MATHEMATICAL MODEL TO ESTIMATE ENERGY CONSUMPTION OF THE BUILDINGS

AT THE SCALE OF DISTRICT

MACHINE LEARNING AND DATA MINING MASTER

HIBA ALQASIR

SUPERVISERS

ECOLE DES MINES:

ASSISTANT PROFESSOR JONATHAN VILLOT

UNIVERSITÉ JEAN MONNET:

PROFESSOR MARC SEBBAN

23 JUIN 2016

ABSTRACT

An understanding of the energy performance in buildings in an entire municipality or an entire district is important for sustainable energy planning strategies that accelerate the energy renovation process in existing buildings that are not energy efficient. Effective policies and incentive schemes to reduce the climate change footprint of buildings require a solid understanding about the current buildings.

In this respect, this report wants to reconstruct the energy consumption Information in buildings based on information that is already available on the web. There are numerous different approaches and methods which can be applied to reconstruct problems. Which one to select is a difficult question, the set of candidate methods is huge. We have chosen to select a set of promising candidate algorithms, which are most appropriate for our dataset Regression, Kriging and Missforest.

The three techniques implement very different approaches to reconstruct problem. In a general manner the results are quite good for the MissForest but not so good for Kriging and for the linear regression. The methodologies described in this report were tested in a medium sized town (Saint-Etienne) in France, and it can be potentially applied in the future to any French city.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Assistant-Professor Jonathan Villot, for giving me the opportunity to develop this Master's Degree internship at "École des Mines de Saint-Étienne", in order to complete the degree of Master in Machine Learning and Data Mining, and for his gentle guidance and invaluable advice.

Special Thanks to Professor Marc Sebban, the coordinator of my Master's program, without him I wouldn't have been able to come to France, I thank him for the time to answer my emails as quickly as possible and to always help me when needed.

Thanks to the people in "École des Mines", and my classmates in the Master's program at "Universté Jean Monnet" for many unforgettable moments in Saint-Étienne.

Lastly and especially, a big "thank you" to my soulmate Alaa Daoud, for his constant support and unwavering encouragement, I am very blessed to have you in my life, thanks for believing in me, loving me unconditionally, and supporting me through all my endeavors.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1. BACKGROUND

Buildings are at the pivotal center of our lives. The characteristics of a building, its design, its look and feel, and its technical standards not only influence our productivity, our well-being, our moods and our interactions with others, they also define how much energy is consumed in and by a building, and how much heating, ventilation and cooling energy is needed to create a pleasant environment.

We know that buildings cause a significant amount of greenhouse gas emissions, mainly CO_2, altering our planet's climate. By renovating buildings to high standards of efficiency we can demonstrate that ambitious climate change mitigation actions and improvements in living quality can go hand in hand.

Yet, the energy performance of our buildings is generally so poor, that the levels of energy consumed in buildings place the sector among the most significant CO_2 emissions sources in Europe [1]. While new buildings can be constructed with high performance levels, the older buildings which are predominantly of low energy performance are in need of renovation work.

1.2. CONTEXT OF THE PROJECT

The building sector and in particular the old buildings that are known for more than 35 years, is the first French energy consumer sector, and second emitter of CO_2, and it is one of the focus areas within the framework of political action at the territorial and national levels.

THE RENOVATION IN TERMS OF ENERGY EFFICIENCY OF OLD BUILDINGS

France has 31 million residential buildings covering an area of more than two billion square meters. Commercial buildings account for more than 900 million square meters. 20 million dwellings were built before the first thermal regulations were introduced in 1975. Highly demanding in energy, these dwellings represent 58% of the housing sector and account for more than 75% of its energy consumption [2]. Their renovation has therefore become a priority.

Today regulatory changes aimed at improving its structural features and systems and theoretical renovation targets are known (700,000 renovations per year), so we should be able to designate a precise and pragmatic strategic plan which will present for a territory the buildings with the highest energy renovation potential. To do this, it is necessary to evaluate and understand energy consumption of buildings at minimum on a specific area. Nevertheless, efforts and support measures granted seem limited and face the complexity of a heterogeneous and multi-stakeholder system.

Recognizing this, many studies appeal to technical and social fields tend to show the existence of brakes impacting the development of energy rehabilitation process. Among these limitations, the existence of a "gap" between theoretical objectives and practical and pragmatic reality is identified.

In this context, the project ADEMOPE (Aide à la Décision pour l'Evaluation Multi-échelles et Open data du Potentiel Energétique d'un territoire) aims to provide limited partners (local and/or its agents) a support tool for the multi-scale and open data of the energy potential of their territory for decision optimization of public policy taking into account local constraints (Constraints actors and local context), in a perspective to reduce impacts (environmental, social, economic) of the building stock.

1.3. OBJECTIVE OF THE INTERNSHIP

The objective of the internship is to develop a computer model to capture the energy consumption of buildings on a large scale, using the data present on the web. This include a study of computer tools to capture, process and gather targeted information, to provide a relevant model to meet the objectives and provide a database compatible with GIS/WebGIS formats to represent the energy cadastre of a city.

The selected models were tested on a specific study area (Saint Etienne district) to assess their limitations. The final model selected will be formalized through a tool reproducibility of the approach and application to other fields or scales of study. In this context, this work falls within the framework of the SADST (Système d'Aide à la Décision des Systèmes Territoriaux) being developed within the research center.

The work program consists of several tasks:

- State of the art and development of a specification;
- Methodological developments;
- Development of a proof of concept tool:
    a. Development of computer models for tracing the information that related to the buildings on the Web.
    b. Application of the model and validation on the study area.
    c. Generalization of the model and integration within the SADST platform.

1.4. STRUCTURE OF THE REPORT

This report includes six chapters in which the conducted work and the results are described.

The State of the art is shown in Chapter 2.

The data is described in Chapter 3, which lists and summaries the two kinds of data, and shows diagrams about them.

Chapter 4 explains the extraction data model, as well as the correlations between different kinds of variables, and describes the methodology for reconstruct energy consumption information.

In Chapter5, the results of the energy consumption obtained from the different approaches are presented, respectively, as well as comparing the models.

Finally, in Chapter 6, the conclusions and further prospects are summarized.

2 STATE OF THE ART

2.1. ENERGY PERFORMANCE CERTIFICATES

The main objective of the energy certification of buildings, which was introduced in the European Union by EPBD – Energy Performance Buildings Directive [3], is to provide clear guidelines for energy performance of buildings to improve the energetic quality of new buildings and existing building stocks.

In France the certificate is DPE which refers to (Diagnostic Energétique ou Diagnostic de Performance Energétique). And it is made on property for the diagnosis of energy or energy performance, it is integrated into the technical diagnostics record (Dossier de Diagnostics Techniques <DDT>)[4]. The main objective of the energy diagnosis is to inform the buyer on the estimated consumption of his future accommodation.

The Energy Diagnostics in practice

The energy performance diagnosis determines the energy consumption for heating, hot water and the cooling system of the building, but not on other uses (lighting, appliances, ventilation, etc.).

Reading the diagnosis will be facilitated by the use of the double label [5]:

1. A label for energy consumption per class, to represent the annual amount of energy consumed or estimated (primary energy) shown in Figure1. The measuring index is the [kWhep/m^2.year].

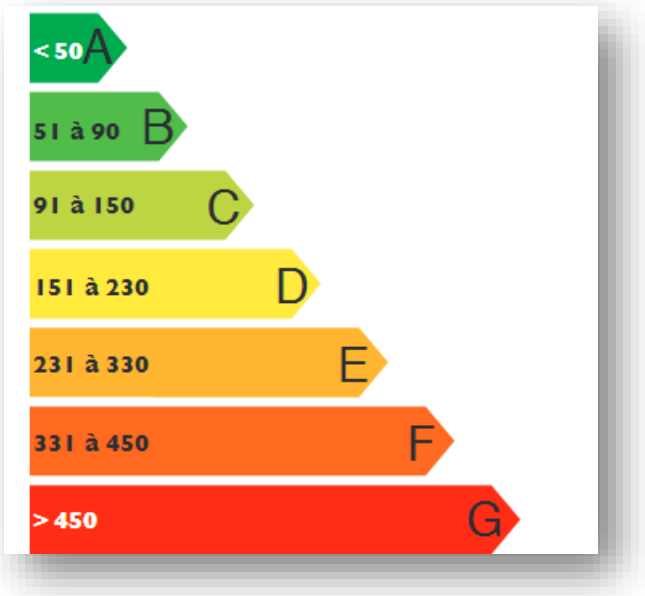


Figure 1 DPE energy consumption per class

2. A label for consumption per year of carbon dioxide emissions. The amount of emissions of greenhouse gases (Les émissions de Gaz à Effet de Serre <GES>) is shown in Figure2. The measurement index is expressed in term of [Kgeq CO_2 /m². year].

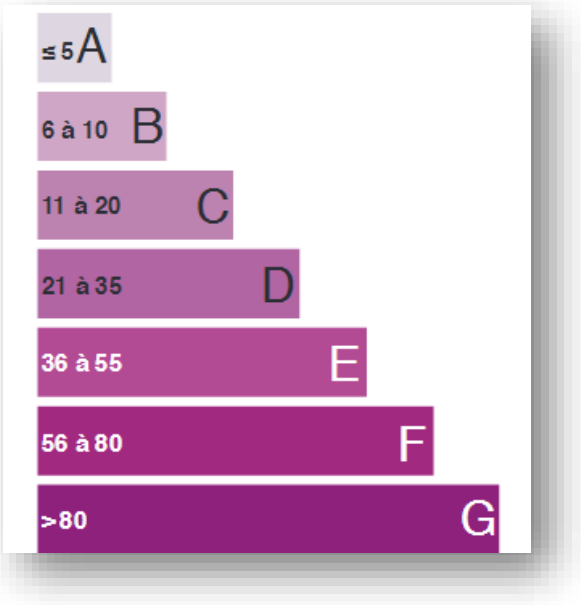


Figure 2 GES the amount of emissions of greenhouse gases

2.2. THE GREEN VALUE OF DWELLINGS IN 2014

The green value is defined in [6] as the increase in economic value generated by the improved energy and environmental performance of a property over another, all their characteristics being equal. It is expressed in terms of "market value". The energy and environmental performance is measured in this study by the mere energy performance (DPE).

