$	$D = 6$	$D = 7$	$D = 8$	$D = 9$	Mean Cond.	Marginal ($S = \beta$)
$S = 0 D$	1.0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	N/A	0.179
$S = 1 D$	0.0	0.197	0.188	0.178	0.182	0.180	0.172	0.191	0.170	0.186	0.183	0.150
$S = 2 D$	0.0	0.428	0.425	0.461	0.441	0.448	0.440	0.424	0.407	0.421	0.433	0.357
$S = 3 D$	0.0	0.283	0.292	0.273	0.286	0.283	0.296	0.296	0.321	0.297	0.292	0.238
$S = 4 D$	0.0	0.091	0.094	0.087	0.091	0.088	0.092	0.088	0.101	0.095	0.092	0.075
$S = 5 D$	0.0	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table C.2: Marginal and conditional probabilities for S (main slope variable) and $S | T$ (T being the roof type variable). **Mean Cond.** is the average conditional probability $\mathbb{P}(S = \beta | T = c)$ for each β value.

\mathbb{P}	$T = 0$	$T = 1$	$T = 2$	$T = 3$	$T = 4$	$T = 5$	Mean Cond.	Marginal ($S = \beta$)
$S = 0 T$	1.000	0.000	0.000	0.000	0.000	0.000	N/A	0.179
$S = 1 T$	0.000	0.210	0.144	0.142	0.189	0.161	0.169	0.150
$S = 2 T$	0.000	0.480	0.353	0.498	0.482	0.468	0.456	0.357
$S = 3 T$	0.000	0.250	0.351	0.295	0.278	0.307	0.296	0.238
$S = 4 T$	0.000	0.059	0.151	0.065	0.050	0.063	0.078	0.075
$S = 5 T$	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table C.3: Marginal and conditional probabilities for T (roof type) and $T | D$ (D being the main aspect variable). **Mean Cond.** is the average conditional probability $\mathbb{P}(T = c | D = \gamma)$ for each c value.

\mathbb{P}	$D = 0$	$D = 1$	$D = 2$	$D = 3$	$D = 4$	$D = 5$	$D = 6$	$D = 7$	$D = 8$	$D = 9$	Mean Cond.	Marginal ($T = c$)
$T = 0 D$	1.0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	N/A	0.179
$T = 1 D$	0.0	0.524	0.528	0.496	0.533	0.527	0.530	0.550	0.461	0.514	0.518	0.427
$T = 2 D$	0.0	0.347	0.361	0.365	0.351	0.350	0.360	0.339	0.390	0.357	0.358	0.293
$T = 3 D$	0.0	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.004	0.003
$T = 4 D$	0.0	0.089	0.075	0.097	0.080	0.085	0.072	0.073	0.102	0.089	0.085	0.069
$T = 5 D$	0.0	0.036	0.032	0.038	0.032	0.034	0.034	0.034	0.042	0.036	0.035	0.029

$$\forall \beta, \forall \gamma, \mathbb{P}(S = \beta | D = \gamma) \approx \mathbb{P}(S = \beta), \text{ so } S \perp D \quad (\text{C.1})$$

$$\forall \beta, \forall c, \mathbb{P}(S = \beta | T = c) \approx \mathbb{P}(S = \beta), \text{ so } S \perp T \quad (\text{C.2})$$

$$\forall c, \forall \gamma, \mathbb{P}(T = c | D = \gamma) \approx \mathbb{P}(T = c), \text{ so } T \perp D \quad (\text{C.3})$$

Eventually, the three variables can be considered to be independent with each other and we can write:

$$\forall \beta, \forall \gamma, \forall c, \mathbb{P}(S = \beta, D = \gamma, T = c) \approx \mathbb{P}(S = \beta) \mathbb{P}(D = \gamma) \mathbb{P}(T = c) \quad (\text{C.4})$$

References

- [1] Adrianus Johannes Maria van Wijk and Jan P Coelingh. *Wind power potential in the OECD Countries*. Department of Science, Technology and Society, Utrecht University, 1993.
- [2] LK Wiginton, Ha T Nguyen, and Joshua M Pearce. “Quantifying rooftop solar photovoltaic potential for regional renewable energy policy”. In: *Computers, Environment and Urban Systems* 34.4 (2010), pp. 345–357.
- [3] Salvador Izquierdo, Marcos Rodrigues, and Norberto Fueyo. “A method for estimating the geographical distribution of the available roof surface area for large-scale photovoltaic energy-potential evaluations”. In: *Solar Energy* 82.10 (2008), pp. 929–939.
- [4] Monique Maria Hoogwijk. “On the global and regional potential of renewable energy sources”. PhD thesis. 2004.
- [5] L Rybach and Th Kohl. “The geothermal heat pump boom in Switzerland and its background”. In: *Proc. International Geothermal Congress*. 2003, pp. 47–52.
- [6] *SCHWEIZERISCHE GESAMTENERGIESTATISTIK 2017*, author=Swiss Federal Office of Energy SFOE.
- [7] Soteris A Kalogirou. “Artificial neural networks in renewable energy systems applications: a review”. In: *Renewable and sustainable energy reviews* 5.4 (2001), pp. 373–401.
- [8] Amit Kumar Yadav and SS Chandel. “Solar radiation prediction using Artificial Neural Network techniques: A review”. In: *Renewable and sustainable energy reviews* 33 (2014), pp. 772–781.
- [9] Cyril Voyant et al. “Machine learning methods for solar radiation forecasting: A review”. In: *Renewable Energy* 105 (2017), pp. 569–582.
- [10] SHI Xingjian et al. “Convolutional LSTM network: A machine learning approach for precipitation nowcasting”. In: *Advances in neural information processing systems*. 2015, pp. 802–810.
- [11] Navin Sharma et al. “Predicting solar generation from weather forecasts using machine learning”. In: *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*. IEEE. 2011, pp. 528–533.
- [12] Kabir Rasouli, William W Hsieh, and Alex J Cannon. “Daily streamflow forecasting by machine learning methods with weather and climate inputs”. In: *Journal of Hydrology* 414 (2012), pp. 284–293.
- [13] Mikhail Kanevski, Vadim Timonin, and Alexi Pozdnukhov. *Machine learning for spatial environmental data: theory, applications, and software*. EPFL press, 2009.
- [14] Ahmet Öztopal. “Artificial neural network approach to spatial estimation of wind velocity data”. In: *Energy Conversion and Management* 47.4 (2006), pp. 395–406.
- [15] Loris Foresti et al. “Learning wind fields with multiple kernels”. In: *Stochastic Environmental Research and Risk Assessment* 25.1 (2011), pp. 51–66.
- [16] Sylvain Robert, Loris Foresti, and Mikhail Kanevski. “Spatial prediction of monthly wind speeds in complex terrain with adaptive general regression neural networks”. In: *International Journal of Climatology* 33.7 (2013), pp. 1793–1804.
- [17] Miloš Marjanović et al. “Landslide susceptibility assessment using SVM machine learning algorithm”. In: *Engineering Geology* 123.3 (2011), pp. 225–234.

- [18] X Yao, LG Tham, and FC Dai. "Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China". In: *Geomorphology* 101.4 (2008), pp. 572–582.
- [19] Biswajeet Pradhan. "A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS". In: *Computers & Geosciences* 51 (2013), pp. 350–365.
- [20] JN Goetz et al. "Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling". In: *Computers & geosciences* 81 (2015), pp. 1–11.
- [21] Dan Assouline, Nahid Mohajeri, and Jean-Louis Scartezzini. "Estimation of Large-Scale Solar Rooftop PV Potential for Smart Grid Integration: A Methodological Review". In: *Sustainable Interdependent Networks*. Springer, 2018, pp. 173–219.
- [22] Dan Assouline, Nahid Mohajeri, and Jean-Louis Scartezzini. "Quantifying rooftop photovoltaic solar energy potential: A machine learning approach". In: *Solar Energy* 141 (2017), pp. 278–296.
- [23] Dan Assouline, Nahid Mohajeri, and Jean-Louis Scartezzini. "Large-scale rooftop solar photovoltaic technical potential estimation using Random Forests". In: *Applied Energy* 217 (2018), pp. 189–211.
- [24] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, NY, USA: 2001.
- [25] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830.
- [26] Vladimir N Vapnik and Alexey J Chervonenkis. "Theory of pattern recognition". In: (1974).
- [27] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.
- [28] Vladimir Naumovich Vapnik. "An overview of statistical learning theory". In: *IEEE transactions on neural networks* 10.5 (1999), pp. 988–999.
- [29] Alex J Smola and Bernhard Schölkopf. "A tutorial on support vector regression". In: *Statistics and computing* 14.3 (2004), pp. 199–222.
- [30] Mervyn Stone. "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the royal statistical society. Series B (Methodological)* (1974), pp. 111–147.
- [31] Cort J Willmott and Kenji Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". In: *Climate research* 30.1 (2005), pp. 79–82.
- [32] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
- [33] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.
- [34] Harris Drucker et al. "Support vector regression machines". In: *Advances in neural information processing systems*. 1997, pp. 155–161.
- [35] Christopher JC Burges. "A tutorial on support vector machines for pattern recognition". In: *Data mining and knowledge discovery* 2.2 (1998), pp. 121–167.
- [36] Vladimir Vapnik. *Statistical learning theory*. 1998. Vol. 3. Wiley, New York, 1998.
- [37] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
- [38] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. "Kernel principal component analysis". In: *International Conference on Artificial Neural Networks*. Springer. 1997, pp. 583–588.
- [39] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. "Kernel k-means: spectral clustering and normalized cuts". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2004, pp. 551–556.

