Residential density classification for sustainable housing development using a machine learning approach

Residential density classification for sustainable housing development using a machine learning approach

N Mohajeri[1], A Walch[2], D Assouline[2], A Gudmundsson[3], A Smith[4], T Russel[4], J Hall[4]

[1] Institute for Environmental Design and Engineering, University College London, UK
[2] Solar Energy and Building Physics Laboratory, EPFL, Lausanne, Switzerland
[3]Department of Earth Sciences, Royal Holloway, University of London, UK
[4]Environmental Change Institute, University of Oxford, Oxford, UK

Abstract. Using Machine Learning (ML) algorithms for classification of the existing residential neighbourhoods and their spatial characteristics (e.g. density) so as to provide plausible scenarios for designing future sustainable housing is a novel application. Here we develop a methodology using a Random Forests algorithm (in combination with GIS spatial data processing) to detect and classify the residential neighbourhoods and their spatial characteristics within the region between Oxford and Cambridge, that is, the 'Oxford-Cambridge Arc'. The classification model is based on four pre-defined urban classes, that is, Centre, Urban, Suburban, and Rural for the entire region. The resolution is a grid of 500 m × 500 m. The features for classification include (1) dwelling geometric attributes (e.g. garden size, building footprint area, building perimeter), (2) street networks (e.g. street length, street density, street connectivity), (3) dwelling density (number of housing units per hectare), (4) building residential types (detached, semi-detached, terraced, and flats), and (5) characteristics of the surrounding neighbourhoods. The classification results, with overall average accuracy of 80% (accuracy per class: Centre: 38%, Urban 91%, Suburban 83%, and Rural 77%), for the Arc region show that the most important variables were three characteristics of the surrounding area: residential footprint area, dwelling density, and number of private gardens. The results of the classification are used to establish a baseline for the current status of the residential neighbourhoods in the Arc region. The results bring data-driven decision-making processes to the level of local authority and policy makers in order to support sustainable housing development at the regional scale.

1. Introduction

Identifying the growth process of urban areas from building to neighbourhood and to city scale is crucial for policy making and sustainable resource management. While some recent studies use image analysis and Machine Learning (ML) to analyse how individual buildings may evolve over time [1 - 3], others use ML to predict how and where urban growth is most likely to happen in the future [4]. For example, a ML methodology has been used to analyse existing patterns and processes of neighbourhood development to understand complex urban processes such as gentrification [4]. Identifying the differences between neighbourhoods at a regional scale [5] and classifying their spatial characteristics [6] is a challenge but provides useful information for future sustainable housing developments.

The 'Oxford-Cambridge Arc' region contains some of the fastest growing and most productive towns and cities in the UK [7]. It provides homes for 3.7 million people [8]. Existing population centres within the Arc are likely to undergo significant future expansion, with corresponding investment in intra-city transport infrastructure and associated high levels of economic growth [7-8]. The demand for housing throughout the Arc is high. The future of the Arc is likely to include a combination of different types of housing developments, that is, expansion of existing settlements as well as entirely new settlements. Part of the expansion is likely to be in the form of densification whereby additional dwellings are built within existing urban areas, primarily on brownfield land and/or permitted greenfield land.

According to the local plans which are prepared by the Local Planning Authority (LPA) it is important that new neighbourhoods harmonise with the existing ones; that is, new dwellings should not be in complete visual or typological contrast with the existing ones. With this in view, the aim of the present study is to develop a Machine Leaning methodology to detect and classify the existing residential neighbourhoods and their spatial characteristics including density within the Arc. The classification results can be used as a baseline when proposing density scenarios for neighbourhoods within each class (Centre, Urban, Suburban, Rural) for future sustainable urban densification.

