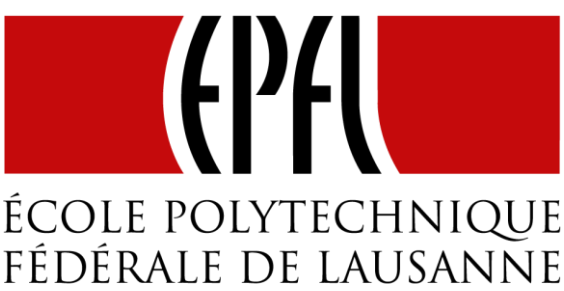
Avertissements au lecteur

Ce document est le produit d'un projet de master réalisé à l'EPFL dans le laboratoire d'Ecohydrologie sur une période de quatre mois. Son contenu n'engage que son auteur et n'est reproductible qu'avec le consentement écrits de ce dernier.

Warning to the reader

This document is the product of a master's project done at EPFL in the laboratory of Ecohydrology during a four month period. Its content commits solely it author and can be reproduced only with the written consent of the latter.



Master Thesis – 2014

Geostatistical modelling for the monitoring of environmental-related endemic diseases:

Application to the distribution of schistosomiasis in Burkina Faso

---

by

LAURENT HAFIZ

Under the direction of :

Dr. ANDREA RINALDO
Professor, ECHO, ENAC – EPFL, Switzerland

and under the supervision of :

Dr. MARIAM SOU
Professor, 2IE, Ouagadougou, Burkina Faso

JAVIER PEREZ SAEZ
PhD, ECHO, ENAC – EPFL, Switzerland

Acknowledgements

This thesis was possible thanks to the collaboration between the EPF Lausanne and the 2IE institute located in Ouagadougou, Burkina Faso.

Starting with the 2IE team, I would like to thanks my advisor Dr.Mariam SOU for her support and investment in the collaboration project, making it possible for me to spend two enriching months in Burkina Faso. Thanks also to those who accompanied me in this collaboration project in Ouagadougou and with whom a friendship is born. These are especially Nafissatou Ganou, David Yameogo, Frederic Campaore and Mohamed Bagayan with whom I shared as much instructive talks as good moments.

For the Swiss team, I would like to thank especially my supervisor Javier Perez Saez to have suggested me to participate to that collaboration, which changed to be a very interesting cultural exchange with the African continent and enriching educational project. Thank you very much for your inspiring advices, continuous support and personal investment. Of course thanks to Professor Andrea Rinaldo for making that project possible. Thank you to Jean-Marc Froehlich for his sympathy and advices; to Tabea Schutter for the mutual support we brought to each other; to Natalie Ceperley and Theophile Mande for their sympathy, advices and beautiful shared moments.

Finally, a special thanks comes to my family, mother, father and brothers; and friends who always have been here to support me all along these studies.

Abstract

The increased interest of reducing the infection rates of neglected tropical diseases like schistosomiasis in the world has raised the necessity of developing epidemiological monitoring techniques, in order to target specific areas where the risk of infection are at highest.

The aim of this project was to produce infection probability maps of the urinary schistosomiasis, caused by the parasite *S.haematobium,* in order to identify high risk zones where targeted interventions could be undertaken in Burkina Faso. These maps were produced thanks to Bayesian analysis techniques, using geostatistical generalized linear models. The predictions were effectuated thanks to 247 community level infection prevalence data collected from the published literature, using environmental predictors as the NDVI, population density, elevation, mean temperature, mean decadal rainfall estimates and a mean dry-season period time.

The predicted results showed that prevalence rates were at highest in the northern part of the country, with a tendency to decrease in a homogeneous way to the South. The absence of heterogeneous covariates, explaining more localized environmental information like distances to water bodies or mobility information, prevented the geostatistical model to explain the local variations in *S.haematobium* prevalence rates. These could be integrated in the model for future works to see their capability to explain heterogeneity in the prevalence rates observations.

Summary

1 Introduction

Schistosomiais is a waterborne parasite-induced disease affecting around 240 millions of people around the world, with more than 700 million of people being at risk (WHO, 2014). The disease is closely correlated with conditions of poverty, poor sanitation and lack of clean water, and is emerging in areas undergoing major water resources development and management (Southgate, 1997).

A global strategy for controlling schistosomiasis has been proposed by the World Health Organization (WHO) and aims to inject preventive chemotherapy, meaning a repeated large-scale administration of an antischistosomal drug prazitanquel to at-risk populations (Schur, 2012). But despite all these efforts undertaken since many years, as shown in Burkina Faso with the National Program started in 2008,schistosomiasis prevalence rates remain globally high. In order to control and perform the fight againstschistosomiasis at a national scale, there is a need of a monitoring of the disease distribution. Furthermore, most prevalence estimates lack empirical modeling, which is necessary in order to obtain accurate averages of infection risks over large geographical regions (Schur, 2012).

The approach of this project is an occasion to satisfy the lack of such modelling. It proposes a method in order to generate infection risk maps that can help directing the programs of disease controlling, by using Bayesian geostatistical models combined with socio-environmental information layers.

This work consists of answering to the following question:

How well can socio-environmental spatial information accurately estimate infection risk probabilities at a national scale through Bayesian geostatistical modelling?

Datasets of prevalence rates combined with environmental layers, mainly derived from satellite imagery, obtained from different sources and surveys, will constitute the necessary information in order to fit and run a geostatistical model for predicting prevalences over the country.

This project will be structured as follow: a first part will describe the theoretical knowledge acquired through documentation reviews. This part will consists of explaining in a more detailed way the theory of schistosomiasis in the world and in West Africa, followed by a description of the tools used to store and process the raw data, and finally by detailing a more detailed overview of the geostatistical model theories used as prediction tools in epidemiological studies. The second part will describe the overall methodology in order to obtain the predicted infection rates. A third part will present and comment the final predicted results. And a final chapter will be dedicated to the discussions and perspectives that discharge from this project, followed by a conclusion.

2 Theoretical background

Before presenting the methodology and the results of the work, it is essential to present a theoretical review of all components of the work process. First of all, it will be necessary to present an overview of schistosomiasis as a neglected tropical disease, its current state over the world and more precisely in West Africa, and its transmission cycle and implications for control. Then, the tools necessary for gathering and construct the spatial data underpinning the geostatistical modelling endeavour will be presented. These tools regroup Database Management Systems (DBMS) and Geographic Information Systems (GIS). Finally, a theoretical review of the geostatistical model will be presented, and which software will be used and why.

2.1 Schistosomiais generalities

2.1.1 Schistosomiais in the world

Schistosomiais is a parasite-induced disease that is also known as bilharzia. The parasite, or shistosome, was first identified by Theodor Bilharz in Egypt in the year 1851.schistosomiasis affects around 240 million people in the world and approximately 700 million people live in endemic areas (WHO, 2014). The infection is majorly present in tropical and subtropical areas (Figure 1Figure 2), generally in poor communities where access to water and sanitation are inadequate. The disease is endemic in 74 countries of Africa, South America and Asia (Boelee, 2006). In Africa, it is estimated that bilharioza is endemic in 46 of the 54 countries of the continent (Van.d.Werf, 2003). In term of morbidity, the disease creates more than 4.5 million DALYs[1] lost (WHO Expert Committee 2002). These numbers show that this disease is a worldwide risk for human health and many of international organization like the World Health Organization considers this desease necessary to be seriously monitored and treated. For definition purposes, the rate of infection, characterized by the total number of cases in a given population at a specific time, is known as prevalence.

[1] DALYs : disability-adjusted life years, an indicator that allows comparison between acute and chronic diseases. Calculated as the sum of years of potential life lost due to premature mortality and the years of productive life lost due to disability

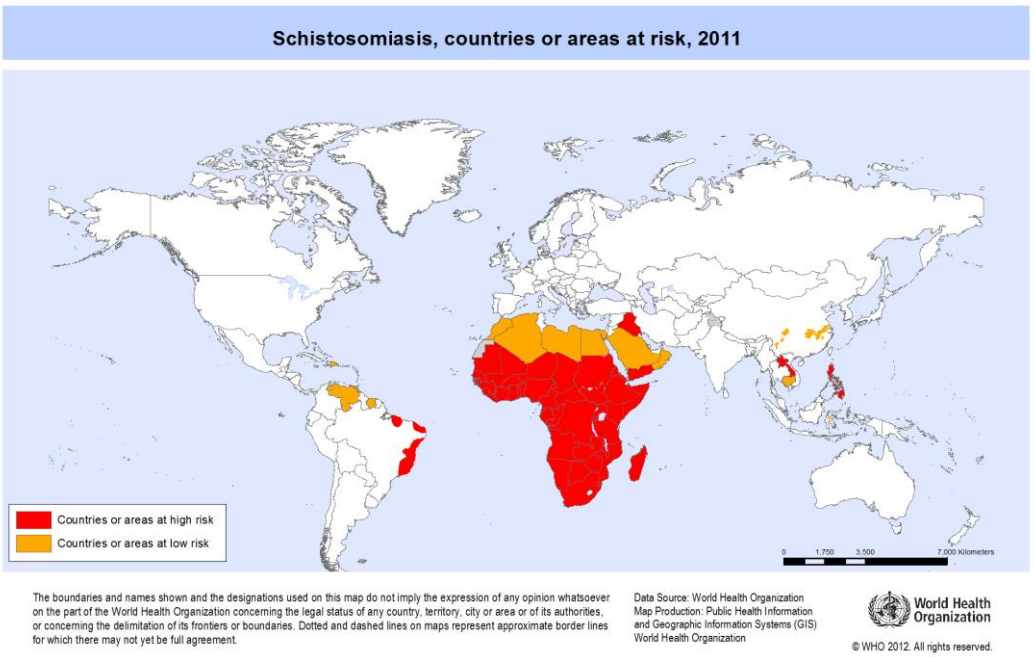


Figure 1 Geographic distribution ofschistosomiasis (WHO, 2012)

There are 19 different species of *Schistosoma* parasites being infectious to either animal, the zoophilic schistosomes, or humans, anthropophilic schistosomes. More precisely, five species can infect humans ; and all of them having similar life cycles involving freshwater snails acting as intermediary hosts, while human represent the final host. The five human-oriented parasite species are (Pearson, 2013):

- *S.haematobium :* it is widely distributed over the African continent and causes urinary tract diseases. Also found in the Middle East, Turkey and India
- *S.mansoni :* also widespread in Africa, but also in the Middle East and other parts of the world like the Western Hemisphere, parts of South America and Carribbean Islands. Responsible for intestinalschistosomiasis.
- *S. japonicum* found in Asia
- *S. mekingo* found in Southeast Asia
- *S.intercalatum* found in Central and West Africa

The cycle of the schisotomas reproduction is a key concept for understanding the close relationship between the environment, hydrological factors, human activities and the disease persistance and spreading. Below on Figure 2, a representation of this cycle is shown, which is common to all schistosoma species. Only the type of parasite, according to a certain species of intermediate host and final host can differ.

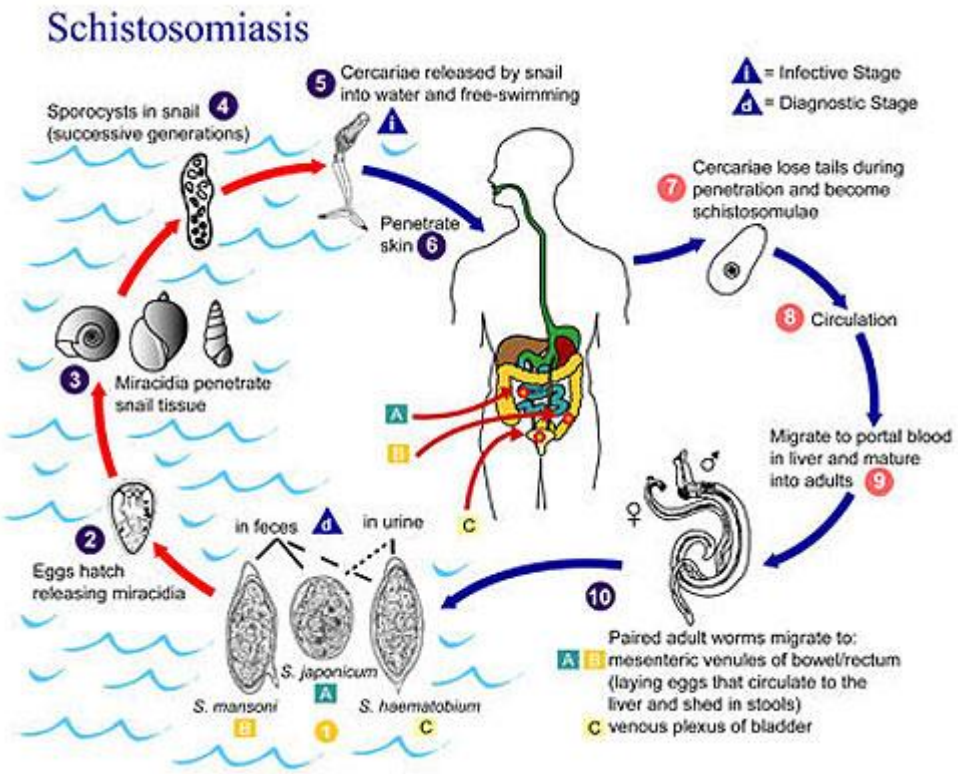


Figure 2: Cycle of Schistosoma reproduction (CDC)

The cycle starts when the human, or animal, final host releases the mature parasite‘s eggs in the water either contained in urine (*S.haematobium*) or in faeces (*S.mansoni*). In the aquatic environment, the eggs evolve until reaching a state of miracidia (step 3), capable of entering in the snail intermediate host. Within the snail, the miracidia progress through two generations of sporocysts to become cercariae (step 4). These cercariae, capable of swimming, are released from the snail and penetrate the skin of the final host. During penetration, the cercariae lose their forked tail and become a so called *schistosomula*. In the final host, these *schistosomula* mature into adults. Then, the paired (male and female) adult worms migrate to the intestinal veins or rectum or to the venous plexus, depending on the species, where they are able to reproduce and produce eggs.

Transmission of the disease can take place in a wide range of habitats, from large lakes and rivers to small seasonal ponds. It has been shown that there are preferential habitats for different intermediary hosts (Poda, 1996) and this could partially explain the different distributions of schistosomiasis, at large scale and small scale.

The treatment of bilharzia has for now a unique solution and consists of a single-day oral treatment with an antihelmintic called praziquantel (Utroska, 1990). The effect of this strongly concentrated pill is to kill most of the schistosomes in the body and can produce relatively strong adverse effects like abdominal pain, diarrhea, headache and dizziness. Praziquantel has been used over the past 30 years to control schistosomiasis in many countries (WHO), like in Egypt or China where mass treatment, combined with preventive chemotherapy, have been applied and have shown significant reduction of prevalences. Nevertheless, these campaigns could still not eradicate the disease and the re-infection risks in many countries (WHO, 2014). Burkina Faso is an example of such country where the disease

is still strongly present, and where other solutions combined with existing ones have to be thought in order to effectively reduce the infection rates.

2.1.2 Schistosomiasis in West Africa Burkina Faso

In West Africa, the distribution of bilharzia is organised in centres of variable endemic levels (Poda, 1996). This diversity is due to a variety of factors like the ecology of the intermediary hosts than the behaviour of the final host in his environment.

Generally in West Africa and in Burkina Faso, the parasite species found in majority are *S.haematobium* and *S.mansoni*, responsible for uro-genital and intestinal bilharzia respectively. The next figure is showing an estimation of the prevalence data of the two different parasites in Africa.

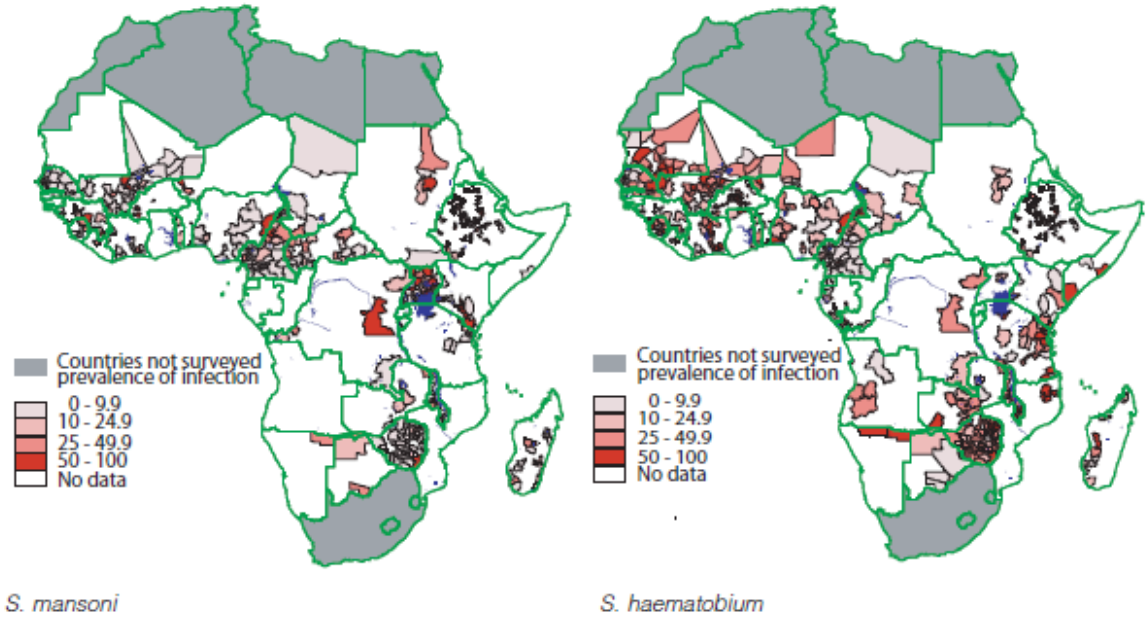


Figure 3 Prevalence ofschistosomiasis caused by *S.mansoni* and *S.haematobium* in sub-Saharan Afica by district *(Source : Simon Brooker, Imperial College School of Medicine, London, 2012)*

In this work, focus will only be put on *S.haematobium* and *S.mansoni*, since they are the only one found in Burkina Faso, and all the diagnostics performed on the population are mainly focused to identify these two species of schistosomes.

The next paragraph will introduce the concepts of databases and spatial information management, being necessary tools for the evolution of this work.

2.2 Database management systems and GIS programs

Since the beginning of the project, it was known that a large amount of spatial and non-spatial data would have to be collected in order to be used in different steps of the work. It was therefore necessary to structure these data and organize them in a way that they can be used, stored, modified as simply and rapidly as possible. Database management systems give a good tool when dealing with

a large number of data shared by different users. Since this work is part of a larger project regrouping different actors, students and collaborators in two different countries (Switzerland and Burkina Faso), and that this project is only in its beginning, it was necessary to construct a database.

Because this project contains large amount of geographical data, the best option in terms of software choice was to use a database system management capable of storing spatial data. The software PostgreSQL was a good candidate for the management system. PostgreSQL is an object-relational database management system (ORDBMS). Object, meaning that it is an object-oriented database model: the objects, classes and inheritances are directly supported in database schemas and in the query language. For PostgreSQL, the query language is SQL, an easy and very intuitive language capable of accessing datas in only a few lines of request. The SQL, literally "Structured Query Language", has been designed for optimizing queries in relational database management system. The general structure of an SQL query is as following: "SELECT [variable] … FROM [a certain schema.table] … WHERE [logical condition applied to that variable]".

In addition of being an open-source software, PostgreSQL has the advantage of having a spatial extension, PostGIS, able to store, manipulate and create spatial objects. This extension is very interesting since it enables to mix non-spatial data to spatial data and make spatial analyses or calculation, e.g. distance. Thanks to integrated spatial functions, it is possible to rapidly and automatically select, calculate, and modify spatial objects through simple SQL queries.

Another advantage of using the PostGIS extension is that the structured geographic information stored in the server can be directly loaded into GIS programs like QuantumGIS (Team, 2014). This software is an open-source geographic information system which provides data viewing, editing and analysis capabilities. Even if the functionalities as spatial analyst are relatively limited, QGIS (TEAM, 2014) is very useful for creating maps, viewing data and easily reaching the whole structured database.

For more advanced geographic operations, the desktop software ArcMap from ESRI's ArcGIS (ESRI, 2011) was used. This GIS is much more efficient for more complex operations, giving a large panel of functionalities. The advantage of using these three programs (PostGIS, QGIS (TEAM, 2014), ArcGIS (ESRI, 2011)) is that the file formats are compatible between each other, and that the connexion to the database is rapid and efficient. This combination of programs are well adapted to the construction of the data necessary to the model building, which is going to be described in the next chapter.

2.3 Geostatistics: a prediction tool

Spatial statistics and analysis are very useful in studying spatial objects in order to qualify or especially to quantify their topological, geometrical or geographical properties. The underlying hypothesis is to consider that a certain event, in our case the probability of being infected, is not randomly distributed over space. Spatial statistics help us identifying spatial processes by extracting patterns and spatial relationships. Geostatistics is one of the main branches of spatial statistics (Cressie, 1993) and deals with modelling and inference of spatially continuous phenomena. It applies concept going from classical probabilistic statistics to Bayesian-based statistical analysis. The theory

of geostatistics is though very large and its applications diverse. In this chapter, the focus will be put on the use of model-based Bayesian inference as a spatial prediction tool.

For a better contextualisation, it is important to remind that these statistical tools are required in order to predict the prevalences of schistosomiasis infection over the whole country. Furthermore, the prediction is based on a model, trying to construct a generalized linear model to link the prevalence to some chosen environmental parameters, and to a certain spatial effect. This spatial effect will appear in the model building and is the "geo" part of the term "geostatistics", differentiating a purely statistical regression to what is tried to be done here: a spatial prediction.

As introduced before, a generalized linear model will be used for the prevalence predictions. It is necessary to redefine more precisely the differences between linear and generalised linear models. Furthermore, the "geostatistical" part of these models will be explained and why they have to be used in this work.

First of all, ordinary linear models take into account that a certain response variable Y is distributed normally with a certain mean μ and variance σ^2 and changes linearly with a weight beta times an observed value X. It implies that a constant change in the predictor is leading to a constant change in the response variable. This is shown in the equations (1):

$$Y_i \sim N(\mu_i, \tau^2)$$

$$\mu_i = \beta X_i$$

The model consists of a simple fitting by finding the optimum slope beta.

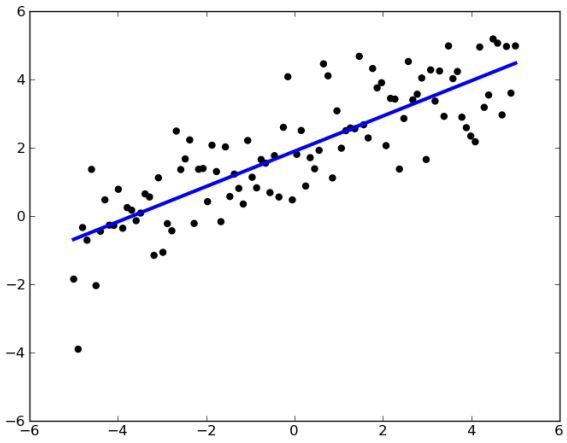


Figure 4 Ordinary Linear Regression

But often the distribution of a certain variable doesn't follow normality, or cannot (for modelling), for many different reasons. For example, when the response variable is the probability of making a yes/no choice, typically a Bernoulli variable, than the ordinary model is not suitable since the response variable has to take values bounded between 0 and 1. Extreme observed values cannot induce higher probabilities than 0 or 1.

