Faculté des sciences

GAM Modeling of the Annual and Longterm Temperature Dependency of the Gas and Electricity Consumption for the Belgian Residential Market

Auteure : Elisa Ribesse
Promoteur : Prof. Christian Ritter
Lecteur·rices : Prof. Rainer von Sachs, Anne De Frenne

Année académique 2020-2021

Master 120 en Statistique Orientation Générale

Remerciements

Je tiens a remercier mes proches: mes parents, mes soeurs et mes amis qui m'ont soutenue tout au long de la scolarité et plus particulièrement dans cette dernière période spéciale.

Un grand merci particulièrement à Stéphanie, à Héloïse et surtout à la famille Grancitelli qui a beaucoup oeuvré pour que ce mémoire se déroule dans les meilleures conditions possibles. Sans eux, la rédaction n'aurait certainement pas été pareille.

Merci à Maxime, Franck, Gudalmatien, Martin, Laura, Laure et Amélie pour la relecture et correction de différents chapitres.

Merci à monsieur Rainer von Sachs, mon lecteur, d'avoir accepté de lire ce mémoire et merci pour ses recommandations.

Merci à madame Anne De Frenne, ma lectrice, d'avoir accepté de lire ce mémoire et merci pour le suivi et les retours tout au long de l'année.

Enfin et surtout, je tiens à remercier monsieur Christian Ritter, le promoteur de ce mémoire grâce à qui il a pu être terminé. Merci pour sa disponibilité lors de réunions, pour son encadrement et son suivi sur l'année, pour ses précieux conseils, pour sa patience et sa bienveillance hors du commun. Merci pour l'aide apportée qui a largement depassé ce qu'on pourrait espérer d'un promoteur.

Contents

Chapter 1

Introduction

How much electricity and gas do Belgians consume in a day? How does this consumption depend on the weather, the month of the year or the holidays?

How much will they consume tomorrow, in one year or in 5 years? And if the weather is very cold or very hot, how will it impact this consumption?

If these are questions that the average Belgian does not ask himself every day, the workers of the gas and electricity industry do. Thanks to this, there is always a supply of electricity and gas on the lines, and the average Belgian can continue not to worry about these interrogations.

If the question is based on issues of human and social behaviour depending on various behavioural and psychological factors (e.g. how sensitive people are to the new GIEC report if it has been overshadowed in the news by the news of Lionel Messi's transfer), this dissertation has one aspiration: to answer it solely from statistics.

The structure of this report is as follows:

First of all, an overview of the data and their context. An exploration of the different variables, an identification of the first behaviours with a visual analysis, carried out thanks to the construction of an interactive dashboard with the help of Shiny library in R. If temperature is the main influence on consumption, how does the temperature/consumption relationship evolve as a function of time and the time of year?

Next, a short theoretical review. The main statistical concept of this thesis is GAM. GAM stands for generalized additive models. The idea is to represent a model by a sum of smooth function bases: splines. A presentation of different types of splines is given.

Then, a look at a powerful and flexible R package : mgcv. This one aims at using GAMs to model data. An explanation of its features and different parameters is made using the consumption data.

This is followed by the construction of two models fitting the data, one for electricity th other for gas. These are then studied in more detail.

What are the observations made from these models?

Finally, how does this impact the prediction and the extrapolation? In order to answer these questions, an other dasboard is constructed.

The report ends with a conclusion setting out its contributions, its limitations and possible avenues for improvement.

Chapter 2

Data

2.1 Context

In Belgium, a challenge for gas and electricity companies is to manage the supply. On the one hand, too much energy produced can lead to waste. On the other hand, sufficient supply must be ensured at all times on the network. There is a reserve which purpose is to compensate for peaks in consumption or failures in power stations or local networks. But this one is very expensive. The bigger the difference between the energy consumed and the energy made available by a supplier, the more it will cost.

The overall consumption of gas or electricity can be divided into different categories. The first category consists of large industrial companies whose consumption is measured in real time. The second category consists of private individuals and other entities that have meters that are read on an annual basis. This paper focuses on this second category only.

Temperature is the main influence on consumption. According to the spf économie [1], in 2016, 74 % of the energy[1] consumed by Belgian households was used for heating. However, the share of heating in household energy consumption fluctuates depending on weather conditions. The rest of the energy consumed by households is mostly used for lighting and electrical appliances (12%), for water heating (12 %) and for cooking (2%). [1]

The need for heating does not influence gas and electricity in the same way. According to a 2016 study on household budgets [1], only 3% of the energy consumption for heating is electric compared to 47% for gas. The two types of energy must therefore be analysed in a different way.

Climatic conditions can also influence the need to heat water or lighting.

In summer, high temperatures can lead to the use of electricity to cool homes. According to the spf économie, only 0.3% of the electricity consumed by Belgian households was used to cool rooms in 2017. However, in the electricity consumption data from 2017, there is an increase for high temperatures.

[1] 74% of all energy sources combined (gas, oil, electricity, propane-butane, biomass,...). However this master thesis focuses only on the consumption of electricity and gas.