2. Data and method

2.1 Data

The following data sources have been used to build the features for the training of the model (Figure 1): (i) A GIS building layer for the Arc, extracted from Ordnance Survey Mastermap (OSMM), which is composed of 1,362,743 residential buildings [8]. (ii) Residential private gardens, also extracted and processed from OSMM. (iii) Street networks, which were extracted and processed from OpenStreetMap. We consider the relations between the urban form and infrastructure systems in our models when selecting the scale and features. For the neighbourhood scale we use a grid (pixel) of the size 500 m × 500 m. The grid covers the entire area of the Arc. The scale of the study is so chosen in order to (i) obtain a generic residential urban form within a given pixel, ideally consisting of a building type (e.g. detached, semi-detached, terraced, and flats), though this was not always possible, and (ii) compute attributes that can affect the infrastructure systems (e.g. green infrastructure). The following attributes have been chosen in order to reflect the design parameters of neighbourhoods and their impacts on the infrastructure systems (e.g. green infrastructure). (1) dwelling geometric attributes (e.g. garden size, building footprint area, building perimeter), (2) street networks (e.g. street length, street density, street connectivity), (3) dwelling density (number of housing units per hectare), (4) building residential types (detached, semi-detached, terraced, and flats), and (5) statistical characteristics (e.g. minimum, maximum, mean) of the above attributes for up to eight surrounding pixels. The surrounding attributes (5) are computed using geospatial tools in combination with the Python programming language).

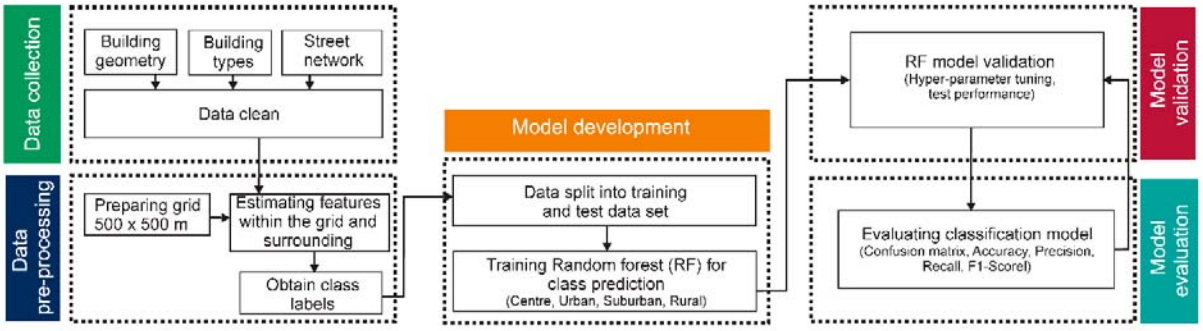

Figure 1. Workflow for classification modelling of residential neighbourhoods

2.2. Machine Leaning (ML) classification method using Random Forest (RF)

ML methods are algorithms that learn patterns from examples (labelled set) in order to make predictions. Random Forests (RF), an ensemble-learning algorithm proposed by Breiman [9], is used in this study for residential urban form classification. The RF algorithm applies the technique of bagging (bootstrap aggregating) to decision-tree learners [10]. The idea of bagging is that multiple trees on random subsets of data are trained with replacement and then use an average in case of regression (or use the majority vote in case of classification) of their outputs to predict the label of a new observation. In this case we would get more accurate results. RF is one of the most popular ML algorithms because (i) ensemble learning generally limits the overfitting of the data, (ii) bootstrapping enables RF to work well on relatively small datasets, (iii) predictors can be trained in parallel, (iv) decision-tree learning enables automatic feature selection, (v) RF does not require much hyper-parameter tuning [9-10], and (vi) RF also provides a feature importance metric. For the RF implementation, Python's scikit-learn package is used.

In order to maximise the performance of the classifier and allow the classifier to generalise well outside the labelled set, we use the following strategy: (i) Separate the labelled set into a training set (75% of the data) and a test set (25% of the data); (ii) Train a model using solely the training set; (iii) Use the trained model to predict output values for pixels in the test set; (iv) Compute a test error by measuring the discrepancy between the predicted outputs and the known labels. The number of total labelled data is 1655 and validation size is 414 samples (25% of labelled data). Most ML models include parameters, called *hyper-parameters*, that must be tuned to obtain the best model possible for a given dataset. The *hyper-parameters* are tuned while training the model, using a procedure called *k-fold cross validation*. In order to measure the performance of the classifier (by measuring the test error), we use Precision (% of correct predictions over all predictions of a class) and Recall (% of values of a class being correctly classified). We also use the accuracy estimation, which is a classical error measure for classification tasks. The accuracy estimation computes the probability of features being well-classified in the test set, using the model built in the training set (Figure 1).