The result of the energy performance diagnosis comes in the form of two labels: energy performance label and climate performance label. The energy label class for housing from A to G in order of decreasing energy consumption; on the climate class label the greener housing is (label A) and most polluting is (label G) in terms of greenhouse gas emissions (GES).

This study presents estimates of the impact of the energy label on housing prices in 2014 alone according to climatic zones and new administrative regions into force on 1 January 2016, and a first reflection on the possibility to include, alongside the energy label. Overall, whether apartments or houses, the french southern regions are characterized by better energy performance than the North, which is not surprising in view of the influence of climate on the value of labels (Figure3 and Figure4).

Using a spatial econometric model allowed to consider the existence of a spatial correlation of real estate data, which, in theory, limit potential biases and almost led to the same estimates more precisely. As a result from this study, we found that there are certain property attributes that are important determinants of price; size, location and type of dwelling being the more obvious ones. We try to determine the factors affecting the price and its relation to energy consumption using this study, to see if there was any link between them and how we can benefit from it, a quick glance at summary statistics reveals that DPE ratings are correlated with several other attributes, notably age. But although growing in importance, energy efficiency was a weaker determinant in comparison.

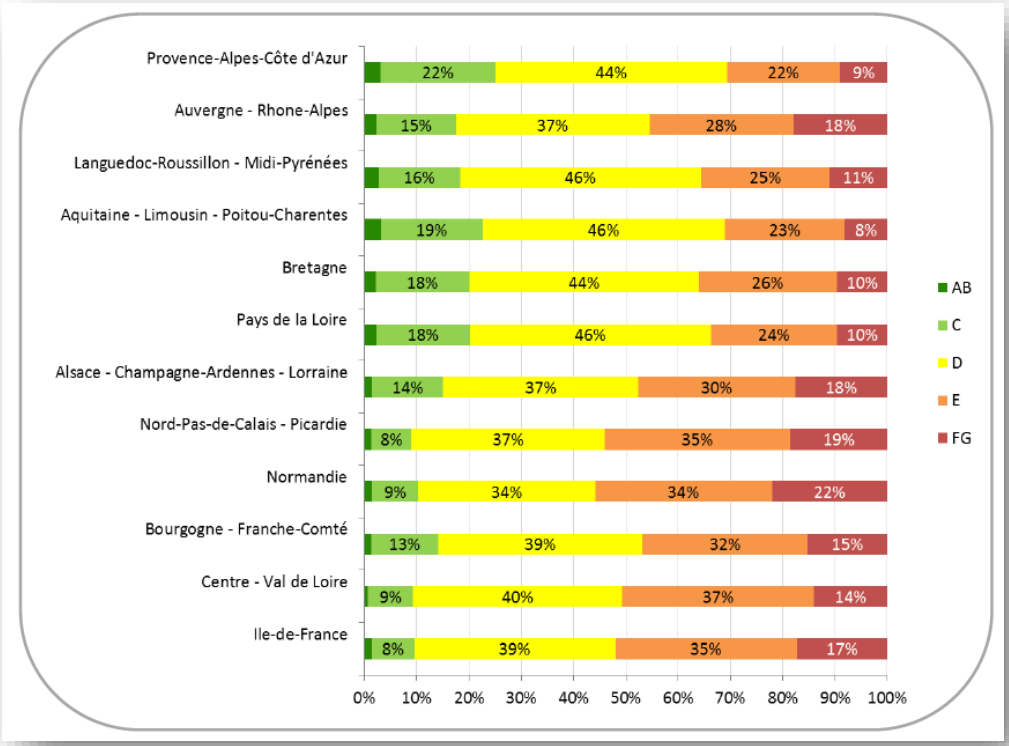


Figure 3 Distribution of energy label by region apartments

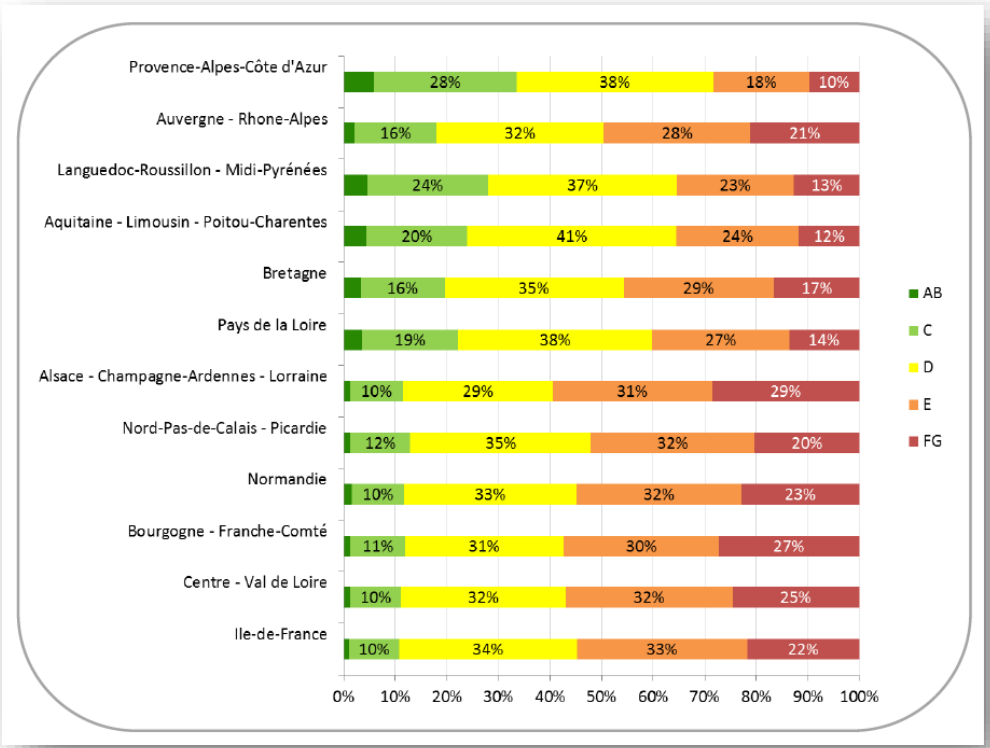


Figure 4 Distribution of the energy label for houses in region

2.3. THE ITALIAN CHALLENGE

Because of the need to widespread retrofit strategy to drastically decrease the energy demand in the existing buildings, Paola Caputo and Giulia Pasetti [7] [8] analyzed the composition of the Italian municipalities and the main features of their building stock. The aim of this research was to create a methodology that leads to quickly implement, and fine tuning later, an energy model and tools for all Italian municipalities.

They investigated the most important barriers that slow down or block local energy refurbishment with particular regard to small and medium Italian municipalities. They checked their approach also by sending a questionnaire to a sample of these municipalities. They suggest how to overcome the barriers related to the first phases of effective energy plans providing a scheme for data collection and suggesting a new Municipal Energy Model. Further, they proposed a reform of the municipal technical office in order to improve the technical competence about energy issues related to the built environment.

There are several methodologies to model energy consumption in the residential sector for space heating and cooling, domestic hot water, and appliances and lighting, as reviewed by Swan and Ugursal (2009). The aim of these methodologies is to evaluate the energy performance of a building stock and assess the potential energy saving using retrofitting measures. This potential saving should be considered as the maximum technical potential, since further analysis are needed in order to clarify how this potential could be achieved and to identify a robust approach to implementing retrofitting measures. The research here presented analyzes the reasons of the inertia of building energy retrofit at municipal level starting from the analysis of what hinders effective energy policies carried out by the public administration and what hinders the involvement of citizens and other stakeholders.

The results of the questionnaire confirm that, in small and medium Italian municipalities, there are several barriers to develop effective municipal energy plans.

It is possible to get a rough estimate of the actual energy consumption and of the energy savings potential after a standard or advanced refurbishment, as shown in Figure5. Moreover, it is possible to get estimate about energy cost per building and per dwelling, as shown in Figure6. The energy cost per building is useful for energy measures carried on the whole building (e.g. thermal insulation of external walls), while the energy cost per dwelling is useful for energy measures carried on the single dwelling (e.g. windows replacement).

The application to a case study demonstrates that is possible to develop the proposed tools in medium municipality, also due to the hybrid approach that includes the use of estimate data.

Thanks to the good data availability of Lombardy Region and also to the GIS plugins developed ad hoc, it is already possible to implement quickly MEM and Energy Scout for Lombardy municipalities.

These are just some of the possible outputs that is possible to obtain through MEM and Energy Scout. Using GIS capabilities more advanced queries and analyses are possible.