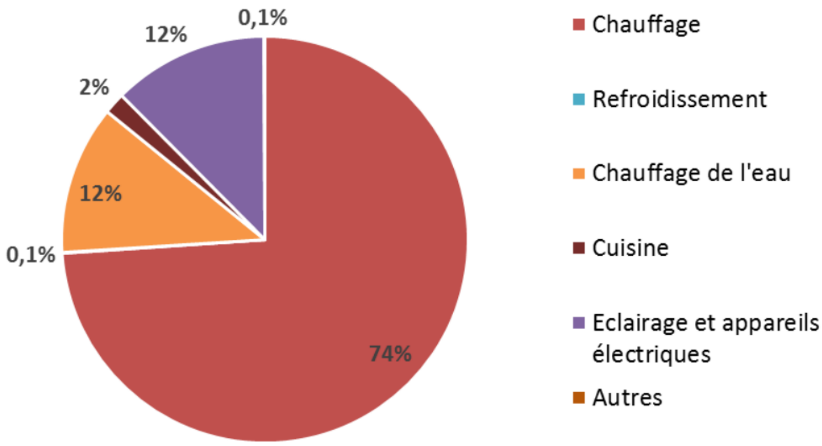

Figure 2.1.1: Parts of energy consumption of the households in 2016

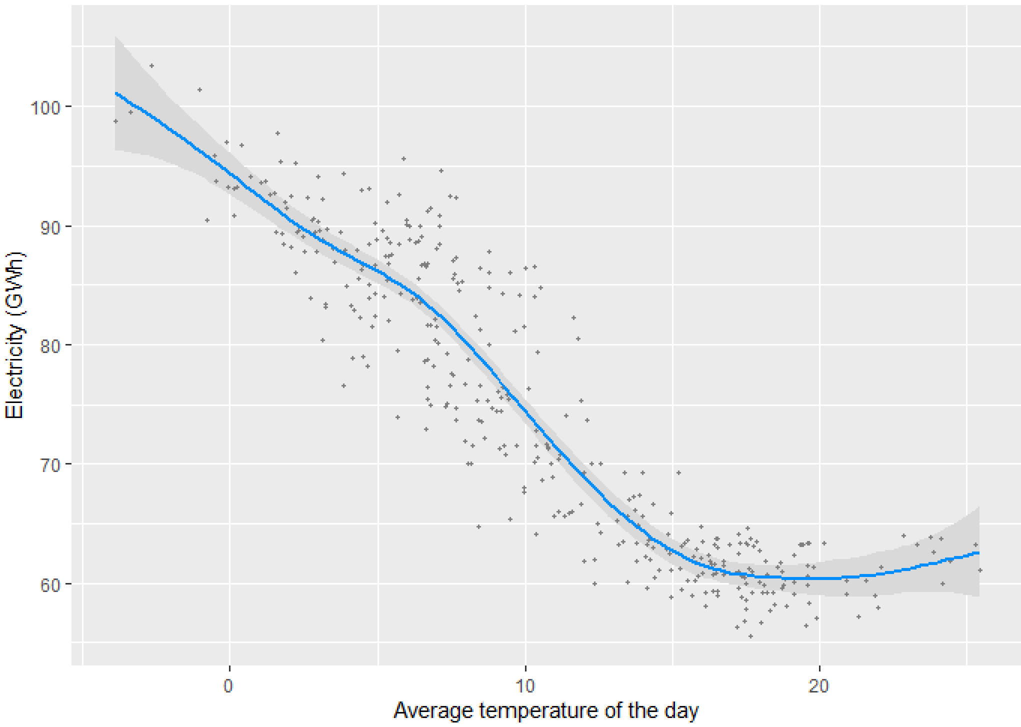

Figure 2.1.2: Total electricity consumption per day in Belgium in 2017. The temperature is an average over the day. The smooth curve is generated by the gam() function in the mgcv package.

Another influence on consumption is the lifestyle pattern. The needs of households during the week differ from those during the weekend, especially for electricity. Consumption during school holidays is also a special case.

The evolution of technologies, especially insulation technologies, can also influence consumption. The data in this report cover the period from 01/01/2010 to 31/01/2020. According to the spf économie (2019)[1], there is an improvement in the energy efficiency of buildings (renovations, insulation...) and electrical appliances. On the other hand, the penetration rate of air conditioning device [2] and electronic devices has increased, mainly laptops and digital decoders. Overall, consumption has been decreasing since 2010.

The idea of this master thesis is to determine the link between temperature and gas and electricity consumption.

2.2 Data

2.2.1 Incoming data

One observation represents one day. Observations range from 0/10/2010 to 23/12/2020.

The variables **"Year"**, **"Month"** and **"Day"** represent the date of each observation.

The variable **"Temp"**: is calculated as the daily average of hourly temperature readings. The highest recorded average temperature is 29.86° (in July 2019) and the lowest is -9.3° (in February 2012).

The variable **"GWhElec"** represents the total daily electricity consumption in Gigawatt-hours (GWh). The first data being collected by hours, every March there is a day with only 23 observations (on the day of the changeover to summer time). To remedy this, the calculation of the consumption per day is done by taking the average consumption of the hourly observations of a day, multiplied by 24. The first data are collected in Kilowatt-hours, they quickly become difficult to grasp, so they are converted into Gigawatt-hours. The highest daily consumption is 120.82 GWh (in January 2010) and the lowest is 53.51 GWh (in July 2019).

The variable **"GWhGas"** represents the total daily gas consumption in Gigawatt-hours (GWh) . As for GWhElec, the calculation of the daily consumption is done by taking the average of the hourly consumption of a day, multiplied by 24. The highest daily consumption is 120.82 GWh (in January 2010) and the lowest is 53.51 GWh (in July 2019).

A house can retain some heat from the previous days, it is therefore relevant to introduce a new variable : **"tempEq"** for equivalent temperature. For one observation at day d: $tempEq(d) = \frac{1}{10}\,Temp(d-2) + \frac{3}{10}\,Temp(d-1) + \frac{6}{10}\,Temp(d)$. This measure is used by the gas industry.

[2] growth rate of 15% in 2019 for Benelux. [5]

There are binary data that help to understand the life rhythm pattern:

- **"Mon"**=1 if the observed day is a Monday.
- **"Tue"**=1 if the observed day is a Tuesday.
- ...
- **"Sun"**=1 if the observed day is a Sunday.
- **"SummerHoliday"**=1 if the observed day is during the school summer holidays (about 61 days per year).
- **"WinterHoliday"**=1 if the observed day is during the professional winter holidays (about 10 days per year).
- **"EasterHoliday"**=1 if the observed day is during the Easter holidays (about 16 days per year).
- **"IsaWeekHoliday"**=1 if the observed day is a weekday holiday (about 8 days per year)

2.2.2 Predictions from another model

There is an existing model that already deals with predicting consumption based on these factors (and others such as brightness or wind power). This model is still under development and is managed by Anne De Frenne and Christian Ritter. It provides hourly and daily predictions.

"GWhEstElec" expresses the total hourly predictions of a day for electricity. Like GWhElec, it is calculated by taking the average of the predictions multiplied by 24.

"GWhDayEstElec" expresses the predicted electricity consumption for a day.