- [40] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. "A practical guide to support vector classification". In: (2003).
- [41] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.
- [42] L. Breiman et al. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984. URL: <https://books.google.ch/books?id=JwQx-WOmSyQC>.
- [43] Gilles Louppe. "Understanding random forests: From theory to practice". In: *arXiv preprint arXiv:1407.7502* (2014).
- [44] Leo Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.
- [45] Yoav Freund, Robert E Schapire, et al. "Experiments with a new boosting algorithm". In: *Icml*. Vol. 96. Citeseer. 1996, pp. 148–156.
- [46] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.
- [47] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.
- [48] Nicolai Meinshausen. "Quantile regression forests". In: *Journal of Machine Learning Research* 7.Jun (2006), pp. 983–999.
- [49] Andy Liaw and Matthew Wiener. "Classification and regression by randomForest". In: *R news* 2.3 (2002), pp. 18–22.
- [50] Fei Tang and Hemant Ishwaran. "Random forest missing data algorithms". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 10.6 (2017), pp. 363–377.
- [51] Alexander Statnikov, Lily Wang, and Constantin F Aliferis. "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification". In: *BMC bioinformatics* 9.1 (2008), p. 319.
- [52] Rich Caruana and Alexandru Niculescu-Mizil. "An empirical comparison of supervised learning algorithms". In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 161–168.
- [53] Joseph O Ogutu, Hans-Peter Piepho, and Torben Schulz-Streeck. "A comparison of random forests, boosting and support vector machines for genomic selection". In: *BMC proceedings*. Vol. 5. 3. BioMed Central. 2011, S11.
- [54] Manish Kumar and M Thenmozhi. "Forecasting stock index movement: A comparison of support vector machines and random forest". In: (2006).
- [55] Miao Liu et al. "Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar". In: *Sensors and Actuators B: Chemical* 177 (2013), pp. 970–980.
- [56] Gijs AM Van Kuik. "The Lanchester–Betz–Joukowsky limit". In: *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology* 10.3 (2007), pp. 289–291.
- [57] David MacKay. *Sustainable Energy-without the hot air*. UIT Cambridge, 2008.
- [58] Roland B Stull. *An introduction to boundary layer meteorology*. Vol. 13. Springer Science & Business Media, 2012.
- [59] Timothy R Oke et al. *Urban climates*. Cambridge University Press, 2017.
- [60] Julieta Silva, Carla Ribeiro, and Ricardo Guedes. "Roughness length classification of Corine Land Cover classes". In: *Proceedings of the European Wind Energy Conference, Milan, Italy*. Vol. 710. 2007, p. 110.

- [61] RW Macdonald, RF Griffiths, and DJ Hall. "An improved method for the estimation of surface roughness of obstacle arrays". In: *Atmospheric environment* 32.11 (1998), pp. 1857–1864.
- [62] Marco Casini. "Small vertical axis wind turbines for energy efficiency of buildings". In: *Journal of Clean Energy Technologies* 4.1 (2016), pp. 56–65.
- [63] Abhishiktha Tummala et al. "A review on small scale wind turbines". In: *Renewable and Sustainable Energy Reviews* 56 (2016), pp. 1351–1371.
- [64] Sandra Eriksson, Hans Bernhoff, and Mats Leijon. "Evaluation of different turbine concepts for wind power". In: *renewable and sustainable energy reviews* 12.5 (2008), pp. 1419–1434.
- [65] TF Ishugah et al. "Advances in wind energy resource exploitation in urban environment: A review". In: *Renewable and sustainable energy reviews* 37 (2014), pp. 613–626.
- [66] AD Peacock et al. "Micro wind turbines in the UK domestic sector". In: *Energy and Buildings* 40.7 (2008), pp. 1324–1333.
- [67] Magdi Ragheb. "Vertical axis wind turbines". In: *University of Illinois at Urbana-Champaign* 1 (2011).
- [68] Andrew R Winslow. "Urban Wind Generation: Comparing Horizontal and Vertical Axis Wind Turbines at Clark University in Worcester, Massachusetts". In: (2017).
- [69] Gerald Müller, Mark F Jentsch, and Euan Stoddart. "Vertical axis resistance type wind turbines for use in buildings". In: *Renewable Energy* 34.5 (2009), pp. 1407–1412.
- [70] Francesco Balduzzi et al. "Feasibility analysis of a Darrieus vertical-axis wind turbine installation in the rooftop of a building". In: *Applied Energy* 97 (2012), pp. 921–929.
- [71] Jorge L Acosta et al. "Performance assessment of micro and small-scale wind turbines in urban areas". In: *IEEE Systems Journal* 6.1 (2012), pp. 152–163.
- [72] DR Drew, JF Barlow, and TT Cockerill. "Estimating the potential yield of small wind turbines in urban areas: A case study for Greater London, UK". In: *Journal of Wind Engineering and Industrial Aerodynamics* 115 (2013), pp. 104–111.
- [73] Rakesh Kumar, Kaamran Raahemifar, and Alan S Fung. "A critical review of vertical axis wind turbines for urban applications". In: *Renewable and Sustainable Energy Reviews* 89 (2018), pp. 281–291.
- [74] PAB James et al. "Implications of the UK field trial of building mounted horizontal axis micro-wind turbines". In: *Energy Policy* 38.10 (2010), pp. 6130–6144.
- [75] Ronit K Singh and M Rafiuddin Ahmed. "Blade design and performance testing of a small wind turbine rotor for low wind speed applications". In: *Renewable Energy* 50 (2013), pp. 812–819.
- [76] William E Glassley. *Geothermal energy: renewable energy and the environment*. CRC Press, 2014.
- [77] Rafid Al-Khoury. *Computational modeling of shallow geothermal systems*. CRC press, 2011.
- [78] Glendon W Gee and James W Bauder. *Particle-size analysis 1. methods of soil analysis*. Soil Science Society of America, American Society of Agronomy, 1986.
- [79] Christopher J Rhodes. "Feeding and healing the world: through regenerative agriculture and permaculture". In: *Science progress* 95.4 (2012), pp. 345–446.
- [80] Henk JL Witte, Guus J Van Gelder, and JD Spitler. "In situ measurement of ground thermal conductivity: a Dutch perspective." In: *Ashrae Transactions* 108.1 (2002), pp. 263–272.
- [81] Carsten Rücker, Thomas Günther, and Florian M. Wagner. "pyGIMLi: An open-source library for modelling and inversion in geophysics". In: *Computers and Geosciences* 109 (2017), pp. 106–123. URL: <http://www.sciencedirect.com/science/article/pii/S0098300417300584>.
- [82] John M Reynolds. *An introduction to applied and environmental geophysics*. John Wiley & Sons, 2011.
- [83] P. Rajeev and J. Kodikara. "Estimating apparent thermal diffusivity of soil using field temperature time series". In: *Geomechanics and Geoengineering* 11.1 (2016), pp. 28–46.