When this type of non-normal situation happens and linearity wants to be kept, meaning finding the β weight, it is necessary to work with generalized linear models (GLM). These GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function. The magnitude of the variance of each measurement can be set as a function of its predicted value. The next equations (2) are defining GLMs:

$$Y_i \sim f(\mu_i; \tau)$$

$$g(\mu_i) = \beta X_i$$

g() is the so called-link function, making it possible to keep a linear relationship between the « generalized response » and the predictor variable X.

Typically, following the example of the Bernoulli process, the response variable of the summed Bernoulli independent variables would follow a Binomial law (function f() in equation 2), and the link function would be a logit function, illustrated as an example in the next figure.

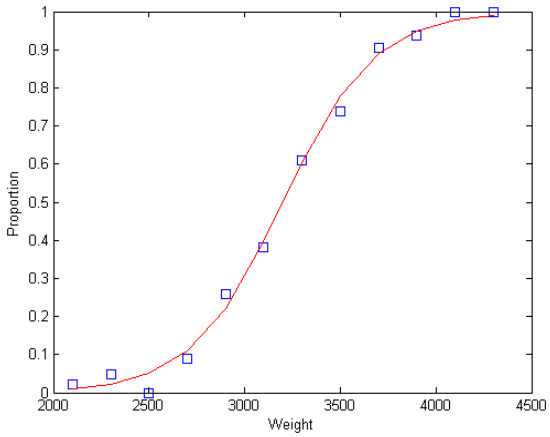


Figure 5 Generalized linear models : logistic function

Secondly, these types of model, e.g. Generaliszed Linear Models, do not take into account the spatiality of a certain variable distribution. Spatiality is obtained by adding to the deterministic component of the predictor a stochastic component with covariance in the random effects. The geostatistical model that is going to be used in this work is shown in the next equations, integrating the spatiality:

$$Y_i|U(s_i) \sim f(\lambda(s_i), \nu) \quad (1)$$

$$g[\lambda(s_i)] = \mu + \beta X(s_i) + U(s_i) \quad (2)$$

$$cov[U(s_i), U(s_j)] = \sigma^2 \rho(h/\Phi; \theta) \quad (3)$$

Having Y the realization (observed values) of a distribution f, e.g Binomial with parameters λ and ν, g() the link function (i.e logistic) of the generalized linear model, μ the intercept, β the regression coefficients, X the covariates (predictors) and h the distance $|s_i - s_j|$. Writing U(s) as the value of a Gaussian random field U at location *s*, the geostatistical model is characterized by the joint multivariate normal distribution:

$$[U(s_1)...U(s_n))]' \sim MVN(0, \Sigma)$$

The covariance, ∑, of two realization of that random field U, is defined by a spatial correlation function ρ(h/ϕ,θ), where h is the distance between two points, ϕ a scaling parameter and θ a spatial parameter. An example of such correlation function is the exponential correlation function ϕ, defined by:

$$\phi(\tau) = \sigma^2 exp(-\alpha|\tau|)$$

Where σ and α are, respectively, the standard deviation and the inverse correlation time of the process. This process is also known as the *Ornstein-Uhlenbeck process.*

Generally, in the case of Gaussian data, the optimization of the model's parameters is done using Maximum Likelihood Estimates (MLE's). In this case, the combination of non-Gaussian data and an unobserved latent variable make the likelihood function intractable and computing the MLE's difficult (Brown, 1996).

A solution for finding the optimum model parameters in non-Gaussian responses is to use Bayesian inference techniques, using for example Markov Chain Monte Carlo (MCMC) algorithms, which has been shown to be the most common method for making statistical inference with Generalized Linear Geostatistical Models (GLGMs) (Brown, 1996). Bayesian inference implies specifying prior distributions for the parameters μ, β, ϕ and σ of equations 1), 2) and 3).

An alternative to MCMC is the Integrated Nested Laplace Approximation (INLA) algorithm (Rue, 2009). While MCMC algorithms can be difficult and require a specialized skill set, it has been shown that INLA is much easier to use and much more computationally time-efficient (Brown, 1996).

For computing and running these models, the statistical software R has been used. More details about the computation and programming will be detailed in the methodology section.

3 Methodology and primary results

In this section, the methodology of the whole project will be described in details, also presenting the intermediate results which were necessary to obtain the final results from the geostatistical model. Two main parts will be underlined here: a part dedicated to data acquisition and preparation on one hand; and model construction and running on the other hand.

3.1 Data acquisition and preparation

In order to run a model, a necessary work has to be dedicated to data acquisition, storing and preparation. Since the aim of the work is to create a prevalence map over the whole country of Burkina Faso, it is a need to correctly choose all the required information. These informations are mainly spatial, and can be separated into two groups: environmental data and disease related data.

Environmental data are necessary in order to construct the covariates for the geostatistical model. These data are also useful when creating maps to have a better visualisation of the countries organization, in terms of transport, population, and administrative zones.

The disease related data will form the observed response variables for the model; these are mainly prevalence data. In this chapter, these prevalence data will be presented in a more detailed way, explaining the sources and how they can be used as best for the model construction.

A first chapter will be dedicated to the database management, in which it will be explained how the data have been imported and organized in the DBMS PostgreSQL.

3.1.1 Database management

The database creation was an important path of the global work. Integrating all the appropriate information in a well-structured way is a necessary step. The main advantages of having all the data stored in one single database are, first, the capability of accessing rapidly the information by queries, using conditions and joining tables together. Secondly, it is a good way of keeping a same structure and organization between all the collaborators that will need to access the information. Because the whole database can be exported as one block, it can be reloaded, and therefore shared between different computers. It could also be possible to upload this database on a shared server, so the modification could be applied to all the users.

PostgreSQL gives also the advantage of separating the data, i.e the tables,into different so-called “Schemas”. The database was therefore separated in four different schema, each one of them representing a different “type” of information. The next table describes the general structure of the database, showing the four different schemas, “Vectors”, “Rasters”, “Schisto” and “Sites”; and their respective tables.

Schemas	*Vectors*	*Rasters*	*Schisto*	*Sites*
Tables	bfa_adm0 bfa_adm1 bfa_adm2 bfa_rivers bfa_water_areas hydrobasins regions_geo reservoirs roads villages water_landsat	dry_season hii landcover landcover_names ndvi popdensity rain rainstd temperature temestd	gntd gntd_classed pncs db_r haema_covariates manso_covariates	eaulioulgou ecoliersliouloou lioulgou panamasso

Table 1 Database organization

In this next paragraph, the different schemas will be detailed for a better understanding of the database organization:

Vectors: This schema contains all the spatial information being of vector type, meaning, *administrative boundaries, river network, water areas, watershed boundaries, dams, roads and villages.*

Rasters: It is important to separate the raster and the vector data because they are structurally very different in the DBMS. Rasters have to be loaded differently because their structure lack of tables and attributes, since raster are images, with only projection and pixel-resolution metadata. More explanation about raster and vector will be given in the next chapters. The function *raster2pgsql* is used in the PostgreSQL terminal in order to load the raster in the database. A typical command line is given here:

C:\Program Files\PostgreSQL\9.3\bin> ***raster2pgsql -s 4326 -I -C -M D:\Users\JMF\Desktop\MasterProject\geodata\Precip\final_clipped_mean_rain.tif -F rasters.rain | psql -d burkina -U postgres***

A deeper explanation of the above code can be found at the official PostGis website[2]

Schisto: This schema contains all the information related to the disease data. It is also essentially spatial information, consisting of mainly two used disease-related "databases", GNTD and PNCS, which will be explained more deeply in the next chapters.

Sites: Finally, this schema contains all the information gathered on site in Burkina Faso during. The tables are related to interviews done to the population of two different villages, Lioulgou and Panamasso.

[2] http://postgis.net/docs/using_raster_dataman.html

3.1.2 Environmental data

The environmental data has been classified in two categories: raster and vector information. As a reminder, the difference between these two types of data is that raster information is composed of pixels, while vectors are composed of path. A raster, having formats like "*:JEPG"* or "*.GIFF",* is a representation of the world as a surface divided into a regular grid of cells. Therefore, rasters are useful in storing data that varies continuously, as a satellite image or an elevation surface. On the other hand, vector information is a representation of the world using points, lines and polygons. They are useful in order to store data that has discrete boundaries, such as country borders, rivers…etc. Both of them are useful for representing spatial information, and their graphical representation can be thought as described in the next figure:

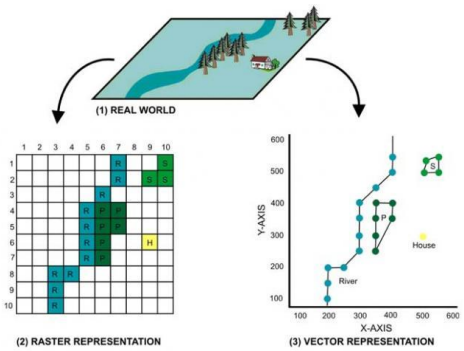


Figure 6 Representation of the world with vector and raster data[3]

All the following figures were produced by the QGIS (TEAM, 2014) print composer.

i) Vector information

In this chapter, a list of the vector raw information obtained for the spatial database creation will be displayed, describing the sources and their utility for the project.

a) Administrative boundaries

As a first step, it was important to obtain the spatial limits of the project. Since the work focuses on Burkina Faso, administrative boundaries are necessary. The next map is representing three different polygon layers:

- The country borders as a red line (*bfa_adm0 in the database*)
- The provinces as coloured polygons (*bfa_adm1 in the database)*
- The departments as black lines *(bfa_adm2 in the database)*

[3] http://www.sigare.net/etudes/tech_donnees.htm

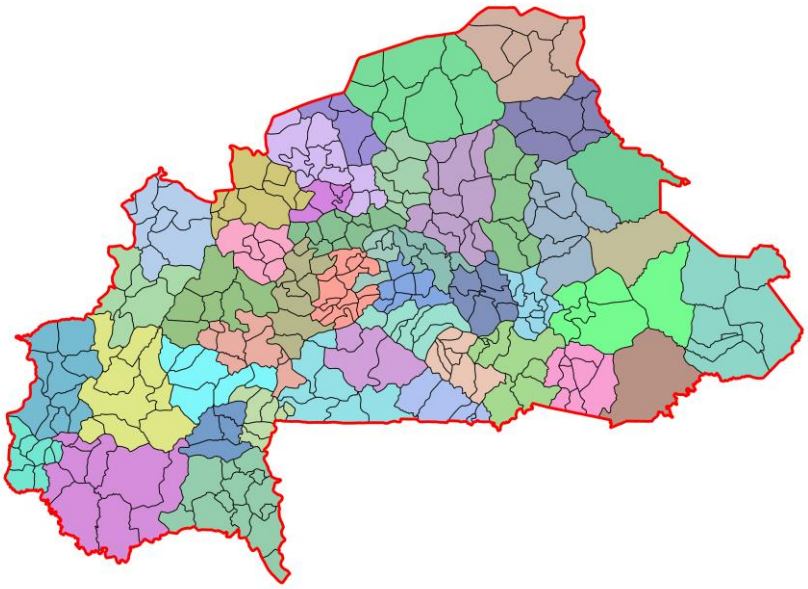


Figure 7 Administrative boundaries of Burkina Faso

These files can be uploaded from the internet. A possible source can be found at this link[4] Not visible in Figure 7, the regions are also important administrative information, which can be observed in the next map:

Figure 8 Regions of Burkina Faso

[4] *http://www.mapmakerdata.co.uk.s3-website-eu-west-1.amazonaws.com/library/stacks/Africa/Burkina%20Faso/index.htm*

b) Hydrological network

(a) Rivers and dams

Since development of schistosomiasis is closely related to the hydrological regime of the infection sites and the development of water retention elements like dams (Poda, 1996), it is relevant to gather spatial information about the hydrological network of Burkina Faso.

In this chapter, three main groups of hydrological data will be presented, which are: a river network obtained from an exterior source; then a database containing precise information about dams and finally water areas obtained by a supervised classification of satellite images realized in this work.

First, the river network is composed of lines, having an attribute informing about perennial or temporal/seasonal regime of the river. The shapefile of the river network of Burkina Faso can be found at the following link[5]

Then, the dams have been obtained thanks to an inventory work (DRAH/RH--EIER-ETSHER-2IE, 2005) and gather relatively precise information of around 1200 dams in the coutry. These information contain for example the year of construction, the volume of retention, quality and type of the dam.

The next map illustrates the river network and the dams, distributed around Burkina Faso.

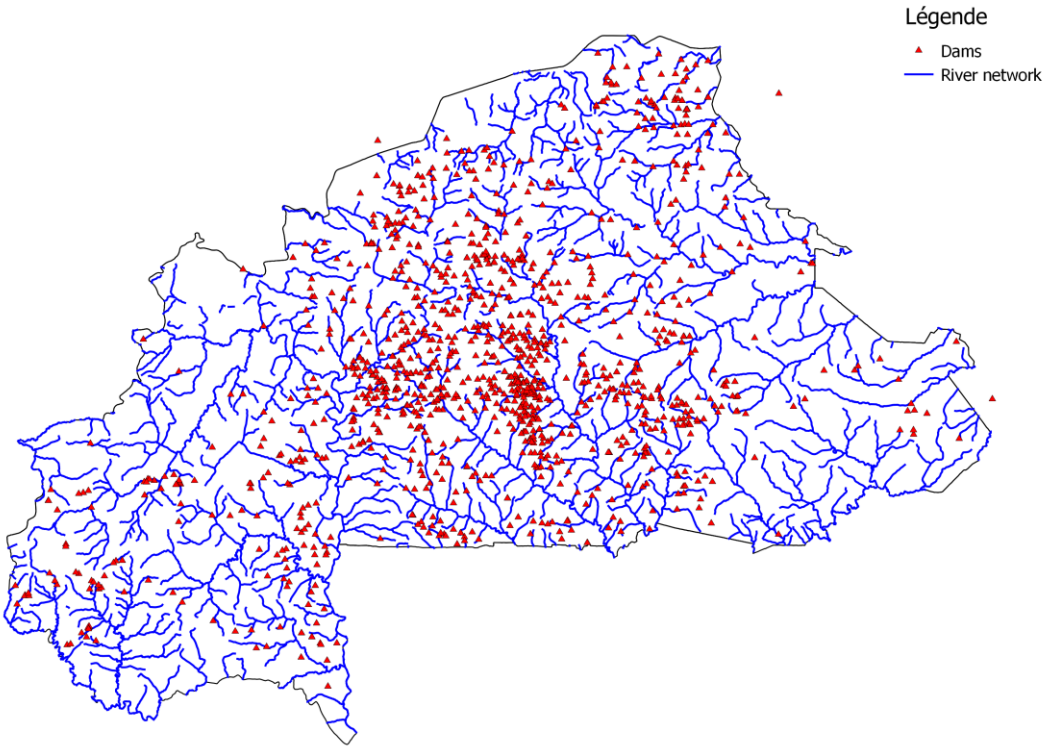


Figure 9 River network and dams

[5] *https://www.humanitarianresponse.info/applications/data/datasets/locations/burkina-faso*

(b) Supervised classification for water areas

As an important part of the hydrological network, retention basins are very common in Burkina Faso and the construction of dams increased since the past fifty years (DRAH/RH - BF /EIER-ETSHER -2IE, 2005). It was therefore interesting to obtain spatial information about these water areas. A commonly used process to obtain such result is to extract information from satellite imagery, using classification methods.

The methodology employed for extracting such land cover information, i.e. presence or absence of water, will be presented in these next paragraphs.

The aim of classification methods in imagery is to assigning observed spectral signatures to distinct classes, in this case the class *Water* and the class *No Water*. Classification techniques differ by the fact that they can be unsupervised or supervised, differing by the absence or presence of user-defined training data. In this case of classification, a supervised method is more appropriate, since the needed result only consists of two groups. Since water has a well differentiated spectral signature, supervising the classification method makes it easier to class a pixel to the wished cluster.

The method that is been used for the classification is the *maximum likelihood classification* (MLC), which has the advantage of being relatively simple to use and is implemented in a wide range of programs. MLC needs an *a priori* set of sampled pixels that are regrouped in the wished classes to classify. Then it proceeds by assigning each pixel of the image to one of these classes. The likelihood is expressed as a function of distance between the values of the spectral bands of the analysed pixel and the mean and covariances of the spectral signatures of the *a priori* defined training set.

The first step of the classification was to acquire the necessary satellite images, covering the whole country. These were found at the *United States Geological Survey* website[6], from the new Landsat 8 satellite, having a resolution of 30 meters. A more detailed description of this satellite and the delivered images can be found at the following website[7]. The satellite covers the whole planet at a period of 16 days. In order to cover the entire surface of Burkina Faso, 18 images had to be used, representing a heavy amount of data, since each image is composed of 11 different spectral bands.

A first filtering of possible images had to be done. Were only selected the images that had a cloud covering less than 5% in order to not miss any water surface that could be hidden. A second filter had to be applied, selecting the 18 necessary images at a relatively same period of the year, since Burkina Faso suffers from important evaporation, where water areas can be absent during dry seasons. Since the rainy season ends around the end of the year, the images were selected between end of January and beginning of February, when water is still present and clouds are absent.

An overview of the Landsat image is shown in Figure 10, showing water areas by selecting only a specific band combination (R: 5-Near InfraRed / G: 6-ShortWave InfraRed / B: 4-Red) that especially highlights the differences between land surface and water.

---

[6] http://earthexplorer.usgs.gov/

[7] *http://landsat.usgs.gov/landsat8.php*.

The next step of preparing the classification was to create composite images for each Landsat image, since the downloaded images are delivered as multispectral independent images. For a rapid execution, the software Matlab and its ToolBox Mapping was used.

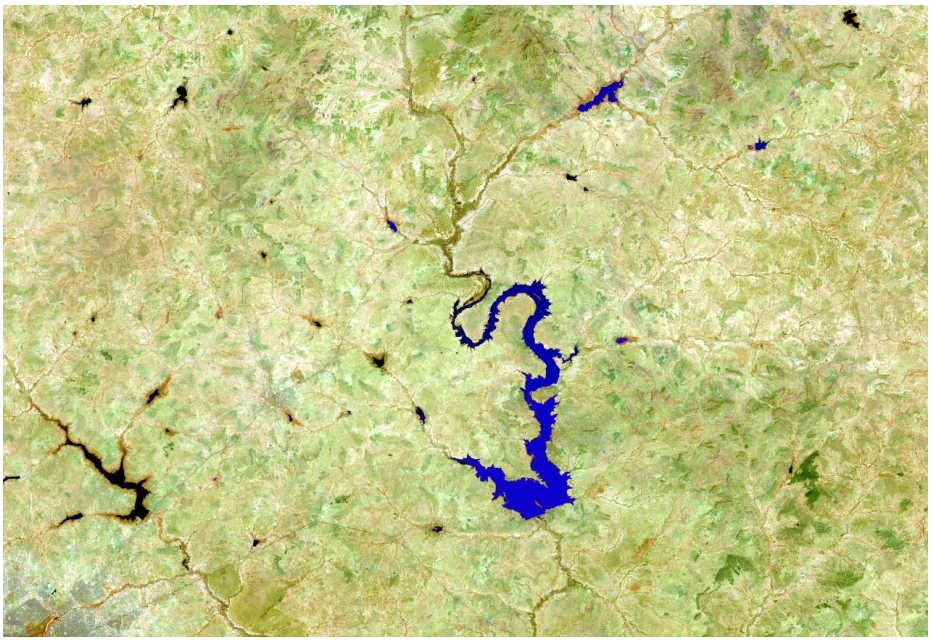


Figure 10 Landsat image with water-highlighting bands configuration

The classification was performed on ArcGIS (ESRI, 2011). A signature file was created by selecting samples of different water areas in order to create a class *Water* and different types of land surface were creating for a class named *Other.* The output of the ArcGIS (ESRI, 2011) MLC function is a raster, having two different possible values for created class in the signature file. A visualisation of this raster is shown in Figure 11, where the water areas are clearly visible.

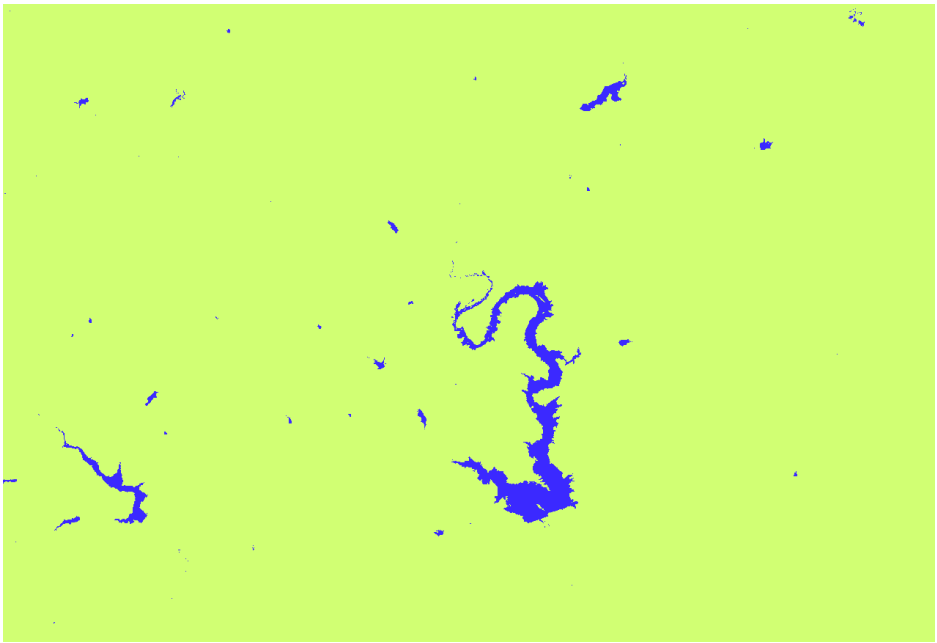


Figure 11 MLC classification raster output *(Green: Others, Blue : Water)*

The final step was to vectorize the obtained raster in order to extract the classified water areas. In order to have a cleaner output vector, a majority filter was applied on the classification raster. This filter enables to replace a pixel value by the value of the majority of the neighbouring pixels. It helps neglecting very small areas, or to smooth bigger areas that could be cut in two for example. After applying that filter, a simple vectorizing function enables to extract the water areas as a multipolygon spatial structure. The result is now a *shapefile* containing all the water areas, and a representation of it is showed in the next figure:

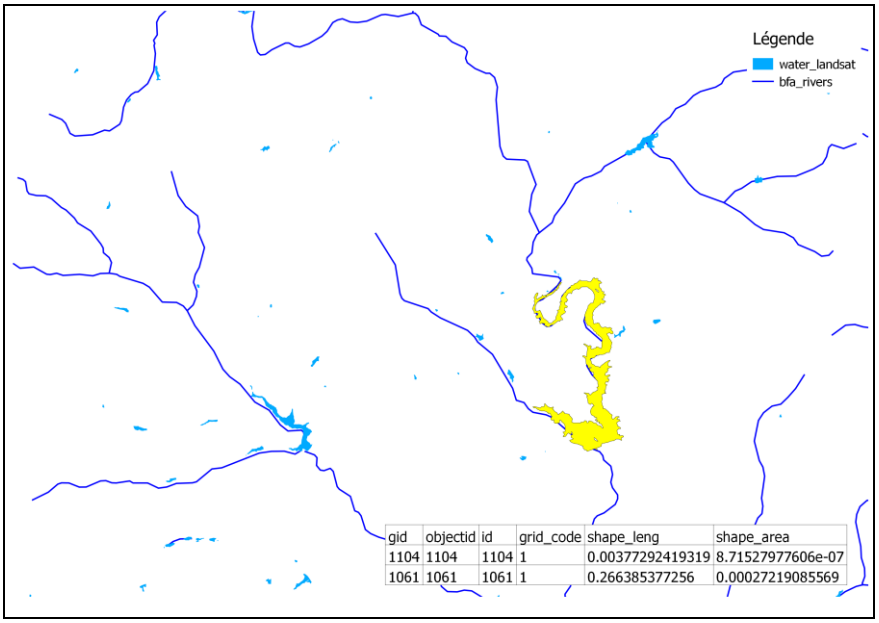


Figure 12 Vectorized water areas

c) Road network

Another vector information contained in the database is the road network in Burkina Faso. These information were obtained by the NASA's Global Roads Open Access Data Set (gROADS), accessible at the following website[8]. These data were actualized during the period between 1980 and 2010, furnishing relatively up-to-date informations.