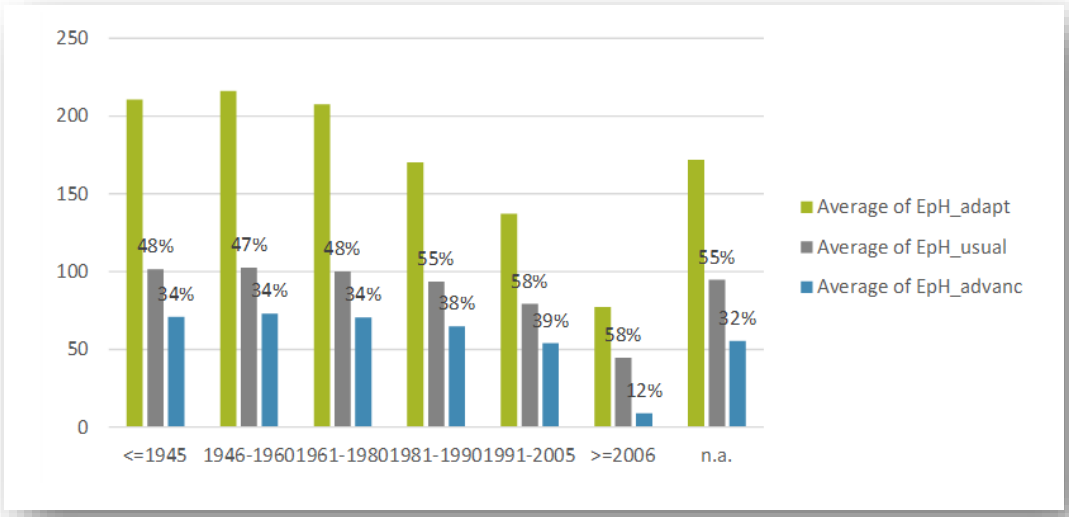


Figure 5 Average Eph_adapt, Eph_usual and Eph advanced [kWh/m²year] for residential building

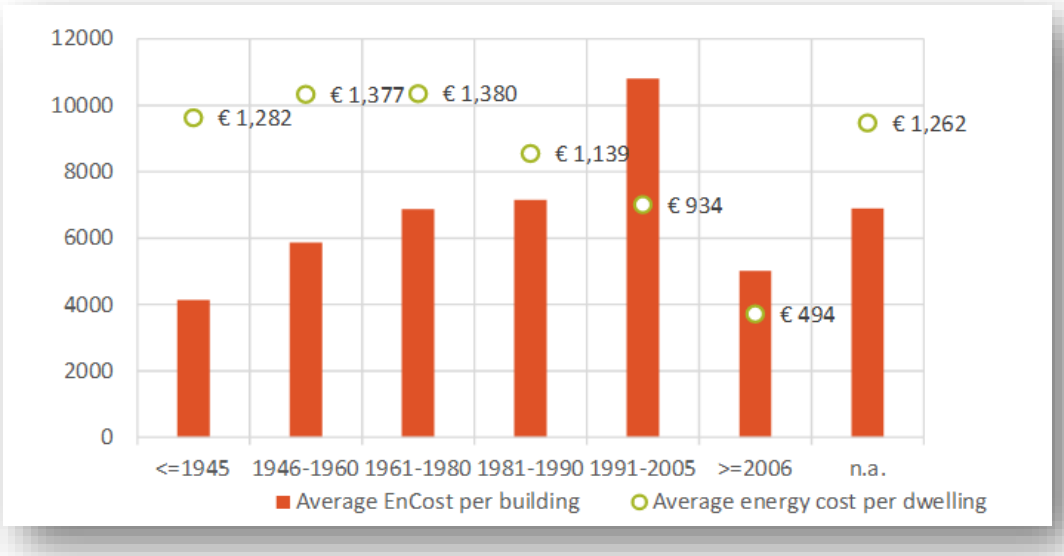


Figure 6 Average energy cost per building and per dwelling

From the technical literature and from our survey we can state that among the local administrations there is a diffuse interest about the energy problem, but there is no real awareness of the possible active role of their own municipality. The data collection seems the most engaging barrier because of the dispersion of data among several municipal offices and other administrative entities and the lack of interoperability among the data collection systems. The need to make data of building stock sharable, accessible and compatible for different scopes and by different offices is very urgent.

This barrier can be overcome with a concerted effort at the national and local levels, for our project we're not in this regard, and we found that the data collection from the Web would be a way faster and provide good data in shorter time.

3 DATA DESCRIPTION

In order to obtain insight into the data, exploratory data analysis was conducted by use of tables and summary statistics.

The housing Information that we have collected are divided into two main sections:

1. Houses;
2. Apartments.

The data provided consisted of 17 variables with 2718 observations from the web, Table1 shows a short description to each variable, its type and its measuring unit.

#	variable	type	measuring unit	description
1	priceRent	number	€	The amount of rent per month
2	priceBuy	number	€	The total buy price
3	surface	number	square meters	Dwelling surface
4	irisName	string	-	The district name of the dwelling, according to INSEE[1]
5	yearOfConstruction	number	year	The year of construction
6	numOfParts	number	part	The number of parts
7	DPE	character	{A,B,C,D,E,F,G}	Energy consumption per class
8	GES	character	{A,B,C,D,E,F,G}	carbon dioxide emissions per class
9	nDPE	number	kWhep/m2.year	The annual amount of energy consumed
10	nGES	number	Kgeq CO_2/m^2. year	The annual amount of carbon dioxide emissions
11	HeatingSystem	string	{collective, individual}	The type of heating system
12	CentralHeating	boolean	Yes/No	Is the heating system central?
13	PoweredBy	string	{gas, fuel, electric }	The material used in the heating system
14	floor	string	{cellar, middle, other, rdc, top}	The floor the dwelling is located in
15	Type	string	{house, apartment }	The type of the dwelling
16	longitude	number	Degree	The Longitude degree of the dwelling
17	latitude	number	Degree	The latitude degree of the dwelling

Table 1 variables description

The annex shows information regarding the missing values in each variable, summary regarding the quantitative variables, and summary regarding the qualitative variables.

[1] INSEE has developed a division of the territory into homogeneous size mesh called IRIS2000 in 1999; an acronym meaning " Ilots Regroupés pour l'Information Statistique " and referring to the target size of 2,000 inhabitants per unit cell. Since then, the IRIS (appellation which now replaces IRIS2000) is the basic building block for the dissemination of infra-municipal data.

4 METHODOLOGY

The objective of this study is to define a tool that creates a comprehensive database of the energy performance of buildings on an urban setting.

Then to create a model that can reconstruct the energy performance from similar database in the future. There are numerous different approaches and methods which can be applied to reconstruct problems. Which one to select is a difficult question, the set of candidate methods is hug.

We have chosen to select a set of promising candidate algorithms, which are most appropriate for our dataset:

1. Regression;
2. Kriging;
3. Missforest.

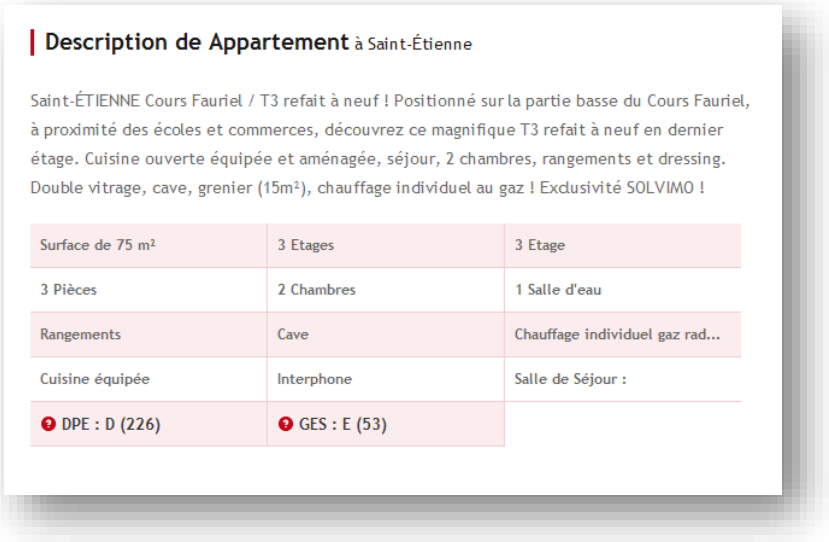
This chapter present and discusses all of them.

4.1. DATA EXTRACTION AND PRE-PROCESSING

4.1.1 Web scraping

The idea here is to extract the housing information from the web, and exactly from housing websites. For this purpose, a tool was developed, this tool takes as input a URL text and the output is a kind of mySQL database, after crawling the website, it parses the html pages in order to record the housing information in an easier way and a more understandable way for the machine.

For example, in Figure7 we can see parts of a page in housing website, it contains important information about one apartment, like the number of parts and rooms, the floor, the surface and of course DPE and GES, in other parts we can find the price and location, so the developed tool catches this information and store it the database.



Description de Appartement à Saint-Étienne

Saint-ÉTIENNE Cours Fauriel / T3 refait à neuf ! Positionné sur la partie basse du Cours Fauriel, à proximité des écoles et commerces, découvrez ce magnifique T3 refait à neuf en dernier étage. Cuisine ouverte équipée et aménagée, séjour, 2 chambres, rangements et dressing. Double vitrage, cave, grenier (15m²), chauffage individuel au gaz ! Exclusivité SOLVIMO !

Surface de 75 m²	3 Etages	3 Etage
3 Pièces	2 Chambres	1 Salle d'eau
Rangements	Cave	Chauffage individuel gaz rad...
Cuisine équipée	Interphone	Salle de Séjour :
DPE : D (226)	GES : E (53)	

Figure 7 Example of housing website

Our web scraping solution is not fully automated system, and it is not able to convert entire web sites into structured information, it is ad-hoc to specific kinds of housing websites structure.

4.1.2 Pre-Processing

The first step in data analysis is to improve data quality.

- ✓ *Noisy Data*

Most parametric statistics, like means, standard deviations, and correlations, and every statistic based on these, are highly sensitive to outliers. And since the assumptions of common statistical procedures, like linear regression, are also based on these statistics, outliers can really mess up the analysis.

In our database we had 2 types of outliers:

1- It is obvious that the outlier is due to incorrectly entered or measured data, in this case we have 2 types:
   - The correction is clear, like 19000 for the year of construction in this case we corrected the value to be 1900.
   - in other hand there are many cases the correction was not that clear, for instance the year of construction is 190, of course it is not correct but we don't know if it is 1900, 1909 or between but because we didn't want to lose the information we just but a random number in the range between 1900 and 1909.