"GWhEstGas" expresses the total hourly predictions of a day for gas.It is calculated by taking the average of the predictions multiplied by 24.

"GWhDayEstGas" expresses the prediction of one day's gas consumption.

2.3 Descriptive analysis: Creation of an interactive dashboard

The focus is on the link between temperature and consumption of the two energies. An interactive dashboard was created in order to better understand this link depending on the variables described above. It was made with the library Shiny in R.

The main graph represents the energy consumption (in GWh) on the y-axis and the temperature (in degree Celsius) on the x-axis (see figure 2.3.1, the one on top left). A panel with many options was created to change the display and the data displayed.

By default, a smoothed curve with a confidence interval is displayed. The user can choose if he wants to keep the confidence interval with the option **"Interval?"** and choose to show the observations or not with the option **"Data?"**.

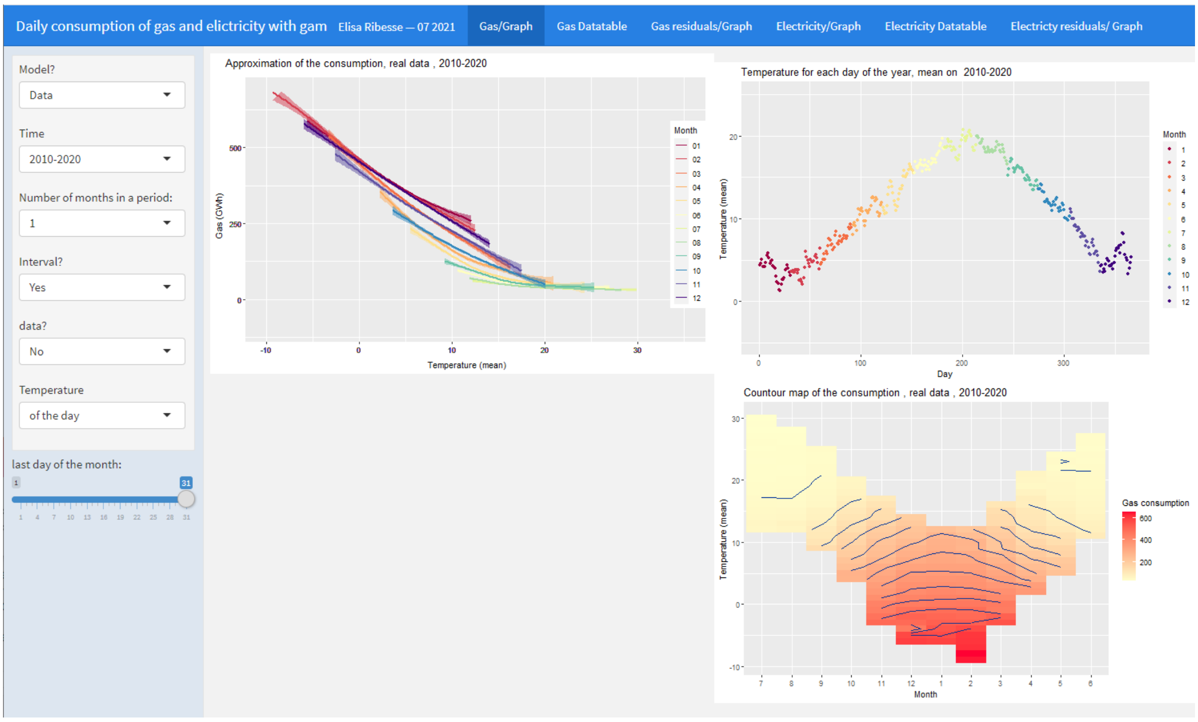

Figure 2.3.1: Overview of the first page of the interactive dashboard

The user can choose to look at the real consumption data "GWhGas" (resp. "GWhElec"), but also at the data predicted by the other model "GWhEstGas" or "GWhDayEstGas" (resp. "GWhEstElec" or "GWhDayEstElec"), by changing the **"Model?"** option. He can choose between the average temperature of the day "Temp" or the equivalent temperature "tempEq" with the option **"Temperature"**.

The years are divided into several time frames: 2010-2012, 2013-2016, 2017-2019 and 2020. The user can choose to look at the data of a particular time frame or to look at the whole data with the option **"Time"**.

The idea is to observe whether the link between temperature and consumption over one year has the same behaviour throughout the year. For this purpose, a breakdown system is provided. With the option **"Number of months in a period"** the user can choose whether he wants to split the year into 12 blocks of 1 month, 6 blocks of 2 months, 4 blocks of 3 months (or seasons) or not to split the year (one block of 12 months). Depending on this division, one or more curves of different colours and a legend are displayed.

As the month breakdown is arbitrary, the **"last day of the month"** option allows the user to choose the day on which the breakdown will be done. (For example, it is possible to split the month depending on the solstice).

The seasons are also an arbitrary construction. This is why when the 3-months split option is selected, new options are displayed: they allow the user to determine the beginning and end of each season. The seasons no longer have to be equal blocks of 3 months each. If required, Winter and Summer can for example be spread over 4 months and Autumn and Spring over 2 months. With the default settings, spring runs from March to May, summer from June to August, autumn from September to November and winter from December to February.

Two other secondary graphs join the previous one. One associates to each day of the year the average temperature recorded on that day over several years. The user can choose between the time frames with **"Time"** and between the temperature or the equivalent temperature with **"Temperature"**. The other one represents the level curves of the consumption depending on the month on x-axis and the Temperature on y-axis. The user can choose the time frames with **"Time"**, the consumption data with **"Model?"**, the temperature type with **"Temperature"** and the day when the month is split with **"last day of the month"**.

The three graphs described above make up the first sheet of the dashboard , see figure 2.3.1. In total, the dashboard consists of 6 sheets, 3 for gas and 3 for electricity.

The second sheet contains an interactive table of consumption data.

The third sheet has a similar structure to the first one. Instead of expressing the energy consumption, it shows the behaviour of the residuals. These are calculated as the difference between the actual consumption "GWhGas" (resp. "GWhElec") and the consumption estimated by another model"GWhEstGas" or "GWhDayEstGas" (resp. "GWhEstElec" or "GWhDayEstElec"). This sheet will be of little use in this thesis but can be used as an indication for the estimation models being developed by Anne De Frenne and Christian Ritter.