[8]. http://sedac.ciesin.columbia.edu/data/set/groads-global-roads-open-access-v1/maps

.

Figure 13 Road network

ii) Raster information

In this chapter, the spatial information being of raster type will be exposed. The next table describes the different used rasters and their respective details and sources. Important linked information is the *Date*, *Temporal Resolution* and *Spatial resolution.* The websites of the downloaded rasters can be found in Annex 1, and the definitions of the data types will be described in the next section :

Data type	Source	Units	Date	Temp. Res.	Spat. Res.
DEM	USGS/GMTED2010	[m]	2010		230 m
NDVI	eMODIS TERRA	[-]	2001-2010	10-day	250 m
RFE	USGS/FWES	mm	1999-2014	10-day	8 km
LST	MOD11C2	°C	2012-2013	8-day	5 km
Landcover	MCD12Q1/Type2(UMD)	classes	2012	-	500 m
HII	Last of the Wild, v2, NASA	[-]	1995-2004		1 km
Population	GRUMPv1	[-]	2000		1 km

Table 1 Sources of raster layers

It is important to notice that for some of these raster informations, i.e. RainFall Estimates (RFE) and LandSurface Temperature (LST), it was necessary to do a pre-treatment in order to calculate means and variances for the available period. The next chapters will describe each raster information one by one.

a) Digital Elevation Model (DEM)

The DEM was available as it is shown in Figure 14, only a *Clipping* function was used to cut out the unnecessary information out of Burkina Faso. This elevation model was developed by a collaboration between the USGS and the NGA, has been named DMTED2010 and replaces the old version GTOPO30, being a dataset of choice for global and continental scale applications (Danielson, 2011).

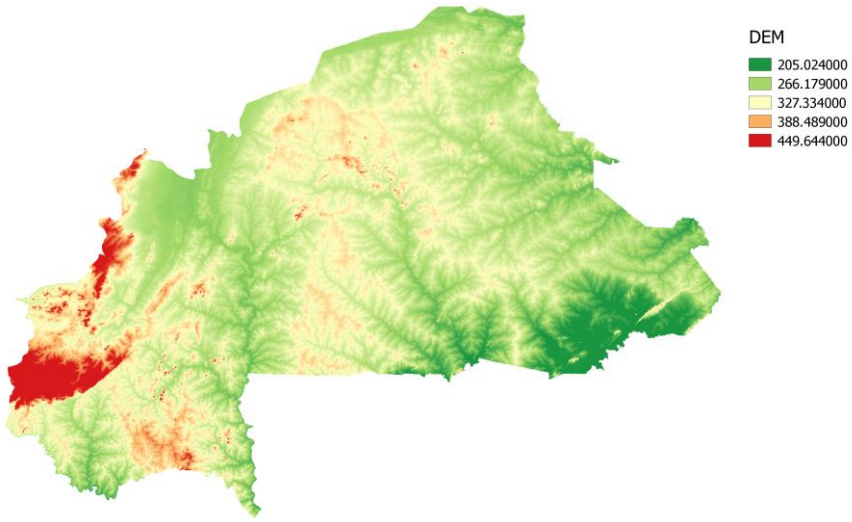


Figure 14 DEM

b) Normalized Difference Vegetation Index (NDVI)

NDVI is a measure of the density of chlorophyll contained in vegetative cover and is defined as (NIR – RED) / (NIR + RED), where NIR is the near-infrared reflectance and RED is the visible-red reflectance. This Index was developed by the U.S Geological Survey (USGS) Earth Resources Observation and Science (EROS), generated from the Moderate Resolution Imaging Spectroradiometer (MODIS). This NDVI raster was calculated from the MODIS L1B Terra surface reflectances, and corrected for molecular scattering, ozone absorption and aerosols using MODIS Science Team algorithms (Swets, 1999). A time series smoothing algorithm developed by Swets et al. (1999) was used to smooth the NDVI composites between the years 2001 to 2010.

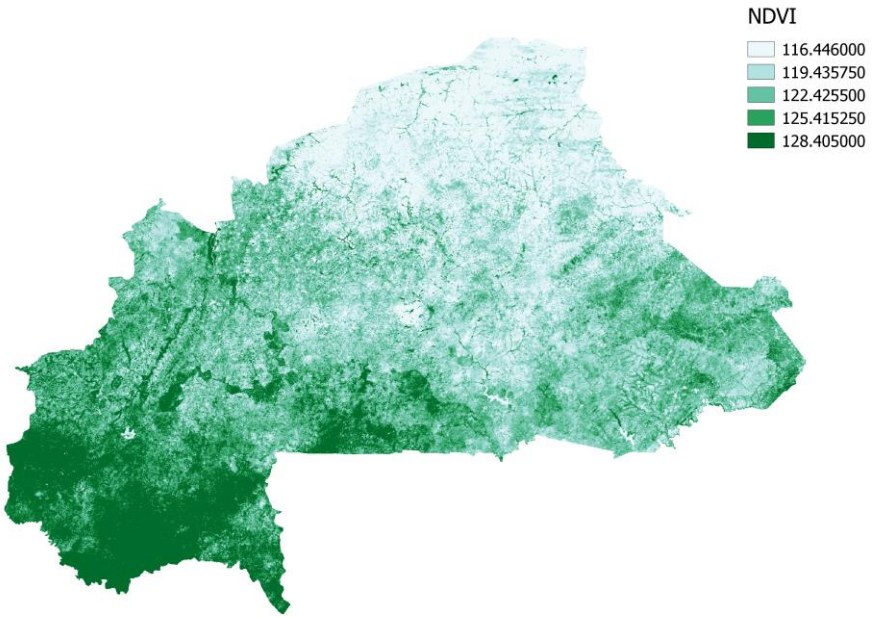


Figure 15 NDVI

c) Rainfall Estimates (RFE)

The estimated precipitation data were obtained from the RFEv2 from NOAA NCEP CPC FEWS Africa TEN-DAY from the Famine Early Warning System (Love, 2002). These data give a 10-day grid of estimated precipitation between the year 1999 and 2014, meaning 522 temporal grids of points. On the IRI Data Library website, it is possible to select a spatial and temporal scale for the data. Only a rectangle surrounding Burkina Faso was downloaded, with the 522 time grids. These data were directly downloaded on Matlab in order to calculate, for each pixel, an annual mean and standard deviation precipitation value. With the Mapping ToolBox from Matlab it was then possible to create a raster from these grids, and to clip them into QGIS (TEAM, 2014). The following figures show the obtained maps:

For the mean values of precipitation:

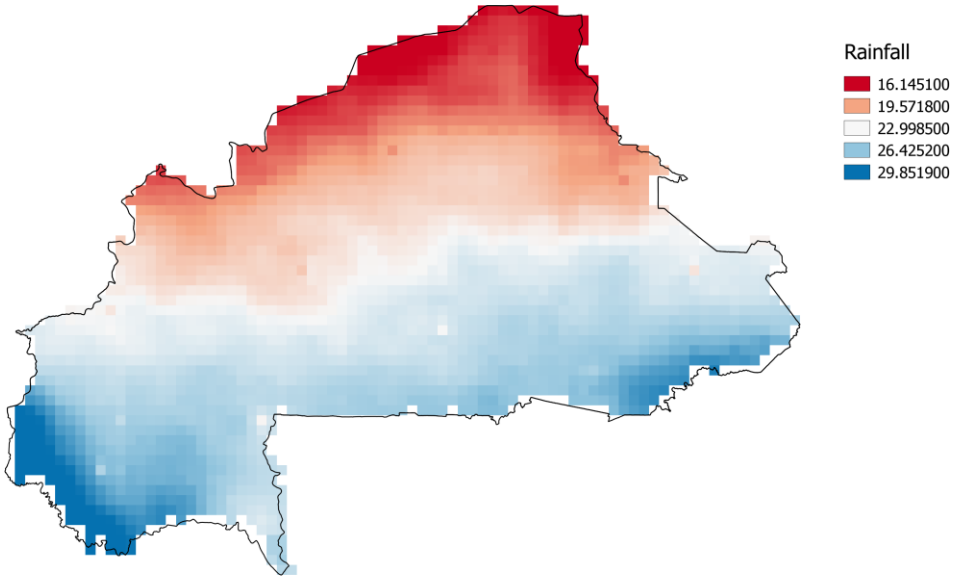


Figure 16 Decadal Precipitation mean

And the standard deviation of precipitation:

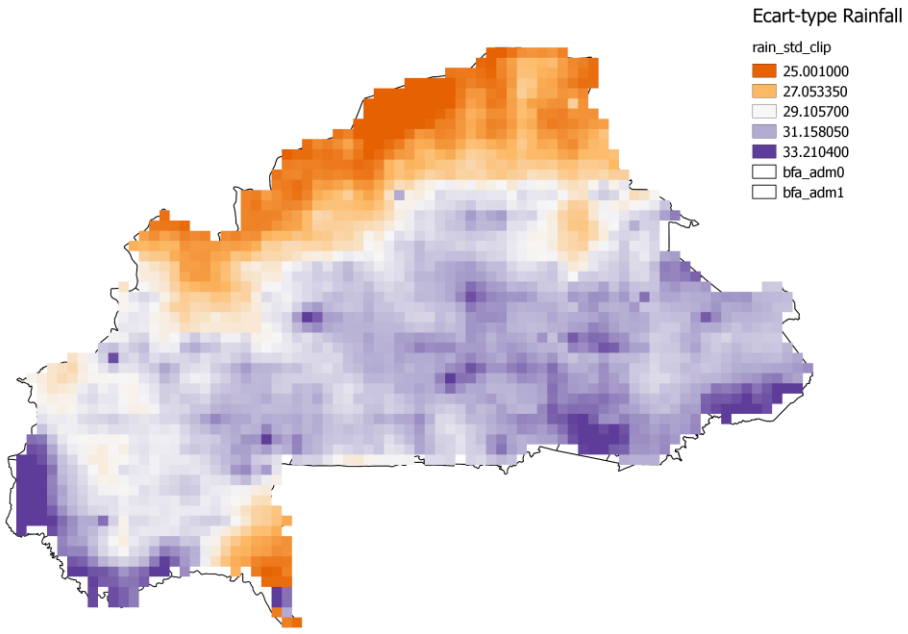


Figure 17 Decadal Precipitation standard deviation

d) LandSurface Temperature (LST)

The temperature informations were taken from the Terra/MODIS V004 & V041 MOD11C2 LST/E Science Data Set. As observed in the figure, the extracted data had missing values in the region of interest (white blanks).

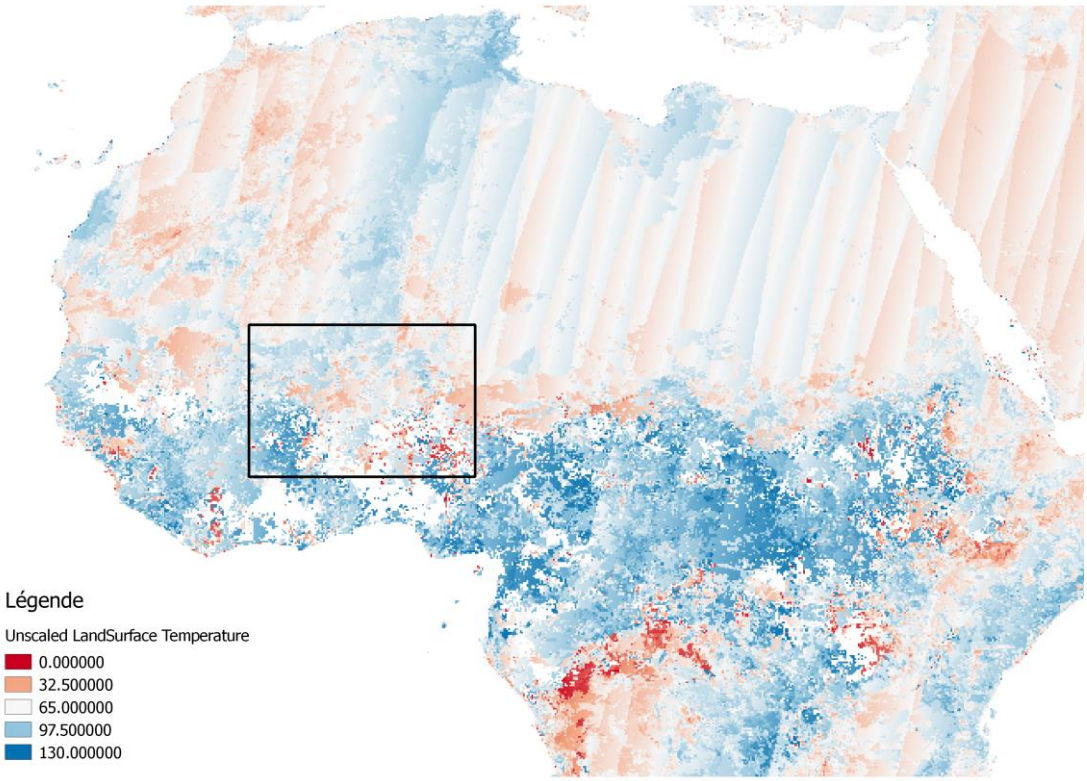


Figure 18 LST over Africa in October 2012

A one year series of monthly temperature was taken between end of 2012 and end of 2013. The interested zone of Burkina Faso (black square in above figure) was clipped from the rest of the world map (see next figure)

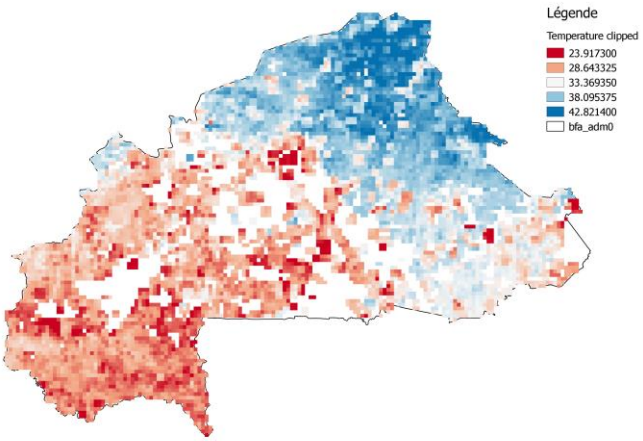


Figure 19 Clipped mean monthlyTemperature raster

Then, in order to fill the missing values, an ordinary kriging has been performed to interpolate the temperature values over the whole country for each month of the year. And finally, a simple mean and variance calculation over the overlapping pixels has been calculated on Matlab to give the next finals rasters. Note that with the ordinary kriging, the pixels were automatically resampled, giving a better resolution to the temperature data.

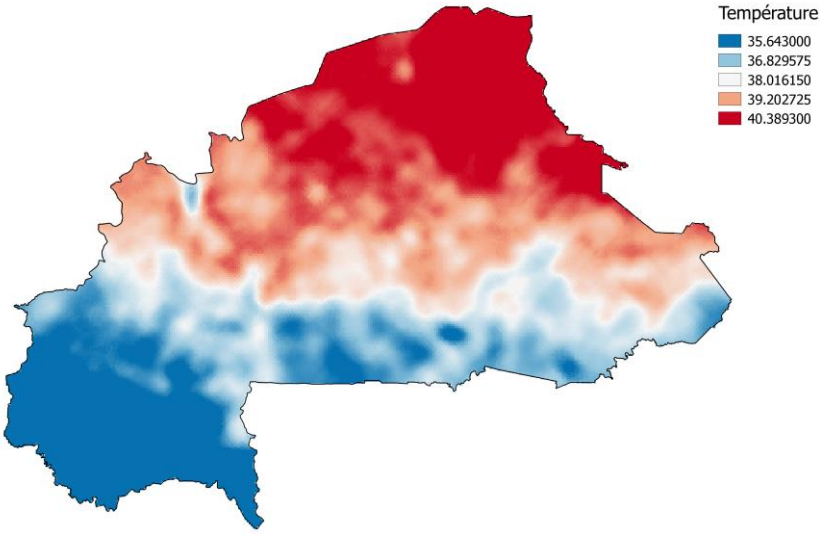


Figure 20 Temperature mean – 2012/2013

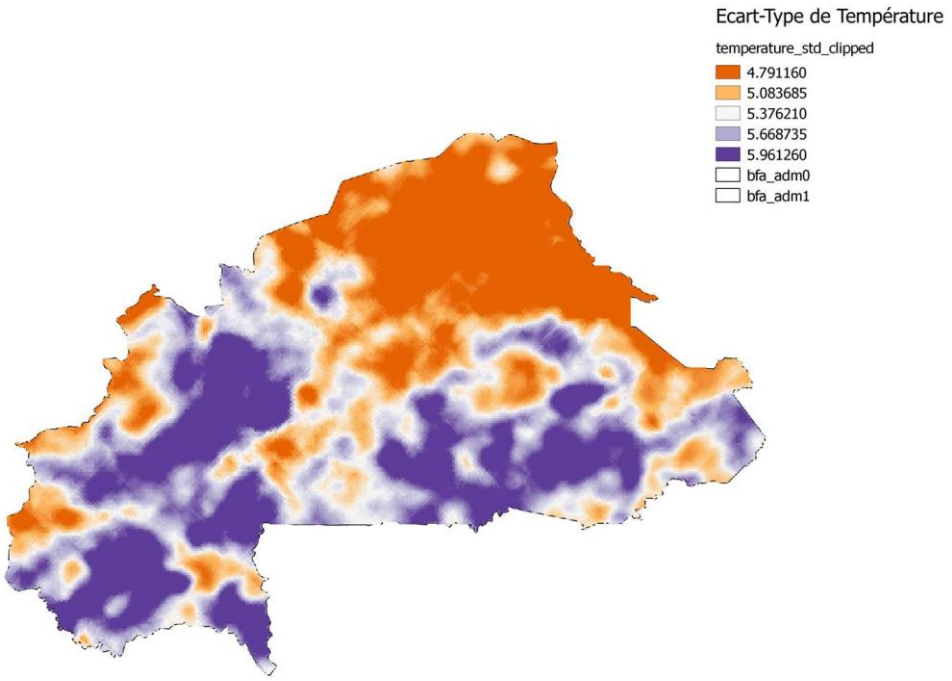


Figure 21 Temperature standard deviation 2012/2013

e) Landcover

This data is taken from the MODIS Land Cover Type products, and contains five classification schemes, describing land cover properties derived from observations spanning a year's input of the satellite acquisitions of Terra- and Aqua-MODIS data (Friedl, 2010). These classification schemes were derived through a supervised decision-tree classification method. For this work, the scheme *Land Cover Type 2* has been selected, also called *University of Maryland (UMD) scheme.* The next figure describes the landcover distribution over Burkina Faso.

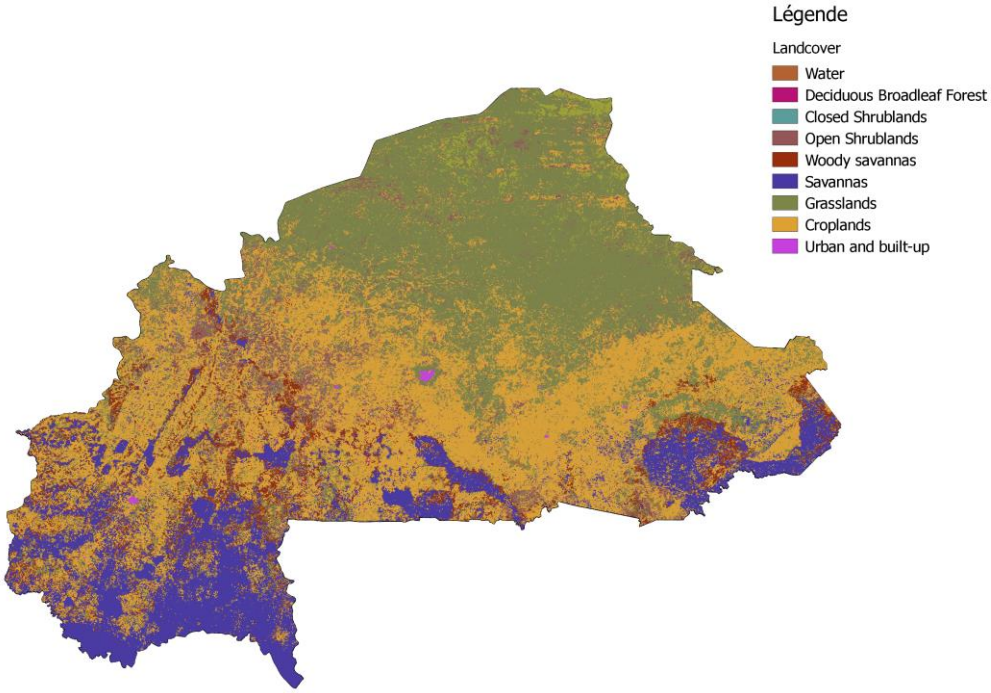


Figure 22 LandCover

f) Human Influence Index (HII)

The HII information is taken from *The Global Influence Index Dataset of the Last of the Wild Project, Version 2, 2005 (LWP-2)*. It's a global dataset of 1-kilometer grid cells, created from nine global data layers covering human population pressure (i.e. population density), human land use and infrastructure (built-up areas, nigttime lights, land use/land cover) and human access (coastlines, roads, railroads, navigable rivers) (WCS, 2005). The next map illustrates the Human Influence Index obtained for Burkina Faso.

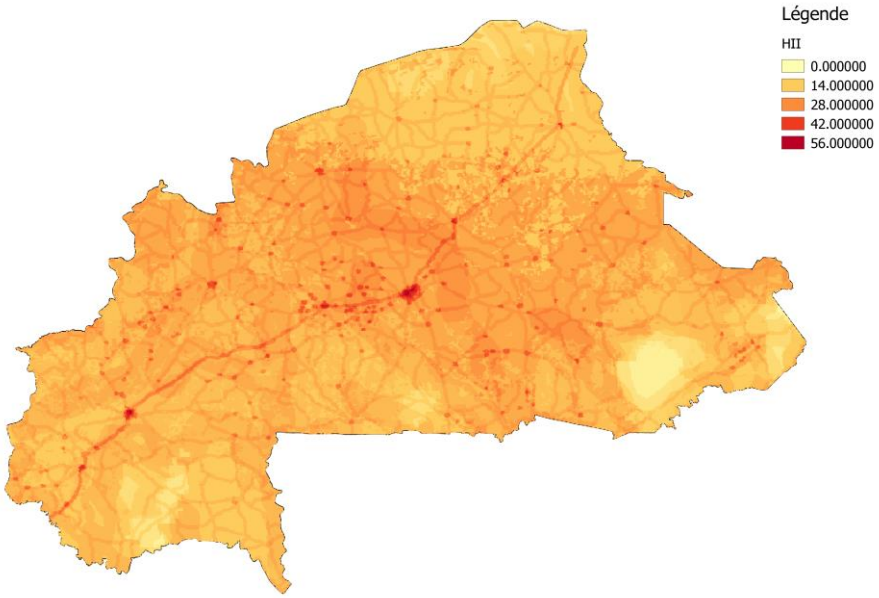


Figure 23 HII

g) Dry season period

In this chapter, a methodology will be proposed for calculating the time of a dry season period for each pixel of a raster, having a decadal precipitation time series for a corresponding pixel. Knowing the "spatial extend" of a rainy season is also important since mean precipitation gives only an information about amount of rain, which is very variable in Burkina Faso, e.g. in the North where precipitation is more intense but during a lower time period.

