BORDEAUX SCIENCES AGRO

Internship report

AGROMET project : Investigating spatial interpolation of temperature using Multiple Linear Regression

Centre de Recherches Agronomiques Wallon (CRA-W)

Rue de Liroux 9, 5030 Gembloux, Belgique

Loïc Davadan

Internship supervisor : Thomas Goossens

Supervisor : Jean-Pierre Da Costa

21/05/2018 — 20/08/2018

Acknowledgements

This is the content of my acknowledgements

Preface

This document is my internship report for Bordeaux Sciences Agro as part of my formation in "Numérique pour l'Agriculture" and my 3-month internship in the CRA-W.

It was completely written with RMarkdown and LaTeX.

Abstract

The European directive 2009/128/CE imposes member-states to set up tools that allow for a more rational use of crop protection products. Among these tools, agricultural warning systems, based on crop monitoring models for the control of pests and diseases are widely adopted and have proved their efficiency. However, due to the difficulty to get meteorological data at high spatial resolution (at the parcel scale), they still are underused. The AGROMET project, led by CRA-W, aims to generate a high spatial resolution network which diffuses interpolated weather data provided by physical weather stations using geostatistical tools. The internship operates in data acquisition and data analysis steps. As a first step, the objective was to collect data which can explained some weather parameters and then, secondly, to integrate them in a benchmark of statistical methods and combinations of variables to identify those which return models with the lowest error. For a matter of time, the internship has been focused on temperature prediction.

La directive européenne 2009/128/CE impose aux états-membres de mettre en place des outils visant à une utilisation rationnelle des produits phytosanitaires. Parmi ces outils, les systèmes d'avertissements agricoles, basés sur des modèles de suivi des maladies ou des ravageurs sont très largement utilisés et ont déjà fait preuve de leur efficacité. Cependant, ils sont encore sous-exploités du fait de la difficulté de disposer d'une information météorologique à haute résolution spatiale (à l'échelle de la parcelle). Le projet AGROMET, dirigé par le CRA-W, a pour but de genérer un réseau de stations virtuelles à haute résolution spatiale qui diffusera des données météorologiques interpolées à partir de données issues de stations physiques et d'outils géostatistiques. Le stage intervient dans la phase d'acquisition des données et d'analyse des données. Dans un premier temps, l'objectif était de récolter des données dont pourraient dépendre certains paramètres météorologiques puis, dans un second temps, de les intégrer dans une analyse comparative des méthodes statistiques et des combinaisons de variables explicatives pour trouver celles qui retournent des modèles avec l'erreur la plus faible. Pour une question de temps, le stage a été ciblé autour de la prédiction de la température.

Abbreviations

- API : Application Programming Interface
- ANN : Artificial Neural Networks
- CRA-W : Walloon agricultural research center
- CRS : Coordinate Reference System
- JSON : JavaScript Object Notation
- MAE : Mean Absolute Error
- RMI : Royal Meteorological Institute
- RMSE : Root Mean Square Error
- OS : Operating System
- SSH : Secure Shell
- WGS84 : World Geodetic System 1984

Glossary

- to nest : *imbriquer*
- late blight : *mildiou*
- wheat septoria : *septoriose du blé*
- rain gauge : *pluviomètre*
- orange midge : *cécidomyie orange du blé*
- leaves wetness : *humidité du feuillage*
- forecast : *prévision*

List of Figures

List of Tables

Table of Contents

Introduction

Use of pesticides and other crop protection products is a topical issue in an environmental and societal context. These products are increasingly criticized for their risks and impacts on human health and environment. Crop monitoring models are developed and their efficiency is well demonstrated. Acting at the right time in plots is inscreasingly possible thanks to these models. In Belgium, the Walloon agricultural resarch centre (CRA-W) is a research centre where a lot of issues are explored to bring solutions.

From May 22nd to August 20th, I did an internship in the (CRA-W). I worked on the AGROMET project which is a project about agrometeorology where the aim is to provide a near real-time hourly gridded datasets of weather parameters at the resolution of 1 km^2 for the whole region of Wallonia characterized by a quality indicator. This project is led by the Farming Systems, Territory and Information Technologies Unit.

The internship has for objective to investigate a spatial interpolation of the temperature using multiple linar regression with the best combination of explanatory variables.

First, the report will present the CRA-W, its organisation, the Unit where I worked and the project. Then, my workflow will be detailed in two parts : the data acquisition and the data analysis through the benchmark. Finally, the results will be interprated and discussed.

Chapter 1

Presentation of the AGROMET project and the CRA-W

1.1 CRA-W and Farming Systems, Territory and Information Technologies Unit

The CRA-W was founded in 1872 and depends on the Regional Goverment of Wallonia. It aims to maintain and develop the scientific excellence and societal usefulness and contributes to sustainable development of the agricultural industry in Wallonia in its economic, ecological and cultural dimension. 120 scientifics are working in the CRA-W on three sites (Gembloux, Libramont and Mussy-la-Ville) representing 300 ha of fields, greenhouses, laboratories and offices. The CRA-W is a place for scientific research but also to provide services in agricultural and agri-food sector keeping a perspective view on the development of agriculture.

The research is divided into 4 main fields where more than 100 projects are permanently in progress. :

- Precision agriculture
- Precision livestock farming
- Risk management
- Understanding products

The CRA-W is divided into 4 departments with 4 research units each :

- Life sciences
- Production et sectors
- Valorisation of agricultural products
- Agriculture and natural environment

The last one has the Unit 11 corresponding to Farming Systems, Territory and Information Technologies Unit where I realized my internship. This Unit develops tools to meet society's new expectations and decision support systems to improve the technico-economic and environmental performance of farming systems. There are actually 28 projects in progress.

The activities of the Unit are the mainly the following :

- Adaptation of agrosystems to global change : definition of references
- Adaptation of agrosystems to global change through bottom-up approaches
- Support to the development of agrosystems in line with territory projects
- Decision support systems and information technologies for the management of multifunctional agriculture
- Spatial information systems for the management of rural areas.

PAMESEB is a non-profit organisation handle by the CRA-W which aims to promote agrometeorology by taking climatic conditions into account in the walloon agriculture. PAMESEB has 30 weather stations in Wallonia. These stations provide measures for ways to fight crop diseases like late blight and wheat septoria. Stations have a local acquisition unit for hourly data recording. The PAMESEB network has an important place in the AGROMET project because it provides data for it.

Each PAMESEB station has 5 basic sensors :

- Temperature sensor
- Relative humidity sensor
- Solar sensor
- Wind sensor
- Rain gauge

1.2 The AGROMET project

1.2.1 Context

The European directive 2009/128/CE imposes member-states to set up tools that allow for a more rational use of crop protection products. Among these tools, agricultural warning systems, based on crop monitoring models for the control of pests and diseases are widely adopted and have proved their efficiency. However, due to the difficulty to get meteorological data at high spatial resolution (at the parcel scale), they still are underused. The use of geostatistical tools (Kriging, Multiple Regressions, ANN, etc.) makes it possible to interpolate data provided by physical weather stations in such a way that a high spatial resolution network (mesh size of 1 km2) of virtual weather stations could be generated.