2- The outlier creates a significant association, we replace the abnormal value with NA, so we didn't lose the observation just because it is an outlier.

- ✓ *Data transformation*

At the beginning in the database we had two numeric variables regarding the floor: the total floor number in the building and the floor order.

We converted these two variables into one nominal variable. Because it will be more useful and valuable if we know the dwelling location in the building than just knowing its number.

Also for the dwelling location we had two numeric variables regarding the location: longitude and latitude, and we converted them into "iris name" nominal variable.

4.1.3 Association/Correlation Between variables

The data provided a number of covariates of interest, in order to select variables that were to be included in modeling, backward variable selection was utilized for each set of outcome and associated exposure(s). Qualitative variables pose a challenge when conducting variable selection as levels of the categorical variables are treated as distinct variables. This often un-fortunately leads to some of the levels of the qualitative variable being dropped in the process of variable selection.

Correlation is a bivariate analysis that measures the strengths of association between two variables. In statistics, the value of the correlation coefficient varies between +1 and -1. When the value of the correlation coefficient lies around ± 1, then it is said to be a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker[2].

[2] http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

Care has to be taken however when computing the correlation matrix as the nature of variables has to be taken into consideration, due to that, we have used 3 kinds of tests are presented below:

1. **Poly-choric correlations between ordinal variables**

For ordinal variables, the popular Pearson product-moment correlation is not advised as the resulting correlation is often artificially deflated; instead poly-choric correlations should be computed. The poly-choric correlation coefficient is a measure of association for ordinal variables which rests upon an assumption of an underlying joint continuous distribution [9]. It is computed by assuming that two ordinal variables represent latent continuous normally distributed variables.

2. **Spearman rank correlation between numeric variables**

In cases where the association is non-linear, the relationship can sometimes be transformed into a linear one by using the ranks of the items rather than their actual values. Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. It was developed by Spearman; thus it is called the Spearman rank correlation [10]. Spearman rank correlation test does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.

3. **Pearson test on quantitative and qualitative data**

Correlations between ordinal and continuous variables may be estimated by computing Person correlations. Pearson correlation coefficient does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level [11]. It is computed as follows; the values of the continuous variable are taken as they are whereas the ordinal variable is assumed to represent an underlying latent normally distributed variable which is used in computing the correlation.

The tables in annex2 shows the test's results, analyses that have been performed using R Studio software.

4.2. REGRESSION

One should probably always include the linear regression analysis, as it is sometimes best, and a standard, widely available procedure.

We used the regression model to see how the energy consumption is related to the predictors, to estimate the unknown effect of changing the variables over the energy consumption.

The linear model is described as Equation1:

$$Y = X\beta + e$$

Equation 1The linear model

Where Y is the output vector, X is the input vector, β is the weighted vector (regression coefficient) obtained from training, and e is the mean square error.

When β is multiplied with X, it will yield output Y' and it will not be the same as Y. The difference in Y and Y' is the mean square error e. Now we have β and e values and these values are then used in Equation2, with a new set of inputs X to obtain the interpolation values Y. This is simple regression analysis [12].

$$\beta = (X^T X)^{-1} X^T Y$$

Equation 2 regression coefficient

To know how close the data are to the fitted regression line we have used the statistical measure R-squared. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or: R-squared = Explained variation / Total variation. R-squared is always between 0 and 1:

- 0 indicates that the model explains none of the variability of the response data around its mean.
- 1 indicates that the model explains all the variability of the response data around its mean.

To examine the quality of the model and observations, we plot the fitted line and observations, and we look at R-squared value.

4.3. KRIGING APPROACH

4.3.1. What is Kriging?

Kriging is defined as a random value interpolation based on nearby observations that are weighted according to spatial covariance values (Kriging-GSM Implementation, in press and Bohling, 2005).

Note that interpolation is estimation of a variable at an unmeasured location from observed values at surrounding locations. Interpolation of parameters is treated as a regionalized variable and is intermediate between a truly random variable and a completely deterministic value. Kriging estimation weights are derived from a covariance matrix and the method employs the concept of the semi-variogram, a function that characterizes the residual components on which the residual estimation of a desired location is dependent. From a covariance model, the estimated variance is minimized. Kriging methods are generally of three kinds: simple, universal and ordinary.

In simple kriging, a particular region is considered from all the available points and is analyzed. In universal kriging, it is applied region-wise and thus the entire available set of points is analyzed. Universal kriging is performed when very large amounts of data or points of measurement are available. Ordinary kriging follows the original development of the method as described in a thesis by Krige (1951). Kriging semi-variograms are of different types depending on the application and these types include spherical, exponential, and Gaussian amongst others [13].

4.3.2. Characteristics of the semi-variogram

The sill is the semi-variance value at which the variogram levels off, as shown in Figure9. The sill is generally 1.0, but may have lesser values. Range is the lag distance at which the variogram reaches the sill. If the variogram does not start from zero, the difference is termed the nugget. Overall sill is the difference between fundamental sill and the nugget. Figure10 illustrates the morphology of spherical, exponential, and Gaussian variograms. From this it should be clear that the spherical variogram reaches the sill earlier than the exponential and Gaussian. So, the spherical variogram is what was considered to employ with the kriging method used here since the equation that reaches the sill earlier yields a better approximation, which our results confirmed [14]. The variogram equations or the covariance function is given by Equation3, Equation4 and Equation5.

$$C(h) = c\left(1 - \exp\left(\frac{-3h}{a}\right)\right)$$

Equation 3 Exponential

$$C(h) = c(\exp\left(\frac{-3h^2}{a^2}\right))$$

Equation 4 Gaussian

$$C(h) = 0.78\left(1 - 1.5\left(\frac{h}{m}\right) + 0.5\left(\frac{h}{m}\right)\right)^3$$

Equation 5 Spherical

Where:

- h is the distance of separation
- m is the maximum distance of separation in the semi-variance C.