The inputs of that calculation are:

- A longitude vector X positioning the point in space (108 points)
- A corresponding latitude vector Y (79 points)
- A time series vector T (decadal series during 15 years, i.e. 3 (3 decades in a month)*12 (month)* 15 (years) ~ 540 values)
- A precipitation three dimension matrix *rain* in [mm/10-days] (X.Y.T, 108x79x540)

The aim of this procedure is to obtain a raster layer, having each pixel corresponding to a certain time, in days, of "dry season period". The next paragraphs will describe how it has been estimated.

Considering seasonality of precipitation, each precipitation pixel in matrix *rain* has been reduced in a twelve month time series corresponding to the mean monthly precipitation for the whole 15 years. The resulting matrix has now a dimension of (108x79x12) and has a mean monthly precipitation value.

The next figure is representing a precipitation series for a pixel at location X=50 and Y=50.

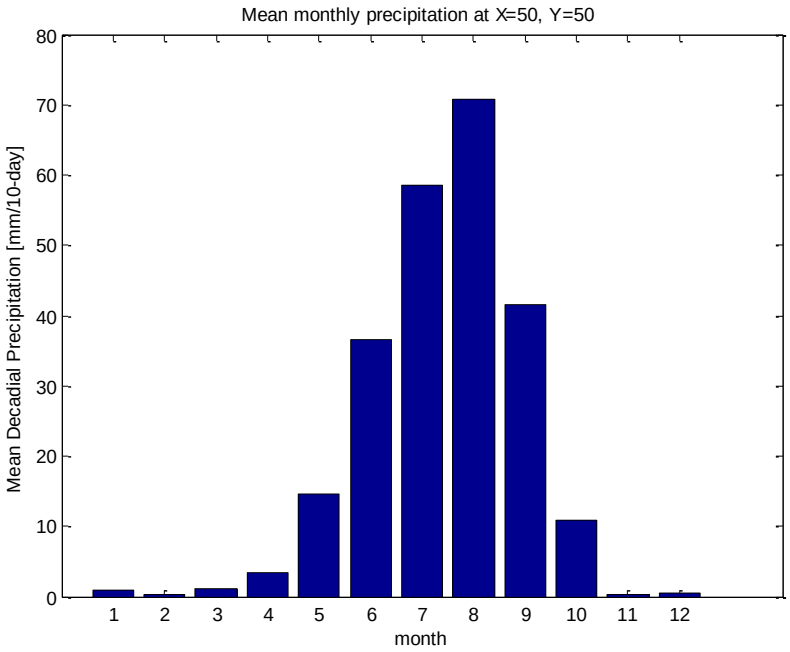


Figure 24 Mean decadial precipitation at X=50,Y=50

The next step was to process a Gaussian regression over each pixel in order to obtain, for each location, a mean μ value and its standard deviation σ of the Gaussian functions. It is important to notice that μ and σ have time dimensions, i.e. months. Intuitionally, we can already link the standard deviation to a certain "wet season" period and the mean to the peak period of the rainy season.

These regression coefficients are represented in the following figures:

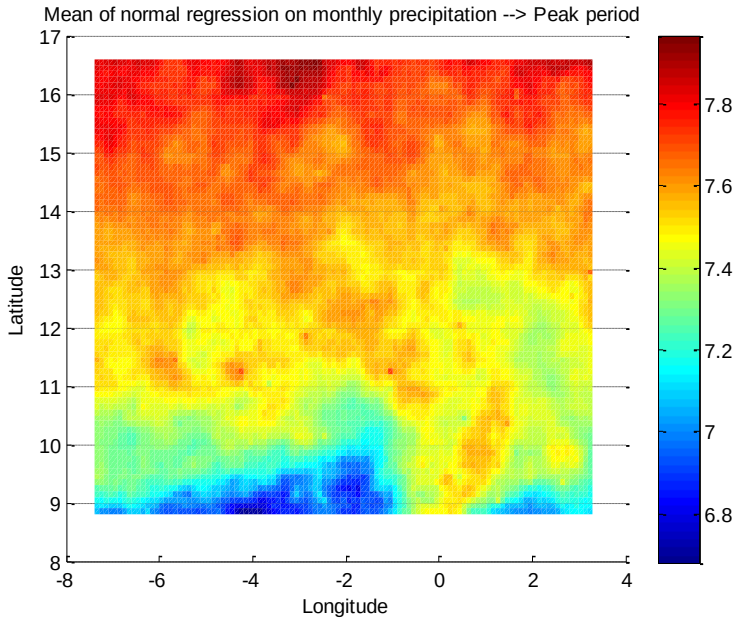


Figure 25 Mean of the normal regression

All the mean values, meaning the mean month of the precipitation time series regression, are situated around month of July, which is the peak of the rainy season in Burkina Faso.

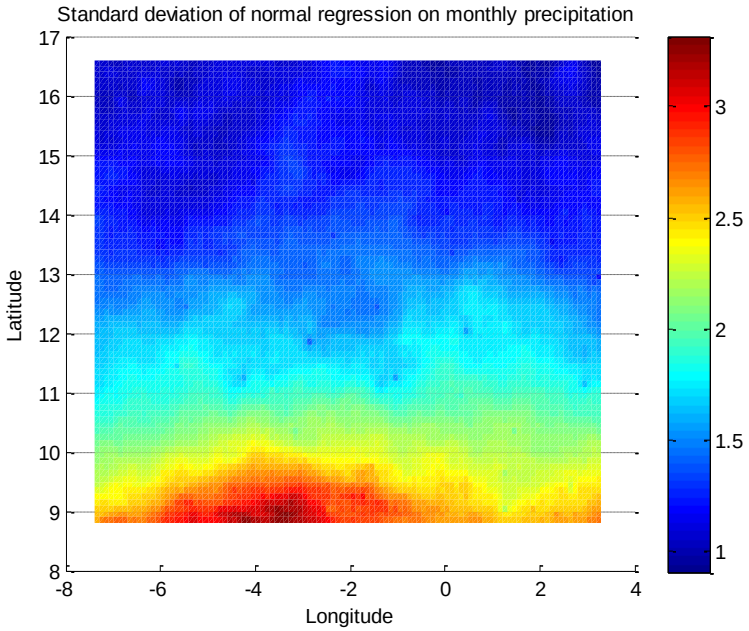


Figure 26 Standard deviation of the normal regression

It is noticeable that the north of Burkina Faso, having a Sahelian (Poda, 1996), very arid climate, has only a precipitation time "extend" of one month, while the south reaches three month, where the climate is Sudano-Guinean (Poda, 1996) which is less dry and milder.

Since the temporal extend of the rain season is the interest for calculating a dry season period, a clustering by K-means has been effectuated on the standard deviation of the Gaussian regressions. It was decided to create six "spatial groups" that have the most similar σ values. The resulting clusters are represented in the next figure.

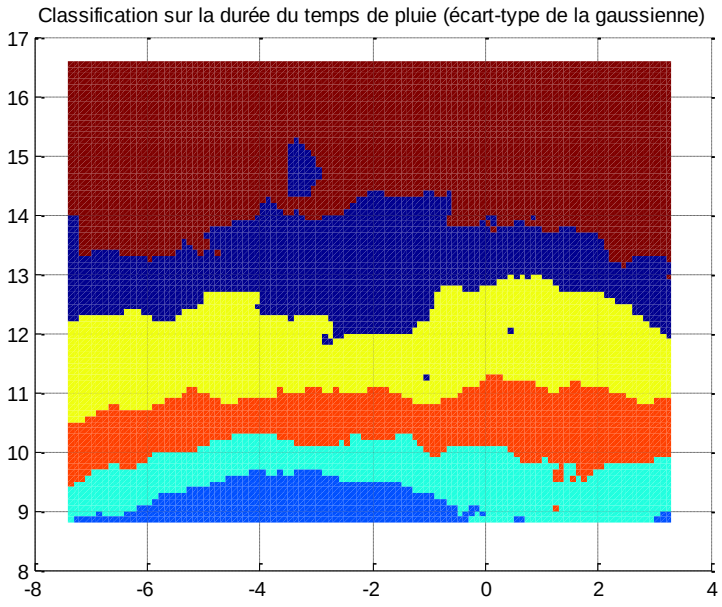


Figure 27 Clusters of the K-means classification on sigma

Next step was to recalculate a mean monthly precipitation over the whole period for the pixel contained in each cluster. A normal regression of these mean monthly precipitation will define a so called “normal rainfall response” for each of the clustered region. These responses are shown in the next figures.

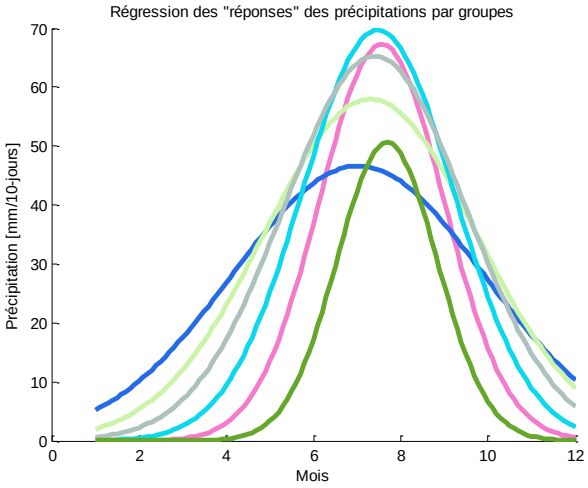
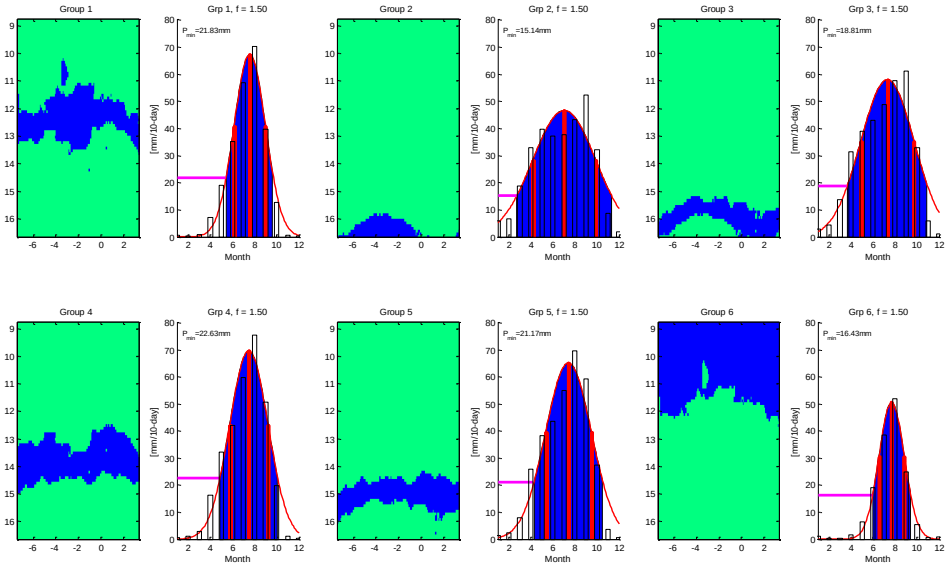


Figure 28 Normal Rainfall Responses for the six clusters

It can be noticed that each cluster, i.e. region, has a different mean, amplitude and standard deviation.

The next step was to define, for each cluster, a “threshold” precipitation value, which will define the minimum precipitation before a period can be considered as the beginning of a rainy season. It was decided to fix this minimum P_{min} value to the precipitation amount corresponding to the standard deviation σ times a factor f = 1.5. This P_{min} value can be observed in the next figures, being obviously different for each of the six clusters.



.

Figure 29 Pmin definition for the six clusters

A zoom of the first group can be visualized in the next figure.

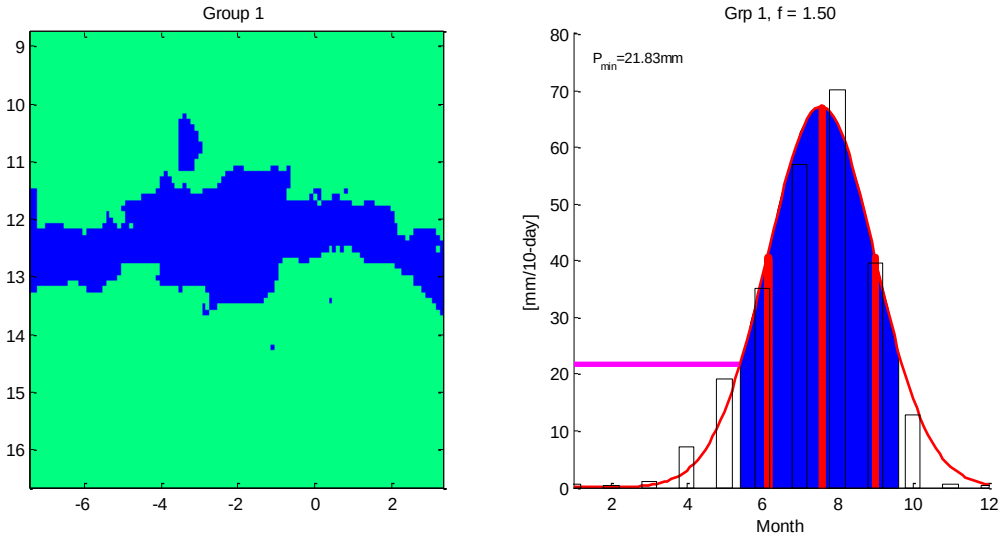


Figure 30 Pmin definition for cluster n°1

In the right figure, red vertical lines represent the mean (middle) and the standard deviation. A rainy season is defined as the blue area, starting at a precipitation equal to $f.\sigma$. The P_{min} value is obtained where the month equals μ- $f.\sigma$. For this cluster, the minimum monthly precipitation defining the start of a rainy season is P_{min} = 21.83 mm/10-day.

The final step was to calculate, for each original pixel, the time of a dry period. The created algorithm considers that if the rainfall sum of 3 consecutive 10-day periods (one month) is lower than three times the P_{min} (which has dimensions of mm/10-day) of its corresponding cluster region, than this month is a dry month (more precisely a 30-day sequence). The final raster is shown in the next figure, showing the estimated dry season time (in days) for each pixel.

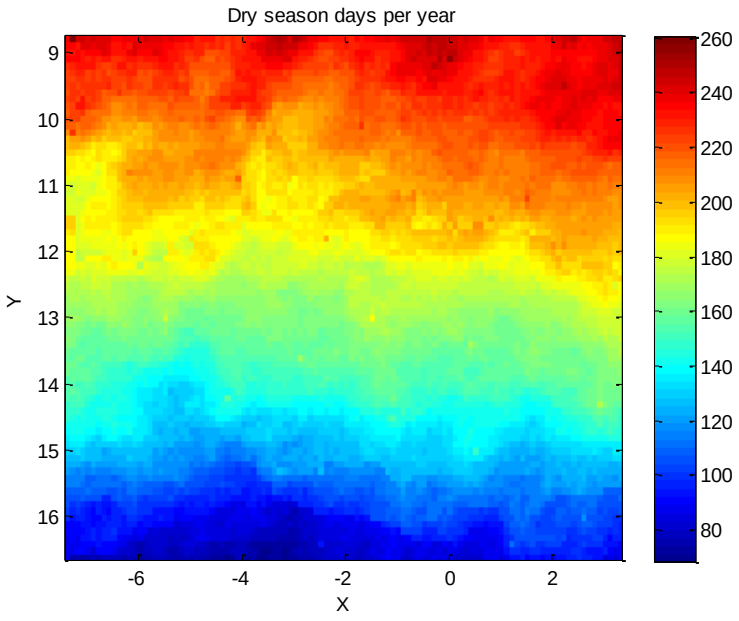


Figure 31 Dry season period [days]

The clipped raster of the dry season time is presented in the next figure:

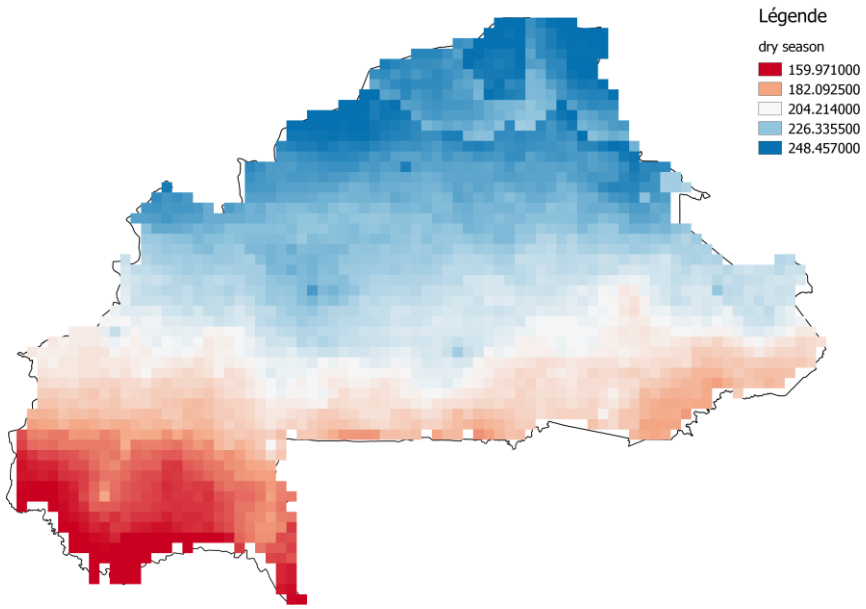


Figure 32 Clipped dry season raster

3.1.3 Disease related data

Now that the environmental informations were presented, this section will describe the obtained and prepared data related to schistosomiasis presence in Burkina Faso. This chapter will consist of describing, on a first hand, the prevalence data obtained from two different sources; and on the other hand information obtained on site in two villages of Burkina Faso during my stay in April 2014.

i) Prevalence data

The data stored in the database containing the information of the schistosomiasis presence and abundance in Burkina Faso was obtained from two different sources. A first source was collected from the "Programme National de lutte Contre la Schistosomiase" (PNCS) conducted by the ministry of Health and a second one from an open-access platform called "Global Neglected Tropical Disease" (GNTD). These two databases will be presented in the next chapters.

a) PNCS

These disease informations were obtained from a national action program against schistosomiasis that has been undertaken since 2004 by the Ministry of Health of Burkina Faso. This database contains a list of 22 cities, where prevalence analyses[9] on the village's population, who was hired from the Ministry's PNCS department to lead his researches. For each village, prevalence data are available between 2008 and 2013. The next figure shows these prevalence data over the country, separated by years of examination.

[9] *M.Bagaya & co.*

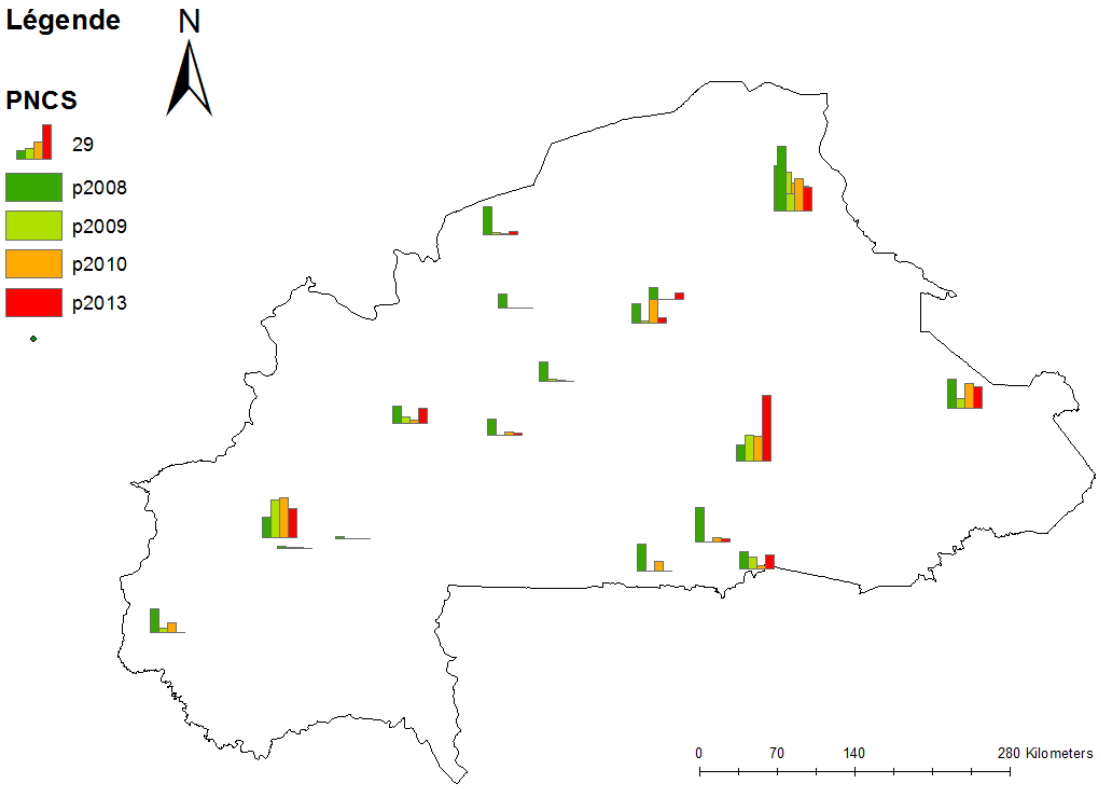


Figure 33 PNCS prevalences

The years 2011 and 2012 do not appear since most of the data during these years are absent and cannot be shown as bars in ARCGIS (ESRI, 2011). These data have several attributes that are listed in the below table:

Attributes
pointname
region
lat
long
parasitename
year
prevalence
num_examined
num_legere
num_moder
num_forte
num_positive

Table 2 Attributes of PNCS database

The attributes *num_positive* and *num_positive* are very important since they will be used in the model building, which are representing the number of examined and positive peoples.

b) GNTD

The second database, named "Global Neglected Tropical Disease" (GNTD) was taken from a web platform[10] which provides compiled historical and contemporary survey data about neglected tropical diseases (NTDs). It consist of an open-access platform in which data can be directly downloaded after possible filtering, for example choosing a specific disease. At moment, around 12'000 unique survey location, mainly *S.mansoni & S.haematobium* in Africa are included in the database. These data were stored by (Vounatsou, 2011) and collaborators and obtained from peer-reviewed publications and 'grey literature'. The next figure shows the data for both parasite species in Burkina Faso. These data contain prevalence information in a time period between 1948 and 2004 and contains about 878 point information.