That is the objective of the AGROMET project. Moreover, some CRA-W's units and other partners are interested in to build models against crop diseases like potato late blight or orange midge which depends on meteorological conditions.

The project was inspired by several academic papers dealing with spatial interpolation of data like *Use of geographic information systems in warning services for late blight* (Zeuner, 2007), *Decision Support Systems in Agriculture : Administration of Meteorological Data, Use of Geographic Information Systems(GIS) and Validation Methods in Crop Protection Warning Service* (Racca *et al.*, 2011) and *Spatial interpolation of ambient ozone concentrations from sparse monitoring points in Belgium* (Hooyberghs, 2006).

1.2.2 Objectives

The project aims to set up an operational web-platform designed for real-time agro-meteorological data dissemination at high spatial (1km2) and temporal (hourly) resolution. To achieve the availability of data at such a high spatial resolution, we plan to "spatialize" the real-time data sent by more than 30 connected physical weather stations belonging to the PAMESEB and RMI networks. This spatialization will then result in a gridded dataset corresponding to a network of 17 000 virtual stations uniformly spread on the whole territory of Wallonia.

These "spatialized" data will be made available through a web-platform providing interactive visualization widgets (maps, charts, tables and various indicators) and an API allowing their use on the fly, notably by agricultural warning systems providers. An extensive and precise documentation about data origin, geo-statistic algorithms used and uncertainty will be also available.

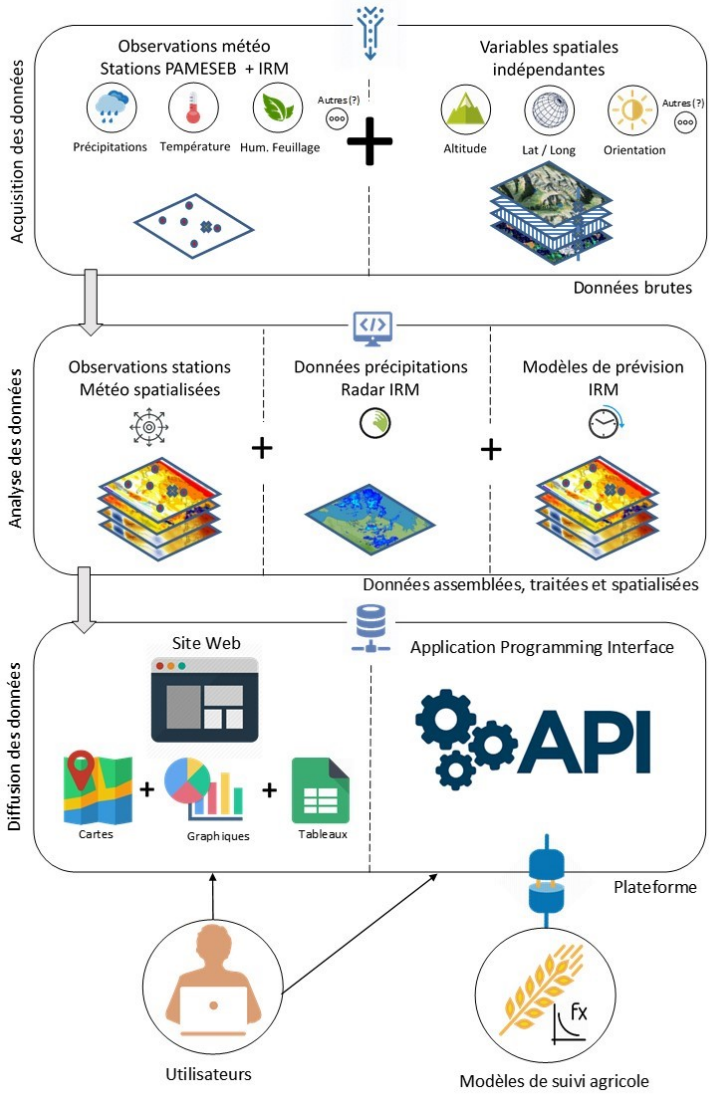


Figure 1.1: First draft of the functional architecture of the platform

Meteorological data wanted to be predict are :

- Temperature (1.5 meters above the ground)
- Relative humidity (1.5 meters above the ground)
- Leaves wetness
- Rainfall will be spatialized from RMI rain radar data.

To predict these variables, known data are used :

- Digital elevation model and its derivatives like aspect and slope
- Solar irradiance
- Other variables discussed to improve the prediction : distance to sea, CORINE land cover...

The Figure 1.1 shows the architecture of the future platform and steps before data diffusion.

1.2.3 Importance of the internship for the project

To predict meteorological variables, several statistical method will be tested. The identified methods at this time are multiple linear regression, ANN and several kriging methods.

Two objectives were clearly defined for my internship. I operated in the data acquisition and data analysis steps with the aim of predicting **temperature** using **Multiple linear Regression**.

First, I have collected data which can be explanatory variables for temperature and organised them to be integrated in the analysis and benchmark.

Then, I ran a benchmark experiment where different desired regression learning algorithms are applied to various regression tasks (i.e. datasets with different combinations of explanatory variables and the target weather parameter) with the aim to compare and rank the combinations of algorithm and used explanatory variables using a cross validation resampling strategy that provides the desired performance metrics. And then, I aggregated, by calculating the mean, all the hourly performance measures to choose the method that globally performs the best. For each desired hourly dataset, I applied the choosen method to build a model to make spatial predictions. The predictions and their uncertainty have been vizualized using maps. Finally, I had to make the predictions available on the platform together with its uncertainty indicator.

For a matter of time, this benchmark is run on two large years (from 2015-11-11 to 2018-06-30) because some data were not available before this period.

Chapter 2

Working environment

2.1 Applications and tools

Working in development field often imposes to be methodic. That's why preparing the working environment is very important.

First, I have installed **Ubuntu GNOME**, a distribution of Linux on my laptop. Indeed, this OS is prefered by developers and open-source addicts thanks to the high contribution to improve distributions. Linux distributions are safer than Windows due to that and it is very easy to automatise a lot of commands. This installation was done thanks to a USB drive with a boot of Ubuntu GNOME.

Once I had Ubuntu installed on my laptop, I used an **ANSIBLE script** to install all the applications I need automatically. Moreover, this script handles the updates of these applications. That is a very useful way to earn some time.

Accessing to servers has to be secured. That's why every developer should have a **SSH** key. This key or token is unique and enable people to access to servers. It is useful to access to Git repositories for example.

GitHub is a hosting service for version control. Its utilisation is very common for developers because their codes are online, the access is public and GitHub enable to handle versions of files. It makes easier the collaborative work on a same code and enable to use codes of other users.