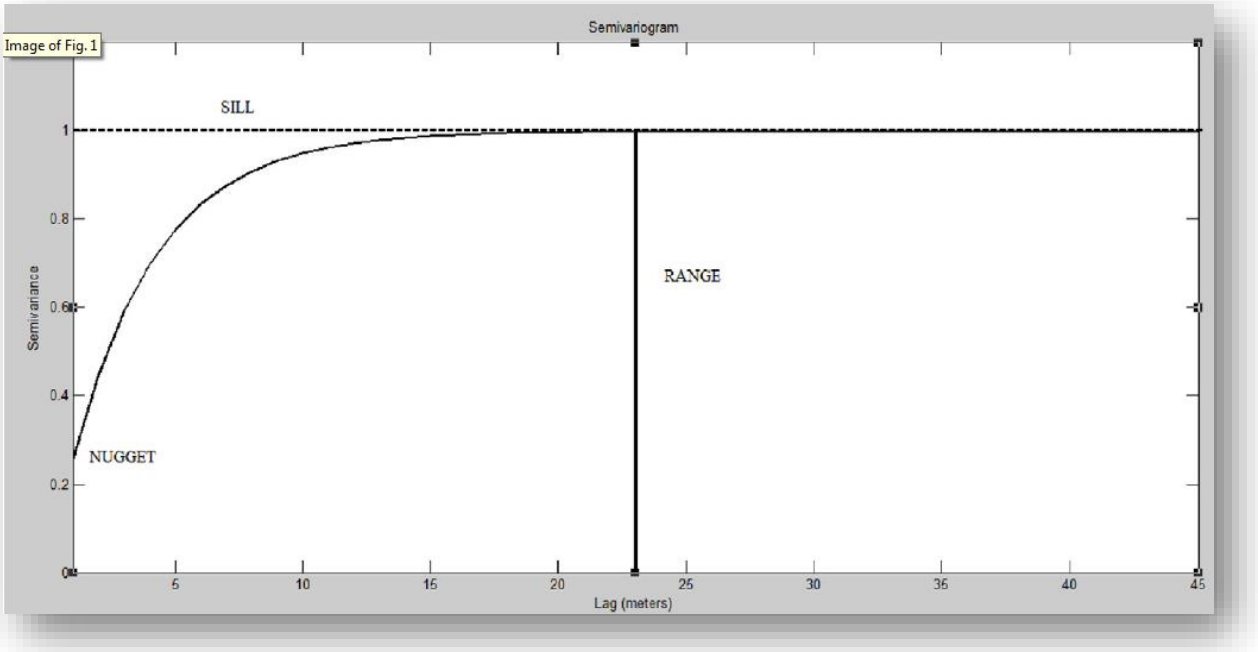


Figure 8 The sill

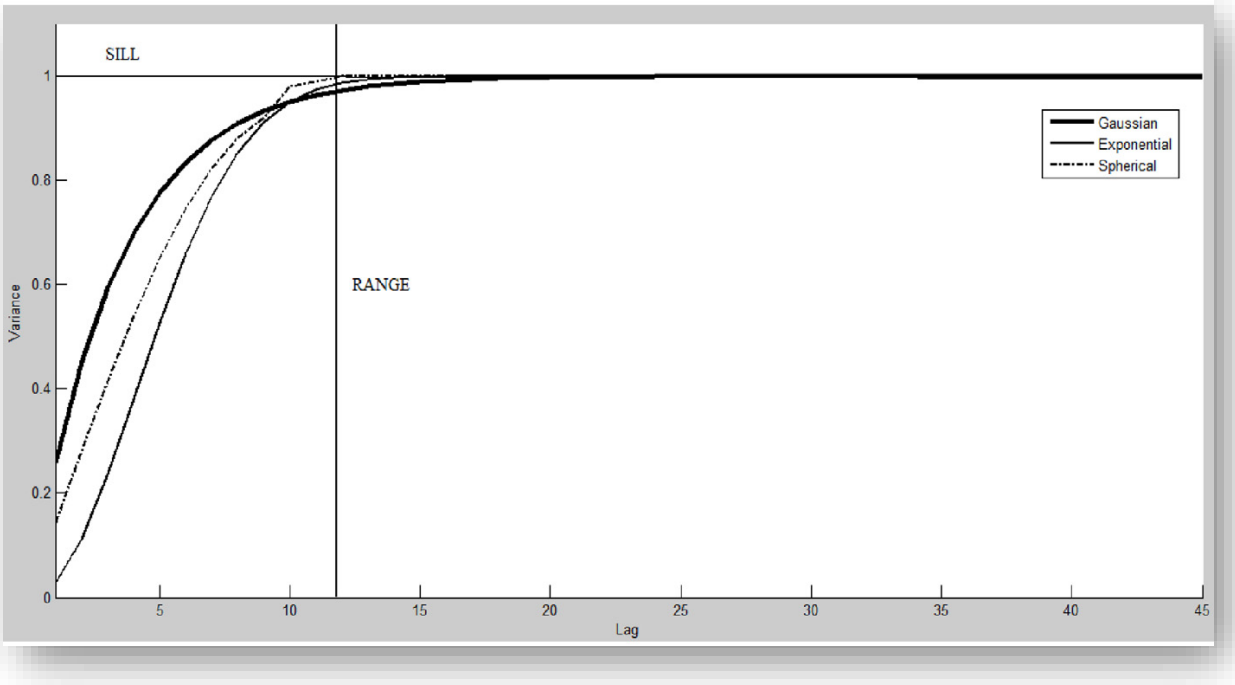


Figure 9 the morphology of spherical, exponential, and Gaussian variograms

4.4. MissForest Method

Modern data acquisition based on high-throughput technology is often facing the problem of missing data. Algorithms commonly used in the analysis of such large-scale data often depend on a complete set. Missing value imputation offers a solution to this problem. However, the majority of available imputation methods are restricted to one type of variable only: continuous or categorical. For mixed-type data the different types are usually handled separately. Therefore, these methods ignore possible relations between variable types.

4.4.1. What is MissForest?

MissForest is a nonparametric method based on a random forest, which can cope with different types of variables simultaneously. It is used to impute missing values particularly in the case of mixed-type data. It can be used to impute continuous and/or categorical data including complex interactions and nonlinear relations. It yields an out-of-bag (OOB) imputation error estimate. Moreover, it can be run parallel to save computation time.

MissForest outperforms other methods of imputation especially in data settings where complex interactions and nonlinear relations are suspected. The out-of-bag imputation error estimates of MissForest prove to be adequate in all settings. Additionally, MissForest exhibits attractive computational efficiency and can cope with high-dimensional data [15].

4.4.2. Out-Of-bag Error

Using the built-in out-of-bag error estimates of random forest we are able to estimate the imputation error without the need of a test set. MissForest could successfully handle missing values, particularly in our data set which includes different types of variables.

estimated OOB imputation error. For the set of continuous variables, the normalized root mean squared error NRMSE and for the set of categorical variables the proportion of falsely classified entries PFC is returned.

The normalized root mean squared error (NRMSE) is defined as Equation6:

$$\sqrt{\frac{mean((X_{true} - X_{imp})^2)}{var(X_{true})}}$$

Equation 6 The normalized root mean squared error

Where:

- X_{true} the complete data matrix
- X_{imp} the imputed data matrix
- 'mean'/'var' being used as short notation for the empirical mean and variance computed over the continuous missing values only.

The proportion of falsely classified (PFC) is computed over the categorical missing values only.

5 RESULTS

After extract data from the web, deal with outliers, we had a dataset with 2718 as a total number of observations, and 2182 observations with a complete information about the energy consumption, and the distribution of the dwelling depending on the consumption class is shown in Figure10.

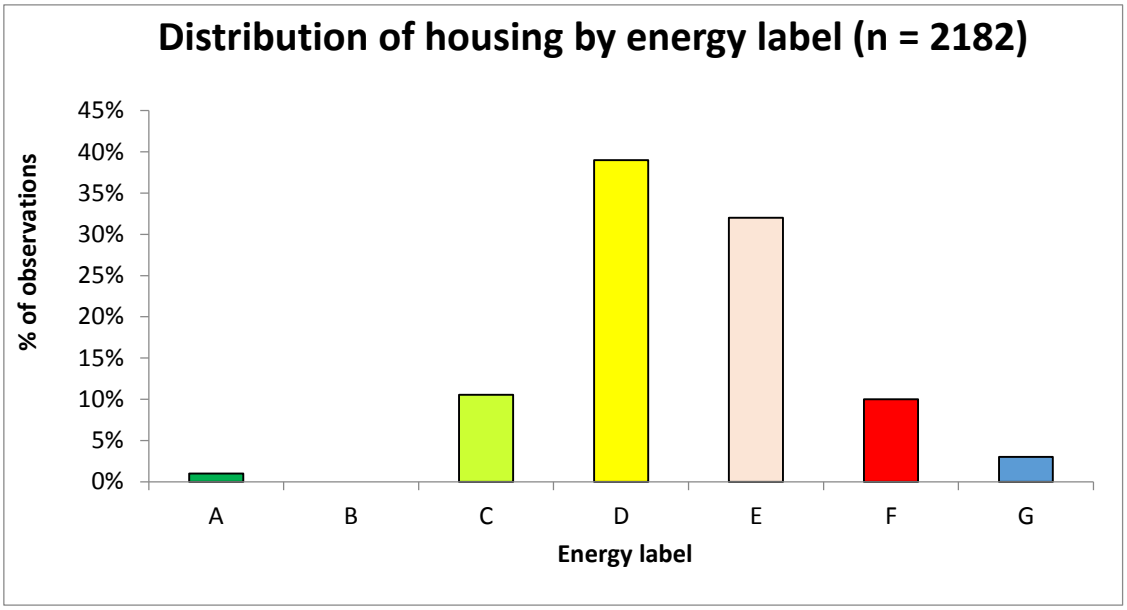


Figure 10 distribution of the dwelling depending on the energy consumption class

Then after study the association and correlation between the variable, we have decided to test the Linear regression, kriging and MissForest for our study. The three techniques implement very different approaches to reconstruct problem.

5.1 THE LINEAR REGRESSION

It provides good results using all possible variables. And it is well known that linear regression models, become unstable when they include many predictor variables relative to the sample size.

Among data-driven algorithms, the regression methods can be considered as direct and explicit procedures. In general, the higher the R-squared is better and the model fits the data. But, we didn't achieve good R-squared values, and the plots explains that we hadn't good linear regression model, and there is no explicit linear relation between the energy consumption and the other variables in the study area.

A simple linear regression was calculated to predict energy consumption for a house or apartment based on each variable, in Table2 we can see some of the R-squared values that achieved as results for the regression analysis, the highest one is calculated based on the surface and the number of parts, but it still very low and doesn't present a good regression.

#	variable	R-squared
1	surface	0.07560509
2	longitude	0.001867809

3	latitude	4.462436e-06
4	yearOfConstruction	0.0177073
5	numOfParts	0.06613442
6	HeatingSystem	0.005062733
7	PoweredBy	0.0470458
8	floor	0.01886318
9	type	0.007948367

Table 2 R-squared values

In Figure11 we can see the simple linear regression was calculated to predict energy consumption for a house or apartment based on its surface.

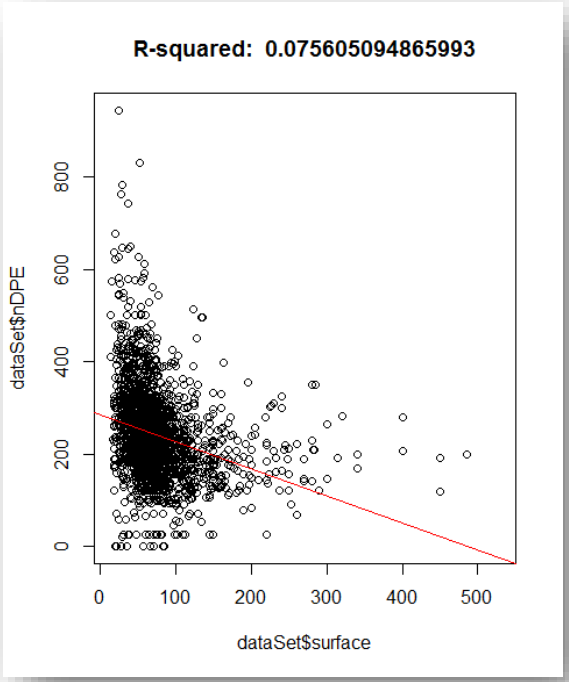


Figure 11 Energy consumption for a house or apartment based on its surface

In annex 3 some other examples are shown, with the according R-Square value.

5.2 KRIGING

Ordinary kriging (OK) has been selected as the method for the exploitation of the spatial variability, as can be observed by means of specific tools, such as the semi-variogram.

All interpolation algorithms (inverse distance squared, splines, radial basis functions, triangulation, etc.) estimate the value at a given location as a weighted sum of data values at surrounding locations. Almost all assign weights according to functions that give a decreasing weight with increasing separation distance. Kriging assigns weights according to a (moderately) data-driven weighting

function, rather than an arbitrary function. In particular, if the data locations are fairly dense and uniformly distributed throughout the study area, we will get fairly good estimates regardless of interpolation algorithm.

We applied ordinary kriging OK to our data, using the spherical semi-variogram, with zero nugget, a sill of 0.78, and a range of 10000, 2182 data points(observations with a complete information about the energy consumption), coordinates summary shown in Table3, distance summary is shown in Table4, data summary is shown in Table5.

In Figuer12 we can see graphical presentation for this information.