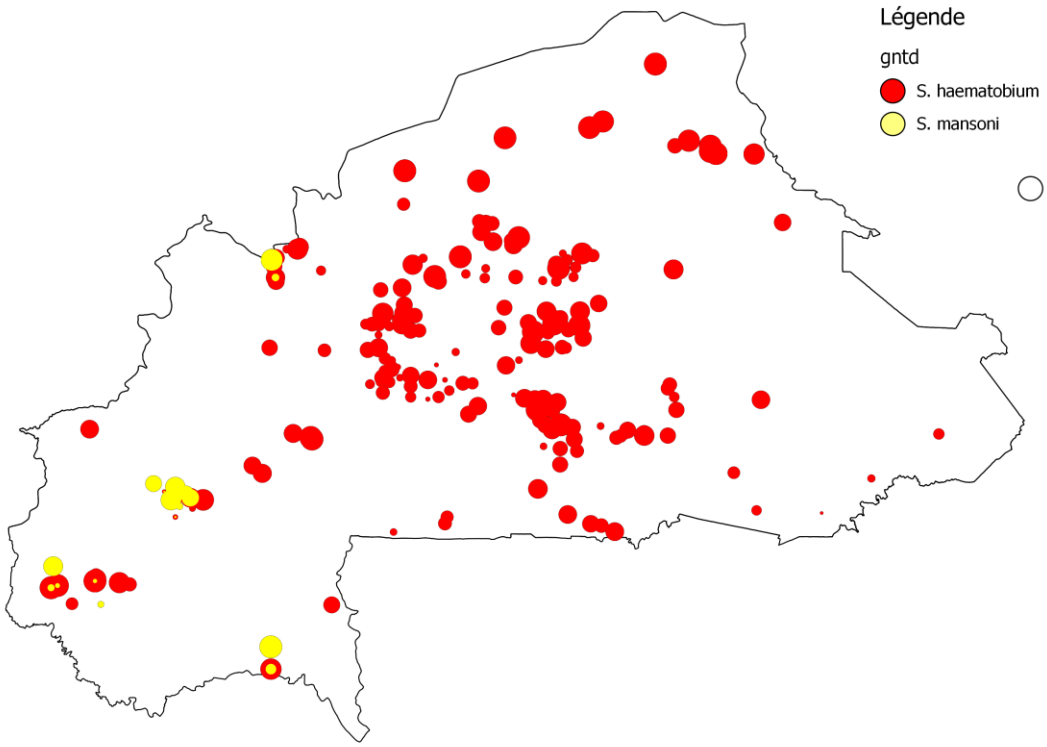


Figure 34 GNTD Database

These data are generally more detailed than the information in the PNCS database. A lot of data include, for example, the sex, age, and method employed during examination. But, since these data are obtained from different sources, not all the information appear in all surveys; a necessary filtering has been necessary for the model construction.

[10] *http://www.gntd.org/login.html*

3.2 Model construction

This part will consist of constructing a generalized linear geostatistical model to the prevalence schistosomiasis data. Recalling the generalized linear geostastistical model defined as the following equations:

$$Y_i|U(s_i) \sim f(\lambda(s_i), \nu)$$
$$g[\lambda(s_i)] = \mu + \beta X(s_i) + U(s_i)$$
$$cov[U(s_i), U(s_j)] = \sigma^2 \rho(h/\Phi; \theta)$$

This model is suitable for these data with *f* being a binomial distribution and *g* the logit link function. The *X(s)* surface is multivariate and will have values of the environmental raster data as presented is the previous chapter. The *X(s)* values will consist of the *rain, temperature, elevation, dry season time* and *NDVI.* The *Landcover* information matrix could not be selected due to software issues with non-numeric data.

The next figure is giving a general overview of the steps realized on the necessary information in order to run the model:

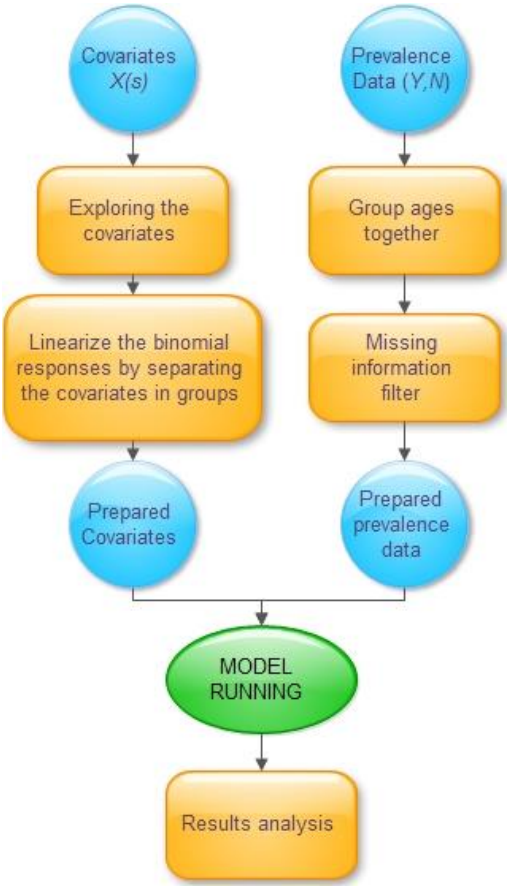


Figure 35 Model building flowchart

Like observed in the above figure, the covariates *X(s)* will be first prepared in order to have a best response of the model. The process is to fit a generalized additive model to see if the covariates are linear. This will be done by creating a binomial-based response variable of the covariates and see where they must be separated into groups in order to have a better linear regression of the model.

3.2.1 Prevalence data preparation

For the prevalence data, it was necessary to regroup each same location data into a same variable. Since most of the data were separated by age classes, the data was regrouped in order to have all the examined and positive cases in one point information. After that, all data having missing necessary information have been removed. These informations were the year of the examination, the number of examined cases, the number of positive cases and the geographical coordinates.

After filtering the *GNTD* database, the data number dropped from 280 to 236, meaning removing about 16% of the data, for the *S.haematobium* species. For the *S.mansoni* species, the number of data passed from 82 to 44 objects, meaning removing about 46% of the data. For the *PNCS* database, only the year of 2013 had information on the number of examined/positive samples and present coordinates. The number of data dropped from 111 to 11 data, meaning removing about 90% of the information; all the information being for the *S.haematobium* species. The problem was that the other year (2008-2013) from this database had only prevalence information in percent, which is not compatible with the binomial assumption of the model responses.

Furthermore, since the prevalence data obtained from the two databases contains information in a large time scale (1948 to 2013), it was interesting to also run the model for different periods, in order to observe a possible evolution of the disease in time. A necessary analysis of the repartition of the data over the years was necessary in order to keep a reasonable amount of data in the separated time periods. The next figure shows the repartition of the data per year, and also the cumulative occurrences of the data for the two different species.

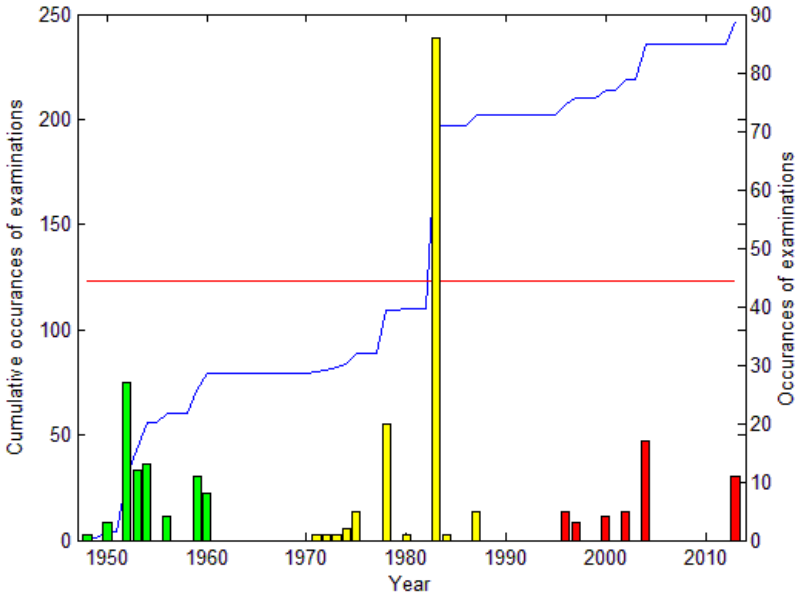


Figure 36 Time repartition of the prevalence data for *S.haematobium*

The three colours define the chosen repartition in three different groups. This clustering is maximizing the time of lacking data between each group and keeping an amount of enough data in order to run the model. The next table is showing the number of data in each of the groups

	GROUP 1	GROUP 2	GROUP 3
Year	*<= 1970*	*1970 - 1990*	*>= 1990*
Number of data	83	135	49

Table 3 Number of data in each group for *S.haematobium*

The next figure shows the spatial repartition of the three different groups, keeping the colour rules for a better graphical understanding. It can be observed that the repartition is relatively homogeneous for the three groups. Nevertheless, group 2 in yellow has a clustered distribution of examinations in the Centre compared to the others and lacks of data in the South East of the country.

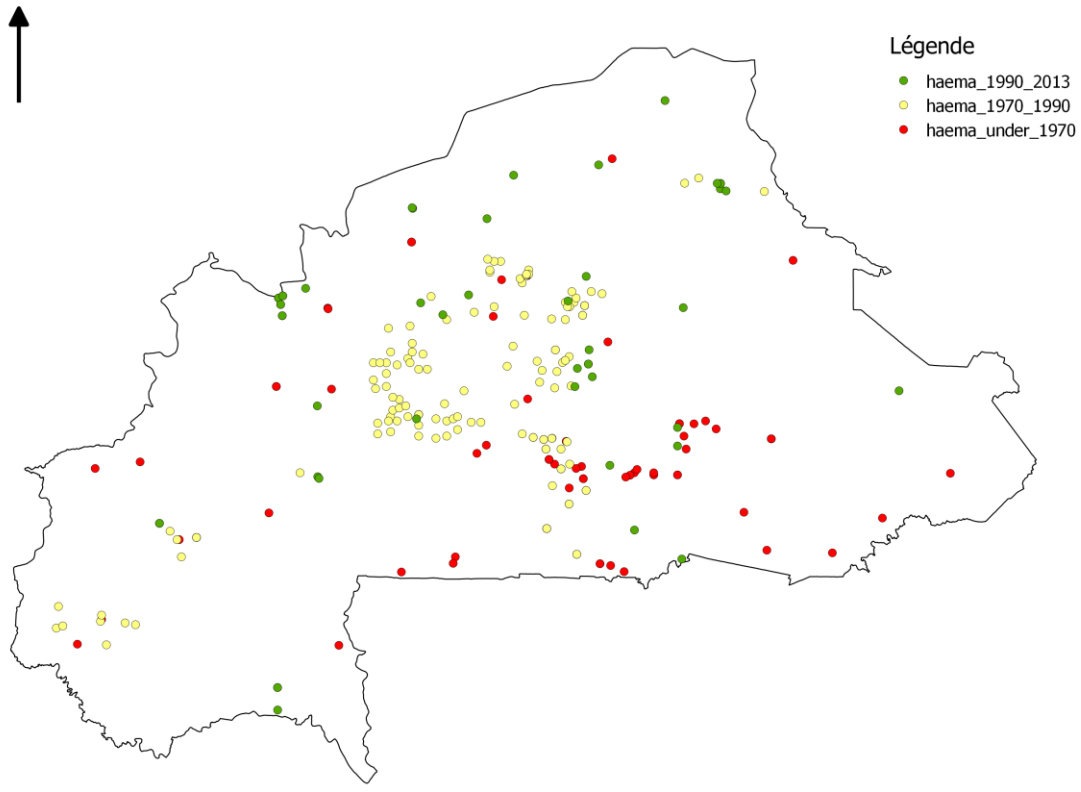


Figure 37 Spatial repartition of the groups for S.haematobium

In order to have a global view on the model results and the efficiency of such time-based clustering, four models will be run for *S.haematobium.* Three will consider time clustering, and another model will be run on the whole dataset, meaning repeating the flowchart in Figure 35 Model building flowchart times.

For the *S.mansoni* species, it was decided not to separate the data in different time groups. Explanation is given below. The next figure shows, as previously, the repartition of the data and the possible interesting clustering that could have been done.

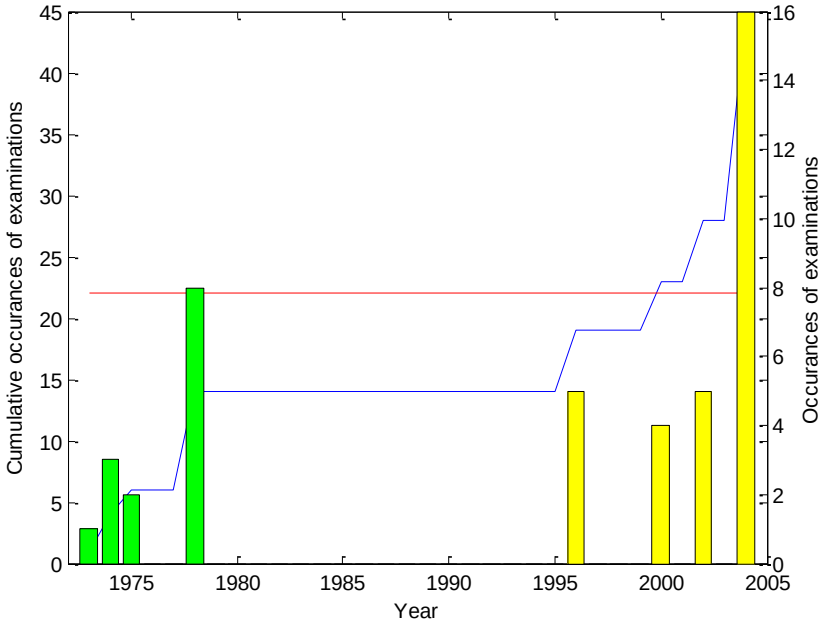


Figure 38 Possible time repartition of the prevalence data for S.mansoni

The gap between the years 1978 and 1996 could have represented a separation for two possible groups (colours). The problem with such a separation is, firstly that the first group would only contain 14 values while the second would have around 35 values, which is an approximated one third two third repartition. Secondly, on the next figure showing the spatial repartition of the two groups, it can be seen that they are covering two different locations of the country, only south for Group 1 and more North for group 2.

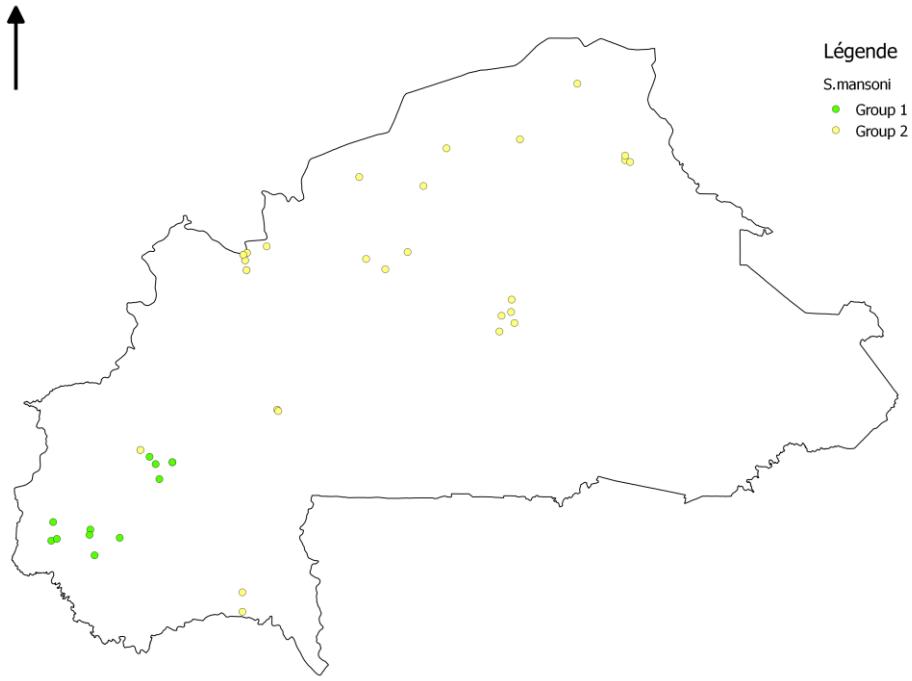


Figure 39 Spatial repartition of the possible groups for *S.mansoni*

For the *S.mansoni* parasite, it was therefore more interesting to effectuate no evolutionary analysis of the prevalence in time. Another argument for this is the relatively small amount of data for this type of schistosomiasis.

3.2.2 Covariates preparation

Now that the prevalence data have been filtered, grouped and prepared, it was also necessary to explore the covariates in order to prepare them for the model running. Beginning from this part of the work, the analyses will be done in the statistical software R. Concretely, the process that will be applied for the preparation of the covariates consists of fitting a generalized additive model to see if the covariates are linear, by creating a response variable for Binomials. Generalized additive models (GAM) are generalized linear model where some chosen predictors x_i, in opposition of being multiplied by a coefficient, are set non-linear by applying them a so-called "smooth" function *f()*. The general formula of a GAM can be expressed as follow:

$$g(\mu) = X\beta + f_1(x_1) + f_2(x_2) + ... + f_n(x_n)$$

where *g* is the "link" function like in a GLM and μ the expected value of the binomial (in our case) response Y of the model. The term *Xβ* represents a matrix multiplication of the intercept and some chosen covariates that are supposed to be linear. The x_j are the other predictor variables (covariates) and the f_j being the so-called 'smooth functions' (Trevor Hastie, 1986). This enables to analyse each covariate's smooth function $y=f(x_i)$, which are generally a sum of partitioned cubic functions. If that function is linear, it could be replaced by a coefficient β (not the same as in the above formula) like in GLM. If not, the covariates can be cut at some value to create two data sets, which show a linear response to the model fitting.

An example of covariates splitting applied in this work is shown in the following paragraphs. First of all, a GAM is being fit to the data, including the N number of examined people, y the number of positive people, and the covariates used in the final model. The code in R can be written as following, using the *mgcv* and *gam* packages:

burkGam = mgcv::gam(response ~ s(rain) + s(temp) + s(dem) + s(dryseas) + s(hii) + s(ndvi) , data=burkprev, family = "binomial")

The variable *response* is a matrix containing the prevalence information N and y. The covariates are the corresponding environmental information located at the examination point. The function *s()* applied to the covariates are the so called smooth functions that are going to be analysed for linearity.

By plotting these smooth functions, we can observe the linearity of the covariate's responses to the generalized model. For temperature in example, for all the years together, the resulting function gives the following figure:

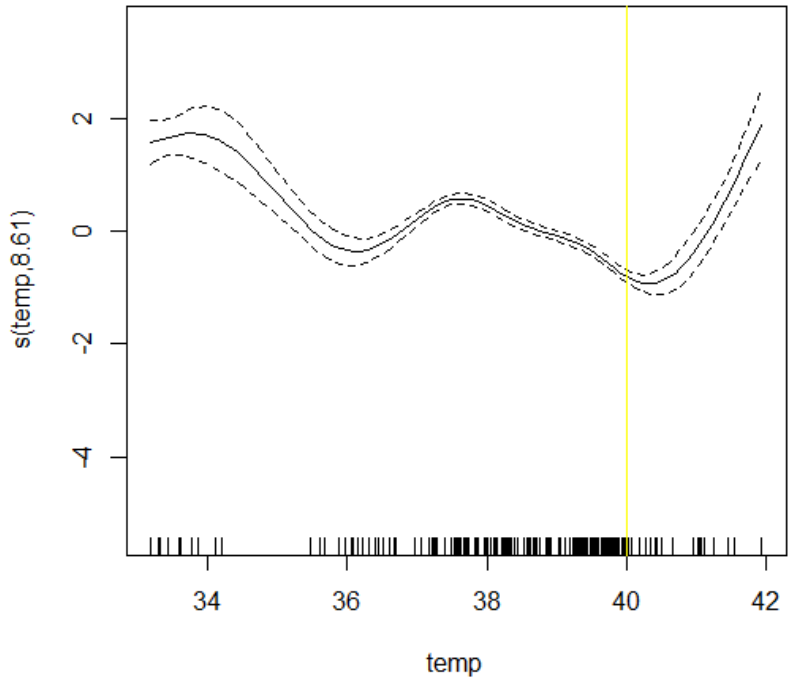


Figure 40 Smooth function of the GAM for temperature variable

The temperature covariate could be done linear by including a changing point at around 40 °C. The covariate has therefore been split between values under and above 40 degrees

3.2.3 Model running

The model is a two-phase process. The first step is the model fitting of the parameters to the observed data, while the second runs the fitted model, i.e. parameters, to the spatial predictors. The R package *geostatsinla* has an integrated function, *glgm(),* that enables, in the same time, to fit and to run the model. This function is performing Bayesian inference for generalized geostatistical models with INLA. For the random spatial effect, the Markov field approximation is used. More generally, this approximation states that each random variable depends on other random variable only through its neighbours (Boykov, 1998).

The *glgm()* function is very easy to use, since the parameters to enter are :

- A object of class *SpatialPointsDataFram* containing the data (Y,N,position)
- A list of the covariates, being of class raster (package *raster*)
- The definition of priors for the coefficients, range, standard deviation or intercept
- The "family" of the model distribution: can be Poisson, or Binomial in this case
- The number of x-directional cells for the final estimated raster

The R-code for the model running is presented:

```
# MODELLING

# Creates a list of the covariates

covList = list(rainmean=a, templow=cLow, temphigh=cHigh,
        demlow=eLow, demhigh=eHigh, dryseas =f ,
        ndvilow=hLow, ndvihigh=hHigh)

# Package geostatsinla : RUNS THE GENERALIZED LINEAR MODEL WITH BINOMIAL DISTRIBUTION

burFit = glgm (burkprev,
      formula = y ~ rainmean+ templow + temphigh + demlow + demhigh + dryseas
      + ndvilow + ndvihigh,
      family = "binomial",Ntrials=burkprev$N, cells = 20, covariates = covList,
      shape = 1 , priorCI = list(sd = c(0.2, 4), range = c(0,500)) )
```

It can be seen in the in the covariates list *covList* contains raster information that has been split for a better linearization, e.g *templow & temphigh*.

It has been chosen that prior list *priorCI* will contain priors for the standard deviation and for the spatial range parameters only, which is enough for the model to run. These priors have been scaled after a first run of the model in order to approach the most probable value.

The output values contain three components:

- A so-called *inla* object containing the results of the called function *inla* which performs a Bayesian analysis of structured additive models using Integrated Nested Laplace approximations

- A *raster* object containing all the layers produced by the model: posterior means, quantiles, standard deviations for the random effects *U(s)* and the predicted values of the link scale *g[λ(s)].* This RasterStack also contains the posterior means of *λ(s),* being the values of inversed logit function, determining the probability of infection.

- A *parameters* object containing a list of the prior and posterior spatial parameters of the model, and a summary of all parameters for each of the covariates.

4 Final results

This section aims to provide and comment the results of the geostatistical model, i.e., the generation of an infection risk map of schistosomiasis over the country of Burkina Faso and the related parameters.

Like precised previously, the model has been run four times, one with all the available data, and three times in order to study a possible evolution of schistosomiasis in time.