For my internship, I need to work with my mentor to code. GitHub is the best solution to that. I created a folder in my laptop to clone all the repositories I need for my work. Then, I have a copy that I can modify and I can send my modifications on GitHub. To clone these repositories, my SSH key was useful.

The AGROMET project for whom I worked has an **API** to store meteorological data from all the stations. An account has been created for my internship. Then, I can get data from the API to test my codes. These data have JSON or GeoJSON format, a open-standard file format derived from JavaScript which is easy for machines to parse and generate and easy to read and write for humans.

Docker is a software for containerize platforms. This container approach has many advantages compares to the use of virtual machines : lightweight, quick and modular.

There are two main reasons to use R in conjunction with Docker. First, it allows you to quickly and easily share your work whatever the OS and R configuration of your collaborators. Second, it allows you to work in an isolated environment. This means that you will never pollute your OS and e.g. run in time-consuming re-installation procedures due to broken configuration. In case of OS crash, simply relaunch your Docker R container with a single command and you are ready to work.

2.2 Reproducible science

The AGROMET project is a public one. That means that the CRA-W and all people involved have to be transparent about the work and the results. That will give to the project more credibility and more reliability.

Transparency is promoted thanks to open science. That means the content and the results of the project will be accessible to others. Indeed, transparency is superior to trust and is an ideal (Munafo, 2017).

In the case of the project, development represents its major part. Today, open science is widely used and tools have been developed for that. That insures to all codes written for the project to be available. As a consequence, anyone will be able to check the code and inspect it. Then, some people can improve codes and increase efficiency of work.

Moreover, this transparency insures that the models built will be completely explicit for people who will use them. That is a proof of the quality of models.

Chapter 3

Data acquisition and preparation

Variables have to be identified to build models. To do that, we need response variables, i.e. variables to predict, and explanatory variables, i.e. variables on which response variables depend.

3.1 Interest variables

The AGROMET project will provide information about weather parameters which are important for some crop diseases. These parameters are temperature, relative humidity, leaves wetness and rainfall. The last one is retrieved from a Dutch company. The others are measured by weather stations from PAMESEB network, data are stored on an API. This API is an intermediary software where we can make requests to get some data. Data extracted from the API need to be transformed to be more manipulable. Functions were wrote to transform these data. For example, convert measures from character to numeric.

3.2 Explanatory variables

A little reminder : I will use multiple linear regression. That means I want to find an equation where a response variable can be modeled from two or more variables. The equation will have the form : $Y = b_0 + b_1.X_1 + b_2.X_2 + ... + b_n.X_n$ where Y is the response variable and X_n your n explanatory variables related to their estimated parameter b_n.

These explanatory variables have been identified from academic papers (Zeuner 2007, Janssen 2011). Two types of explanatory variables can be discriminate : static variables and dynamic variables.

3.2.1 Static variables

Land cover

All PAMESEB weather stations are sited in agricultural or herbaceous areas. That is a way to reduce errors about measures. However, the environment of each station can be different and can have an impact on measures. For example, a station could have a different behaviour if a forest is near its area or if an artificial surface (road, construction) is near it.

CORINE land cover is an inventory updated every 6 years by **Copernicus**, the European Union's Earth Observation Programme. These data can also be found on the **Belgian geo-portal**. CORINE Land Cover has been already used to make a spatial interpolation of air pollution (Janssen *et al.* 2011).

CORINE Land Cover is divided in 47 different land covers. 26 of them are in Wallonia. However, 26 land covers is too much to be integrated as explanatory variables. These land covers have been grouped in 5 classes we juged relevant :

- **Agricultural areas** : areas where crops can be tall
- **Herbaceous vegetation** : cleared areas like pastures and grasslands
- **Artificial areas** : roads, rails and constructions where anthropogenic material can impact temperature
- **Forest** : large areas providing shadow and cold
- **Water bodies** : areas like river, lake, wetlands and bogs. Finally, this class has been removed because of the fact that no stations are located near a water body

After a long data preparation with transformation of CRS from WGS84 to Belgian Lambert 2008 and conversion in different forms of Spatial objects (Vector and Raster), data were completely manageable.

Finally, data are recovered at stations positions with buffers. These buffers have a radius of 100 meters for physical stations and 500 meters for virtual stations (because each station covers 1 km^2). The Table 3.1 below shows the structure of the data frame where each station identified by an ID has the percentage of cover for each class.

Buffers of 500 meters radius and grid cells of 1km^2 were compared but buffers were chosen because they have more relevance when they are compared to physical stations where only buffers were computed.

Digital Terrain Model

In the same way as land cover, the terrain model could have an impact on temperature of the environment. These variables have been integrated in the models made by Zeuner *et al.* (2007) and the relevance has been demonstrated several times.

Elevation data have been recovered for Wallonia from **NASA's SRTM** providing a high-resolution (90 meters) topographic data. Then, slope, aspect and roughess of terrain have been calculated with spatial libraries from R. These data are very large and data processing is very long because resolution is high.

Table 3.1: Distribution of land covers around physical stations

sid	crops	artificial	forest	herbaceous
1	63.85818	4.265755	0.00000	31.83038
4	64.11932	35.834990	0.00000	0.00000
7	75.28137	0.000000	0.00000	24.67295
9	99.95431	0.000000	0.00000	0.00000
10	69.91902	0.000000	30.03530	0.00000
13	89.50912	0.000000	10.44519	0.00000

3.2.2 Dynamic Variables

Solar irradiance

In the same way as temperature is a dynamic variable, explanatory variables can be dynamic. In the case of temperature, we can be interested in solar irradiance. Indeed, solar irradiance has an impact on climate changes (Dewitte *et al.* 2004).

Data are recovered from **EUMETSAT**, the European Organisation for the Exploitation of Meteorological Satellites. They are produced every 30 minutes and expressed in W/m^2. These data are aggregated in hourly data and they are stored on a API of AGROMET.

Data are available from 2015-11-11. As a consequence, we can not build models from data before this date.

In parallel with that, PAMESEB stations also measure solar irradiance. But only 27 stations are useable.

Temperature forecasts

The AGROMET project is supported by RMI, the Belgian equivalent of Météo France. As a partner, RMI will provide temperature forecasts based on their own algorithms. These data will be integrated as explanatory variables to build models.

At the time of my internship, these data were not available.

3.3 Data organisation

Once all the data are available, an important task is to organize them to realise the modeling. This organisation needs to respond to a methodic approach. Indeed, to reduce computation time, structure of data have to be optimised.

The objective is to build models for each hour and compute its related error.