	Coord1	**Coord2**
Min	4539805	4330470
Max	4547305	4457235

Table 3 Coordinates summary in OK

min	**max**
3.38883e-05	1.26883e+05

Table 4 Distance summary in OK

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
1.0	185.0	227.0	240.2	282.8	944.0

Table 5 Data summary in OK

In Figure13 we can see the distribution of the dwelling depending on the energy consumption class, after applying OK on 2718 points as a test dataset.

The predicted values summarized in Table6, and the variance ranged between 0.069 and 10.54.

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
33.51	201.30	232.00	243.60	279.20	515.20

Table 6 predicted nDPE values in OK

It is clear that most of the dwelling are on class E or D, but if we compare with original dataset extracted from the Web, we can see that the two profiles are very similar and this is an indicator that our results are good.

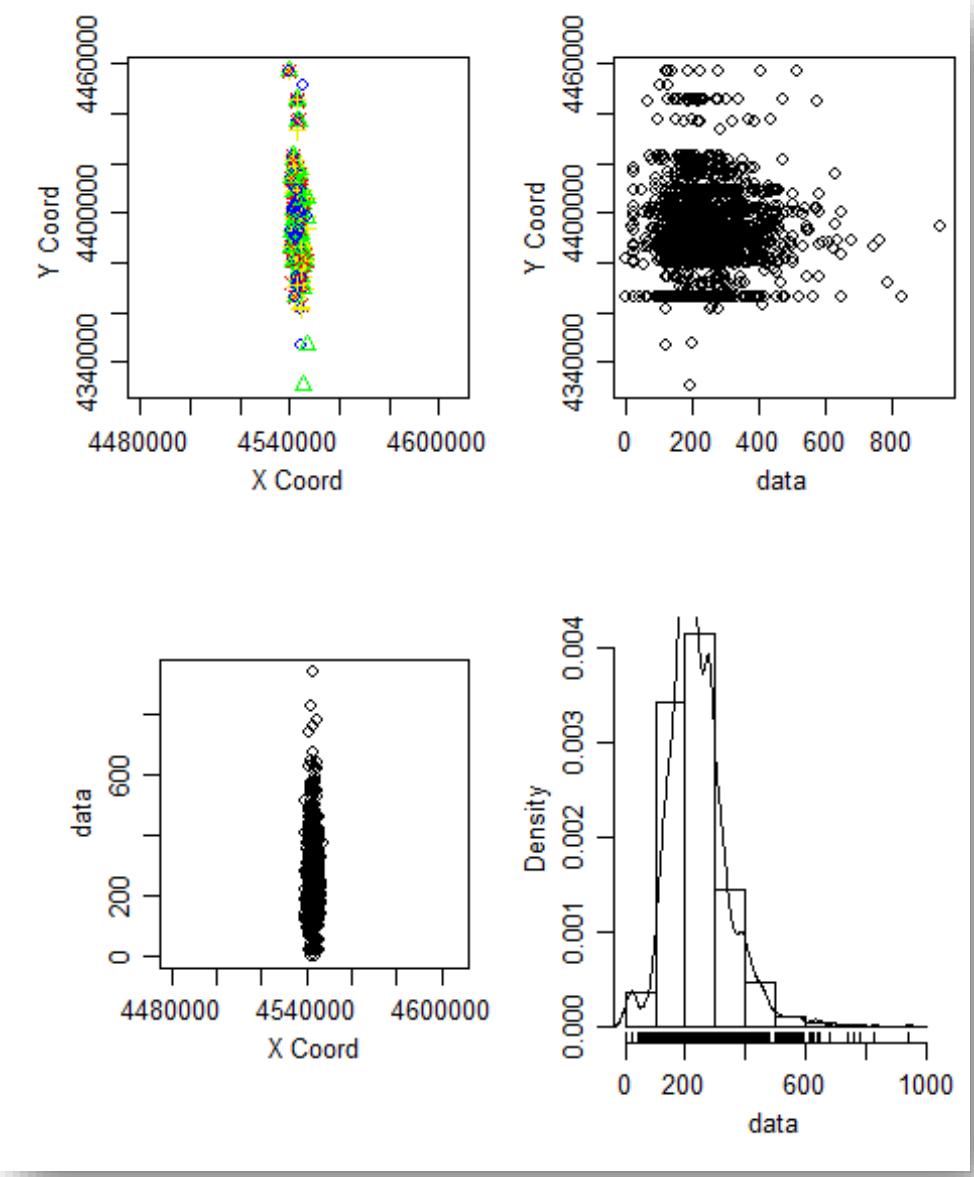


Figure 12 data used in ordinary kriging method

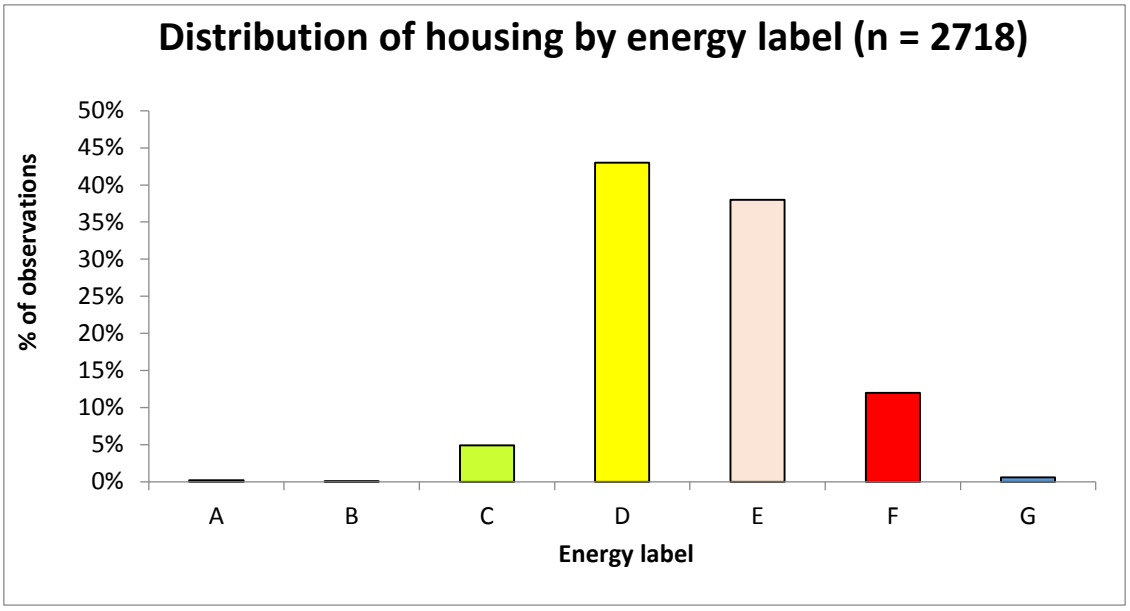


Figure 13 distribution of the dwelling depending on the energy consumption class (Kriging results)

5.3 MissForest

In many cases where the previous methods did badly, the decision-tree methods did relatively. We have decided to include MissForest method in our study, as a method based on a random forest.

It imputed the missing values in our dataset that includes complex interactions and nonlinear relations. It yielded an out-of-bag (OOB) imputation error estimate less than 0.5 for the numeric variables and less than 0.05 for the categorical variables. The predicted values summarized in Table7, and the out-of-bag (OOB) imputation error estimate as following: 0.421 for NRMSE, and 0.039 for PFC.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.0	190.0	251.0	245.8	280.0	944.0

Table 7 predicted nDPE values in MissForest

And the distribution of the dwelling depending on the energy consumption class, is shown in Figuer14.

Here we can see that most of the dwelling are on class E or D also but the percentage in E is higher, so we can say that Kriging results is better because it is more similar to the original dataset.

But in general the energy consumption was reconstructed good for both of Kriging and MissForeast, we had good profiles and similar to the original dataset that was extracted from the Web.

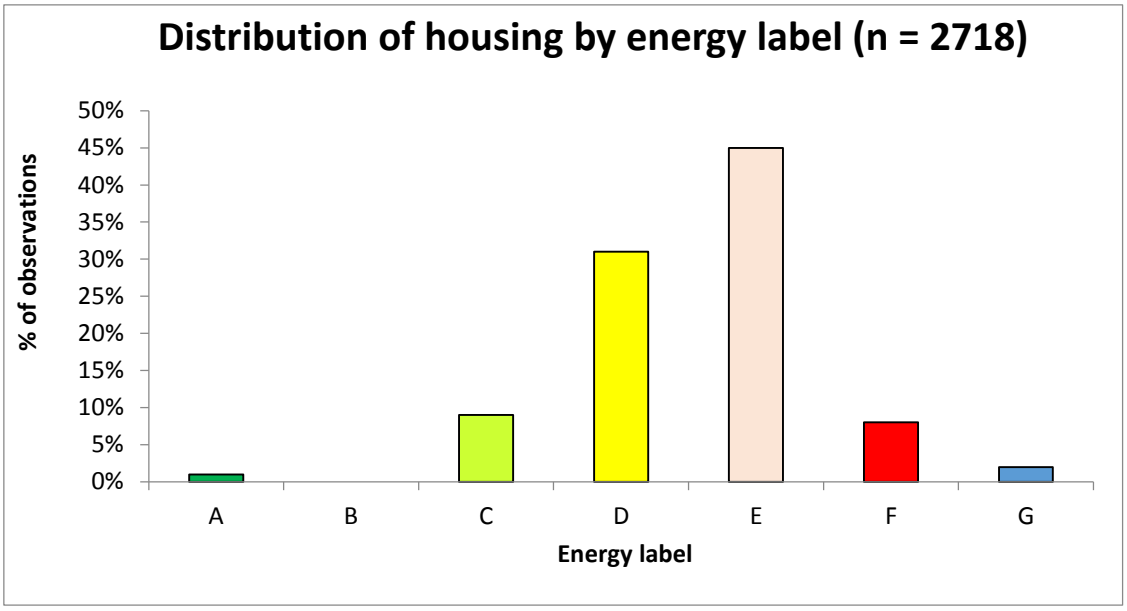


Figure 14 distribution of the dwelling depending on the energy consumption class (MissForest results)

6 CONCLUSIONS AND FUTURE PROSPECTS

As we saw in this report, this internship was about extract data from web and reconstructs the energy consumption.