3. Results

A combination of GIS for spatial data processing and ML for classification (using a RF algorithm) are used to identify the variability of urban-form typologies within the existing cities and towns in the Arc region. We built our model based on four pre-defined classes, that is, Centre, Urban, Suburban, and Rural in order to characterise the residential urban forms within these classes. The residential urban forms and their spatial characteristics are classified based on Centre (38%), Urban (91%), Suburban (83%), and Rural (79%) within the Arc region with 80% general RF model accuracy (Table 1). The values inside parenthesis show the percentage values of a class being correctly classified (recall). The importance of the variables is also presented in Figure 2. Among the 24 features, the most important variables were three characteristics of the surrounding area (i.e. the eight neighbouring 500 m × 500 m pixels): building footprint area, dwelling density, and number of private gardens.

Table 1. Model performance results for testing data for each class. The diagonal numbers of the table (marked in bold) show the number of pixels (size 500 m × 500 m) belonging to each class that are correctly classified as part of that class. 80% is overall accuracy for the entire dataset.

Confusion matrix for predicted classes						
		Centre	Urban	Suburban	Rural	Recall per class (%)
True class	Centre	**18**	20	8	1	38
	Urban	2	**127**	11	0	91
	Suburban	1	13	**108**	8	83
	Rural	0	0	20	**77**	79
Precision per class (%)		86	79	73	90	Accuracy (80%)