The first step consists of grouping data. Static and dynamic variables are grouped in a data frame. Then, to reduce time computation and to prepare the integration of the data frame for the modeling, there is a way to nest data frames with the library purrr. In this way, it is possible to have one single row for each hour but every row contains data frames inside. The Figure 3.1 shows how it looks. This nested data

Table 3.2: Example of nested data frame in a row corresponding to 2016-05-19 15:00:00

altitude	slope	aspect	roughness	crops	artificial	forest	herbaceous	ens	tsa	X	Y
473.6300	3.392046	211.6529	12.917668	63.85818	4.265755	0.00000	31.83038	348	12.4	721240.2	568849.6
345.7340	1.908891	162.7127	9.446796	64.11932	35.834990	0.00000	0.00000	389	12.8	714221.2	543453.8
348.8835	2.611751	165.0716	10.269126	75.28137	0.000000	0.00000	24.67295	779	14.1	750500.7	550825.6
497.7260	1.823958	137.0710	8.165029	99.95431	0.000000	0.00000	0.00000	916	13.5	753130.6	581716.9
389.9188	6.510637	310.6517	22.573696	69.91902	0.000000	30.03530	0.00000	916	14.9	734687.9	580969.6
259.7389	1.669001	288.5355	6.330072	89.50912	0.000000	10.44519	0.00000	774	15.2	641664.8	588814.6

frame is a efficient way to manipulate many sub-tables at once.

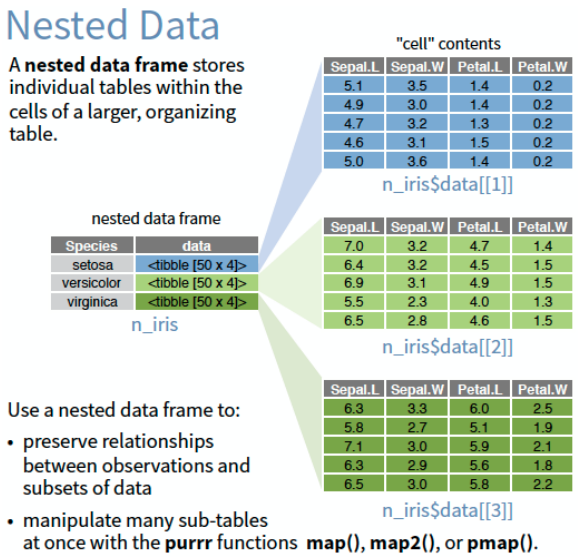


Figure 3.1: Structure of a nested data frame

In the case of the project, the nested data frame contains one row for each hour which has a data frame containing data from each station at this time. In the Table 3.2, there is a preview of the data frame contained into each row.

Chapter 4

Modeling with machine learning methods

Once the dataset is ready, the next step is to model predictions of temperature. To do that, machine learning is used through R.

4.1 Principle of machine learning

4.1.1 Definition

Machine learning is the idea that there are generic algorithms that can tell you something interesting about a set of data without you having to write any custom code specific to the problem. Instead of writing code, you feed data to the generic algorithm and it builds its own logic based on the data. In other words, Machine learning is a subset of deep learning or Artificial Intelligence that provides an ability to "learn" with data.

There are 2 types of machine learning : supervised and unsupervised learning.

4.1.2 Supervised Machine Learning

In practice, most of machine learning uses supervised learning.

From *machinelearningmastery.com* :

> Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output : Y = f(X).
> The goal is to approximate the mapping function so well that when you have new input data (x), you can predict the output variables (Y) for that data.
> It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process

In this learning, the algorithm tries to learn from examples we give to it and then it returns a model of prediction. Classification and regression are supervised machine learning.

This is the case of the AGROMET project where regression models are used.

4.1.3 Unsupervised machine learning

Although that is not the subject, unsupervised learning consists in algorithms which have to find themselves interesting structures in the data. It differs from supervised learning because correct answers are not given to the machine and it has to find answers itself. Association and clustering are unsupervised machine learning.

For example, take 400 pictures of cats and 400 of dogs. While supervised learning will give these pictures with the answer to the machine to make it to find a way to discriminate cats and dogs, unsupervised learning will give only the pictures and the machine will have to find itself differences between cats and dogs separating them in two groups. Obviously, the machine will not know that one group is corresponding to cat and the other to dog because we did not labelled them but the machine will be able to distinguish them like two separate entities.

4.2 Machine learning approach in the AGROMET project

The objective is to predict weather parameters (temperature, relative humidity, leaves wetness). We use data from meteorological stations and from other sources like EUMETSAT for solar irradiance and COPERNICUS for land cover.

Here is our approach :

We choose a weather parameter to predict, temperature for example. It is our **target**.

Then, we define our **explanatory variables**, i.e. the parameters we want to build our model. These variables are our **task** and it contains target data.

In our case, we will use several statistical methods to build our model like kriging, multiple linear regression model or neural networks. Those which are chosen to build a model are called **learners** and will be compared.

To measure performance of our predictions, we need to use a **cross-validation resampling strategy**. Several methods exist but we will use the Leave-One-Out cross-validation method. It consists to establish model based on every samples except one which will be the sample where the model is tested to compute the error and then doing it again as many times as the number of samples. The error measured for our models will be the RMSE and MAE.

The entire methodology is detailed in the *Appendix A* (Spatialization methodology).

4.3 Machine Learning in R

All of our work is done on RStudio. It is a very powerful open-source software for R. Moreover, a R package which provides the infrastructure to run machine learning is available. This package **mlr** is very complete to build models, make predictions and evaluate performances.

Machine learning in R offers a common and simplified interface for all statistical methods implemented in the package. With this package, run a benchmark with several statistical methods can be done on data from a period. This benchmark returns a lot of informations.

It is possible to compare statistical methods through measures of error like RMSE or MAE. Comparing several combinations of explanatory variables is also possible.

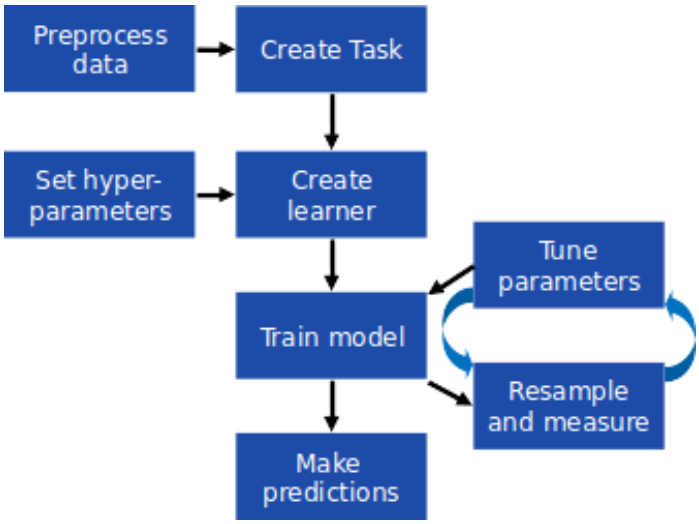


Figure 4.1: mlr workflow

The Figure 4.1 shows the general workflow.