- [84] J. Busby. "Determination of thermal properties for horizontal ground collector loops". In: *Proceedings of the World Geothermal Congress 2015, Melbourne, Australia, 19-25 April 2015*. 2015.
- [85] S. Hurley and R.J. Wiltshire. "Computing thermal diffusivity from soil temperature measurements". In: *Computers & Geosciences* 19.3 (1993), pp. 475–477.
- [86] W.R. van Wijk. *Physics of plant environment*. John Wiley & Sons, 1963.
- [87] J. E. Carson. "Analysis of soil and air temperatures by Fourier techniques". In: *Journal of Geophysical Research* 68.8 (1963), pp. 2217–2232.
- [88] H.S. Carslaw and J.C. Jaeger. *Conduction of heat in solids*. Clarendon Press, Oxford, 1959.
- [89] Devendra Narain Singh, Sneha J Kuriyan, and K Chakravarthy Manthena. "A generalised relationship between soil electrical and thermal resistivities". In: *Experimental Thermal and Fluid Science* 25.3-4 (2001), pp. 175–181.
- [90] S Sreedeeep, AC Reshma, and DN Singh. "Generalized relationship for determining soil electrical resistivity from its thermal resistivity". In: *Experimental Thermal and Fluid Science* 29.2 (2005), pp. 217–226.
- [91] Yusuf Erzin et al. "Artificial neural network models for predicting electrical resistivity of soils from their thermal resistivity". In: *International Journal of Thermal Sciences* 49.1 (2010), pp. 118–130.
- [92] Burkhard Sanner. "Shallow geothermal energy". In: (2001).
- [93] Burkhard Sanner et al. "Current status of ground source heat pumps and underground thermal energy storage in Europe". In: *Geothermics* 32.4-6 (2003), pp. 579–588.
- [94] Eloisa Di Sipio and David Bertermann. "Factors influencing the thermal efficiency of horizontal ground heat exchangers". In: *Energies* 10.11 (2017), p. 1897.
- [95] F TINTI, BMS GIAMBASTIANI, and M MASTROICCO. "Types of Geo-exchanger Systems for Underground Heat Extraction". In: ().
- [96] Babak Dehghan, Altug Sisman, and Murat Aydin. "Parametric investigation of helical ground heat exchangers for heat pump applications". In: *Energy and Buildings* 127 (2016), pp. 999–1007.
- [97] Zeyu Xiong, Daniel E Fisher, and Jeffrey D Spitler. "Development and validation of a Slinky™ ground heat exchanger model". In: *Applied Energy* 141 (2015), pp. 57–69.
- [98] Angelo Zarrella, Antonio Capozza, and Michele De Carli. "Analysis of short helical and double U-tube borehole heat exchangers: A simulation-based comparison". In: *Applied Energy* 112 (2013), pp. 358–370.
- [99] Angelo Zarrella and Michele De Carli. "Heat transfer analysis of short helical borehole heat exchangers". In: *Applied Energy* 102 (2013), pp. 1477–1491.
- [100] Angelo Zarrella, Antonio Capozza, and Michele De Carli. "Performance analysis of short helical borehole heat exchangers via integrated modelling of a borefield and a heat pump: A case study". In: *Applied Thermal Engineering* 61.2 (2013), pp. 36–47.
- [101] Hassen Boughanmi et al. "Thermal performance of a conic basket heat exchanger coupled to a geothermal heat pump for greenhouse cooling under Tunisian climate". In: *Energy and Buildings* 104 (2015), pp. 87–96.
- [102] G Masson, Izumi Kaizuka, and Carlotta Cambie. "Snapshot of global photovoltaic markets 2018". In: *Report IEA PVPS T1-33:2018* 1 (2018).
- [103] A. Angstrom. *On the computation of global radiation from records of sunshine*. URL: <http://www.scopus.com/inward/record.url?eid=2-s2.0-0000522693&partnerID=tZ0tx3y1>.
- [104] AA El-Sebaii et al. "Global, direct and diffuse solar radiation on horizontal and tilted surfaces in Jeddah, Saudi Arabia". In: *Applied Energy* 87.2 (2010), pp. 568–576.
- [105] R.B. Benson et al. "Estimation of daily and monthly direct, diffuse and global solar radiation from sunshine duration measurements". In: *Solar Energy* 32.4 (Jan. 1984), pp. 523–535. URL: <http://www.scopus.com/inward/record.url?eid=2-s2.0-0021183853&partnerID=tZ0tx3y1>.

- [106] CJT Spitters, HAJM Toussaint, and J Goudriaan. "Separating the diffuse and direct component of global radiation and its implications for modeling canopy photosynthesis Part I. Components of incoming radiation". In: *Agricultural and Forest Meteorology* 38.1-3 (1986), pp. 217–229.
- [107] Wanxiang Yao et al. "New models for separating hourly diffuse and direct components of global solar radiation". In: *Proceedings of the 8th International Symposium on Heating, Ventilation and Air Conditioning*. Springer. 2014, pp. 653–663.
- [108] Muhammad Iqbal. *An introduction to solar radiation*. Elsevier, 2012.
- [109] SA Klein. "Calculation of monthly average insolation on tilted surfaces". In: *Solar energy* 19.4 (1977), pp. 325–329.
- [110] Pauli Andersen. "Comments on "calculations of monthly average insolation on tilted surfaces" by SA Klein". In: *Solar Energy* 25.3 (1980), p. 287.
- [111] Marko Gulin, Mario Vašak, and Mato Baotic. "Estimation of the global solar irradiance on tilted surfaces". In: *17th International Conference on Electrical Drives and Power Electronics (EDPE 2013)*. 2013, pp. 334–339.
- [112] Colienne Demain, Michel Journée, and Cédric Bertrand. "Evaluation of different models to estimate the global solar radiation on inclined surfaces". In: *Renewable Energy* 50 (2013), pp. 710–721.
- [113] Ali Mohammad Noorian, Isaac Moradi, and Gholam Ali Kamali. "Evaluation of 12 models to estimate hourly diffuse irradiation on inclined surfaces". In: *Renewable energy* 33.6 (2008), pp. 1406–1412.
- [114] Gh A Kamali, I Moradi, and A Khalili. "Estimating solar radiation on tilted surfaces with various orientations: a study case in Karaj (Iran)". In: *Theoretical and applied climatology* 84.4 (2006), pp. 235–241.
- [115] PG Loutzenhiser et al. "Empirical validation of models to compute solar irradiance on inclined surfaces for building energy simulation". In: *Solar Energy* 81.2 (2007), pp. 254–267.
- [116] A Maria Gracia and T Huld. "Performance comparison of different models for the estimation of global irradiance on inclined surfaces". In: *EUR-Scientific and Technical Research series* (2013).
- [117] B Liu and R Jordan. "Daily insolation on surfaces tilted towards equator". In: *ASHRAE J. (United States)* 10 (1961).
- [118] Pericles S Koronakis. "On the choice of the angle of tilt for south facing solar collectors in the Athens basin area". In: *Solar Energy* 36.3 (1986), pp. 217–225.
- [119] YQ Tian et al. "Estimating solar radiation on slopes of arbitrary aspect". In: *Agricultural and Forest Meteorology* 109.1 (2001), pp. 67–74.
- [120] V Badescu. "3D isotropic approximation for solar diffuse irradiance on tilted surfaces". In: *Renewable Energy* 26.2 (2002), pp. 221–233.
- [121] JW Bugler. "The determination of hourly insolation on an inclined plane using a diffuse irradiance model based on hourly measured global horizontal insolation". In: *Solar Energy* 19.5 (1977), pp. 477–491.
- [122] Ralph C Temps and KL Coulson. "Solar radiation incident upon slopes of different orientations". In: *Solar energy* 19.2 (1977), pp. 179–184.
- [123] Thomas M Klucher. "Evaluation of models to predict insolation on tilted surfaces". In: *Solar energy* 23.2 (1979), pp. 111–114.
- [124] Richard Perez et al. "A new simplified version of the Perez diffuse irradiance model for tilted surfaces". In: *Solar energy* 39.3 (1987), pp. 221–231.
- [125] Cort J Willmott. "On the climatic optimization of the tilt and azimuth of flat-plate solar collectors". In: *Solar Energy* 28.3 (1982), pp. 205–216.
- [126] John E Hay. "Calculation of monthly mean solar radiation for horizontal and inclined surfaces". In: *Solar Energy* 23.4 (1979), pp. 301–307.