2.4 Descriptive analysis of data: Gas

Looking at the relationship between temperature and gas consumption on figure 2.4.1, two trends can be distinguished: a quasi-linear decreasing relationship for lower temperatures and then a very slightly decreasing (almost constant) linear relationship for higher temperatures. The break between these two trends is around 16° and 50 GWh. At ∼ -8° the consumption reaches ∼650 GWh. The first slope has an inclination of about -25GWh/°[3].

If one looks at the same data but presented by season, on figure 2.4.2, it can be seen that the bend at 16° was artificial. Indeed, there is no longer 1 curve with 2 trends but 4 curves with 3 distinct trends. The winter curve is close to a straight line with an inclination of about -22GWh/°. The autumn and spring curves have the same behaviour: a quasi-linear relationship for the lowest temperatures with a slope of about -25GWh/° and then a softer curve with a second degree behaviour. Finally, in summer, the slope is almost flat for the highest temperatures and a little steeper for temperatures between 10° and 15°.

Is this phenomenon logical? Let's take an example: for a day with a particularly low average temperature of 12° in Summer, the Belgians would consume about 80 GWh of gas

[3]Approximated slopes are roughly deduced by looking at the graphs.

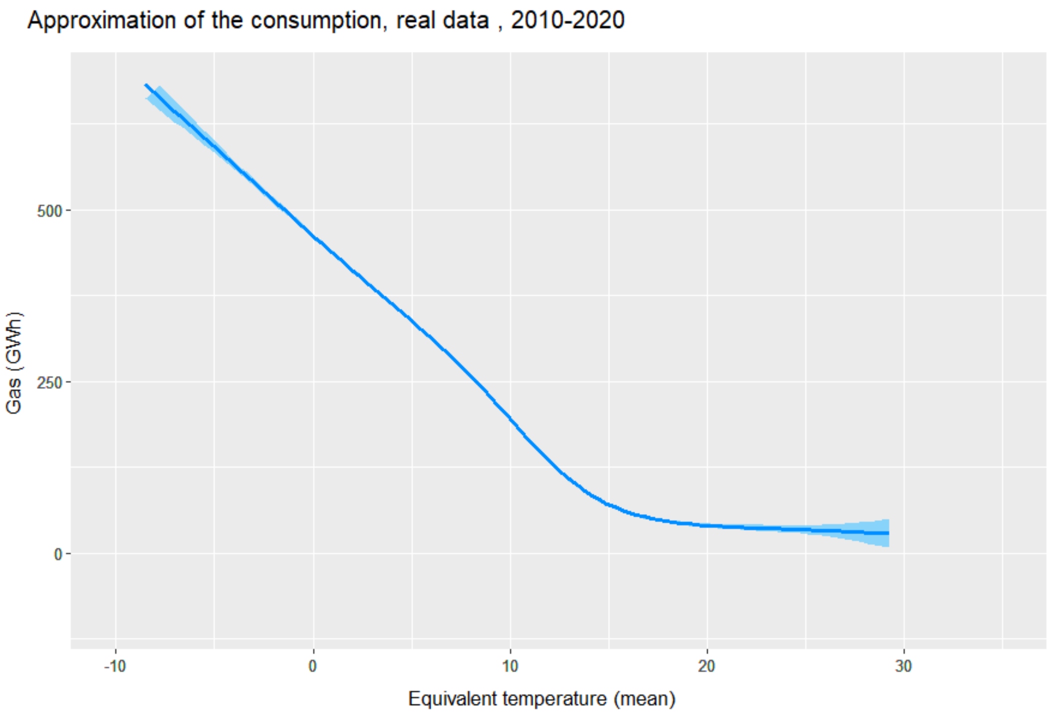

Figure 2.4.1: Approximation of Belgian Gas consumption as a function of temperature (calculated using the equivalent temperature formula "tempEq"). The year is not split.

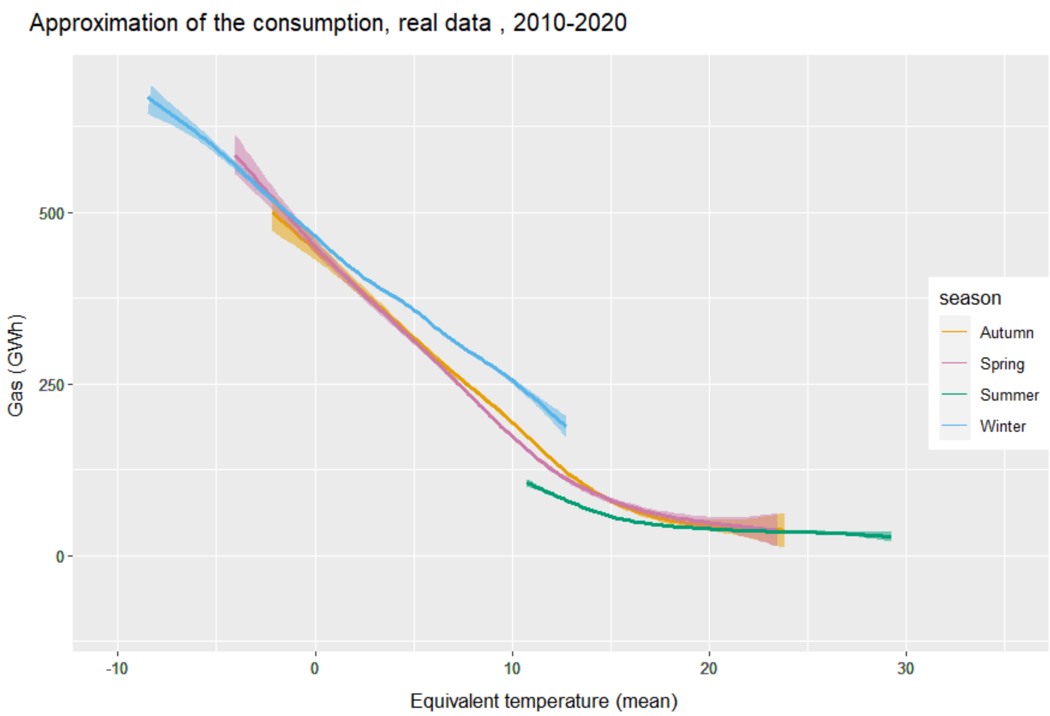

Figure 2.4.2: Approximation of Belgian gas consumption as a function of temperature (calculated based on the equivalent temperature formula "tempEq"). The year is split by season with default months.

compared to ∼ 190GWh for a very hot day in Winter which would also have an average temperature of 12°. This seems logical since in Winter the boilers are running at full capacity, whereas in Summer many of them are switched off.