Recalling the geostatistical model as following,

$$Y_i|U(s_i) \sim f(\lambda(s_i), \nu) \quad (1)$$

$$g[\lambda(s_i)] = \mu + \beta X(s_i) + U(s_i) \quad (2)$$

$$cov[U(s_i), U(s_j)] = \sigma^2 \rho(h/\Phi; \theta) \quad (3)$$

Having *f* a binomial function, the model will firstly fit the parameters β, ϕ and σ, respectively being the covariate's regression coefficients, the spatial *range,* and the spatial *standard deviation*. It has to be precised that the covariance function *ρ* of the random Gaussian field values *U(s)*, is set as a Matérn covariance function, where the shape coefficient θ is set to 1. It is also necessary to recall that Bayesian inference needs a *prior* definition of the *range* and *standard deviation.* These *priors* will converge to the definition of an optimal mean value of the coefficients through the INLA algorithm.

i) *S.haematobium* modelling – No time division, whole dataset

For the whole dataset of the parasite *S.haematobium* parasite prevalence, the model is giving the infection probability map as presented in Figure 41. The values represent the expected *λ(s)* values in the equation *(2)* of the model, meaning that these results are taken as the inverse of the logit function *g()* fitted values. In other words, as being the probability infection rates, these values represent prevalence rates.

A first clear observation is the North-South prevalence gradient which reflects the climatic gradient of rainfall and temperature. These two covariates, or one of them, are expected to have significant regression coefficients. The regression coefficients are presented in Table 4, and it can be observed that, effectively, the covariate *temphigh*, being the split upper value of the temperature raster, has a high positive value of 0.826, explaining the high prevalence in the North of country. Another interesting parameter is the mean rainfall, showing a negative relatively high coefficient value of -0.156. This can be correlated to the temperature, since decreasing temperature is linked with increasing rainfall.

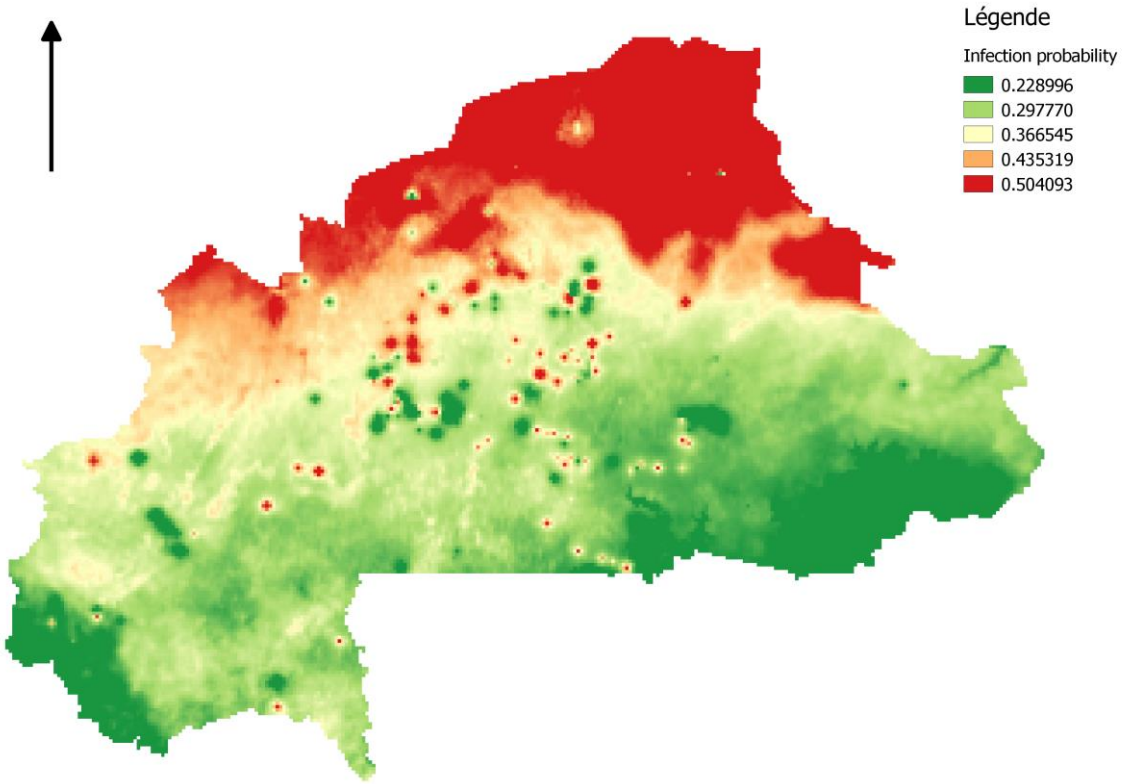


Figure 41 Probability of infection *S.haematobium*, whole dataset

	mean	sd	0.025quant	0.975quant
(Intercept) μ	0.753	6.999	-13.237	14.294
rainmean	-0.156	0.140	-0.431	0.120
templow	-0.064	0.139	-0.345	0.203
temphigh	**0.826**	0.483	-0.145	1.756
demlow	0.029	0.044	-0.059	0.116
demhigh	0.001	0.004	-0.007	0.008
dryseas	0.009	0.025	-0.040	0.060
ndvilow	0.050	0.138	-0.223	0.322
ndvihigh	0.034	0.021	-0.008	0.076
Range ϕ	**0.073**	0.019	0.041	0.114
Sd σ	1.339		1.177	1.490

Table 4 Regression coefficients μ, β, σ and ϕ

An interesting observation is the presence of localized spots of high prevalence in regions of low prevalence, and the inverse case. This must be explained by the spatial noise induced by the Gaussian random field *U(s)*. Effectively, these *U(s)* values, co-varying accordingly to the Matérn function and both the *range* and *standard deviation* parameters, are influenced by the neighbouring observed values.

The next figure is a zoomed representation of Figure 41 in a region where these located spots are present, also showing the observed prevalence rates used in the model fitting.

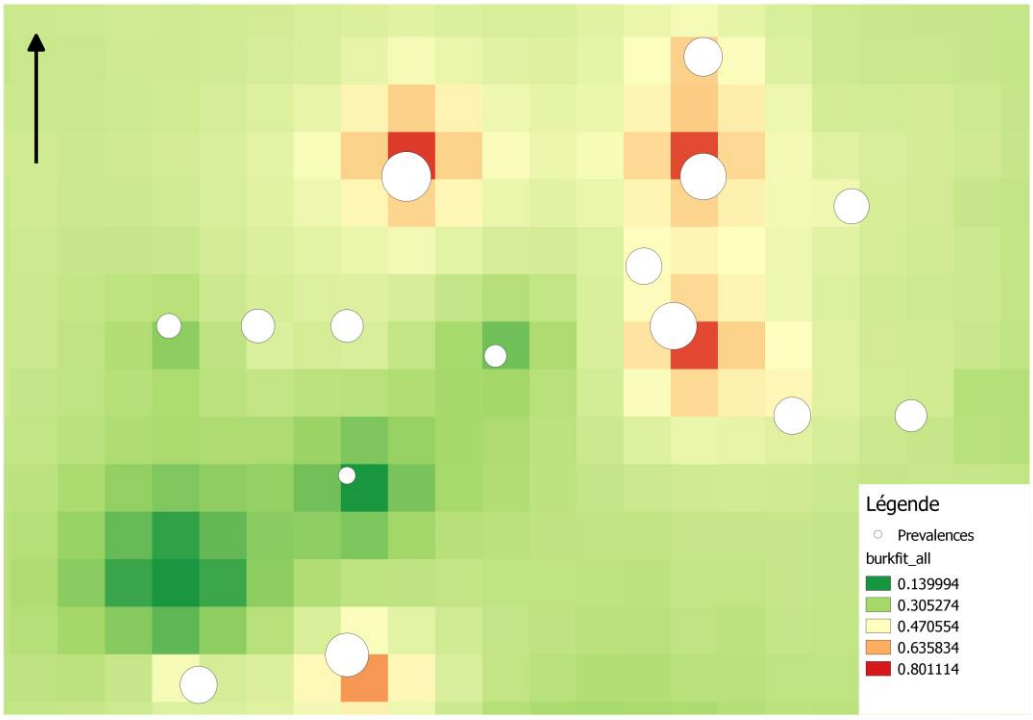


Figure 42 Zoomed prevalence fitted map of *S.haematobium*

It can be observed how high prevalence values, represented as bigger dots on the above Figure 42, have a local effect on the neighbouring pixels. These high modelled prevalence rates are decreasing rapidly, since the GLGM imposes in this region low prevalence values due to the regression coefficients β multiplied by the covariates. This decrease is induced by the covariance function of the Gaussian field, and the "speed" of this decrease is induced by the *range* parameter ϕ. The distance "scale" of decrease is, in this model, the parameter ϕ multiplied by the length of the pixel, which is approximately 3 km. As observed in Table 4, the *range* parameter equals 0.073, which gives an influence distance range of approximately 3000 [m] * 0.073 ~ 220 meters. This explains the rapid decrease of these high points influence in the more global opposite field of prevalence. The model is forced to give a small influence to the spatial "noise", due to the very local variations of prevalence observed in the dataset; and of course the global climatic and environmental covariates are unable to explain such local variation, that can also be temporally varying.

Recalling that the *range* and *standard deviation* parameters have been approximated through Bayesian inference, these *prior* and *posterior* values can be observed in the next graphs. Through the INLA approximation process, it is observed that these parameters converged to an optimal value and variance, collected in Table 4.

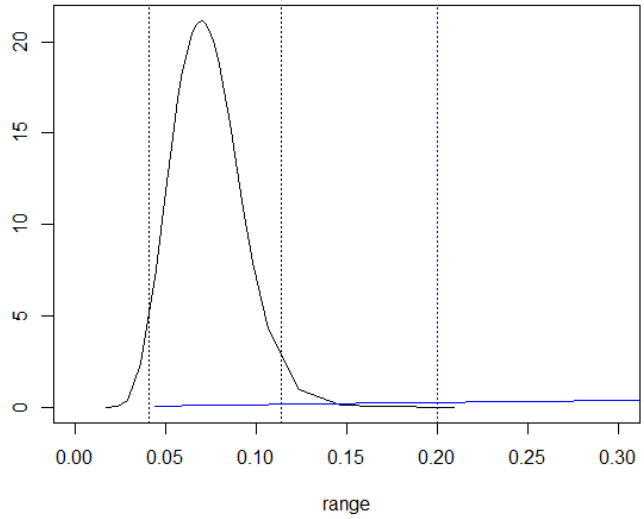


Figure 43 *Range* ϕ prior (blue) and posterior (black) values

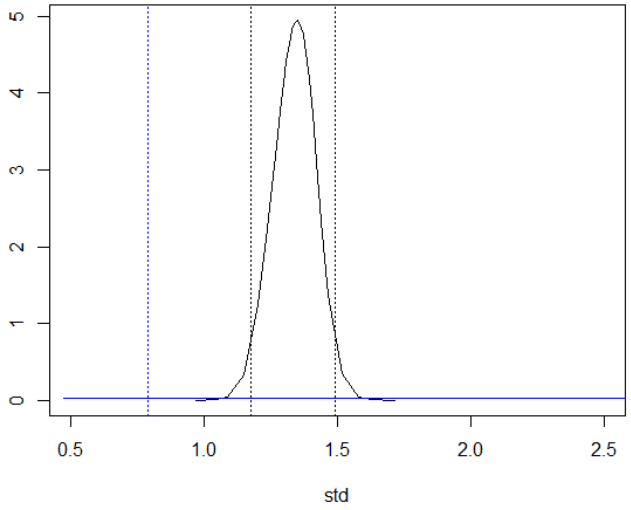


Figure 44 *Standard deviation* σ prior (blue) and posterior (black) values

Another interesting result provided by the *geostatsinla package*'s function *glgm()* is the raster result of the posterior means of the Gaussian random space, i.e. *E[U(s)|Y]*. This map, represented in Figure 45, gives an idea of the spatial noise intensity induced by the fitted Gaussian field. It can be observed that this influence is very localized, like explained above, and that these local variations in prevalence intensity forced the range parameter to be so small, giving few influence at a larger scale.

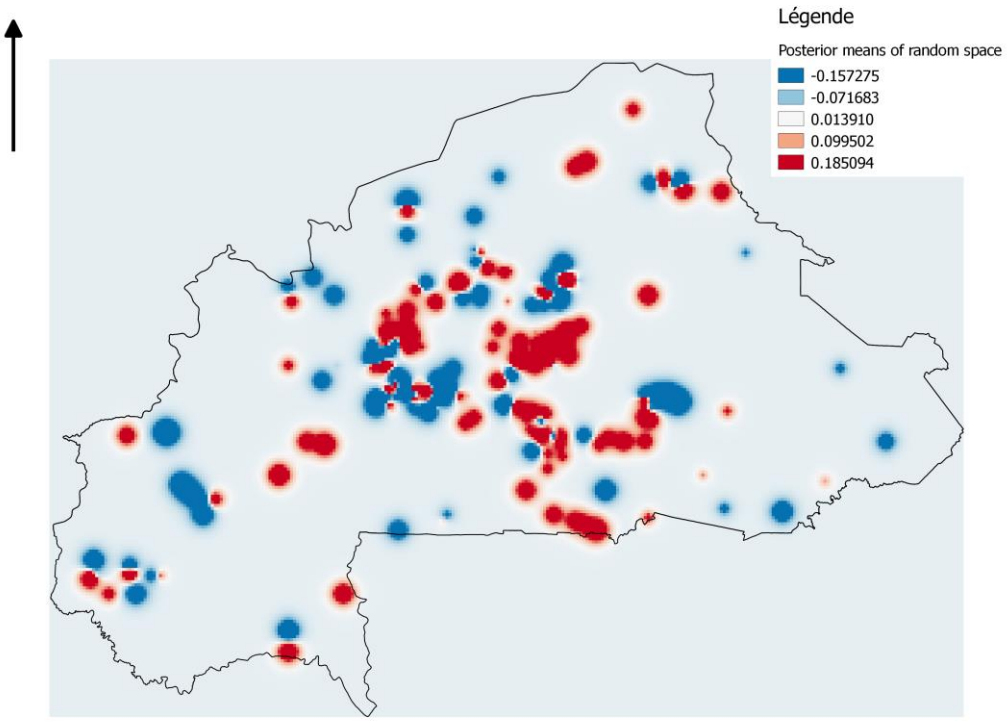


Figure 45 Posterior means of the random space E[U(s)|Y]

ii) *S.haematobium* modelling – 1948-1970

In this section, the same process will be applied for the first group of the previously temporal division of the data, meaning fitting the geostatistical model to a temporal scale covering the years between 1948 and 1960. The next figure shows the modelled infection probability:

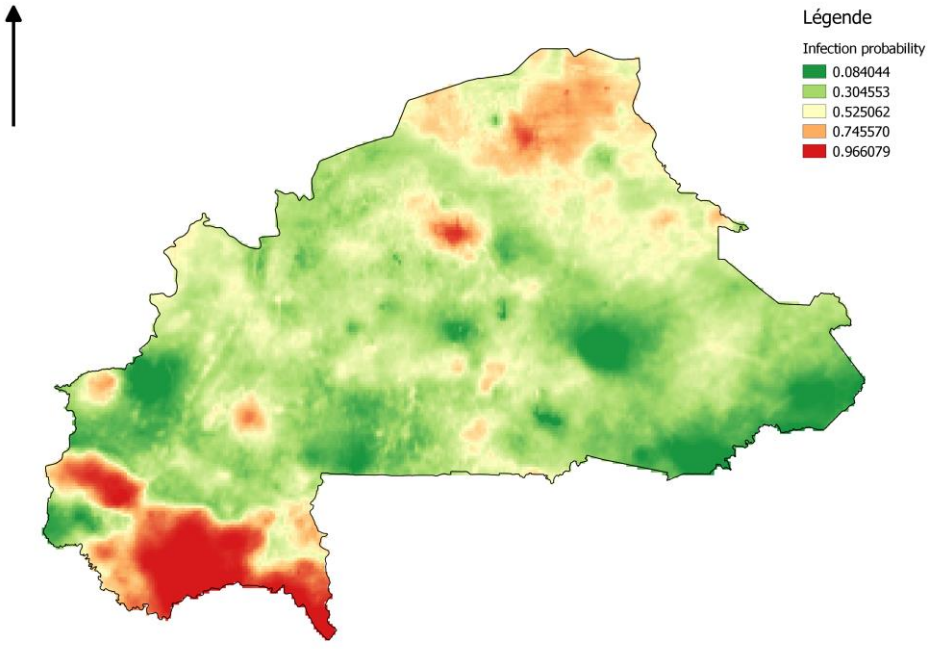


Figure 46 Infection probabilities - *S.haematobium* - 1948-1960

A first interesting information is the presence of a very high infection probability in the south of the country, information that is differing strongly with the whole dataset map previously presented in i). As it can be seen in Table 5, the upper split temperature coefficient *templow* has a very high negative value of -3.42, which is up to more than ten times than other coefficients. The difference with the whole dataset model is that the temperature is negative, showing more prevalence for higher values of temperature, which was the opposite previously.

	mean	sd	0.025quant	0.975quant
(Intercept)	-2.770	7.307	-17.794	10.987
rainlow	-0.495	0.267	-1.018	0.036
rainhigh	-0.950	0.942	-2.828	0.894
templow	**-3.418**	1.542	-6.446	-0.346
temphigh	0.729	0.351	0.049	1.434
dem	0.004	0.007	-0.009	0.018
dryseas	-0.105	0.036	-0.177	-0.033
ndvi	0.106	0.043	0.025	0.194
range	**0.524**	0.137	0.310	0.842
sd	1.564	NA	1.173	2.097

Table 5 Regression coefficients μ, β, σ and ϕ (Group 1)

The explanation of this is a bad choice in the splitting of the temperature values, *templow* & *temphigh.* According to the next figure, cutting the temperature raster at a value of 34.5 °C is leaving only two values of prevalence in the *templow* variable. These two values where temperature is low are showing high prevalence rates, which are modifying the linear response of the "smooth" functions as shown in Figure 47.

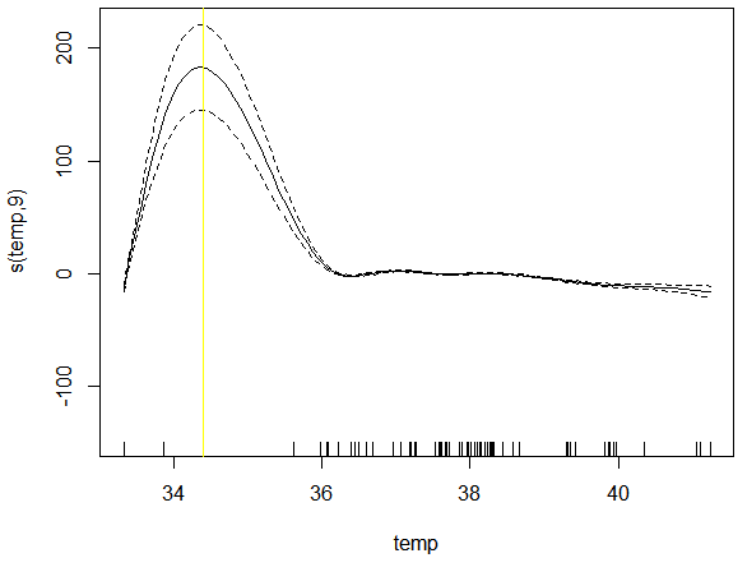


Figure 47 Temperature raster cutting at 34.5°C

The next figure shows the posterior means of the random space for this time group. The range value is eight times bigger than the whole dataset's range, meaning that the spatial noise has a much greater influence in terms of distance to the observed infection sites. This is illustrated by the bigger size of the influence areas showed in the figure.

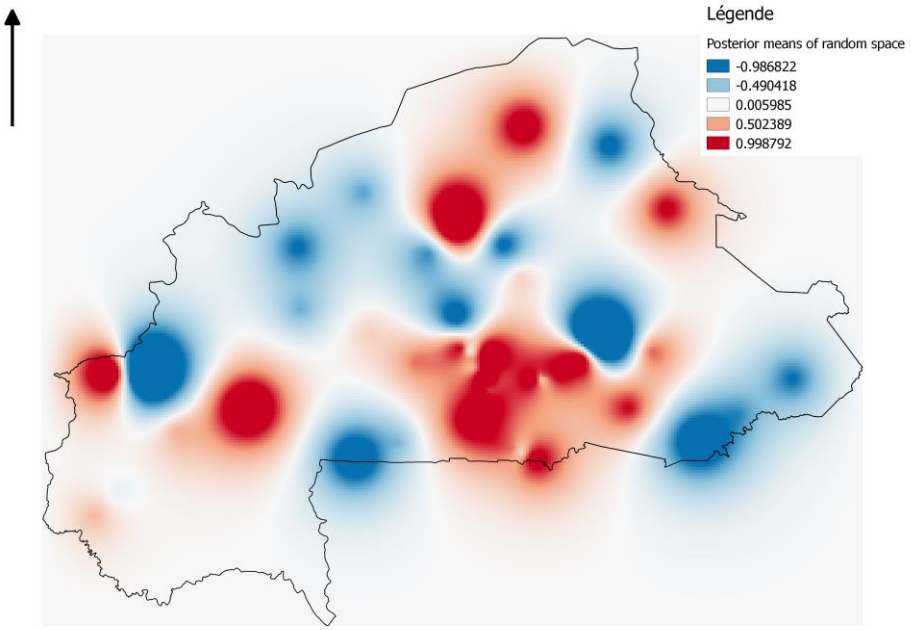


Figure 48 Posterior means of the random space E[U(s)|Y] (Group 1)

The model has been run again with a temperature covariate that has not been split in two low and high groups. The infection probability map is shown in the next figure:

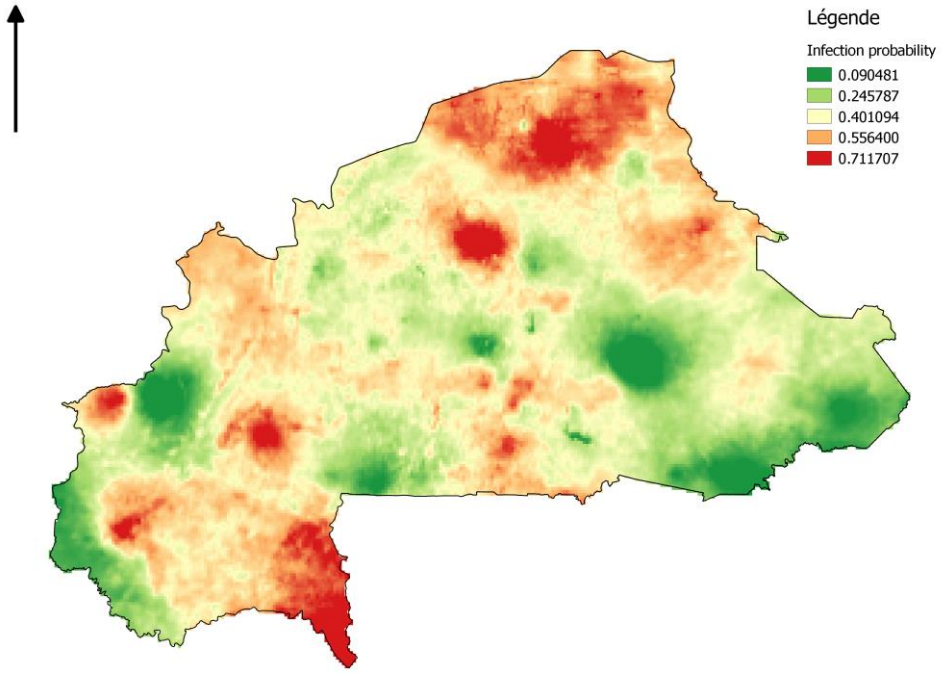


Figure 49 Infection probability map - *S.haematobium* - *Group 1* - *corrected*

The related coefficients values of this new model are shown in the next table:

Colonne1	mean	sd	0.025quant	0.975quant
(Intercept)	-14.060	14.771	-44.245	13.951
rainlow	-0.553	0.282	-1.105	0.009
rainhigh	-0.965	0.998	-2.952	0.988
temp	**0.316**	0.322	-0.305	0.965
dem	0.001	0.007	-0.013	0.015
dryseas	-0.091	0.037	-0.163	-0.018
ndvi	0.110	0.044	0.027	0.201
range	0.611	0.160	0.362	0.984
sd	1.744	NA	1.300	2.355

Table 6 Regression coefficients μ, β, σ and ϕ (Group 2)

The temperature coefficient *temp* has now a unique value of 0.316, giving back the positive correlation between temperature and infection risk. This is illustrated in the preceding Figure 49, where the infection probabilities in the South are now lower than with the model having two temperature covariates. Since the two or three only available observed values in the South show an important infection, the model is forcing this area to elevate the probability of infection.