- [127] Arvid Skartveit and Jan Asle Olseth. "Modelling slope irradiance at high latitudes". In: *Solar Energy* 36.4 (1986), pp. 333–344.
- [128] Arvid Skartveit and Jan Asle Olseth. "A model for the diffuse fraction of hourly global radiation". In: *Solar Energy* 38.4 (1987), pp. 271–274.
- [129] DT Reindl, WA Beckman, and JA Duffie. "Evaluation of hourly tilted surface radiation models". In: *Solar Energy* 45.1 (1990), pp. 9–17.
- [130] John A Duffie and William A Beckman. "Solar engineering of thermal processes". In: (1980).
- [131] Christian Gueymard. "An anisotropic solar irradiance model for tilted surfaces and its comparison with selected engineering algorithms". In: *Solar Energy* 38.5 (1987), pp. 367–386.
- [132] Ursula Eicker. *Solar technologies for buildings*. John Wiley & Sons, 2006.
- [133] "Solar Thermal Markets in Europe - Trends and Market Statistics 2013". In: (). Accessed: 2018-12-12.
- [134] *International Energy Agency IEA (2014a). Technology Roadmap: Solar Photovoltaic Energy*. OECD/IEA, 2014.
- [135] Matthew P Lumb et al. "GaSb-Based Solar Cells for Full Solar Spectrum Energy Harvesting". In: *Advanced Energy Materials* 7.20 (2017), p. 1700345.
- [136] Z Zhou and M Carbajales-Dale. "Assessing the photovoltaic technology landscape: efficiency and energy return on investment (EROI)". In: *Energy & Environmental Science* 11.3 (2018), pp. 603–608.
- [137] Pius Hüsler and Nora Farrag. "National Survey Report on PV Power Applications in Switzerland 2013". In: *IEA-PVPS Programme-NSRs for Switzerland* (2014).
- [138] *Global Wind Statistics 2017*, link = http://gwec.net/wp-content/uploads/vip/GWEC_PRstats2017_EN-003_FINAL.pdf, note = Accessed: 2018-11-15.
- [139] MAURIZIO Cellura et al. "Wind speed spatial estimation for energy planning in Sicily: Introduction and statistical analysis". In: *Renewable energy* 33.6 (2008), pp. 1237–1250.
- [140] M Cellura et al. "Wind speed spatial estimation for energy planning in Sicily: a neural kriging application". In: *Renewable energy* 33.6 (2008), pp. 1251–1266.
- [141] Dimitrios Mentis et al. "Assessing the technical wind energy potential in Africa a GIS-based approach". In: *Renewable Energy* 83 (2015), pp. 110–125.
- [142] MH Soulouknga et al. "Analysis of wind speed data and wind energy potential in Faya-Largeau, Chad, using Weibull distribution". In: *Renewable Energy* 121 (2018), pp. 1–8.
- [143] Jose A Carta, Penelope Ramirez, and Sergio Velazquez. "A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands". In: *Renewable and sustainable energy reviews* 13.5 (2009), pp. 933–955.
- [144] Nazli Yonca Aydin, Elcin Kentel, and Sebnem Duzgun. "GIS-based environmental assessment of wind energy systems for spatial planning: A case study from Western Turkey". In: *Renewable and Sustainable Energy Reviews* 14.1 (2010), pp. 364–373.
- [145] Pilar Díaz-Cuevas et al. "Developing a wind energy potential map on a regional scale using GIS and multi-criteria decision methods: the case of Cadiz (south of Spain)". In: *Clean Technologies and Environmental Policy* (2018), pp. 1–17.
- [146] Mostafa Mahdy and AbuBakr S Bahaj. "Multi criteria decision analysis for offshore wind energy potential in Egypt". In: *Renewable Energy* 118 (2018), pp. 278–289.
- [147] Peter Enevoldsen and Finn-Hendrik Permien. "Mapping the Wind Energy Potential of Sweden: A Sociotechnical Wind Atlas". In: *Journal of Renewable Energy* 2018 (2018).
- [148] Beata Sliz-Szkliniarz and Joachim Vogt. "GIS-based approach for the evaluation of wind energy potential: A case study for the Kujawsko-Pomorskie Voivodeship". In: *Renewable and Sustainable Energy Reviews* 15.3 (2011), pp. 1696–1707.

- [149] Ma Lei et al. "A review on the forecasting of wind speed and generated power". In: *Renewable and Sustainable Energy Reviews* 13.4 (2009), pp. 915–920.
- [150] Antonino Marvuglia and Antonio Messineo. "Monitoring of wind farms? power curves using machine learning techniques". In: *Applied Energy* 98 (2012), pp. 574–583.
- [151] Aoife M Foley et al. "Current methods and advances in forecasting of wind power generation". In: *Renewable Energy* 37.1 (2012), pp. 1–8.
- [152] Da Liu et al. "Short-term wind speed forecasting using wavelet transform and support vector machines optimized by genetic algorithm". In: *Renewable Energy* 62 (2014), pp. 592–597.
- [153] Can Wan et al. "Probabilistic forecasting of wind power generation using extreme learning machine". In: *IEEE Transactions on Power Systems* 29.3 (2014), pp. 1033–1044.
- [154] DA Fadare. "The application of artificial neural networks to mapping of wind speed profile for energy application in Nigeria". In: *Applied Energy* 87.3 (2010), pp. 934–942.
- [155] Younes Noorollahi, Mohammad Ali Jokar, and Ahmad Kalhor. "Using artificial neural networks for temporal and spatial wind speed forecasting in Iran". In: *Energy Conversion and Management* 115 (2016), pp. 17–25.
- [156] Stefano Grassi, Fabio Veronesi, and Martin Raubal. "Satellite remote sensed data to improve the accuracy of statistical models for wind resource assessment". In: *Proceedings of EWEA Conference 2015, Porte de Versailles Pavillon 1, Paris, France, 17-20 November 2015*. European Wind Energy Association (EWEA 2015). 2015.
- [157] Daniel R Drew, Janet F Barlow, and Siân E Lane. "Observations of wind speed profiles over Greater London, UK, using a Doppler lidar". In: *Journal of Wind Engineering and Industrial Aerodynamics* 121 (2013), pp. 98–105.
- [158] Keith M Sunderland, Gerald Mills, and Michael F Conlon. "Estimating the wind resource in an urban area: A case study of micro-wind generation potential in Dublin, Ireland". In: *Journal of Wind Engineering and Industrial Aerodynamics* 118 (2013), pp. 44–53.
- [159] Edward Ng et al. "Improving the wind environment in high-density cities by understanding urban morphology and surface roughness: a study in Hong Kong". In: *Landscape and Urban planning* 101.1 (2011), pp. 59–74.
- [160] B Wang et al. "Cross indicator analysis between wind energy potential and urban morphology". In: *Renewable Energy* 113 (2017), pp. 989–1006.
- [161] Malcolm A Heath, John D Walshe, and Simon J Watson. "Estimating the potential yield of small building-mounted wind turbines". In: *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology* 10.3 (2007), pp. 271–287.
- [162] Alex Kalmikov et al. "Wind power resource assessment in complex urban environments: MIT campus case-study using CFD Analysis". In: (2010).
- [163] Teresa Simões and Ana Estanqueiro. "A new methodology for urban wind resource assessment". In: *Renewable Energy* 89 (2016), pp. 598–605.
- [164] An-Shik Yang et al. "Estimation of wind power generation in dense urban area". In: *Applied energy* 171 (2016), pp. 213–230.
- [165] Muhammad Rizwan Awan, Fahid Riaz, and Zahid Nabi. "Analysis of conditions favourable for small vertical axis wind turbines between building passages in urban areas of Sweden". In: *International Journal of Sustainable Energy* 36.5 (2017), pp. 450–461.
- [166] Sara Louise Walker. "Building mounted wind turbines and their suitability for the urban scale? A review of methods of estimating urban wind resource". In: *Energy and Buildings* 43.8 (2011), pp. 1852–1862.
- [167] Gabriela Seiz and Nando Foppa. "National Climate Observing System of Switzerland (GCOS Switzerland)". In: *Advances in Science and Research* 6.1 (2011), pp. 95–102.