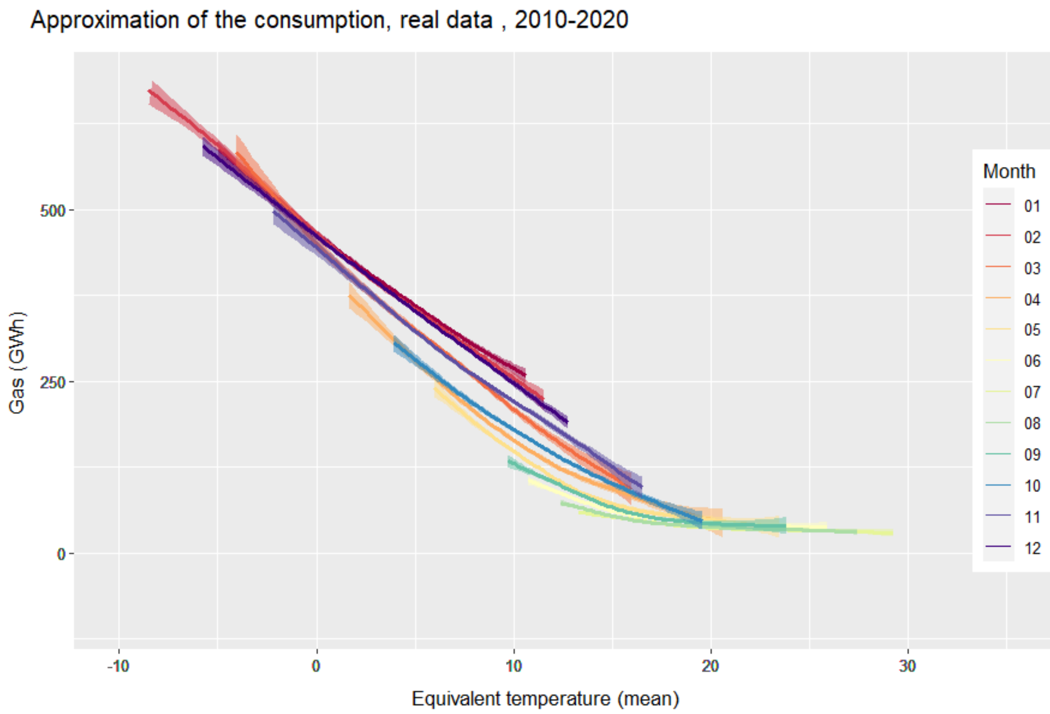

Figure 2.4.3: Approximation of Belgian gas consumption as a function of temperature (calculated based on the equivalent temperature formula "tempEq"). The year is split by month.

To better understand the impact of the time of year on the temperature/gas consumption relationship, one can look at a breakdown by month, as shown in figure 2.4.3.

In this figure, it can be seen that the curves for the months of December, January and February have the same behaviour. They are represented by a quasi-linear curve. Then, the months of November and March also have the same behaviour. They have a quasi-linear curve with a steeper slope than the previous one. The five following months progressively adopt the tendency of a curve of the second degree, with a positive concavity more and more weak. The curve for July and August months is almost flat at the warmest temperatures. September and October curves progressively go back to the shape of the curves of the colder months.

Looking at the contour map of the consumption, one can see the influence of the period of the year on the temperature/consumption relationship. This influence can be seen by looking at the blue contour lines.

A flat curve means that there is no dependence on the month. This is the case for temperatures about 2°. Other curve indicate that for the same global temperature, the consumption is higher in the Winter months.