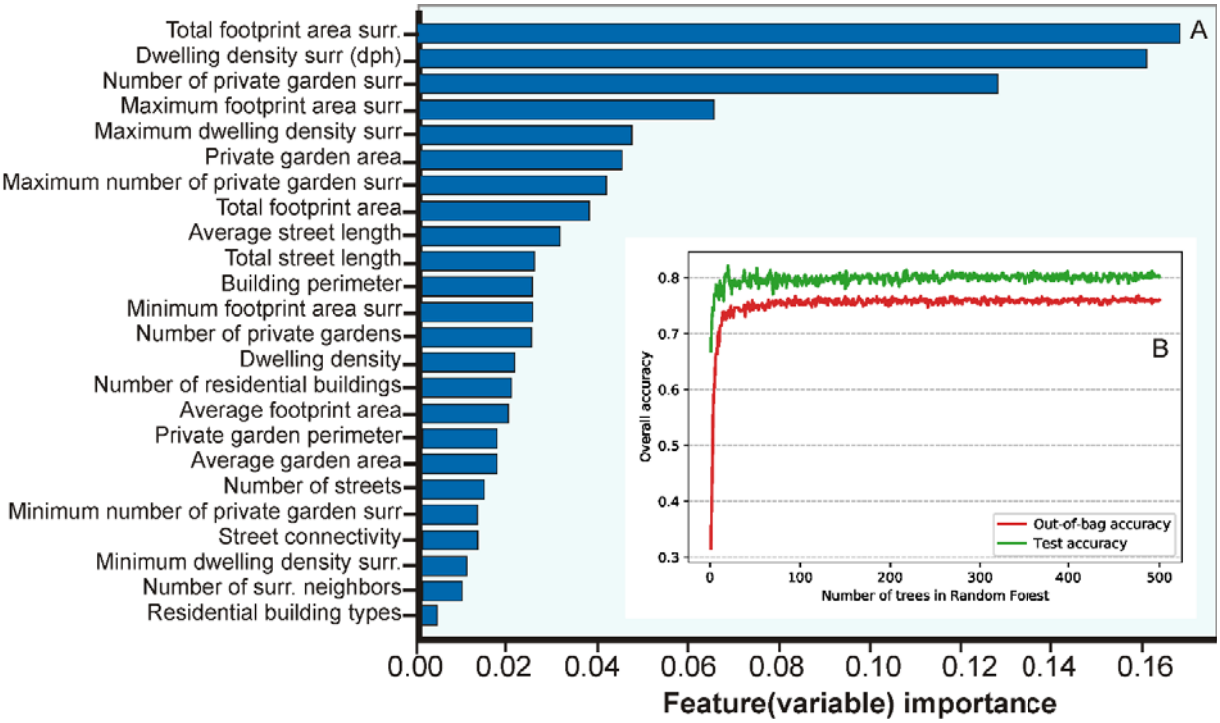

Figure 2. (A) Variable (feature) importance. The graph shows the importance of each variable during the RF training for the classification of urban forms. Surr. in the figure refers to the surroundings. (B) the graph show the Out-of-bag accuracy and test accuracy for 500 trees in Random Forest.

The result of the classification, based on the classes Centre, Urban, Suburb, and Rural, for a neighbourhood size of 500 m × 500 m in the Arc is visualised in Figure 3A. Dwelling density is an important variable (feature) in the classification (Figure 2); for example, for district councils when making plans for future housing development of the local area, particularly as regards densification. The box plots in Figure 3B show the average dwelling density in five local authority districts for each class in Oxfordshire. They also show the median and the 5th, 25th, 75th, and 95th percentiles. Of the five local authority districts, Oxford City has the highest average dwelling density in all residential urban form classes. In Oxford City, there is also a decreasing trend from the Centre (dwelling density of 43) and Urban (dwelling density of 30) to Suburban (dwelling density of 27) and Rural (dwelling density of 24). Cherwell and the Vale of White Horse have lower dwelling densities than Oxford (Cherwell being slightly higher of the two), and both also show a density decrease from Centre to Rural. Of the five districts, South Oxfordshire and West Oxfordshire have the lowest average dwelling densities, and also differ in having higher densities in the Urban class (25 and 24 respectively) than the Centre class (21 and 22).

Figure 4 shows the average dwelling density for 26 local authority districts in the Arc region for each residential urban form class. The broken lines in blue show the 5 local authority districts in Oxfordshire (Figure 3B). While there are some fluctuations in the variation in dwelling densities, most districts show a decrease in dwelling density from Centre to Rural. Of all the districts, the dwelling density in the Centres is highest in Luton (51) followed by Oxford City and Northampton (both 43). Oxford and Cambridge have the highest dwelling density in the Urban class (both 30), Suburban (27 in Oxford and 28 in Cambridge), and Rural (24 in Oxford and 27 in Cambridge). Of all the districts in the Arc, Chiltern has the lowest dwelling density with 12 in the Urban class, and 10 in Suburban and Rural.