This package has an important community which improves it regularly. A lot of statistical methods are implemented from other packages like *tgp*, *kknn* or *DiceKriging*.

Chapter 5

Results and discussion

5.1 Benchmark

5.1.1 Methodology

To realize the benchmark, data from 2015-11-11 00:00:00 to 2018-06-30 00:00:00 were used. The objective is to run a benchmark on 5 years of data but at this moment, solar irradiance data from EUMETSAT were not available before this date. Here, the dataset has 23089 hours.

The learners were defined with filter methods, i.e. the same statistical method was applied to different combinations of explanatory variables. For my internship, I only used **Multiple Linear Regression**.

The Table 5.1 shows the different combinations used and compared.

Every computation is done for each hour. As a consequence, the combination of explanatory variables is not unique for each hour but depends on a condition checked every time.

Performances were measured with a Leave-One-Out cross-validation resampling strategy.

The benchmark took about 30 hours, i.e. 3 hours per method. Computations are very long and results are very large. Each method represents more than 1 Gigabyte of data.

Table 5.1: Combination of explanatory variables used

Statistical Method	**ID**	**Explanatory variables**
Multiple Linear Regression	lm.Long.Lat	Longitude & Latitude
Multiple Linear Regression	lm.Long.Lat.Elev	Longitude & Latitude & Elevation
Multiple Linear Regression	lm.SolIrr+1bestVar	Solar Irradiance & best variable based on an hourly linear correlation computation
Multiple Linear Regression	lm.SolIrr+2bestsVar	Solar Irradiance & 2 bests variables based on an hourly linear correlation computation
Multiple Linear Regression	lm.SolIrr+3bestsVar	Solar Irradiance & 3 bests variables based on an hourly linear correlation computation
Multiple Linear Regression	lm.2bestsVar	2 bests variables based on linear correlation computation for every hour
Multiple Linear Regression	lm.3bestsVar	3 bests variables based on linear correlation computation for every hour
Multiple Linear Regression	lm.4bestsVar	4 bests variables based on linear correlation computation for every hour
Multiple Linear Regression	lm.Vars.r>0,5	Variables with a linear correlation greater than 0.5
Multiple Linear Regression	lm.Vars.r>0,3	Variables with a linear correlation greater than 0.3

5.1.2 Comparison of methods

Once benchmark results are available, comparison of methods is possible. This comparison is based on the error of measures. In our case, RMSE and MAE were computed.

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_j - \widehat{y}_j|$$

RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It is the square root of the average of squared differences between prediction and actual observation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_j - \widehat{y}_j)^2}$$

They both express average model prediction error in units of the variable of interest. Both metrics can range from 0 to ∞ and are indifferent to the direction of errors. They are negatively-oriented scores, which means lower values are better.

Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is more useful because large errors are particularly undesirable in the project. (Chai, 2014)

From the 10 methods compared, MAE and RMSE are computed and compared. The results are shown in the Figure 5.1. In this case, they both have the same behaviour. Multiple linear regression using coordinates to build models has a large error, the model is too simplified. Multiple linear regression using explanatory variables whose their linear correlations with temperature is greater than 0.3 has an error larger than the other methods too, this filter method is too flexible to return a valid model. A few methods have a similar error. In particular, that is the case when too many variables are chosen to build models.

On the three best methods, one is better than the others. That is the model built from an equation using longitude, latitude and altitude as explanatory variables. Tests realized on two months of data have already shown that altitude is a powerful explanatory variable. The two other methods are based on the computation of the linear correlations with temperature, with or without solar irradiance as mandatory variable have a similar error. However, they are more interesting because the equation is dynamic throughout hours and, in this way, the model is adapted to the evaluated hour.

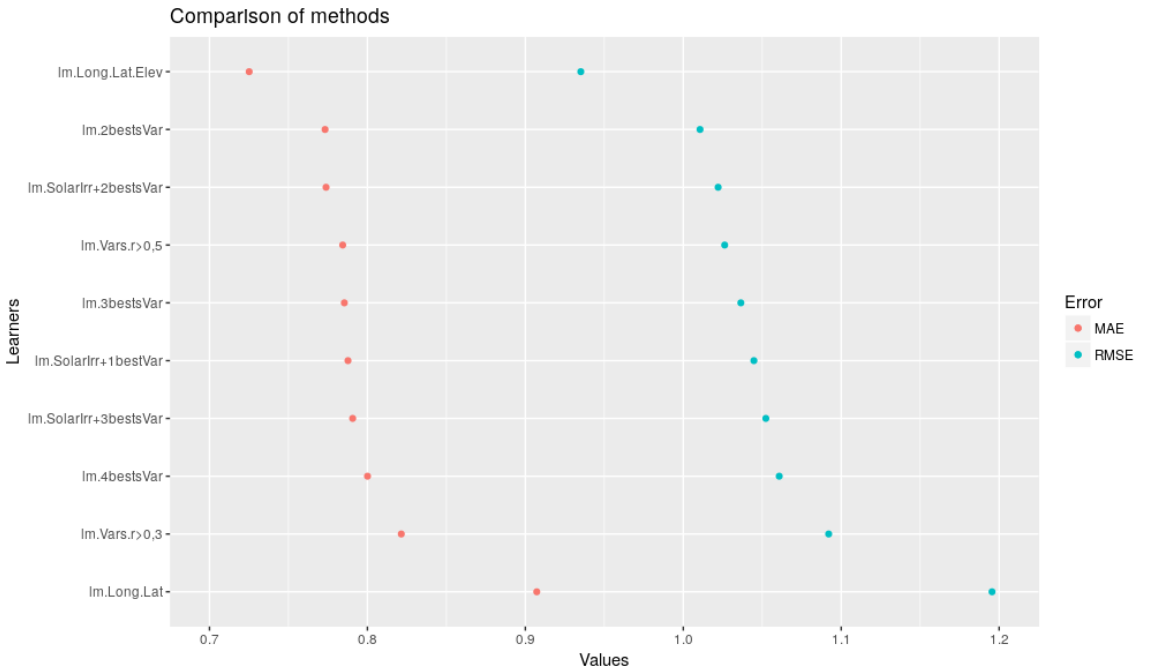


Figure 5.1: Errors (RMSE and MAE) of methods

Errors are between 0.72 and 0.91 for MAE and between 0.93 and 1.20 for RMSE. These errors should be near zero. Both of MAE and RMSE are expressed in degrees such as temperature. An error of 1 degree is relatively important and has to be taken into consideration.

Performances of methods can be compared computing their rank for each hour. The Figure 5.2 compares five of the bests methods :

- the 2 variables with the best linear correlation with temperature
- longitude, latitude and elevation
- solar irradiance and the 2 variables with the best linear correlation with temperature

This barchart corroborates the precedent graph. The method based on coordinates and elevation is widely better than the two others which are more similar but with a relevant difference.