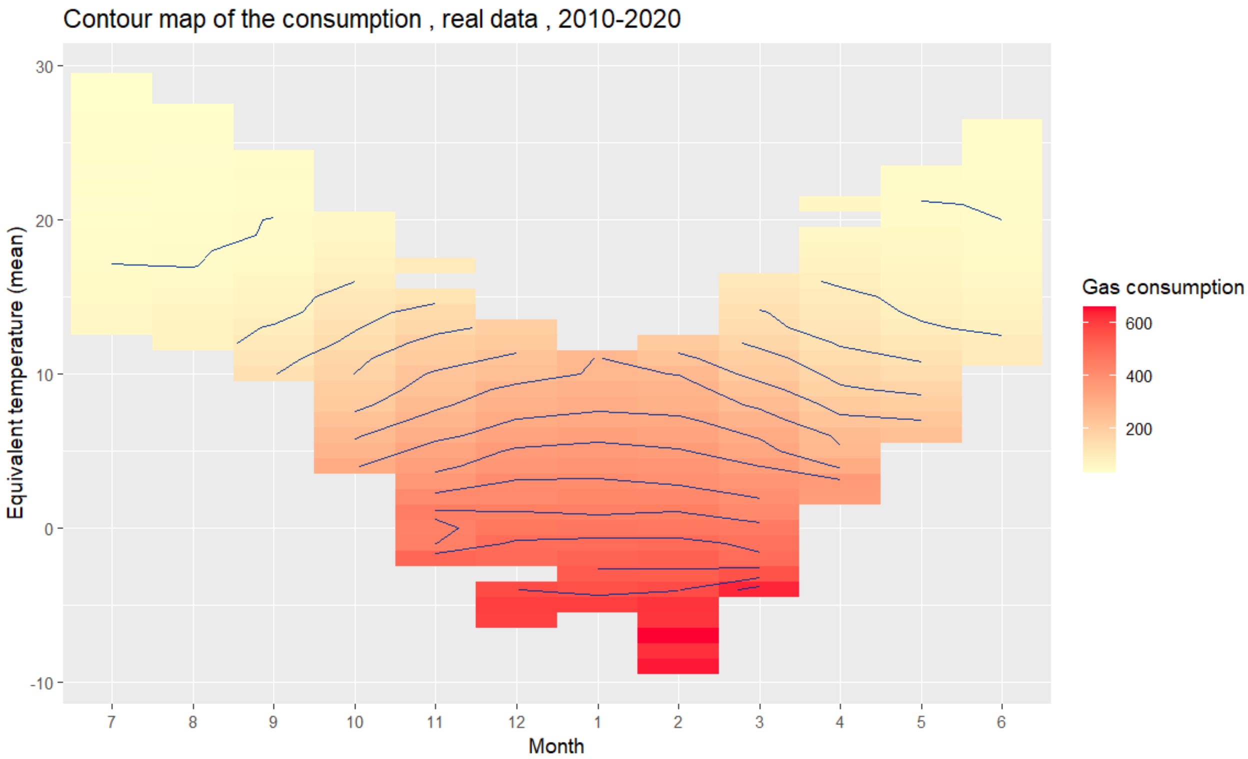

Figure 2.4.4: Level curves for Belgian gas consumption as a function of temperature (calculated based on the equivalent temperature formula "tempEq") and month of the year. In order to better observe the behaviour of the curves during cold periods, the winter months are located in the centre of the graph and the summer months at the ends.

For very low temperatures, the tendency seems to be reversing. But this result should be treated with caution as there is little data for such low temperatures, and they were mainly recorded in December and February.

The contour lines on the left of the graph are steeper than on the right. This suggests that autumn and spring do not have an exactly symmetrical pattern.

2.5 Descriptive analysis of data: Electricity

Let's look at the relationship between temperature and electricity consumption on figure 2.5.1. As for the gas, the curve initially behaves like a straight line at low temperatures. Its slope is about -2.5 GWh/°.

For higher temperatures it behaves like a function of the second degree. The break between the two shapes occurs at about 15° and 65 GWh. The electricity consumption rises for the warmest temperatures.

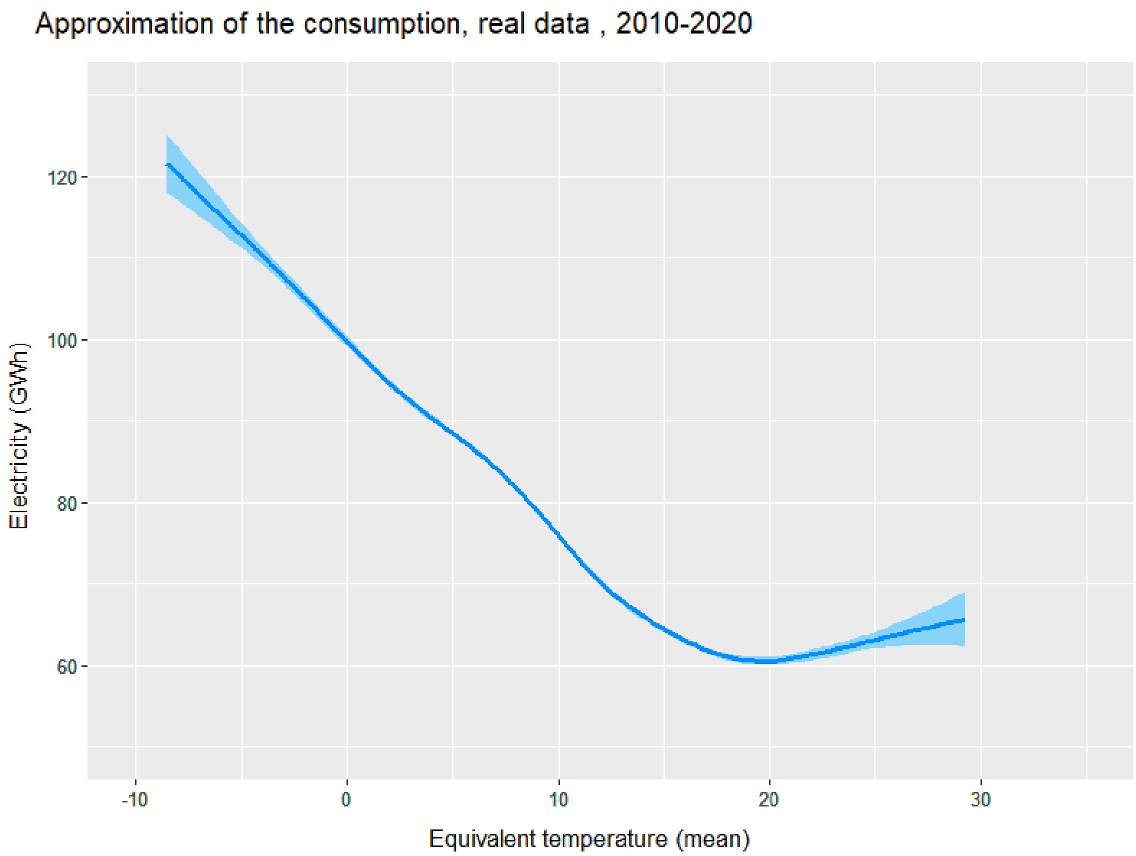

Figure 2.5.1: Approximation of Belgian electricity consumption as a function of temperature (calculated based on the equivalent temperature formula "tempEq"). The year is split by season with default months.

As for gas, the behaviour can be divided into several trends depending on the month of the year or the season (see appendix 9.2 for graphs).

By comparing the graphs 2.5.2 and 2.5.3 we can see the evolution of the temperature/consumption relationship with the years.

Globally, the consumption decreases for the same temperature. Let's take an example: for an average temperature of 10°:

- The months of January-February have a consumption ∼ 90 GWh in 2010-2012 against ∼78 GWh in 2017-2019.
- The months of March-April have a consumption ∼78 GWh in 2010-2012 against ∼70 GWh in 2017-2019.
- The months of May-June have a consumption ∼72 GWh in 2010-2012 against ∼68 GWh in 2017-2019.