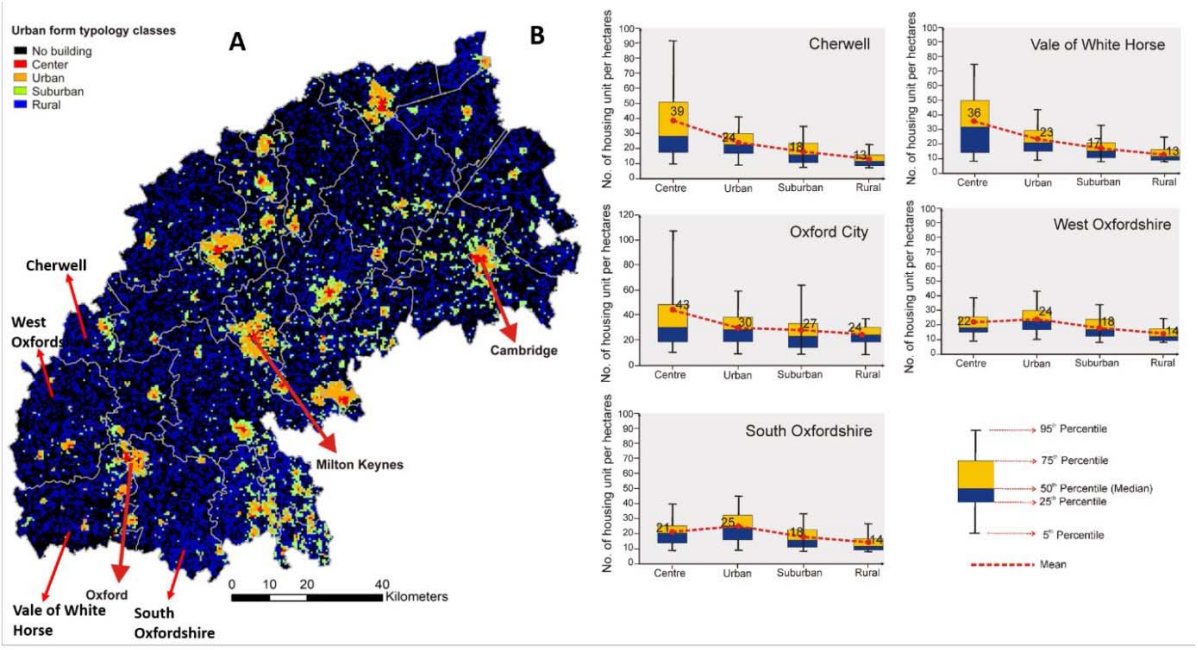

Figure 3. (A) Classification of residential neighbourhoods in the Arc, based on Centre, Urban, Suburb, and Rural classes, for a neighbourhood size of 500 m × 500 m. (B) Box plots show the dwelling density (number of housing units per hectare) of five local authority districts for each class in Oxfordshire.

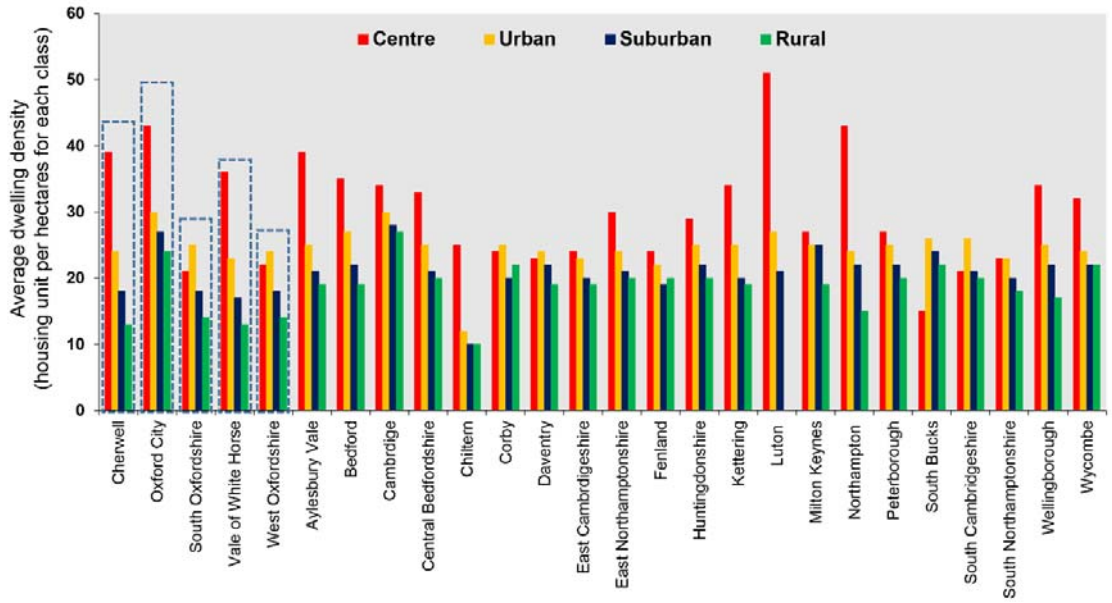

Figure 4. Average dwelling densities for the 26 local authority districts in the Arc region for each residential urban form class. The broken blue lines show the five local authority districts in Oxfordshire (see Figure 3B)