Table 5.2: Models with their equations

	Datetime	Equation	BestVar1	BestVar2	RMSE	MAE
16135	2017-09-13 06:00:00	T = 14.656695 + -0.008887.altitude + 1e-06.X	altitude	X	0.5674956	0.4107285
16136	2017-09-13 07:00:00	T = 14.722147 + -0.00574.altitude + 0.012719.ens	altitude	ens	0.6122121	0.4529545
16137	2017-09-13 08:00:00	T = 15.676696 + -0.004304.altitude + -0.011732.herbaceous	altitude	herbaceous	0.9375293	0.7287999
16138	2017-09-13 09:00:00	T = 13.701668 + -0.012835.herbaceous + 0.00415.ens	herbaceous	ens	1.2198397	0.9671947
16139	2017-09-13 10:00:00	T = 14.987154 + -0.001473.altitude + -0.009431.herbaceous	altitude	herbaceous	1.0206276	0.8224733
16140	2017-09-13 11:00:00	T = 14.131936 + -0.003737.altitude + 0.005707.ens	altitude	ens	0.8312792	0.5979420
16141	2017-09-13 12:00:00	T = 11.852128 + -0.002653.altitude + 0.012421.ens	altitude	ens	0.9727891	0.7488237
16142	2017-09-13 13:00:00	T = 22.557946 + 0.003638.ens + -1.4e-05.Y	ens	Y	1.2605875	0.9444467
16143	2017-09-13 14:00:00	T = 15.441184 + 0.000484.altitude + -0.015099.herbaceous	altitude	herbaceous	1.0796645	0.8760026
16144	2017-09-13 15:00:00	T = 10.730857 + -0.003947.altitude + 9e-06.Y	altitude	Y	0.9526830	0.7777128

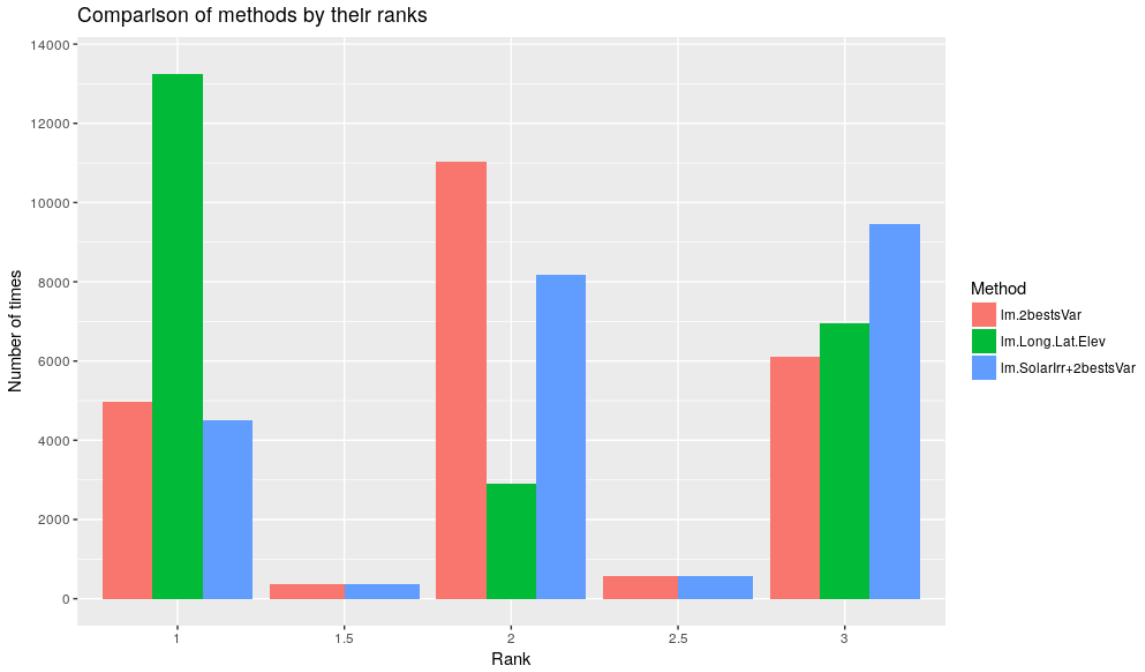


Figure 5.2: Comparison of methods by rank

For each hour, the equation of the model is computed. They can be extracted from the benchmark. The Table 5.2 shows the structure of the equations in the case where explanatory variables can be different from one hour to another (lm.2bestsVar). The table also shows what are the bests variables and display the related error.

5.1.3 Visualization

These models can be observed on maps. For that purpose, functions building maps have been made with ggplot2 library from R for static maps and leaflet library for interactive maps.

Models built from physical stations data are applied to the 1 km^2 grid cells. Then, the temperature is mapped with a color palette similar to the one of RMI. Class breaks are based on quantiles of temperature values. Standard error is computed for each cell and it is shown on the map with a white layer which has different levels of transparency according to the error. A large standard error is related to an opacity

and vice-versa.

The Figure 5.3 shows an output for one hour based on the method where explanatory variables are Solar irradiance and the 2 variables with the best linear correlation with temperature. To build this map, some objects are needed : an object containing data (temperature and standard error) for the grid and a spatial vector object containing boundaries of Wallonia, but also the name of the variable to display. Then, some conditions can be chosen, like the display of the layer containing error, the display of the legend for error, the way to build the legend and its classes. Some arguments enable to customize the map with titles and comments. The function is thus reusable for other usages. For example, the function builds maps spatializing hydric deficit in Wallonia.

The figure has 2 maps. The one on the left is the map with spatialization of temperature, the right one has the layer with error on the temperature layer. It is possible to see regions where errors are more important.

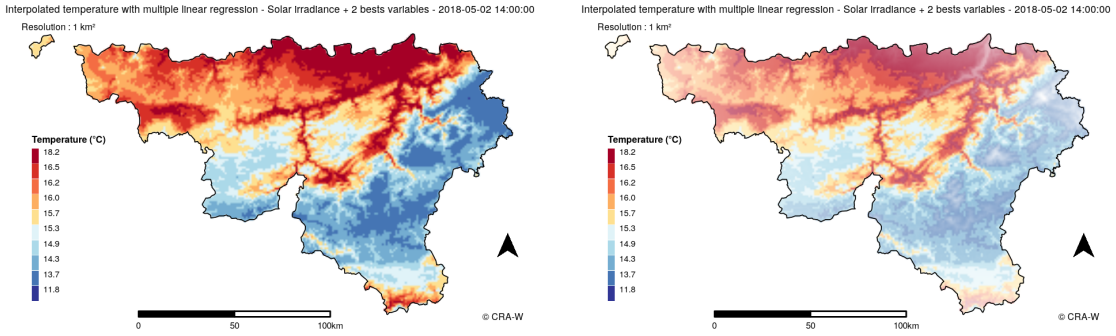


Figure 5.3: Example of an output for 2018-05-02 14:00:00 (left : without error ; right : with error)