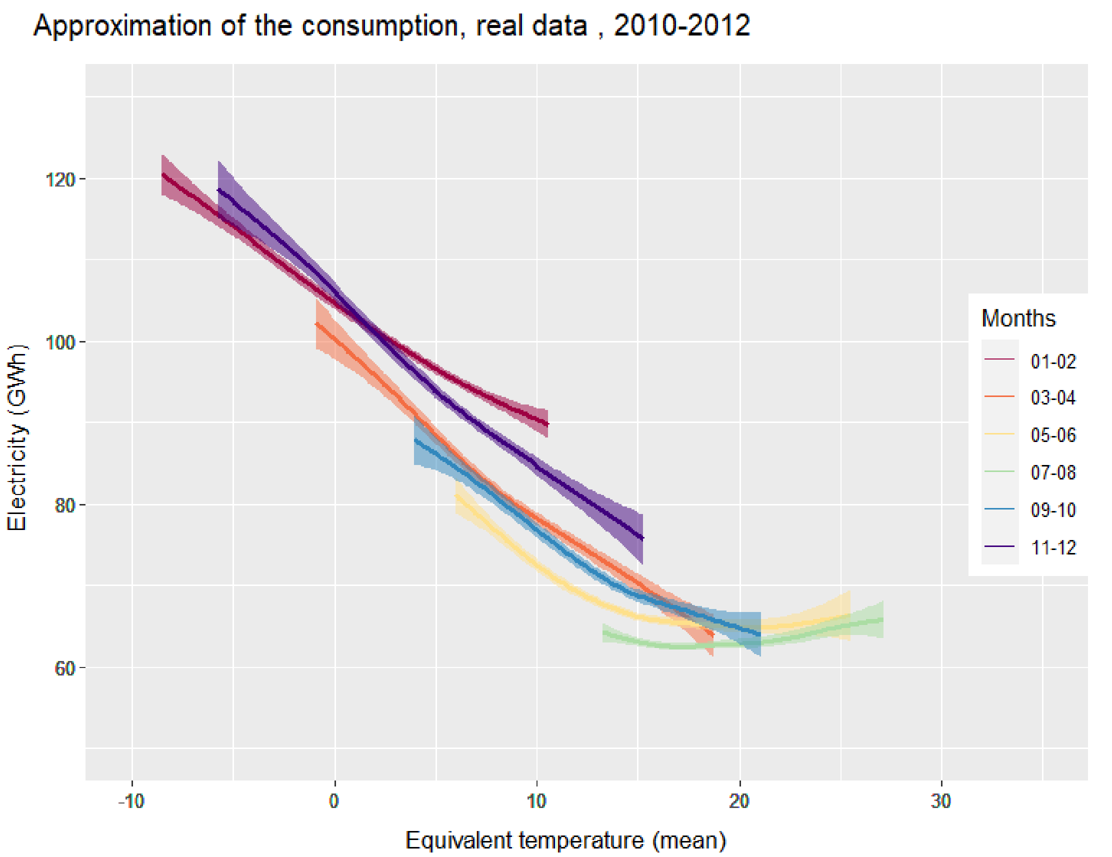

Figure 2.5.2: Approximation of Belgian electricity consumption as a function of temperature (calculated based on the equivalent temperature formula "tempEq"), for years 2010,2011 and 2012. The year is split into periods of 2 months.

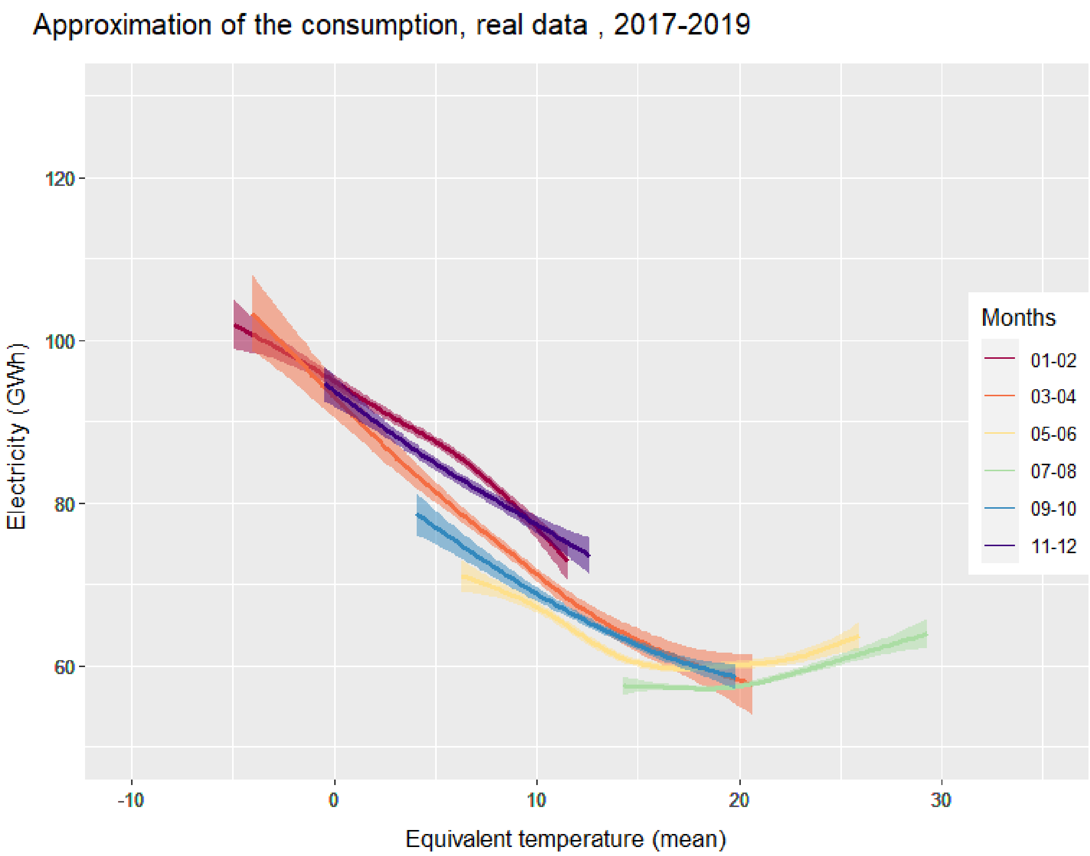

Figure 2.5.3: Approximation of Belgian electricity consumption as a function of temperature (calculated based on the equivalent temperature formula "tempEq"), for years 2017,2018 and 2019. The year is split into periods of 2 months.