The next figure is showing the standard deviation of the predicted probabilities, in other words it corresponds to the uncertainty of the prediction. It can be seen that around the observed values, logically, the uncertainty is very low, since they are input values of the model. It can be observed that in the South part, where observed values are less, that the uncertainty is high, meaning that the model is not capable of precisely predict an infection probability with the available information. This is also confirmed by the high standard deviation of the *Interecpt* shown in

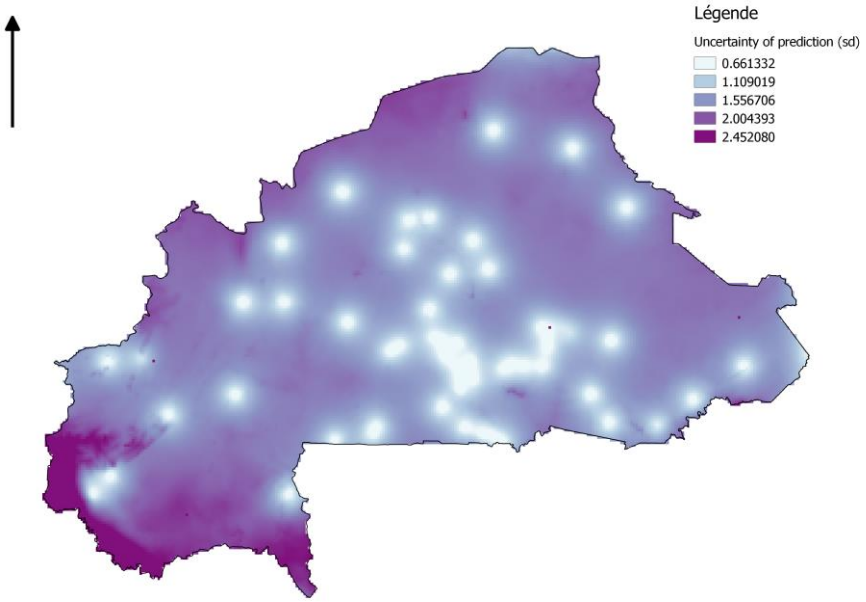


Figure 50 Uncertainty map - *S.haematobium* - *Group 1*

iii) *S.haematobium* modelling – 1970-1990

The second time group considers the period of years between 1970 and 1990. The infection probability map is illustrated in the next figure.

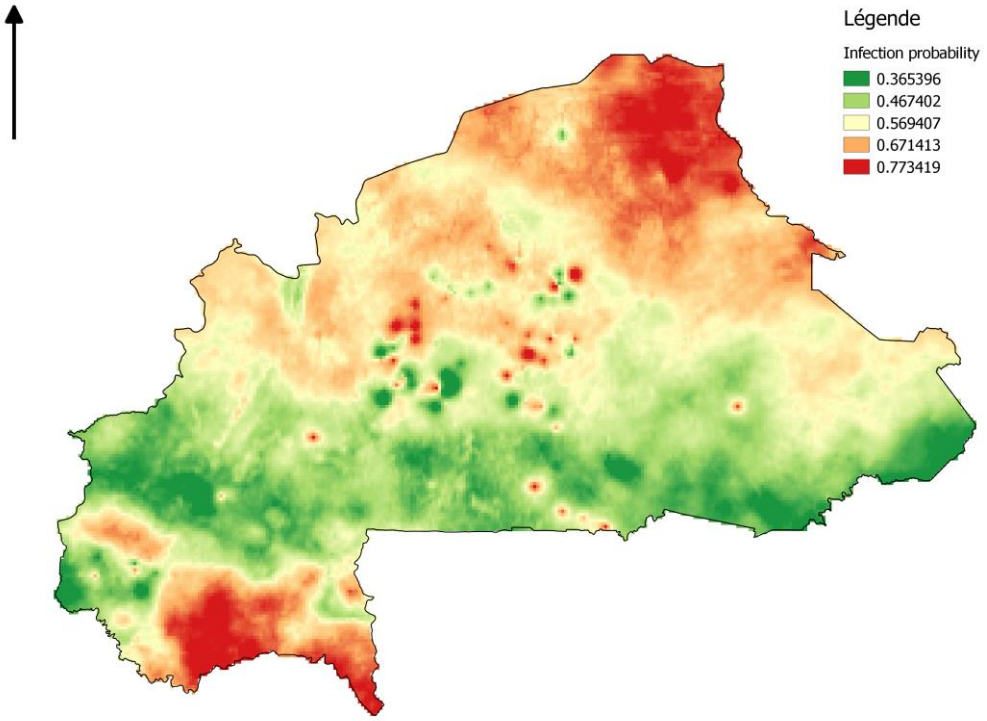


Figure 51 Infection probability map - *S.haematobium* - Group 2

The next tables gives the fitted coefficients and parameters of the geostatistical model.

	mean	sd	0.025quant	0.975quant
(Intercept)	-3.183	6.858	-16.909	10.0778324
rainlow	-0.159	0.296	-0.739	0.42602271
rainhigh	-0.296	0.296	-0.899	0.26881121
templow	-0.680	0.502	-1.709	0.26734773
temphigh	0.495	0.290	-0.071	1.07330731
dem	-0.004	0.005	-0.013	0.00535221
dryseas	-0.031	0.037	-0.104	0.04115081
ndvi	0.057	0.032	-0.005	0.12156731
range	0.132	0.034	0.078	0.21026265
sd	1.265	NA	1.031	1.534

Figure 52 Regression coefficients μ, β, σ and ϕ (Group 2)

It can be also observed that the *templow* covariate has a negative coefficient, meaning that the observed prevalence in the South of the country has relatively high values. This information is confirmed by the presence of a relative high number of observations with relatively strong prevalence in the South as it is showed in the next Figure 53 of the posterior means. One more time, the spatial noise is very localized, confirmed by a low range value of 0.132

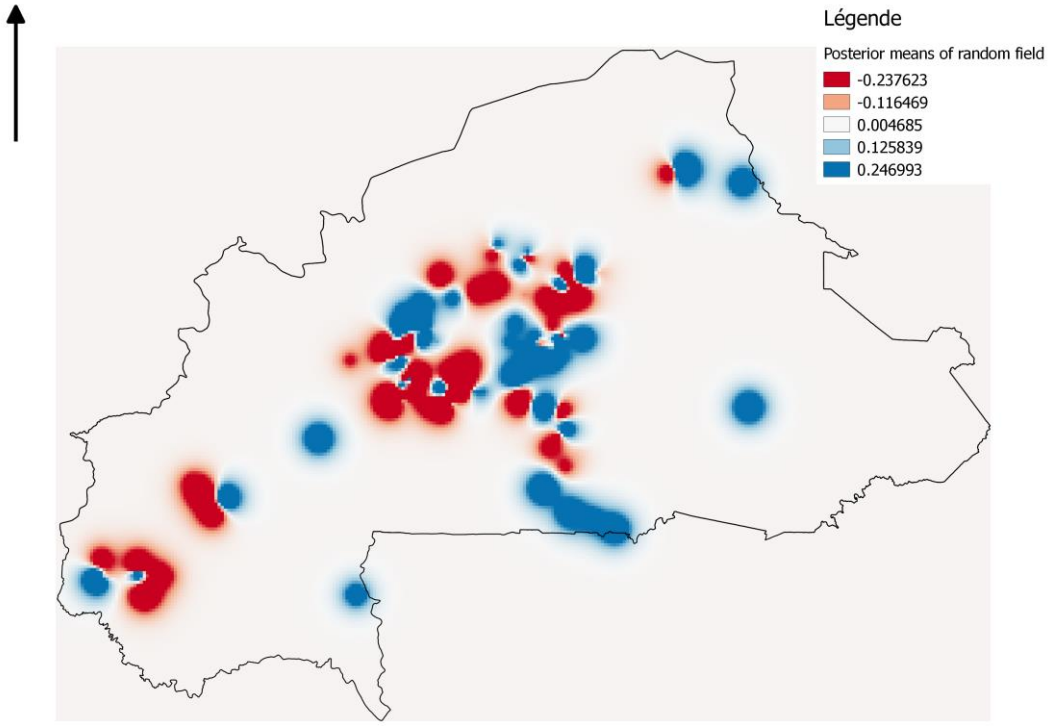


Figure 53 Posterior means of the random space E[U(s)|Y] (Group 2)

The uncertainty map is illustrated in the next figure.

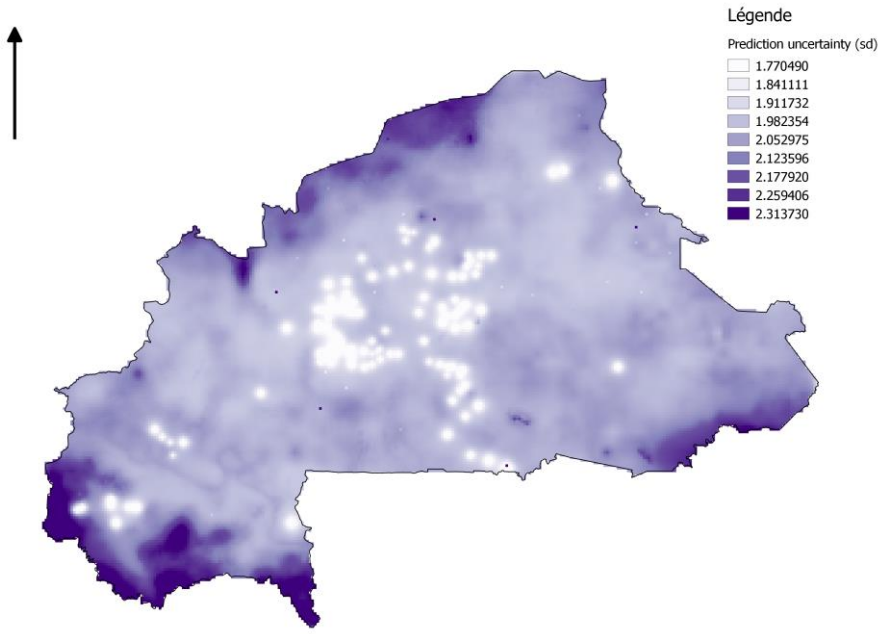


Figure 54 Uncertainty map - *S.haematobium* - *Group 2*

Uncertainties are again at higher level at the southern part of the country and the limits of the country.

iv) *S.haematobium* modelling – 1990-2013

For the last period covering the years between 1990 and 2013, the model results are presented in the next figures and table.

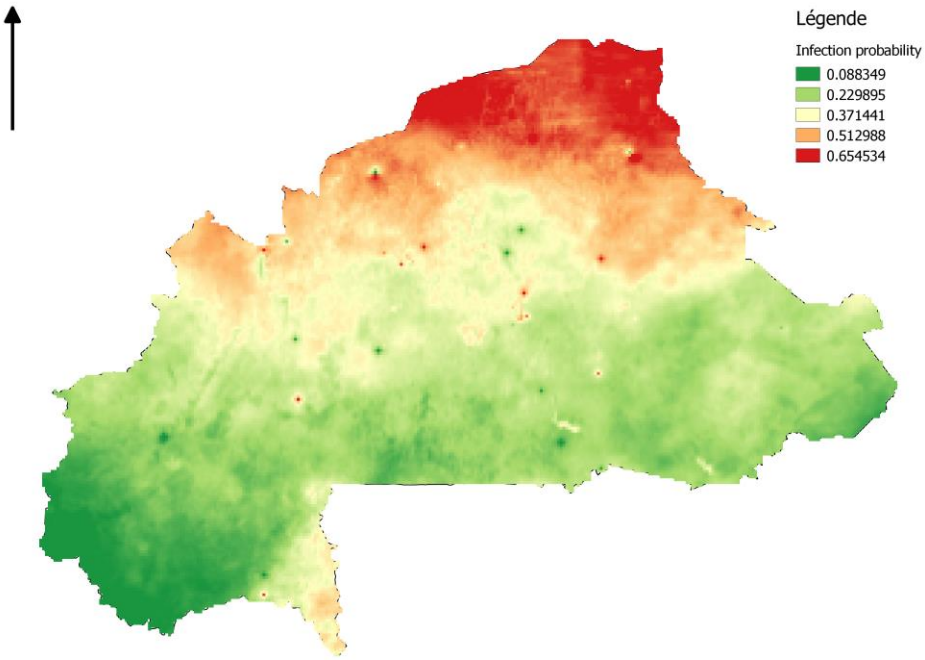


Figure 55 Infection probability map - S.haematobium - Group 3

The above infection probability map shows a North to South decreasing rate of schistosomiasis. It is pretty similar to the first model result where the whole dataset is used. Coefficient shows a positive correlation to the temperature and a negative correlation to rain covariates, which forces the model to generate such spatial gradient in the infection risk probabilities.

Colonne1	mean	sd	0.025quant	0.975quant
(Intercept)	21.114	22.970	-24.406	66.34873433
rain	-0.460	0.376	-1.203	0.284281674
temp	0.144	0.370	-0.585	0.879823821
dem	-0.002	0.011	-0.023	0.020082005
dryseas	-0.050	0.072	-0.192	0.092020707
ndvi	-0.069	0.064	-0.196	0.058111313
range	0.072	0.028	0.034	0.142984901
sd	1.802	NA	1.402	2.330305662

Table 7 Regression coefficients μ, β, σ and ϕ (Group 3)

The spatial *range* has a value of 0.072, which is small. The spatial noise, like the preceding modelling, results as being very localized as it is illustrated in Figure 56.

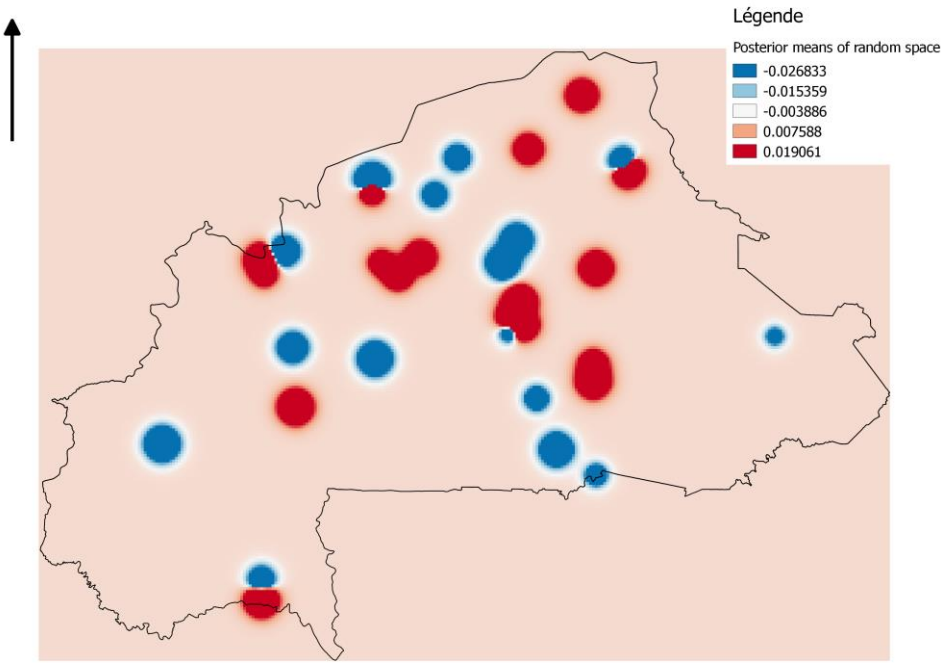


Figure 56 Posterior means of the random space E[U(s)|Y] (Group 3)

The following uncertainty map illustrates a maximum uncertainty in the South West part of the country, where the elevation is the highest. Since the elevation is relatively constant elsewhere in Burkina Faso, and the regression parameter of the DEM is very close to zero, the model is possibly unable to explain the change in prevalence in this region. It can be supposed that the covariate elevation is unable to explain any variance in the infection risk.

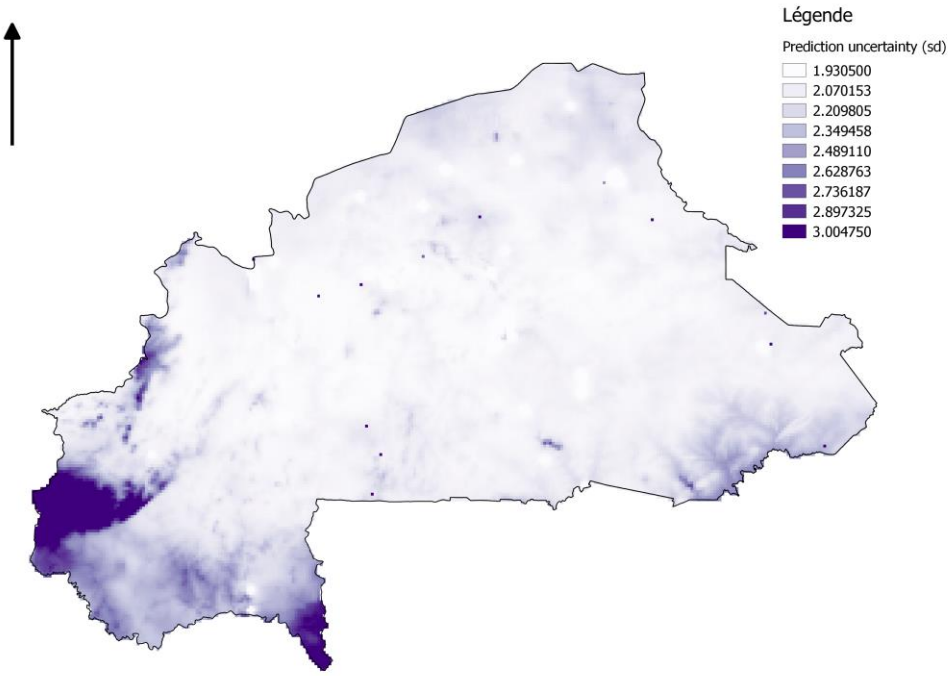
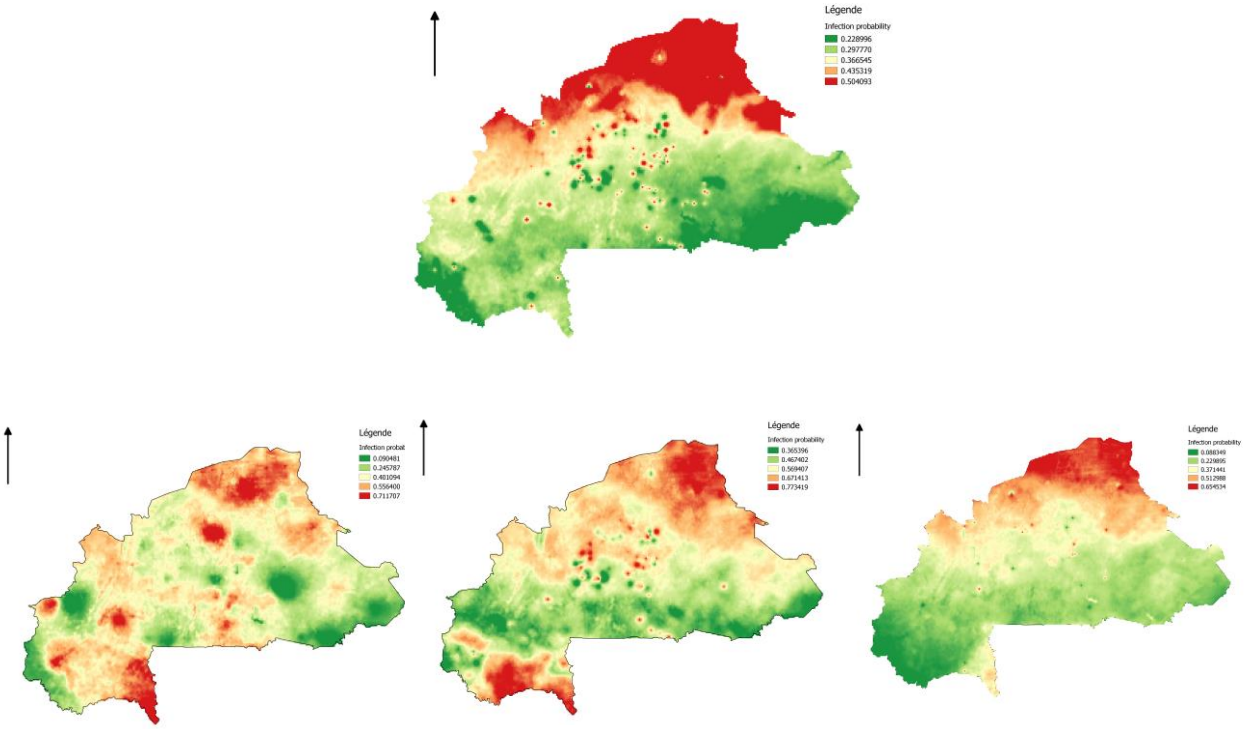


Figure 57 Uncertainty map - *S.haematobium* - *Group 3*

v) Comparison of the previous results

In this section, the four model results will be overviewed, compared and summarized. The next figure gives the four infection probability maps that were previously commented.

Figure 58 Infection probability maps of the whole period (top) and group 1 to 3 (bottom left to right)



A common result between the four maps is the modelled high prevalence rate in the North of the country that stays constant. This is confirmed by the regression coefficient β of the rain and temperature covariates, being always correlated to a stronger infection risk in the hot and dry climate of this region.

Low infection probabilities are generally found in the centre and South of the country. Nevertheless, for the two first groups, regions with high infection rates can be observed in the South of the country. The observed values of high prevalence in the South part for these groups forced the model to increase the risk of infection by either increasing the coefficient for the *templow* covariate, or simply by forcing the spatial noise or other covariates to generate high prevalence rates in this region.

It seems more adapted to conclude the constant presence of high prevalence rate in the North of the country than to try to give a temporal decrease of the schistosomiasis prevalence in the South. A first reason for this is the previously shown infection risk uncertainties in the South, which lowers the significance of the modelled results. Secondly, the absence of observed high prevalence rates for the third group doesn't explain the real absence of prevalence, and oppositely. However, in order to interpret these results correctly, it is necessary to discuss about the model results in regard to the available data. This will be given in the next chapter.

5 Discussions and Perspectives

This chapter aims to discuss point by point about the whole work presented previously and to explain the pros and contras of the different steps undertaken in this project. Possible perspectives in order to perform or continue the present work will also be taken in review.