- [168] Eric W Weisstein. “Moore neighborhood”. In: *From MathWorld—A Wolfram Web Resource*. <http://mathworld.wolfram.com/MooreNeighborhood.html> (2005).
- [169] William C Skamarock and Joseph B Klemp. “A time-split nonhydrostatic atmospheric model for weather research and forecasting applications”. In: *Journal of Computational Physics* 227.7 (2008), pp. 3465–3485.
- [170] Dasaraden Mauree, Nadège Blond, and Alain Clappier. “Multi-scale modeling of the urban meteorology: Integration of a new canopy model in the WRF model”. In: *Urban Climate* 26 (2018), pp. 60–75.
- [171] Amy Bowen et al. *Small wind turbine testing results from the National Renewable Energy Lab*. Tech. rep. National Renewable Energy Lab.(NREL), Golden, CO (United States), 2009.
- [172] *Schweizerische Elektrizitätsstatistik 2017*, author=Swiss Federal Office of Energy SFOE, link = http://www.bfe.admin.ch/themen/00526/00541/00542/00630/index.html?lang=en&dossier_id=00765, note = Accessed: 2018-12-16.
- [173] Paul Denholm et al. “Land-use requirements of modern wind power plants in the United States”. In: *Golden, CO: National Renewable Energy Laboratory* (2009), p. 57.
- [174] Johan Meyers and Charles Meneveau. “Optimal turbine spacing in fully developed wind farm boundary layers”. In: *Wind Energy* 15.2 (2012), pp. 305–317.
- [175] Jose Zayas et al. “Enabling wind power nationwide”. In: *US Department of Energy* (2015).
- [176] Alessandro Casasso and Rajandrea Sethi. “Assessment and mapping of the shallow geothermal potential in the province of Cuneo (Piedmont, NW Italy)”. In: *Renewable Energy* 102 (2017), pp. 306–315.
- [177] D Bertermann et al. “Modelling vSGPs (very shallow geothermal potentials) in selected CSAs (case study areas)”. In: *Energy* 71 (2014), pp. 226–244.
- [178] John Lund et al. “Geothermal (ground-source) heat pumps-a world overview”. In: (2004).
- [179] David Beamish. “The bedrock electrical conductivity map of the UK”. In: *Journal of Applied Geophysics* 96 (2013), pp. 87–97.
- [180] Eloisa Di Sipio et al. “Subsurface thermal conductivity assessment in Calabria (southern Italy): a regional case study”. In: *Environmental earth sciences* 72.5 (2014), pp. 1383–1401.
- [181] Soteris A Kalogirou et al. “Artificial neural networks for the generation of a conductivity map of the ground”. In: *Renewable Energy* 77 (2015), pp. 400–407.
- [182] Joris Ondreka et al. “GIS-supported mapping of shallow geothermal potential of representative areas in south-western Germany—Possibilities and limitations”. In: *Renewable Energy* 32.13 (2007), pp. 2186–2200.
- [183] Verein Deutscher Ingenieure. “VDI-Richtlinie 4640: Thermische Nutzung des Untergrundes—Blatt 2: erdekoppelte Wärmepumpenanlagen”. In: *Beuth Verlag, Berlin* (2001).
- [184] Alejandro García-Gil et al. “GIS-supported mapping of low-temperature geothermal potential taking groundwater flow into account”. In: *Renewable Energy* 77 (2015), pp. 268–278.
- [185] Kerry Schiel et al. “GIS-based modelling of shallow geothermal energy potential for CO2 emission mitigation in urban areas”. In: *Renewable energy* 86 (2016), pp. 1023–1036.
- [186] Antonio Galgaro et al. “Empirical modeling of maps of geo-exchange potential for shallow geothermal energy at regional scale”. In: *Geothermics* 57 (2015), pp. 173–184.
- [187] Alessandro Casasso and Rajandrea Sethi. “G. POT: A quantitative method for the assessment and mapping of the shallow geothermal potential”. In: *Energy* 106 (2016), pp. 765–773.
- [188] Teppo Arola et al. “Mapping the low enthalpy geothermal potential of shallow Quaternary aquifers in Finland”. In: *Geothermal Energy* 2.1 (2014), p. 9.
- [189] Alistair Allen, Dejan Milenic, and Paul Sikora. “Shallow gravel aquifers and the urban ‘heat island’ effect: a source of low enthalpy geothermal energy”. In: *Geothermics* 32.4-6 (2003), pp. 569–578.

- [190] Ke Zhu et al. "The geothermal potential of urban heat islands". In: *Environmental Research Letters* 5.4 (2010), p. 044002.
- [191] Teppo Arola and Kirsti Korkka-Niemi. "The effect of urban heat islands on geothermal potential: examples from Quaternary aquifers in Finland". In: *Hydrogeology Journal* 22.8 (2014), pp. 1953–1967.
- [192] Jaime A Rivera, Philipp Blum, and Peter Bayer. "Increased ground temperatures in urban areas: Estimation of the technical geothermal potential". In: *Renewable energy* 103 (2017), pp. 388–400.
- [193] J. M. Andújar Márquez, Miguel Ángel Martínez Bohórquez, and Sergio Gómez Melgar. "Ground Thermal Diffusivity Calculation by Direct Soil Temperature Measurement. Application to very Low Enthalpy Geothermal Energy Systems". In: *Sensors* 16.3 (2016), p. 306.
- [194] David Bertermann et al. "Thermomap-an open-source web mapping application for illustrating the very shallow geothermal potential in europe and selected case study areas". In: *Eur Geotherm Congr, Pisa* (2013), pp. 1–7.
- [195] D Bertermann, H Klug, and L Morper-Busch. "A pan-European planning basis for estimating the very shallow geothermal energy potentials". In: *Renewable Energy* 75 (2015), pp. 335–347.
- [196] Miles S Kersten. "Thermal properties of soils". In: (1949).
- [197] Mikhail Kanevski and Michel Maignan. *Analysis and modelling of spatial environmental data*. Vol. 6501. EPFL press, 2004.
- [198] Sajid Hussain and Ali AlAlili. "A hybrid solar radiation modeling approach using wavelet multiresolution analysis and artificial neural networks". In: *Applied Energy* 208 (2017), pp. 540–550.
- [199] S Alessandrini et al. "An analog ensemble for short-term probabilistic solar power forecast". In: *Applied energy* 157 (2015), pp. 95–110.
- [200] Siwei Lou et al. "Prediction of diffuse solar irradiance using machine learning and multivariable regression". In: *Applied Energy* 181 (2016), pp. 367–374.
- [201] Muhammed A Hassan et al. "Exploring the potential of tree-based ensemble methods in solar radiation modeling". In: *Applied Energy* 203 (2017), pp. 897–916.
- [202] Nils André Treiber, Justin Heinermann, and Oliver Kramer. "Wind power prediction with machine learning". In: *Computational Sustainability*. Springer, 2016, pp. 13–29.
- [203] Justin Heinermann and Oliver Kramer. "Machine learning ensembles for wind power prediction". In: *Renewable Energy* 89 (2016), pp. 671–679. URL: <http://www.sciencedirect.com/science/article/pii/S0960148115304894>.
- [204] Najeebullah et al. "Machine Learning based short term wind power prediction using a hybrid learning model". In: *Computers & Electrical Engineering* 45 (2015), pp. 122–133. URL: <http://www.sciencedirect.com/science/article/pii/S0045790614001876>.
- [205] Bikash Joshi et al. "Rooftop detection for planning of solar PV deployment: a case study in Abu Dhabi". In: *International Workshop on Data Analytics for Renewable Energy Integration*. Springer, 2014, pp. 137–149.
- [206] Soteris A Kalogirou et al. "Artificial neural networks for the generation of geothermal maps of ground temperature at various depths by considering land configuration". In: *Energy* 48.1 (2012), pp. 233–240.
- [207] Graeme Beardsmore et al. "A Bayesian inference tool for geophysical joint inversions". In: *ASEG Extended Abstracts* 2016.1 (2016), pp. 1–10.
- [208] Graeme Beardsmore. *Data fusion and machine learning for geothermal target exploration and characterization*. Tech. rep. Technical report, National ICT Australia Limited (NICTA), Australia, 2014.
- [209] J Jenness. "Tools for graphics and shapes: Extension for ArcGIS". In: *Jenness Enterprises* (2011).
- [210] Bertrand Dumont and Dominique Chapellier. *INVENTAIRE DES SONDAGES ELECTRIQUES DE SUISSE (Publication nr. 42)*. Accessed: 2018-05-23. Swiss Geophysical Commission.