This figure shows the output based on a model depending on the method where explanatory variables are Solar irradiance and the 2 variables with the best linear correlation with temperature, the equation of the model is the following one :

$$T = 15.59716 + -0.00629 \times Elevation + -0.00197 \times Herbaceous + 0.00206 \times SolarIrradiance$$

These maps show whether the models are relevant. The *Appendix B* shows maps made with all methods for one hour. These maps show differences in the relevance

of each model. For example, the model depending on longitude and latitude is very simplistic compared to the others. The other models are more similar but show that some of them are more reliable because the error is smaller. The better model is the second one on the first row. It is corresponding to the the model depending on longitude, latitude and elevation. For this hour, the better model is the following :

$$T = -9.375042 + -0.010431 \times Elevation + 1.02e{-}05 \times Longitude + 1.59e{-}05 \times Latitude$$

Longitude and latitude are expressed in meters because the CRS used is Belgian Lambert 2008. That is why their coefficients are about 1e-05. For information, values of the coordinates are about 6e06.

5.2 Discussion

As a reminder, the objective of the project is to provide weather predictions. These predictions will feed decision support tools to monitor crop diseases like potato late blight and to operate in fields at the right times.

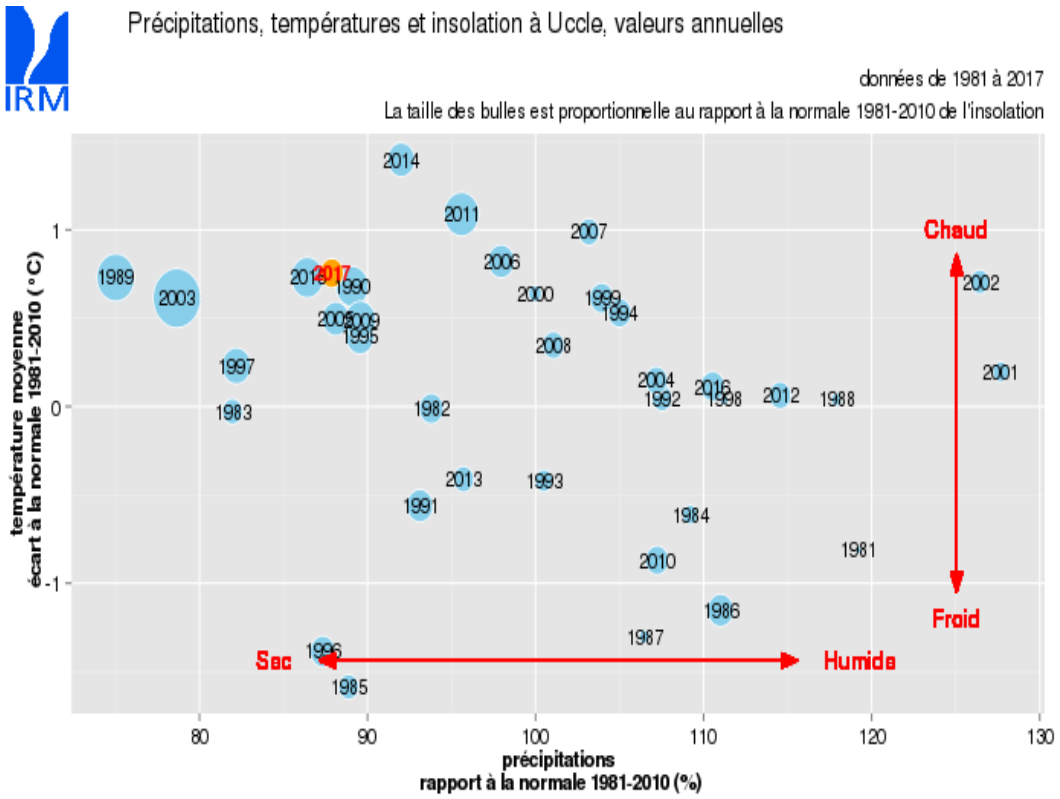


Figure 5.4: Precipitations, temperatures and insolation, annual values

Results have some limits that have to be discussed. The first limit is the period used to build models. Indeed, models were built with data from 2 and a half years. This period could be too short to be relevant. Moreover, this period is not necessarily

representative of a mean period. According to RMI, 2015, 2017 and 2018 are hot and dry years compared to normal year, 2016 were a wet year. This is shown on the Figure 5.4.

Models were built with some explanatory variables but they may be insufficient. Adding new variables like temperature predictions from RMI could improve models. Only 27 stations are used to build these models, adding RMI stations network could be a solution to improve models too. However, weather stations from different networks have differences in their measures, checking interoperability is very important for that and making corrections is essential.

Multiple linear regression is the only statistical method used to build models, but going forward, other methods will be compared like ANN and different kriging methods. The major constraint will be time computation which is relatively long. There is a lot of possible combinations of explanatory variables to compare, choices must be done because of the time computation. These choices sometimes can be subjective and not based on scientific literature.

Beyond these limits, other points can be discussed. For example, solar irradiance data are provided by EUMETSAT and by PAMESEB stations. These data have to be compared because both sources are used. Data from stations are used to build models, data from EUMETSAT for spatialization. The comparison shows a correlation around 0.95. As a consequence, there is nothing wrong with it.

With the aim to provide data for agronomic utilisation, there is an interest to compare mean error observed in Wallonia and this error in agricultural areas to be sure of the accuracy of predictions. That has not be done yet.

Assuming that no other models will be better than those presented, there will be a question of transparency for the project. Indeed, as a public project within the scope of public agencies, the work that is done has to be transparent. In that way, models will be clearly presented. If one static model is chosen for predictions, this will be clear and transparent. But if the chosen method is a dynamic one, using computation every hour, the method used for each hour could be different, and that will have to be mentioned.

Conclusion

At the end of this internship, the progress of the project is great. Principal data have been recovered to be integrated as explanatory variables and first models have been built. Now, the routine is ready to build further models. First results are encouraging because models are relatively relevant and they can be improved adding new explanatory variables and using other statistical methods. Indeed, comparing models with a lot of different methods will be possible to find the best method, i.e. the method with the smallest error (RMSE).

The project will end in 2020, that leaves time for building better models and developing decision support tools for crop monitoring. Therupon, the project will have an importance and some foreign organisations are already interested by the project (Germany and Slovenia).

Beyond the technical skills I developed during my internship, as the utilisation of R language to interpolate spatial data and manipulating them, the discovery of developer applications and tools like Docker or Ubuntu, I also developed skills in collaborative work, in particular with GitHub. It was a very enriching experience.

Outside the AGROMET project, I worked for other projects. I contributed to the report on drought in this summer 2018 in Wallonia, upgrading graphics and building a template of the map of Wallonia. I also contributed to provide data from the API of the project to co-workers for other projects.