This chapter will be structured as following: a first paragraph will be dedicated to the data in general, getting into the data management processes and their quality and quantity. A second part will expose a discussion about the model construction and running, and finally a last paragraph will try to put in relationship the data and the model, discussing about the difficulties and the possible perspectives of such model building.

5.1 About the data

5.1.1 Data management and processing

The data management was an important part of this work. Effectively, dealing with big amount of data, moreover when combining spatial and non-spatial information, can be very time-consuming if not managed in an organized, centralized way. The use of the software Postgres with its spatial extension PostGIS has proved to be very efficient for the data storage, treatment, processing and exchange. The possibility to store raster information was also a big advantage. Postgres enables the rapid calculations and analysis between spatial and non-spatial data, e.g. calculating minimum distances, means of prevalence rates over regions, etc., in a very easy and efficient way.

PostGIS is also effective in the fact that the database server can be easily connected with the GIS QuantumGIS (Team, 2014), which is very efficient to rapidly visualize the data in a Geographic Information System.

It was also necessary to use more functional GIS programs like ArcGIS (ESRI, 2011) for more heavy and complex data treatment, e.g. when working with the satellite images and the supervised classification.

For the statistical treatment of spatial data and modelling, the software R has shown to be very efficient, proposing a large variety of packages in order to support and run spatial information. These packages, like *raster* for the image information or *geostatsinla* for the modelling, enable to spare a lot of time since they propose a very simple and intuitive way for their applications.

Finally, when treatment of a large dataset, for example calculating monthly means from daily raster information (i.e. rain, temperature or dry season time), Matlab is an excellent mathematical tool. The possibility to use the spatial *Mapping Toolbox* made it possible to exchange data between different software without changing their formats

The combination of specific software was necessary in order to rapidly, efficiently access and treat the data and these programs proved to be very efficient in realizing the needed tasks.

5.1.2 Data quality and quantity

The dataset used in this project, presented in part 3.1.2, has the advantages of having a good quality. Effectively, an effort was put into acquiring the data from a wide panel of different sources. These informations showed to be recent, and generated by reliable sources, e.g. for the temperature data obtained from the USGS/NASA research program and others (see 3.1.2).

Nevertheless, more useful information could still be introduced in the database and used for the model building in future works. For example, the supervised classification effectuated on the satellite imagery which generated vectors objects of water areas was not used in the model. It would have been interesting to calculate a raster set containing information about the minimum distance to such water area, river or dam. An intent of calculating such a raster was done but not terminated and was therefore not presented in the methodology. The main idea was to create a small enough grid of points (distance of about 3km to each other) over the whole country, and to calculate the distance of each point to the closest water object. The next step would have been to interpolate these points and generating a raster of distances. But this process was very time consuming, about three days of processing, and the results presented some mistakes. Because of the schedule it was decided to leave this idea. But having such information would have been very useful in order to predict infection risks since water is an essential part of the schistosomiasis reproduction and maintenance cylcle.

Since humans are final host of the disease reproduction, it would have been also interesting to obtain data on the mobility of the population. An idea of generating such spatial data would be, for example, to calculate the distances to a city or to a road, where mobility is presumed to be high. The displacement of infected population induces the displacement of parasites, creating new sites of infection risks. This information should be taken into account for a future work.

The information related to either distance to water bodies or mobility can be obtained from any available sources, like through the internet for example. Nevertheless, other informations have to be obtained on site. For example, it would have been very interesting to obtain data concerning the sanitary state, access to drinking water across the country, as means of provinces for example, which could be linked to the risk of infection distribution. This project aimed to acquire such data by staying on site in Ouagadougou, Burkina Faso, for a duration of two months. It was unfortunately very difficult to obtain such data due to administrative barriers and lack of time. But these type of information are interesting in such model building process and should definitely be introduced if obtained.

Another interesting axis of study could be the acquisition of data related to the intermediate hosts, e.g. the ecology of the water areas, the dynamics of the snails mixed with hydrological informations. Since the hosts are necessary components of the disease distribution and maintenance, they constitute an interesting axis of research for modelling the infection risks.

To conclude, the data obtained certainly constitute a good basis for the model establishment, but they should surely be mixed with more disease-related information, since schistosomiasis has a complex cycle influenced by both human activities in relationship with water and the ecology of the intermediate hosts.

5.2 About the model

A lot of models can be used in order to estimate a probability of infection, but the literature, i.e. (Clements, 2009) has proven that the use of Bayesian geostatistical models (see 2.3) are efficient when dealing with epidemiological predictions. Since the infection observed data present values of positive and examined number of people, a combination of a generalized geostatistical model to an assumed binomial distribution of infected people seems to be an effective method to obtain valuable predictions. These models integrate environmental covariates and enable to estimate how and which of these covariates have an influence on the disease distribution.

The simplicity of use of the *geostatsinla* only requires inputs of covariates as rasters, a definition of a prior distribution of parameters, a simple regression formula and the size of the resulting predicted raster. Unfortunately, some of the raster presented in 3.1.2 could not be used due to computational difficulties, such as the Human Influence Index (HII) and the Landcover information. The HII could have been an interesting set since it describes human oriented information, which is more heterogeneous over space.

Another interesting point of this work was to use General Additive Models (GAM), which enabled to observe the linearity of the covariate values when predicting the infection probability. This was a good tool to choose whether or not, and where to split the covariates in order to enhance the linearity for a better linear fitting. Nevertheless, it was showed in 4ii) that a bad choice of the value at which to split the covariate could significantly modify the predictions results. Another inconvenient of splitting a covariate is that it is adding every time a new predictor; and having too much covariates could lead to an over parameterization of the model.

5.3 Model/data combination

In this section will be discussed more deeply about the model related to the used data. It will be seen how the model works with the obtained predictors and how it could be improved.

Generally, like observed in the results (see 4v)), the resulting are concordant to the idea that the infection probabilities are positively varying gradually from North to South, the variable Y_i being relatively strongly correlated to the temperature and the rain gradient across the country. However, as it is shown in the next Figure 59, these two predictors (*rain & temp)* are highly correlated. It can also be observed that the dry season time *dryseas* and the *ndvi* covariates are also relatively well correlated to the temperature. As it is shown in 3.1.2ii), *dryseas, temperature and rain* are gradually varying through space, in a relative North-South direction, and they are also pretty much correlated. Furthermore, as shown in Figure 60 which describes the local autocorrelation of the average prevalence rates over the Burkina Faso's provinces, high values of infection risks are likely to be distributed in the northern part of the country, and the low values of infection risks in the South.

It is therefore expectable to have a heterogeneous repartition of the predicted prevalence rates from North to South since the fitting is using gradually correlated regression predictors. Furthermore, the posterior means of the random field, also characterized as *spatial noise*, play an important role since they compensate the variance that is not explained by the homogeneous covariates. These *U(s)* values can therefore be seen as "spatial compensators of the model regression". For example, in the

covariates used in the model wouldn't be homogeneous, the spatial noise would have compensate the North South prevalence rates by augmenting the *U(s)* values in the North. In the Figure 42, it can be seen how high prevalence observations are located in a region where the linear regression coefficients impose a low prevalence rate. In order to reduce the impact of such "non-normal" observations (according to the linear regression), the range parameter, defining the influence distance of the noise, is reduced to a small value. By having more heterogeneous covariates (e.g. distance to water, sanitation, mobility), the linear regression could have been less homogeneously distributed, adapting the spatial noise differently.

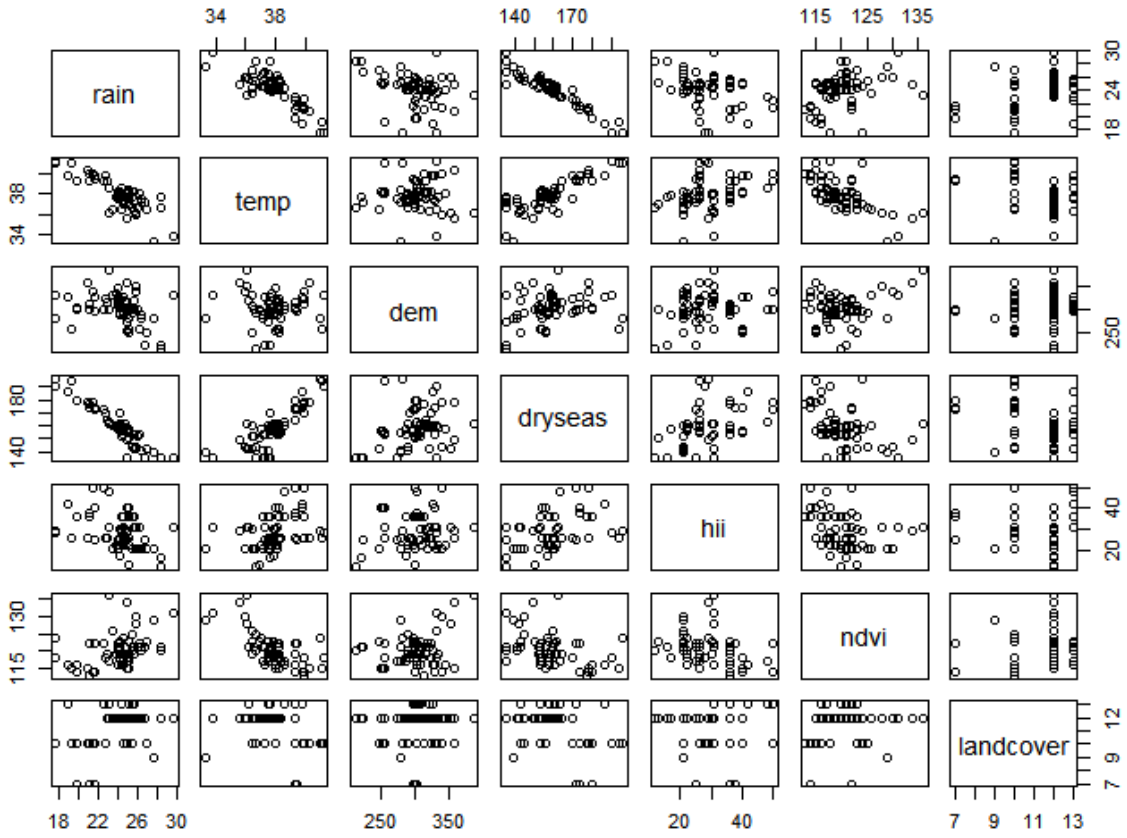


Figure 59 Correlation plots of the model's covariates evaluated at the *S.haematobium* observed points

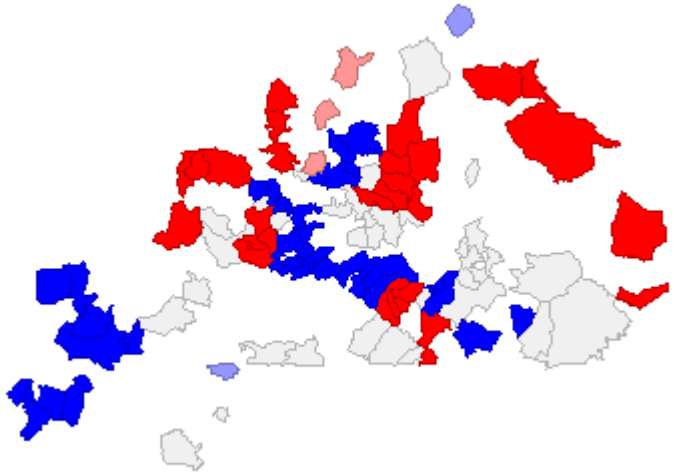


Figure 60 Local Indicators of Spatial Associations (LISA) for the province scale averaged prevalences (High-High clusters in red, Low-Low in Blue)

Finally, it would be interesting to discuss about the temporal evolution analysis of the infection risk maps presented in 4v). Cutting the dataset into three temporal groups is analogically reducing the number of observed points by three. Some regions, like the southern part of the country in Figure 56, end up having very few observation points (two in this case with low and high prevalence rates). This absence of enough data is increasing the uncertainty of prediction of the model. It is therefore more accurate to leave all the data and make a regression over the whole period of time, in order to reduce the uncertainties and to give a more global overview of the schistosomiasis distribution over the country. A condition for starting temporal analysis would be to have more spatially dispersed observations in order to predict efficiently the distribution of risks.

6 Conclusion

Schistosomiasis is a complex disease, evolving processes that imply the displacement of the parasite between two different hosts, depending on specific environmental and ecological conditions for its survival. Geostatistical models only are not capable of explaining the distribution and evolution of the disease across the country, and it's not the aim of this work, but they are a useful tool in order to interpolate and predict, in a global tendency, the spatial repartition of the probability of infection. These stochastic tools enable to predict uncertainties to the predicted values and to correlate the schistosomiasis infection rates to some global environmental conditions, for example to temperature as it has been shown.

This project, by proposing a methodology for the generation of infection probability maps, shows that the use of Bayesian geostatistics can be well adapted if enough care is accorded to the choice of the model predictors and to a good interpretation of the results. Available databases, like the open-source GNTD platform, can provide a large amount of prevalence data covering a large period of time and are a very useful source for making those models possible. Unfortunately, heterogeneously distributed covariates were not used in this work in order to locally explain the variation in prevalence rates; these localized variation being compensated by the spatial noise of the model.

Further work could though be applied in order to perform the model outputs. For example, it would be very interesting to integrate more heterogeneous, localized data to the model, such as sanitation-based information, or distances to water bodies, etc. Also, the combination of homogeneous and heterogeneous covariates seems to be the best option since they enable to predict the large spatial trends (e.g. North-South distribution), and maybe the local disparities in prevalence rates by introducing heterogeneous predictors. Additional work could focus on deciding if taking less homogeneous data would alter or not the model predictions and uncertainties, since lot of the used covariates are spatially correlated.

However, this work showed that the use of specific tools and programs, combined with an efficient data management procedure, and a sufficient theoretical background, enables to generate results that can be used for assisting disease fighting programs.

7 References

Boelee, Madsen. 2006. *Irrigation and Schistosomiasis in Africa - Ecological Aspects.* 2006.

Boykov, Yuri. 1998. *Markov Random Fields with Efficient Approximations.* 1998.

Brown, Patrick E. 1996. *Model-Based Geostatistics the Easy Way.* Toronto : Journal of Statistical Software, 1996.

Clements. 2009. *Use of Bayesian geostatistical prediction to estimate local variations in Schistosoma haematobium infection in western Africa.* 2009.

Cressie, N. 1993. *Statistics for spatial data - revised edition.* Wiley, New York : s.n., 1993.

Danielson, J.J. 2011. *Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010).* 2011. Open-File Report 2011-1073.

DRAH/RH - BF /EIER-ETSHER -2IE. 2005. *Retenues d'eau du Burkina Faso - Inventaire.* 2005.

ESRI. 2011. *ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.* 2011.

Friedl, M. 2010. *MODIS Collection 5 global land cover : Alorith refinements and characterization of new datasets.* s.l. : Remote Sensing Environment, 2010. 114, 168-182.

Harvey, Francis. 2008. *A Primer of GIS : Fundamental Geoagraphic and cartographic concepts.* 2008.

Herman, A. 1997. *Objectively Determined 10-Day African Rainfall Estimates Created dor Famine Warly Warning Systems. Int. J. Remote Sensing.* 1997. 18, 2147-2150.

Love. 2002. *The Climate Prediction Center Rainfall Algorithm Version 2.* 2002.

Montresor. 1998. *Guidelines for the evaluation of soil-transmitted helminthiasis and schistosomiasis at community level.* 1998.

Pearson, Richard D. 2013. *Schistosomiasis.* 2013. http://www.merckmanuals.com/professional/infectious_diseases/trematodes_flukes/schistosomiasis.html.

Poda, J-N. 1996. *Distribution spatiale des hotes intermédiaires des schistosomes au Burkina Faso: Facteurs influençant la dynamique des populations.* Ouagadougou : s.n., 1996.

Rue, Martino, Chopin. 2009. *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.* s.l. : Journal of the Royal Statistical Society B, 2009. 71(2), 319.392.

Schur. 2012. *Determining Treatmen Needs at Different Spatial Scales Using Geostatistical Model-Based Risk Estimates of Schistosomiasis.* 2012.

Southgate. 1997. *Schistosomiasis in the Senegal River basin: before and after the construction of the dams at Diama, Senegal and Manantali, Mali and future prospects.* 1997.

Swets, D.L., Reed. 1999. *A weighted least-squares approach to temporal NDVI smoothing.* Portland, Oregan : s.n., 1999.

Team, QGIS Development. 2014. *QGIS Development Team, <YEAR>. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org.* 2014.

Trevor Hastie, Robert Tibshirani. 1986. *Generalized Additive Models.* 1986.

Utroska. 1990. *An Estimate of global needs for Praziquantel within Schistosomiasis Control Programmes.* s.l. : WHO, 1990.

Van.d.Werf. 2003. *Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa.* s.l. : Acta Tropica, 2003. 86 (2-3): 125-139.

Vounatsou, Penelope. 2011. *GNTD Database.* 2011.

Wan, Zhengming. 2006. *MODIS Land Surface Temperature Products User's Guide.* Santa Barbara : ICESS, University of California, 2006.

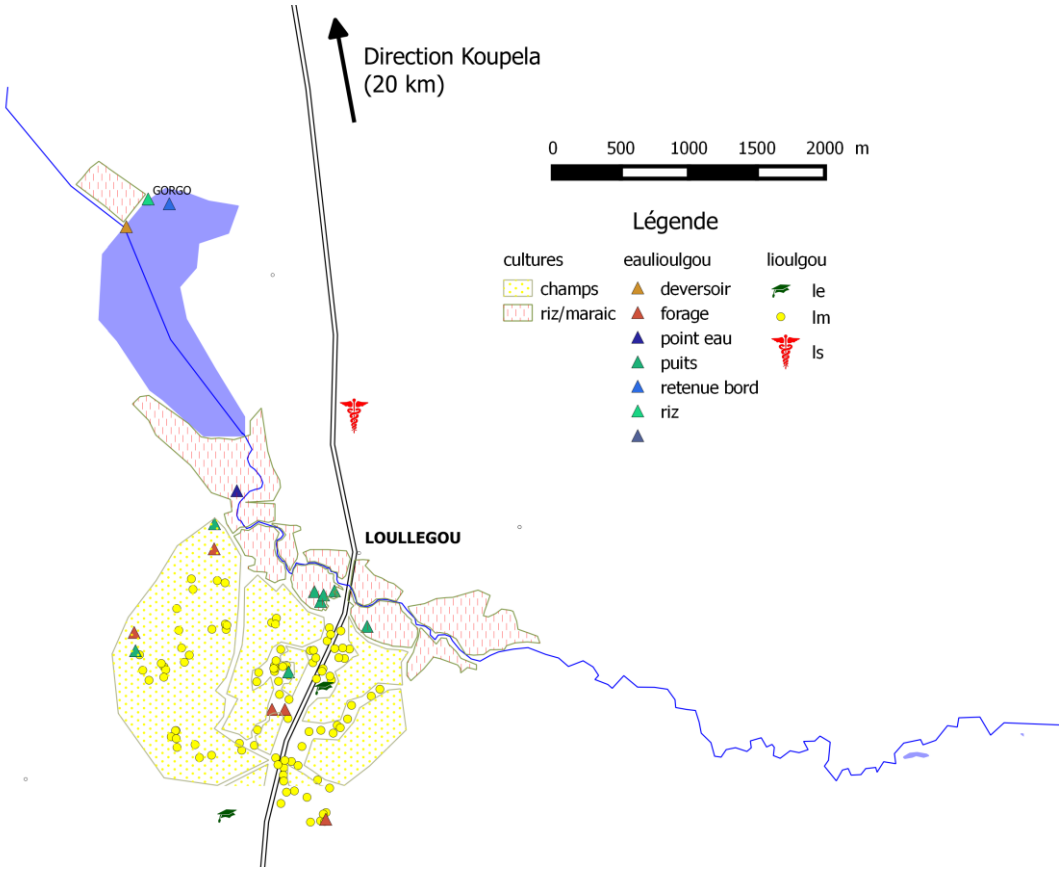
WCS. 2005. *Last of the Wild Project, Version2, 2005 (LWP-2): Global Human Influence Index (HII) Dataset.* Colombia University : s.n., 2005.

WHO. 2014. *Schistosomiasis, a major public health problem.* s.l. : http://www.who.int/schistosomiasis/en/, 2014.

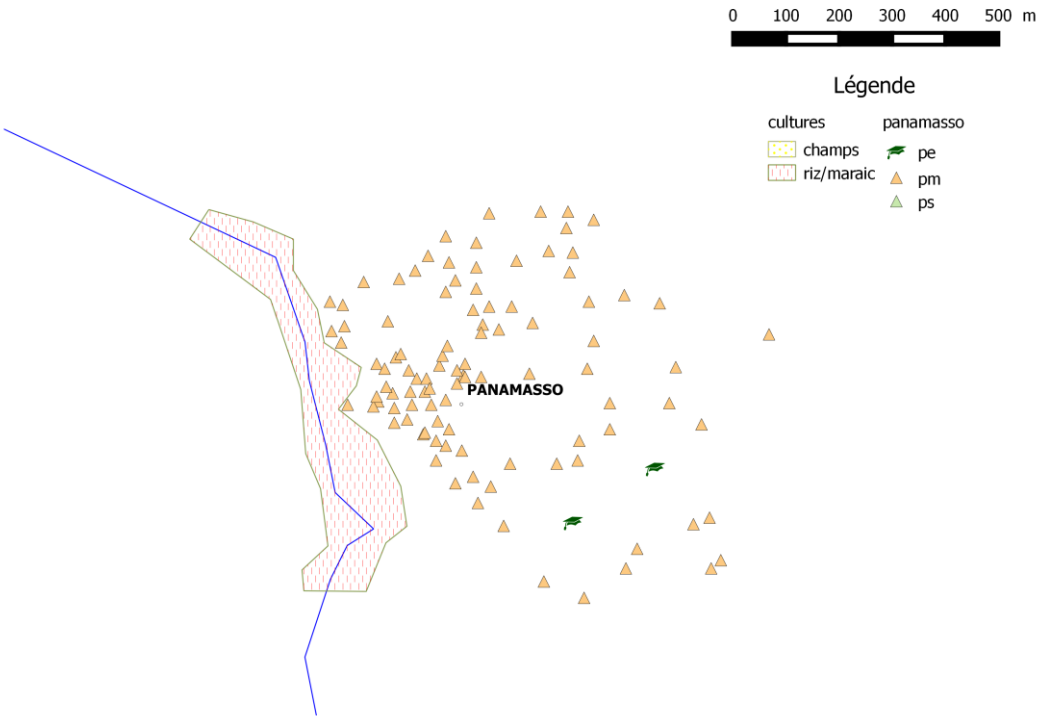
8 Annexes

RFE	http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP/.CPC/.FEWS/.Africa/.TEN-DAY/.RFEv2/
LST	https://lpdaac.usgs.gov/products/modis_products_table/mod11c2
Land cover	https://lpdaac.usgs.gov/products/modis_products_table/mcd12q1
NDVI	http://earlywarning.usgs.gov/fews/africa/web/readme.php?symbol=zd
DEM	https://lta.cr.usgs.gov/GMTED2010
HII	http://sedac.ciesin.columbia.edu/data/set/wildareas-v2-human-influence-index-geographic/
population	http://sedac.ciesin.columbia.edu/data/set/grump-v1-population-density

Annex 1 : Sources of rasters : websites



Annex 2 Lioulgou map



Annex 3 Panamasoo map