- [211] D. Pahud. "Geothermal energy and heat storage". In: *Scuola Universitaria Professionale della Svizzera Italiana* (2002).
- [212] A Kemmler et al. "Analyse des schweizerischen Energieverbrauchs 2000-2017 nach Verwendungszwecken". In: *Bern: Bundesamt für Energie* (2018).
- [213] Aneta Strzalka et al. "Large scale integration of photovoltaics in cities". In: *Applied Energy* 93 (2012), pp. 413–421.
- [214] H Wittmann et al. "Identification of roof areas suited for solar energy conversion systems". In: *Renewable Energy* 11.1 (1997), pp. 25–36.
- [215] *International Energy Agency IEA (2002). Potential for Building Integrated Photovoltaics. IEA-PVPS Task 2002.* OECD/IEA, 2002.
- [216] Rebecca R Hernandez, Madison K Hoffacker, and Christopher B Field. "Efficient use of land to meet sustainable energy needs". In: *Nature Climate Change* 5.4 (2015), p. 353.
- [217] Jihui Yuan et al. "A method to estimate the potential of rooftop photovoltaic power generation for a region". In: *Urban Climate* 17 (2016), pp. 1–19.
- [218] Georgios Mavromatidis, Kristina Orehounig, and Jan Carmeliet. "Evaluation of photovoltaic integration potential in a village". In: *Solar Energy* 121 (2015), pp. 152–168.
- [219] Paulina Wegertseder et al. "Combining solar resource mapping and energy system integration methods for realistic valuation of urban solar energy potential". In: *Solar Energy* 135 (2016), pp. 325–336.
- [220] Marco Berg and Markus Real. *Road Map Renewable Energies Switzerland: An Analysis with a view to harnessing existing potentials by 2050.* Swiss Academy of Engineering Sciences (SATW), 2007.
- [221] John Byrne et al. "A review of the solar city concept and methods to assess rooftop solar electric potential, with an illustrative application to the city of Seoul". In: *Renewable and Sustainable Energy Reviews* 41 (2015), pp. 830–844.
- [222] Luis Ramirez Camargo et al. "Spatio-temporal modeling of roof-top photovoltaic panels for improved technical potential assessment and electricity peak load offsetting at the municipal scale". In: *Computers, Environment and Urban Systems* 52 (2015), pp. 58–69.
- [223] Marcel Suri et al. "Potential of solar electricity generation in the European Union member states and candidate countries". In: *Solar energy* 81.10 (2007), pp. 1295–1305.
- [224] PR Defaix et al. "Technical potential for photovoltaics on buildings in the EU-27". In: *Solar Energy* 86.9 (2012), pp. 2644–2653.
- [225] Paul Denholm and Robert Margolis. "Supply curves for rooftop solar PV-generated electricity for the United States". In: (2008).
- [226] Jay Paidipati et al. "Rooftop photovoltaics market penetration scenarios". In: *Navigant Consulting, Inc., for NREL: February* (2008).
- [227] Ran Vardimon. "Assessment of the potential for distributed photovoltaic electricity production in Israel". In: *Renewable Energy* 36.2 (2011), pp. 591–594.
- [228] Sophie Pelland and Yves Poissant. "An evaluation of the potential of building integrated photovoltaics in Canada". In: *Proceedings of the SESCOI 2006 Conference, submitted.* 2006.
- [229] J Ordóñez et al. "Analysis of the photovoltaic solar energy capacity of residential rooftops in Andalusia (Spain)". In: *Renewable and Sustainable Energy Reviews* 14.7 (2010), pp. 2122–2130.
- [230] Anthony Lopez et al. *US renewable energy technical potentials: a GIS-based analysis.* Tech. rep. NREL, 2012.
- [231] Jaroslav Hofierka and Ján Kaňuk. "Assessment of photovoltaic potential in urban areas using open-source solar radiation tools". In: *Renewable energy* 34.10 (2009), pp. 2206–2214.

- [232] Luca Bergamasco and Pietro Asinari. “Scalable methodology for the photovoltaic solar energy potential assessment based on available roof surface area: Application to Piedmont Region (Italy)”. In: *Solar Energy* 85.5 (2011), pp. 1041–1055.
- [233] Jinqing Peng and Lin Lu. “Investigation on the development potential of rooftop PV system in Hong Kong and its environmental benefits”. In: *Renewable and Sustainable Energy Reviews* 27 (2013), pp. 149–162.
- [234] Julieta Schallenberg-Rodríguez. “Photovoltaic techno-economical potential on roofs in regions and islands: the case of the Canary Islands. Methodological review and methodology proposal”. In: *Renewable and Sustainable Energy Reviews* 20 (2013), pp. 219–239.
- [235] Laura Romero Rodríguez et al. “Assessment of the photovoltaic potential at urban level based on 3D city models: A case study and new methodological approach”. In: *Solar Energy* 146 (2017), pp. 264–275.
- [236] S Suzuki, Masakazu Ito, and Kosuke Kurokawa. “An analysis of PV resource in residential areas by means of aerial photo images”. In: *Proceedings of the 22nd European Photovoltaic Solar Energy Conference*. 2007, pp. 3–7.
- [237] Fayez Tarsha-Kurdi et al. “Model-driven and data-driven approaches using LIDAR data: Analysis and comparison”. In: *ISPRS Workshop, Photogrammetric Image Analysis (PIA07)*. 2007, pp. 87–92.
- [238] Cláudio Carneiro, Eugenio Morello, and Gilles Desthieux. “Assessment of solar irradiance on the urban fabric for the production of renewable energy using LIDAR data and image processing techniques”. In: *Advances in GIScience*. Springer, 2009, pp. 83–112.
- [239] Cláudio Carneiro et al. “Digital urban morphometrics: automatic extraction and assessment of morphological properties of buildings”. In: *Transactions in GIS* 14.4 (2010), pp. 497–531.
- [240] Miguel C Brito et al. “Photovoltaic potential in a Lisbon suburb using LiDAR data”. In: *Solar Energy* 86.1 (2012), pp. 283–288.
- [241] Niko Lukač et al. “Buildings roofs photovoltaic potential assessment based on LiDAR (Light Detection And Ranging) data”. In: *Energy* 66 (2014), pp. 598–609.
- [242] James Gooding, Rolf Crook, and Alison S Tomlin. “Modelling of roof geometries from low-resolution LiDAR data for city-scale solar energy applications using a neighbouring buildings method”. In: *Applied energy* 148 (2015), pp. 93–104.
- [243] A Verso et al. “GIS-based method to evaluate the photovoltaic potential in the urban environments: The particular case of Miraflores de la Sierra”. In: *Solar Energy* 117 (2015), pp. 236–245.
- [244] Caleb Phillips et al. “A data mining approach to estimating rooftop photovoltaic potential in the US”. In: *Journal of Applied Statistics* (2018), pp. 1–10.
- [245] Ronnen Levinson et al. “Solar access of residential rooftops in four California cities”. In: *Solar Energy* 83.12 (2009), pp. 2120–2135.
- [246] Ha T Nguyen and Joshua M Pearce. “Incorporating shading losses in solar photovoltaic potential assessment at the municipal scale”. In: *Solar Energy* 86.5 (2012), pp. 1245–1260.
- [247] M Bouzerdoum, A Mellit, and A Massi Pavan. “A hybrid model (SARIMA–SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant”. In: *Solar Energy* 98 (2013), pp. 226–235.
- [248] Zeynab Ramedani et al. “Potential of radial basis function based support vector regression for global solar radiation prediction”. In: *Renewable and Sustainable Energy Reviews* 39 (2014), pp. 1005–1011.
- [249] Makbul AM Ramli, Ssennoga Twaha, and Yusuf A Al-Turki. “Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi Arabia case study”. In: *Energy conversion and management* 105 (2015), pp. 442–452.
- [250] Philippe Lauret et al. “A benchmarking of machine learning techniques for solar radiation forecasting in an insular context”. In: *Solar Energy* 112 (2015), pp. 446–457.

- [251] Shuai Li, Hongjie Ma, and Weiyi Li. "Typical solar radiation year construction using k-means clustering and discrete-time Markov chain". In: *Applied Energy* 205 (2017), pp. 720–731.
- [252] Adnan Sözen et al. "Use of artificial neural networks for mapping of solar potential in Turkey". In: *Applied Energy* 77.3 (2004), pp. 273–286.
- [253] Nahid Mohajeri et al. "A city-scale roof shape classification using machine learning for solar energy applications". In: *Renewable Energy* 121 (2018), pp. 81–93.
- [254] C Steinmeier. *CORINE Land Cover 2000/2006 Switzerland*. Swiss Federal Institute for Forest, Snow and Landscape Research WSL, 2013.
- [255] *Federal Statistical Office*, link = <https://www.bfs.admin.ch/bfs/en/home.html>, note = Accessed: 2018-01-29.
- [256] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.
- [257] Ian Jolliffe. "Principal component analysis". In: *International encyclopedia of statistical science*. Springer, 2011, pp. 1094–1096.
- [258] Michael Geiger et al. "A web service for controlling the quality of measurements of global solar irradiation". In: *Solar energy* 73.6 (2002), pp. 475–480.
- [259] *Meteoswiss*, link = <http://www.meteosuisse.admin.ch/home.html?tab=overview>, note = Accessed: 2018-01-29.
- [260] *Swisstopo*, link = <https://www.swisstopo.admin.ch>, note = Accessed: 2018-01-29.
- [261] *SITG*, link = http://ge.ch/sitg/sitg_catalog/sitg_donnees, note = Accessed: 2018-01-29.
- [262] P Hüsler. *National Survey Report of PV Power Applications in Switzerland-2015*. International Energy Agency (IEA) PVPS. 2016.
- [263] Niko Lukač et al. "Rating of roofs? surfaces regarding their solar potential and suitability for PV systems, based on LiDAR data". In: *Applied energy* 102 (2013), pp. 803–812.
- [264] Brian Goss et al. "Irradiance modelling for individual cells of shaded solar photovoltaic arrays". In: *Solar Energy* 110 (2014), pp. 410–419.
- [265] Marylène Montavon. "Optimisation of urban form by the evaluation of the solar potential". In: (2010).
- [266] *Sonnendach*, link = <http://www.uvek-gis.admin.ch/BFE/sonnendach>, note = Accessed: 2018-01-29.
- [267] *International Energy Agency IEA (2011)*. *Solar Energy Perspectives*. OECD/IEA, 2011.
- [268] *International Energy Agency IEA (2014b)*. *Energy Technology Perspectives 2014*. OECD/IEA, 2014.
- [269] Daniel Klauser. *Solarpotentialanalyse für Sonnendach.ch, Schlussbericht*. Bundesamt für Energie BFE, 2016.
- [270] Dan Assouline, Nahid Mohajeri, and Jean-Louis Scartezzini. "Building rooftop classification using random forests for large-scale PV deployment". In: *Earth Resources and Environmental Remote Sensing/GIS Applications VIII*. Vol. 10428. International Society for Optics and Photonics. 2017, p. 1042806.
- [271] M Gutschner et al. "Potential for building integrated photovoltaics". In: *IEA-PVPS Task 7* (2002).
- [272] Maya Chaudhari, Lisa Frantzis, and Tom E Hoff. "PV grid connected market potential under a cost breakthrough scenario". In: *Navigant Consulting, Inc. Retrieved on September 16* (2004), p. 2010.
- [273] L Frantzis, S Graham, and J Paidipati. "California rooftop photovoltaic (PV) resource assessment and growth potential by county". In: *Navigant Consulting, California Energy Commission PIER Final Project Report CEC-500-2007-048* (2007).