4. Conclusion and future work

We have developed a ML methodology to classify the residential urban forms based on the classes Centre, Urban, Suburban, and Rural for the Arc region between the cities of Oxford and Cambridge in the UK. For the study, we use a neighbourhood size of 500 m × 500 m. Among the most important variables (features) in the classification model is the dwelling density (number of housing units per hectare) for the surrounding

pixels. The surrounding pixels play an important role in the classification indicating that residential urban forms cannot be classified individually but should be seen in the context of the surrounding neighbourhoods. The results show the average dwelling density for the Arc region for each class is as follows: Centre 36, Urban 25, Suburban 19, Rural 14. The minimum dwelling density target set by [11] for new urban developments is 32 dwellings per hectare. The cities of Oxford and Cambridge have the highest dwelling densities (in all classes) of all the local authority districts.

As to further work on this topic, the results of the present classification can be used to forecast the likely dwelling density (number of housing units per hectare) for the densification of different brownfield lands within the Arc region for the different urban classes under a business as usual scenario where new developments aim to match the existing density. To estimate the housing capacity, we plan to take into account the future demand for housing, the likely population growth, and the characteristics of surrounding neighbourhoods. The location and the size of the brownfield land for each local authority district has recently been published by the National Housing Federation in UK and is freely accessible through the website of each local authority. We will propose different dwelling density scenarios from low to high density, with higher densities having the potential to minimise loss of 'natural capital' assets such as productive farmland and woodlands, so as to estimate the impact of different year 2050 housing scenarios in the Arc region. We plan to use a natural capital scoring approach to compare the impact of different scenarios for accommodating the planned new housing capacity by 2050, including through densification of brownfield lands, expansion of existing settlements or establishment of new settlements at different dwelling densities.

Acknowledgements
The research described in this paper was funded by the Alan Turing Institute as part of the UKRI Strategic Priorities Fund programme on AI for Science and the Government.

References

[1] Hussain M, Chen D. 2018. *Building-level change detection from large-scale historical vector data by using direct and a three-tier post-classification comparison*. In: Gervasi O. et al. (eds), Computational Science and Its Applications – ICCSA 2018. ICCSA 2018. Lecture Notes in Computer Science,**.10962**. Springer, Cham. https://doi.org/10.1007/978-3-319-95168-3_20
[2] Moosavi, V., 2017. *Urban Morphology Meets Deep Learning: Exploring Urban Forms in One Million Cities, Town and Villages across the Planet*, http://arxiv.org/abs/1709.02939.
[3] Hecht, R., Herold, H., Meinel, G. and Buchroithner, M. 2013. *Automatic derivation of urban structure types from topographic maps by means of image analysis and machine learn*ing. In: Buchroithner, M. et al. (Eds.): 26th International Cartographic Conference. PAGES?
[4] Reedes J, De Souza J, Hubbard P, 2019. *Understanding urban gentrification through machine learning*. Urban Studies **56** 922–942.
[5] Schirmer P M, and Axhausen KW. 2015. A multiscale classification of urban morphology. Journal of Transport and Land Use 9 (1 SE–Articles). doi:10.5198/jtlu.2015.667.
[6] Hargreaves A J. 2015. *Representing the dwelling stock as 3D generic tiles estimated from average residential density.* Comput Environ Urban Syst **54** 280–300.
[7] 5th Studio, 2017. Final report. *Cambridge, Milton Keynes and Oxford future planning options Project.*
[8] ITRC. 2020. *A Sustainable Oxford-Cambridge Corridor? Spatial analysis of options and futures for the Arc*. Infrastructure Transitions Research Consortium.
[9] Breiman L, 2001. *Random Forests.* Machine Learning **45** 5-32.
[10] Breiman L, 1996. *Bagging predictors*. Machine Learning **24** 123-140. doi:10.1007/BF00058655.
[11] Sarkar C, Webster, C., Gallacher, J. E. J. 2017. *Association between adiposity outcomes and residential density: a full-data, cross-sectional analysis of 419 562 UK Biobank adult participants*. Lancet Planetary Health **1** e277–e288.