Appendix A

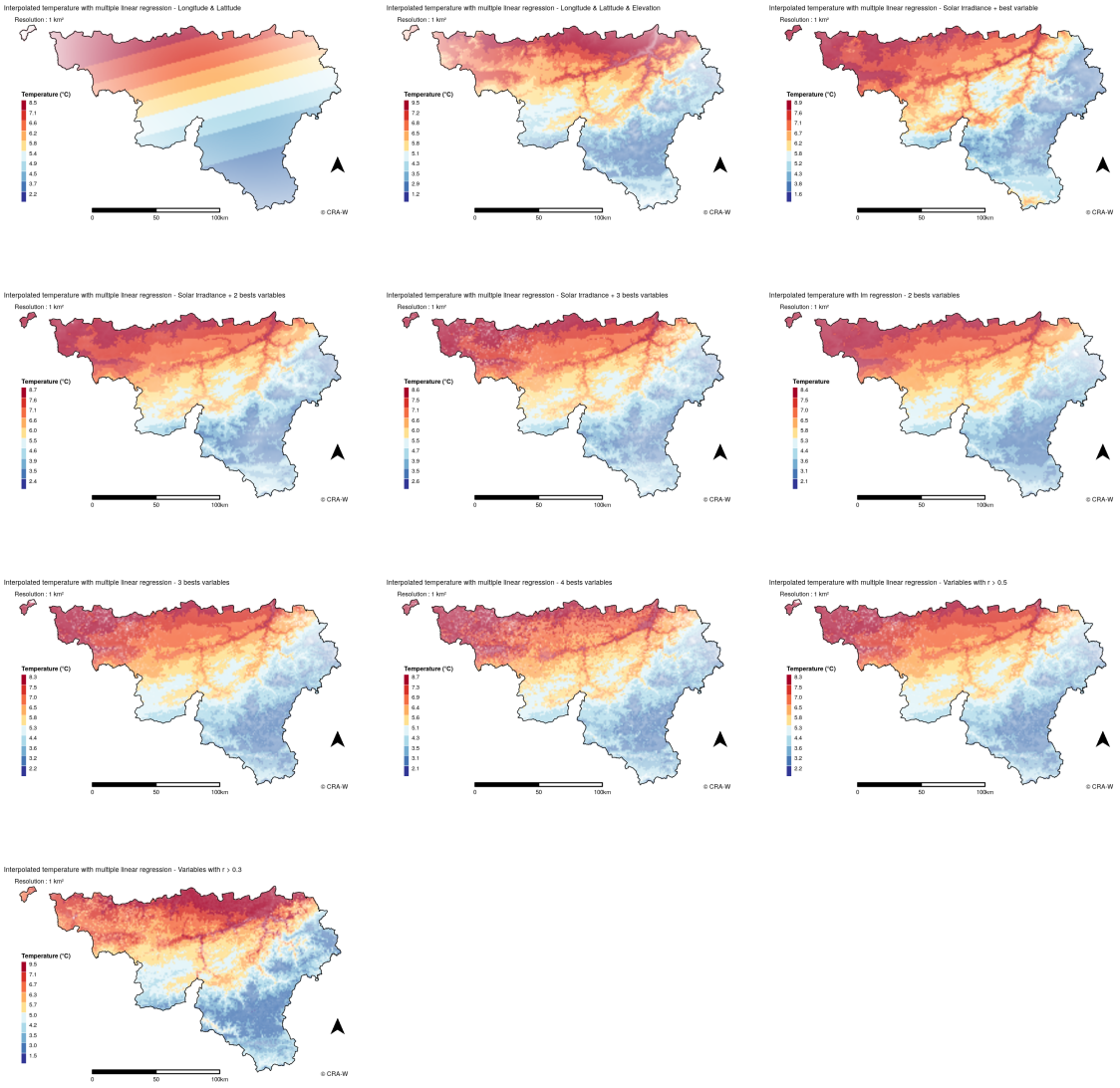
Resources on AGROMET and my work

- European directive 2009/128/CE : https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32009L0128
- Technical details about PAMESEB stations : https://www.pameseb.be/meteo_intro/infos_techniques.html
- Spatialization methodology available here : https://pokyah.github.io/agrometeor-methodo-spatial/assets/uml_images/spatialization-methodology.svg
- All my codes are available on my Github account : https://github.com/ldavadan
- My personal blog, made with Blogdown : ldavadan.github.io
- Contribution to "Crop and Grassland conditions in early August" report : http://www.cra.wallonie.be/fr/etat-des-cultures-et-des-prairies-en-ce-debut-du-mois-daout-2018

Appendix B

Outputs with different methods

Methods from left to right and top to bottom follow the order presented in Table 5.1. Models built for 2018-03-07 14:00:00.



Appendix C

Additional resources

- Ubuntu GNOME : https://ubuntugnome.org/
- ANSIBLE : https://www.ansible.com/
- SSH : https://www.ssh.com/
- GitHub : https://github.com/
- API : https://medium.freecodecamp.org/what-is-an-api-in-english-please-b880a3214a82
- Docker : https://www.docker.com/what-docker
- Copernicus : https://land.copernicus.eu/pan-european/corine-land-cover/view
- Belgian Geoportal : https://www.geo.be/#!/catalog/details/bcd19aa9-c320-4116-971b-6e4376137f13?l=en
- NASA's SRTM : https://lta.cr.usgs.gov/SRTM
- EUMETSAT : https://landsaf.ipma.pt/en/products/longwave-shortwave-radiation/dssf/
- Machine Learning Mastery Blog : https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
- Machine Learning in R : http://mlr-org.github.io/mlr/index.html

References

Chai, T., & Draxler, R. (2014). Root mean square error (rmse) or mean absolute error (mae)? – Arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 1247–1250.

Dewitte, S., & others. (2004). Measurement and uncertainty of the long-term total solar irradiance trend. *Solar Physics*, 209–216.

Hooyberghs, J., & others. (2006). Spatial interpolation of ambient ozone concentrations from sparse monitoring points in belgium. *J. Environ. Monit.*, 1129–1135.

Janssen, S., & others. (2008). Spatial interpolation of air pollution measurements using corine land cover data. *Atmospheric Environment, Volume 42, Issue 20*, 4884–4903.

Munafo, M., & others. (2017). A manifesto for reproducible science. *Nature Human Behaviour Volume 1, Article Number: 0021.*

Racca, P., & others. (2011). Decision support systems in agriculture : Administration of meteorological data, use of geographic information systems (gis) and validation methods in crop protection warning service. *Efficient Decision Support Systems - Practice and Challenges from Current to Future*, 331–354.

Zeuner, T., & Kleinhenz, B. (2007). Use of geographic information systems in warning services for late blight. *Bulletin OEPP/EPPO*, 327–334.