- [274] Alberto Bocca et al. "Estimating photovoltaic energy potential from a minimal set of randomly sampled data". In: *Renewable Energy* 97 (2016), pp. 457–467.
- [275] Pieter Gagnon et al. "Rooftop solar photovoltaic technical potential in the United States: A detailed assessment". In: *National Renewable Energy Laboratory (NREL), Technical report* (2016).
- [276] R Kassner et al. "Analysis of the solar potential of roofs by using official lidar data". In: *Proceedings of the International Society for Photogrammetry, Remote Sensing and Spatial Information Sciences, (ISPRS Congress)*. 2008, pp. 399–404.
- [277] Ha T Nguyen et al. "The application of LiDAR to assessment of rooftop solar photovoltaic deployment potential in a municipal district unit". In: *Sensors* 12.4 (2012), pp. 4534–4558.
- [278] Ayseguel Tereci et al. "Energy saving potential and economical analysis of solar systems in the urban quarter Scharnhauser Park". In: *ISES Solar World Congress 2009, 11-14.10. 2009, Johannesburg, South Africa*. 2009.
- [279] P Redweik, Cristina Catita, and Miguel Brito. "Solar energy potential on roofs and facades in an urban landscape". In: *Solar Energy* 97 (2013), pp. 332–341.
- [280] Mesude Bayrakci Boz, Kirby Calvert, and Jeffrey RS Brownson. "An automated model for rooftop PV systems assessment in ArcGIS using LIDAR". In: *AIMS Energy* 3.3 (2015), pp. 401–420.
- [281] *swissBUILDINGS3D 2.0*, link = <https://shop.swisstopo.admin.ch/fr/products/landscape/build3D2>, note = Accessed: 2018-01-29.
- [282] Xi Zhang et al. "Learning from synthetic models for roof style classification in point clouds". In: *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM. 2014, pp. 263–270.
- [283] *Swissgrid*, link = <https://www.bfs.admin.ch/bfs/fr/home/registres/registre-batiments-logements/acces-donnees-services.html>, note = Accessed: 2018-01-29.
- [284] Yu Ru, Jan Kleissl, and Sonia Martinez. "Storage size determination for grid-connected photovoltaic systems". In: *IEEE Transactions on Sustainable Energy* 4.1 (2013), pp. 68–81.
- [285] Johannes Weniger, Tjarko Tjaden, and Volker Quaschnig. "Sizing of residential PV battery systems". In: *Energy Procedia* 46 (2014), pp. 78–87.
- [286] *STATPOP*, link = <https://www.bfs.admin.ch/bfs/en/home/statistics/population/surveys/statpop.html>, note = Accessed: 2018-01-29.
- [287] Carolin Strobl et al. "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1 (2007), p. 25.
- [288] M Montavon et al. *Solurban Project, Solar Utilisation Potential of Urban Sites*. Tech. rep. 2005.
- [289] Anthony A Clifford. "Multivariate error analysis: a handbook of error propagation and calculation in many-parameter systems". In: (1973).
- [290] Kai O Arras. *An introduction to error propagation: derivation, meaning and examples of equation $CY = FX CX FXT$* . Tech. rep. ETH Zurich, 1998.
- [291] J Kircher. "Data Analysis Toolkit 5: Uncertainty Analysis and Error Propagation". In: *University of California Berkeley Seismological Laboratory* (2001).
- [292] GBM Heuvelink et al. "Propagation of error in spatial modelling with GIS". In: *Geographical information systems* 1 (1999), pp. 207–217.
- [293] Pietro Florio et al. "Assessing visibility in multi-scale urban planning: A contribution to a method enhancing social acceptability of solar energy in cities". In: *Solar Energy* 173 (2018), pp. 97–109. URL: <http://www.sciencedirect.com/science/article/pii/S0038092X18307230>.

- [294] Furkan Dincer. “The analysis on photovoltaic electricity generation status, potential and policies of the leading countries in solar energy”. In: *Renewable and Sustainable Energy Reviews* 15.1 (2011), pp. 713–720.
- [295] John Edward Burns and Jin-Su Kang. “Comparative economic analysis of supporting policies for residential solar PV in the United States: Solar Renewable Energy Credit (SREC) potential”. In: *Energy Policy* 44 (2012), pp. 217–225.
- [296] Robert Carl Pietzcker et al. “Using the sun to decarbonize the power sector: The economic potential of photovoltaics and concentrating solar power”. In: *Applied Energy* 135 (2014), pp. 704–720.
- [297] IDAWEB, *link = <https://gate.meteoswiss.ch/idaweb>, note = Accessed: 2018-01-29.*
- [298] N. Das et al. *SMAP/Sentinel-1 L2 Radiometer/Radar 30-Second Scene 3 km EASE-Grid Soil Moisture, Version 1.*
- [299] SoDa, *link = <http://www.soda-pro.com>, note = Accessed: 2018-01-29.*
- [300] geodata4edu, *link = <http://geodata4edu.ethz.ch>, note = Accessed: 2018-01-29.*
- [301] RegBL, *link = <https://www.bfs.admin.ch/bfs/fr/home/registres/registre-batiments-logements/acces-donnees-services.html>, note = Accessed: 2018-01-29.*
- [302] NABODAT. Accessed: 2018-05-23.

Dan Assouline

Avenue d'Ouchy 75, 1006 Lausanne, Switzerland ◇ Email: dan.assouline@epfl.ch

Swiss number: +41 79 951 48 60

EDUCATION

Ph.D. in Energy/Civil Engineering

Expected in March 2019

Ecole Polytechnique Federale de Lausanne (EPFL)

Thesis topic: Machine Learning (ML) and Geographic Information Systems (GIS) for large scale spatio-temporal mapping of renewable energy potential.

Focus: Take advantage of the growing availability for spatial environmental and energy data, the processing abilities of GIS and the modelling power of traditional physical models to extract examples for multiple variables of interest, and combine them with the predictive power of Machine Learning algorithms to estimate these variables at unknown locations at the large scale of a country and convert them in potential values. The renewable energy systems of interest include solar PV panels on rooftops, shallow geothermal heat pumps and micro wind turbines.

MS in Civil and Environmental Engineering (Double Degree with ESTP)

May 2013

University of California, Berkeley

Major : Civil Systems, Overall GPA: 3.75/4

Minors : Transportation Engineering, Data Science

MS in Civil Engineering

May 2013

Ecole Speciale des Travaux Publics (ESTP), Paris, France

BS in Applied Mathematics ("Classes Preparatoires aux Grandes Ecoles")

2007-2010

Lycee Saint Louis, Paris, France

French Baccalaureate, with Very High Honors

July 2007

Lycee Sainte Marie, Meaux, France

WORK AND TEACHING EXPERIENCE

Graduate Teaching Assistant

September 2014 - Present

Ecole Polytechnique Federale de Lausanne (EPFL)

Lausanne, Switzerland

- TA in "Building Physics I and II". Introduction to classical physics for architects, particularly relevant within buildings : Thermodynamics, acoustics, hydrodynamics, photometry etc.