Indeed, a lot of work has been handled. All the work had a common objective: improve the reconstructed data by using deferent techniques and approaches, and provide tools to help in this purpose. And this allowed me to put in application some data analysis techniques acquired during the master in the real word. And in the other hand, discover new things and work with other people.

In a general manner the results are quite good for Kriging, good for MissForest and quite disappointing for the regression.

Finally, at the time this report is written, the internship is not over yet, and I still have two months to work in École des Mines. I am unfortunately not able to present in this report all the work.

As a future work, we plan to improve our methods and try to obtain better and more accurate results, and we will test DiceKriging R package, that used more variables than the geostatistic package. Also depending on this work, we will publish a paper about the energy consumption for renewal buildings.

Longer-term future work includes to apply the methods on other French cities.

REFERENCE

[1] Francesca Cappelletti, Tiziano Dalla Mora, Fabio Peron, Piercarlo Romagnoni and Paolo Ruggeri, 2015, Building renovation: which kind of guidelines could be proposed for policy makers and professional owners?

[2] Buildings Performance Institute Europe (BPIE), October 2011, Europe's buildings under the microscope.

[3] Giuliano Dall'O', Annalisa Galante and Marco Torri, January 2012, A methodology for the energy performance classification of residential building stock on an urban scale.

[4] ADEME, April 2016, Le Diagnostic de Performance Énergétique.

[5] JOSEP MARIA RIBAS PORTELLA, 2012, Bottom-up description of the French building stock, including archetype buildings and energy demand.

[6] DINAMIC, octobre 2015, La valeur verte des logements en 2014.

[7] Paola Caputo, Giulia Pasetti, Overcoming the inertia of building energy retrofit at municipal level: The Italian challenge.

[8] Giulia Pasetti, January 2016, Stimulate energy renovation of the building stock: Policies and tools at municipal scale.

[9] Joakim Ekstrom, A Generalized definition of the poly-choric correlation coefficient.

[10] MEI paper on Spearman's rank correlation coefficient, December 2007.

[11] Jan Hauke, Tomasz Kossowski, 2011, Comparison of values of Pearson's and Sprerman correlation coefficients on the same sets of data.

[12] Rajesh Guntaka, Harley R. Myler, 2014, Regression and kriging analysis for grid power factor estimation.

[13] Jian Wen, Huizhu Yang, Guanping Jian, Xin Tong, Ke Li, Simin Wang, 2016, Energy and cost optimization of shell and tube heat exchanger with helical baffles using Kriging met model based on MOGA.

[14] Edzer Pebesma, 2016, The meuse data set: a brief tutorial for the gstat R package.

[15] Daniel J. Stekhoven and Peter B¨uhlmann, October 2011, MissForest - nonparametric missing value imputation for mixed-type data.

ANNEXES

ANNEX 1 DATA SUMMARY

Table8 shows information regarding the missing values in each variable.

#	variable	NA count	NA Percentage
1	priceRent	1378	50%
2	priceBuy	1353	49%
3	surface	31	1%
4	irisName	158	5%
5	yearOfConstruction	1701	62%
6	numOfParts	3	0.1%
7	DPE	528	19%
8	GES	614	22%
9	nDPE	528	19%
10	nGES	614	22%
11	HeatingSystem	826	30%
12	CentralHeating	953	35%
13	PoweredBy	1013	37%
14	floor	1482	54%
15	Type	0	0%

Table 8 Missing Values

Table9 shows summary regarding the quantitative variables.

	Min	1st Qu	Median	Mean	3rd Qu	Max	NA's
priceRent	214.0	352.0	445.0	472.6	545.0	1595.0	1378
pricBuy	21000	59000	86000	118006	145000	985000	1353
surface	13.00	49.00	66.00	75.84	87.00	530.00	31
longitude	4.330	4.382	4.388	4.390	4.400	4.457	8
latitude	45.40	45.42	45.43	45.43	45.44	45.47	5
yearOfConstruction	1850	1930	1960	1952	1973	2016	1701
numOfParts	1	2	3	3	4	20	3

nDPE	1.00	184.00	226.00	240.51	282.00	944.00	528
nGES	1.00	29.00	45.00	46.51	61.00	350.00	614

Table 9 quantitative variables summary

Tables from 10 to 15 and figures from 15 to 20 show summary regarding the qualitative variables.

Heating system summary		
	Count	**Percentage**
collective	878	46%
individual	1014	54%

Table 10 Heating system summary

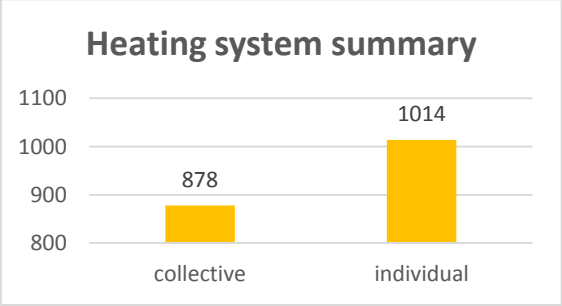


Figure 15 Heating System summary

CentralHeating summary		
	Count	**Percentage**
no	251	14%
yes	1514	86%

Table 11 CentralHeating

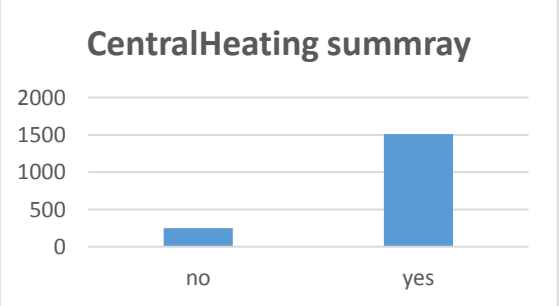


Figure 16 CentralHeating summary

GES summary		
	Count	**Percentage**
A	52	02%
B	89	04%
C	185	09%
D	378	18%
E	743	35%
F	502	23%
G	155	07%

Table 12 GES summary

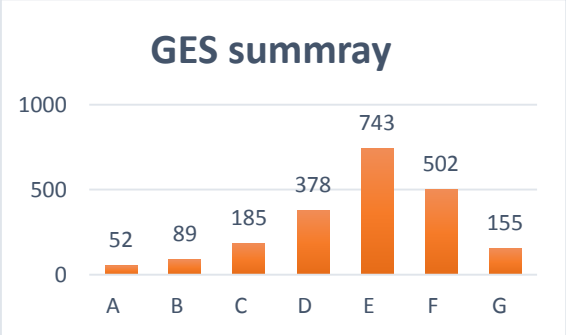


Figure 17 GES summray

DPE summray		
	Count	**Percentage**
A	38	02%
B	20	01%
C	254	11%
D	862	39%
E	716	33%
F	231	11%
G	69	03%

Table 13 DPE summray

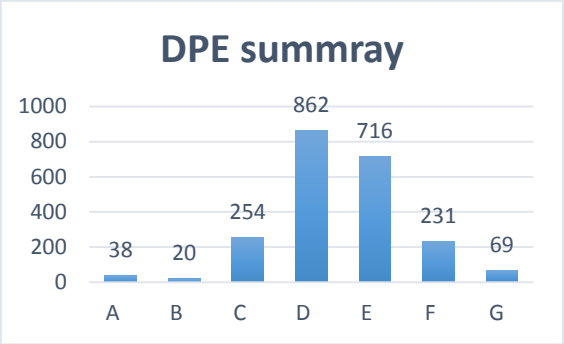


Figure 18 DPE summray

PowerdBy summray		
	Count	**Percentage**
electric	337	20%
fuel	63	04%
gas	1305	76%

Table 14 PowerdBy summray

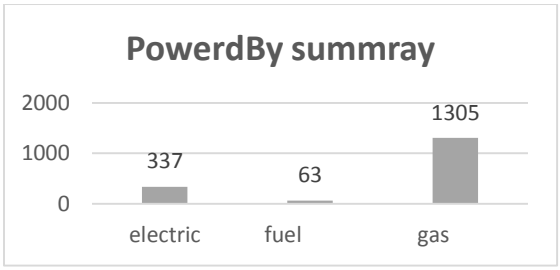


Figure 19 PowerdBy summray

Floor summray		
	Count	**Percentage**
cellar	3	0.1%
middle	668	33%
rdc	282	13%
top	283	13%

Table 15 Floor summray

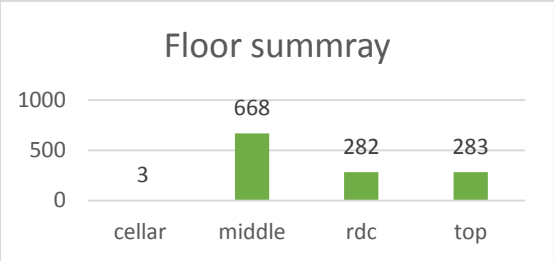


Figure 20 Floor summray

ANNEX 2 ASSOCIATION/CORRELATION TESTS RESULTS

Poly-choric correlations between ordinal variables

Correlations, Standard Errors and P-values for Tests of Bivariate Normality are shown in the following tables:

	Floor	Heating System	PowerdBy	GES	DPE	irisName
floor	1					
Heating System	0.2534	1				
PowerdBy	-0.06564	-0.492	1			
GES	0.009356	-0.4187	0.6513	1		
DPE	0.09986	-0.1211	-0.1987	0.5176	1	
irisName	-0.01662	0.02212	0.07244	0.04874	-0.00533	1