Graduate Student Instructor (GSI)

August 2013 - August 2014

University of California, Berkeley

Berkeley, CA

- 100% (during the summer)GSI appointment in "STATW21: Introduction to Probabilities and Statistics for Business majors"
- 50% GSI appointment in "STAT20: Introduction to Probabilities and Statistics"
- 25% GSI appointment in "CE93: Engineering Data Analysis" (Statistics and Probabilities for engineers)
- 25% GSI appointment in "STAT21: Introduction to Probabilities and Statistics for Business majors"

Data Assistant

June 2013 - August 2013

Pacific Earthquake Research Center (PEER)

Richmond, CA

- Processed and analyzed seismic data with R for the “NGA-East Project”, aiming to develop a new ground motion characterization (GMC) model for the Central and Eastern North-American region.
- Worked on various methodologies to handle and further analyze the data [R].

Graduate Student Instructor (GSI)

University of California, Berkeley

January 2013 - May 2013

Berkeley, CA

- 25% GSI appointment in “CE93: Engineering Data Analysis”

Project Manager Assistant

B.E.G. Ingenierie

June 2012 - August 2012

Troyes, France

- Supervised a whole area of the site (a supermarket in construction) under the site manager.
- Wrote reports on the evolution of the site each week.
- Attended reunions with construction companies each week and made the link between companies and Project Manager.

RESEARCH EXPERIENCE

Doctoral Researcher

Ecole Polytechnique Federale de Lausanne (EPFL)

September 2014 - Present

Lausanne, Switzerland

- Worked on data mining for renewable energy mapping for the SCCER (Swiss Competence Centers for Energy Research) project for Future Energy Efficient Buildings & Districts (FEEB&D), in parallel with the doctoral thesis.
- Collected, processed and analyzed various sources of large data from multiple domains including spatial vector and raster data, meteorology time series data and energy data.
- Used on a day-to-day basis various Machine Learning algorithms, with a focus on fast and easily tunable methods (e.g. Support Vector Machines and Random Forests), with different types of data and for multiple spatio-temporal estimations.
- Used Geographic Information Systems regularly for pre or post-processing steps, notably using the ArcGIS software.
- Acquired deep theoretical and practical knowledge in Machine Learning, Geographic Information Systems, and renewable energies, with a focus on solar, geothermal and wind energy.
- Gained experience in communication while regularly summarizing and presenting results within the project and to other researchers and representatives.
- Gained experience in writing, reading and reviewing academic content.

TECHNICAL SKILLS

Languages and tools (years of experience in [.])

	Proficient	Experience
Computer Languages	Python[6 yrs]	Matlab[3 yrs], R[3 m.], Java[1 yr], Caml[3 yrs], HTML[1 yr]
Protocols & APIs		XML, JSON, Google API, Twitter API
Databases		MySQL, MongoDB
Softwares	ArcGIS[4 yrs]	AUTOCAD, BIOGEME, ANTLR

Used Python extensively, particularly scientific, ML and data-related libraries including NumPy, SciPy, Pandas, Scikit-learn, h5py, Matplotlib, Seaborn; and geospatial libraries such as GeoPandas and arcPy.

Hard Skills

- **Spatial regression**, specifically weather, energy and environmental variables. Use of fast and easily tunable algorithms, including Support Vector Machines and Random Forests. [ArcGIS/Python]
- **Raster Classification** (Roof shape) using LiDAR data. Used Random Forests. [ArcGIS/Python]
- **Uncertainty estimation** using Quantile Regression Forests. [Python]
- **Time series analysis** of ground motion signals and ground temperature data. [R/Python]
- **Clustering** and visualization of a million tweets data using the Twitter API and MongoDB. Used DBSCAN and MiniBatch K-means. [Python]
- **Spatial prediction** and simulation of rain fields. Used Gaussian Processes. [Python]
- **Object classification** (cars) based on aerial images. Used Support Vector Machines. [Python]
- **Classification** of human activities with smart phone accelerometer data. Used various classifiers: SVM, RF, k-Nearest Neighbors, Naive Bayes. [Python]

Soft Skills

- Presentation to non-technical audiences and representatives.
- Regular reporting of technical content for non-technical readers.

PUBLICATIONS

Journal papers

Mohajeri, N., Gudmundsson, A., Kunckler, T., Upadhyay, G., Assouline, D., Kämpf, J. H., & Scartezzini, J. L. (2019). A solar-based sustainable urban design: The effects of city-scale street-canyon geometry on solar access in Geneva, Switzerland. *Applied Energy*, 240, 173-190.

Mohajeri, N., Assouline, D., Guiboud, B., Bill, A., Gudmundsson, A., & Scartezzini, J. L. (2018). A city-scale roof shape classification using machine learning for solar energy applications. *Renewable Energy*, 121, 81-93.

Assouline, D., Mohajeri, N., & Scartezzini, J. L. (2018). Large-scale rooftop solar photovoltaic technical potential estimation using Random Forests. *Applied Energy*, 217, 189-211.

Mohajeri, N., Assouline, D., Gudmundsson, A., & Scartezzini, J. L. (2017). Effects of city size on the large-scale decentralised solar energy potential. *Energy Procedia*, 122, 697-702.

Assouline, D., Mohajeri, N., & Scartezzini, J. L. (2017). Quantifying rooftop photovoltaic solar energy potential: A machine learning approach. *Solar Energy*, 141, 278-296.

Mohajeri, N., Upadhyay, G., Gudmundsson, A., Assouline, D., Kämpf, J., & Scartezzini, J. L. (2016). Effects of urban compactness on solar energy potential. *Renewable Energy*, 93, 469-482.

Conference proceedings, abstracts, posters

Assouline, D., Mohajeri, N., Gudmundsson, A., & Scartezzini, J. L. (2018, July). Combining Fourier Analysis and Machine Learning to Estimate the Shallow-Ground Thermal Diffusivity in Switzerland. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (pp. 1144-1147). IEEE.

Assouline, D., Mohajeri, N., & Scartezzini, J. L. (2017, October). Building rooftop classification using random forests for large-scale PV deployment. In *Earth Resources and Environmental Remote Sensing/GIS Applications VIII* (Vol. 10428, p. 1042806). International Society for Optics and Photonics.

Assouline, D., Mohajeri, N., & Scartezzini, J. L. (2017, April). Random Forests (RFs) for Estimation, Uncertainty Prediction and Interpretation of Monthly Solar Potential. In *EGU General Assembly Conference Abstracts* (Vol. 19, p. 6693). [Poster presentation]

Mohajeri, N., Assouline, D., Guiboud, B., & Scartezzini, J. L. (2016). Does roof shape matter? solar photovoltaic (PV) integration on building roofs. In *Expanding Boundaries-Proceedings of the International Conference on Sustainable Built Environment (SBE)* (No. EPFL-CONF-220006).

Mohajeri, N., Gudmundsson, A., Kunckler, T., Upadhyay, G., Assouline, D., Kämpf, J. H., & Scartezzini, J. L. (2016). How street canyon configuration control the accessibility of solar energy potential: Implication for urban design. In *Proceedings of the 36th International Conference on Passive and Low Energy Architecture* (No. CONF).

Mohajeri, N., Gudmundsson, A., Upadhyay, G., Assouline, D., & Scartezzini, J. L. (2015). Neighbourhood morphology and solar irradiance in relation to urban climate. In *9th International Conference on Urban Climate jointly with 12th Symposium on the Urban Environment* (No. EPFL-CONF-210331).

Assouline, D., Mohajeri, N., & Scartezzini, J. L. (2015). A machine learning methodology for estimating roof-top photovoltaic solar energy potential in Switzerland. In *Proceedings of International Conference CISBAT 2015 Future Buildings and Districts Sustainability from Nano to Urban Scale* (No. EPFL-CONF-213375, pp. 555-560). LESO-PB, EPFL.

Assouline, D., Mohajeri, N., & Scartezzini, J.L., 2015. Machine learning for large scale rooftop photovoltaic potential estimation - Swiss case. Poster presentation at *Machine Learning Summer School in Kyoto University*, 23 August-4 September, Kyoto, Japan.

Book chapters

Assouline, D., Mohajeri, N., & Scartezzini, J. L. (2018). Estimation of Large-Scale Solar Rooftop PV Potential for Smart Grid Integration: A Methodological Review. In *Sustainable Interdependent Networks* (pp. 173-219). Springer, Cham.

LANGUAGES

French	Mother tongue
English	Fluent
Spanish	Basic