Table 16 Correlations between ordinal variables

	Floor	Heating System	PowerdBy	GES	DPE
Heating System	0.04359				
PowerdBy	0.04987	0.04967			
GES	0.03747	0.03732	0.02761		
DPE	0.03767	0.04454	0.04521	0.02438	
irisName	0.03845	0.04455	0.04786	0.03649	0.037

Table 17 Standard Errors:

	Floor	Heating System	PowerdBy	GES	DPE
Heating System	0.006107				
PowerdBy	0.001319	3.182e-12			
GES	0.006057	4.226e-09	1.201e-23		
DPE	8.633e-05	0.0001756	4.759e-06	1.335e-121	
irisName	0.0005054	0.0004266	0.009882	0.2576	0.7912

Table 18 P-values for Tests of Bivariate Normality

Spearman rank correlation between numeric variables

The correlation matrix R, the matrix of the number of observations used in analyzing each pair of variables N, and the p-values corresponding to the significance levels of correlations are shown in the following tables.

	price	surface	longitude	latitude	Year	NumOfParts	nGES	nDPE
price	1.00	0.75	-0.06	-0.04	0.24	0.73	-0.02	-0.29
surface	0.75	1.00	-0.06	0.03	0.09	0.89	-0.05	-0.36
longitude	-0.06	-0.06	1.00	-0.16	0.03	-0.03	0.07	0.06
latitude	-0.04	0.03	-0.16	1.00	-0.17	-0.02	-0.09	0.00
Year	0.24	0.09	0.03	-0.17	1.00	0.12	-0.11	-0.17
NumOfParts	0.73	0.89	-0.03	-0.02	0.12	1.00	-0.01	-0.28
nGES	-0.02	-0.05	0.07	-0.09	-0.11	-0.01	1.00	0.50
nDPE	-0.29	-0.36	0.06	0.00	-0.17	-0.28	0.50	1.00

Table 19 correlation between numeric variables

	price	surface	longitude	latitude	Year	NumOfParts	nGES	nDPE
price	2716	2686	2708	2711	1016	2713	2103	2189
surface	2686	2687	2679	2682	1014	2685	2090	2174
longitude	2708	2679	2710	2710	1015	2707	2096	2182
latitude	2711	2682	2710	2713	1016	2710	2099	2185
Year	1016	1014	1015	1016	1017	1016	894	914
NumOfParts	2713	2685	2707	2710	1016	2715	2104	2189
nGES	2103	2090	2096	2099	894	2104	2104	2100
nDPE	2189	2174	2182	2185	914	2189	2100	2190

Table 20 number of observations used in analyzing each pair of variables

	price	surface	longitude	latitude	Year	NumOfParts	nGES	nDPE
price		0.0000	0.0024	0.0462	0.0000	0.0000	0.4579	0.0000
surface	0.0000		0.0020	0.0740	0.0056	0.0000	0.0275	0.0000
longitude	0.0024	0.0020		0.0000	0.2787	0.0799	0.0019	0.0094
latitude	0.0462	0.0740	0.0000		0.0000	0.4343	0.0000	0.8544
Year	0.0000	0.0056	0.2787	0.0000		0.0001	0.0015	0.0000
NumOfParts	0.0000	0.0000	0.0799	0.4343	0.0001		0.5041	0.0000
nGES	0.4579	0.0275	0.0019	0.0000	0.0015	0.5041		0.0000
nDPE	0.0000	0.0000	0.0094	0.8544	0.0000	0.0000	0.0000	

Table 21 p-values corresponding to the significance levels of correlations

Pearson test on quantitative and qualitative data

The correlation matrix, standard Errors and P-values for tests of bivariate normality are shown in the following tables.

	Floor	Heating System	PowerdBy	GES	DPE	irisName
Price	-0.06409	-0.06209	0.1527	-0.03428	-0.2838	
surface	-0.03235	0.02021	0.2313	-0.02012	-0.3279	0.04428
longitude	0.004262	-0.08237	0.02145	0.06131	0.07507	0.01112
latitude	0.03918	0.1601	-0.1428	-0.1309	-0.01716	-0.3683
Year	-0.1175	-0.2928	-0.04906	-0.05757	-0.04611	
NumOfParts	-0.05962	-0.07972	0.2685	0.0312	-0.2843	0.06482
nDPE	-0.01672	-0.2929	0.4187	0.8593	0.4399	
nGES	0.1058	-0.04305	-0.1608	0.4084	0.9259	

Table 22 correlation matrix between quantitative and qualitative data

	Floor	Heating System	PowerdBy	GES	DPE	irisName
Price	0.03336	0.03337	0.03272	0.03346	0.03081	
surface	0.03364	0.03366	0.03187	0.03366	0.03006	0.01994
longitude	0.03356	0.03333	0.03354	0.03343	0.03337	0.01987
latitude	0.03347	0.03266	0.03284	0.03295	0.03351	0.01612
Year	0.04987	0.04625	0.05044	0.0504	0.05046	
NumOfParts	0.03338	0.03329	0.03109	0.03347	0.0308	0.01975
nDPE	0.03349	0.03063	0.02764	0.008774	0.02703	
nGES	0.03313	0.03344	0.03264	0.02792	0.004791	

Table 23 Standard Errors

	Floor	Heating System	PowerdBy	GES	DPE	irisName
Price	1.459e-260	0	0	4.261e-201	1.591e-134	
surface	2.283e-139	9.6949e-278	0	2.056e-132	4.136e-60	0
longitude	3.468e-125	1.723e-271	0	3.442e-114	4.73e-50	0
latitude	2.055e-134	4.318e-278	0	5.518e-124	2.007e-56	0

Year	1.441e-57	1.288e-120	1.426e-149	1.799e-58	1.291e-26	
NumOfParts	4.511e-150	2.187e-285	0	7.792e-140	4.749e-70	0
nDPE	3.305e-122	2.054e-268	0	3.156e-182	4.008e-97	
nGES	8.607e-127	5.065e-274	0	1.519e-190	1.104e-104	

Table 24 P-values for Tests of Bivariate Normality

ANNEX 3 REGRESSION RESULTS

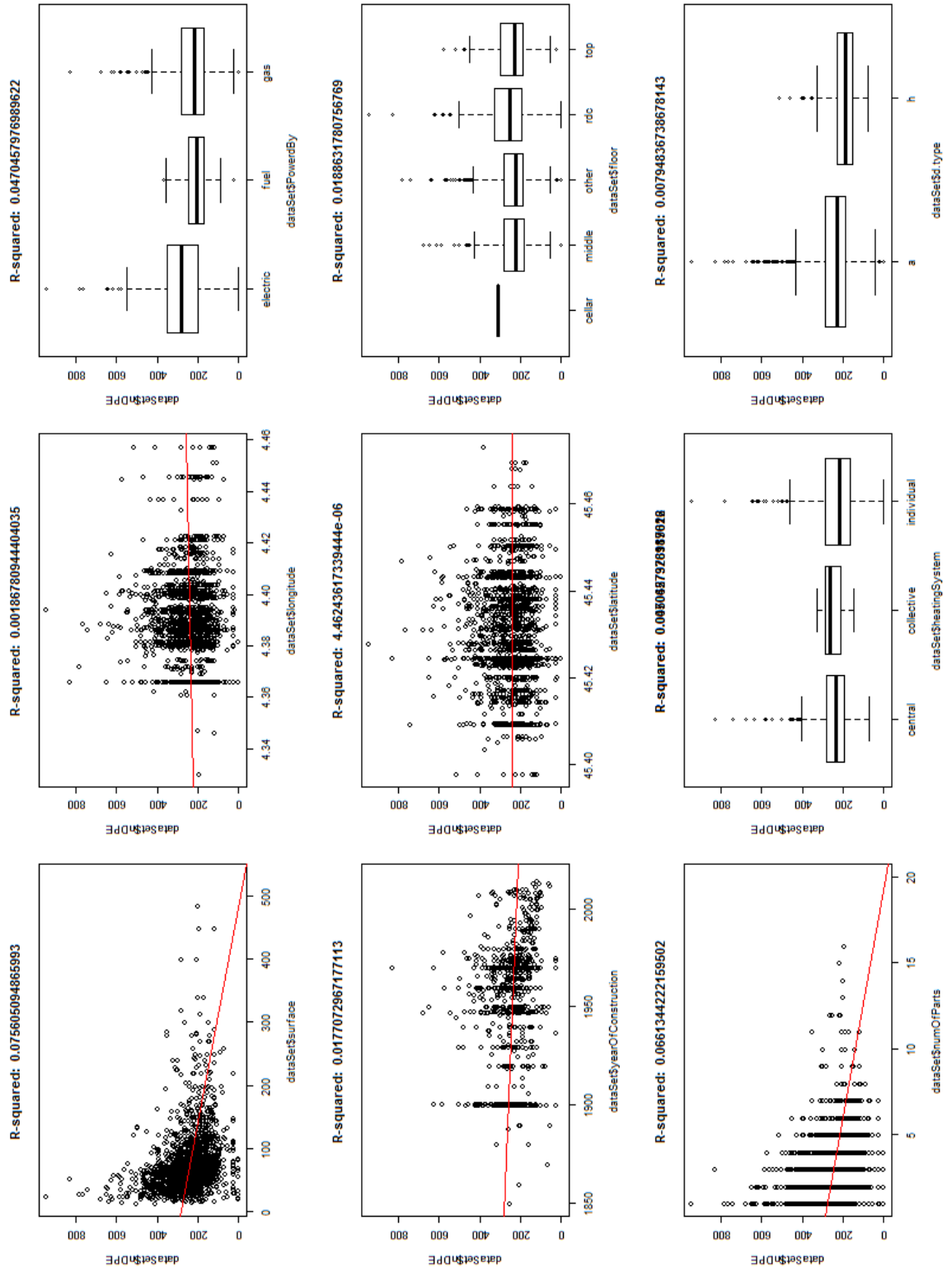


Figure 21 Example on regression results