- The months of September-October have a consumption ∼77 GWh in 2010-2012 against ∼69 GWh in 2017-2019.

- The months of November-December have a consumption ∼85 GWh in 2010-2012 against ∼78 GWh in 2017-2019.

There is no data at 10° for the months of July-August.

On the other hand, for the highest temperatures, the slope is steeper in 2017-2019 than in 2010-2012. This indicates an amplification of the phenomenon of consuming more at higher temperatures is amplified.

The influence of the period of the year on the temperature/consumption relationship seems to decrease. Indeed, the different curves are closer to each other on the graph 2.5.3.

2.6 Summary

The main factor influencing gas or electricity consumption is temperature. Overall, the colder it is, the higher the consumption. However electricity consumption can increase with high temperatures, probably due to air conditioning expansion [5].

Equivalent temperature is a measure used by the gas industry. It allows to take into account the influence of the previous days' temperatures.

Other factors such as the period of year or the time can also play a role. The pace of life can also influence consumption.

The relationship between temperature and gas consumption can differ depending on the time of year. With equivalent temperatures, Belgians consume more gas during a winter day than during a summer day. This phenomenon is also observed if we split the year into months.

Gas consumption has not changed much in recent years.

The relationship between temperature and electricity consumption may differ depending on the time of year. For equivalent temperatures, Belgians consume more electricity during a winter day than during a summer day.

Over the years, the overall electricity consumption seems to decrease. However the electricity consumption on high temperature days is increasing.

The smooth functions on the data and confidence intervals are made from the gam() function of the mgcv package.

This function has the particularity of being a happy medium between interpretability and flexibility. Various options are available to adjust several parameters. The idea is to find a model that takes into account covariates (like temperature or time of year) in a single model for both energies, with relevant parameters.

Before carrying out a range exploration, some theory is needed.

Chapter 3

Theory

3.1 GAMs

GAM stands for generalized additive model.

3.1.1 GLMs

GLM stands for generalized linear model. It is a flexible generalization of ordinary linear regression that allows for the response variable to have an error distribution other than the normal distribution.

A generalized linear model has the basic structure

$$g(\mu_i) = X_i\beta, \tag{3.1.1}$$

where g is a smooth monotonic function called "link function",
$\mu_i \equiv \mathbb{E}(Y_i)$,
β is a vector of unknown parameters,
Y_i are independent and

$$Y_i \sim \text{ some exponential family distribution.}$$

3.1.2 GAMs

A generalized additive model is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates. [8]

It has the basic structure

$$g(\mu_i) = A_i\theta + \sum_j f_j(\bar{x}_{ji}) \tag{3.1.2}$$

where g is a smooth monotonic function called link function,
$\mu_i \equiv \mathbb{E}(Y_i)$,
A_i is a row of the model matrix for any strictly parametric model components,
θ a vector of unknown parameters,

f_j are unknown smooth functions of (possibly vector) covariate x_{k_j} ,
and

$$Y_i \sim \text{ some exponential family distribution.}$$

An example is

$$g(\mu_i) = X_i\theta + f_1(x_{1i} + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + ... \tag{3.1.3}$$

The idea is to approximate each function with a basis expansion:

$$f(x) = \sum_{k=1}^{K} \beta_k b_k(x)$$

with b_k a spline.

3.2 Splines

Let the data be n pairs (x_i, y_i). The idea is to find a smooth looking function that fits the data. In regression problems , using polynomial bases with global representations have undesirable side effects: they tend to be erratic at the boundaries of the data and each observation affects the entire curve. To remedy this, splines are introduced. They are piecewise polynomials joined together to make a single smooth curve.

3.2.1 One predictor variable

Definition 3.1. *Let $a = t_0 < t_1 < ... < t_n = b$. A spline $S(t)$ of degree m on $[a, b]$ is a function of class $C^{m-1}[a, b]$ such that the restriction of S at each interval $[t_i, t_{i+1}]$ is a polynomial of degree $\leq m$. Let $S^m(a, b)$ be the vector space of splines of degree k over $[a, b]$.*

3.2.1.1 B-Splines

To define a B-spline basis, one needs a non decreasing sequence $t := (t_i)$, called knot sequence .

The B-splines of order 1 for this knot sequence are the functions

$$B_{i1}(t) := X_i(t) := \begin{cases} 1, & \text{if } t_i \leq t < t_{i+1} \\ 0, & \text{otherwise.} \end{cases} \tag{3.2.1}$$

The X_i vanish outside the interval $[t_i, t_{i+1}]$

From these first-order B-splines, we obtain higher order B-slpines by recurrence:

$$B_{ik} := w_{ik}B_{i,k-1} + (1 - w_{i+1,k})B_{i+1,k-1} \tag{3.2.2}$$

with

$$w_{ik}(t) := \begin{cases} \frac{t-t_i}{t_{i+k-1}-t_i}, & \text{if } t_i \neq t_{i+k-1} \\ 0, & \text{otherwise.} \end{cases} \tag{3.2.3}$$

$$\tag{3.2.4}$$

The B_{ik} vanish outside the interval $[t_i, t_{i+k}]$

After k-1 steps of the recurrence, one obtains b_{jk} which are polynomials of degree k-1 such that:

$$B_{ik} = \sum_{j=i}^{i+k-1} b_{jk} X_j, \tag{3.2.5}$$

For example if k=3 and t:

$$B_{i3} = w_{i3} w_{i2} X_i + (w_{i3}(1 - w_{i+1,2}) + (1 - w_{i+1,3}) w_{i+1,2}) X_{i+1} + (1 - w_{i+1,3})(1 - w_{i+2,2}) X_{i+2}$$

where w_{ik} are linear polynomials.

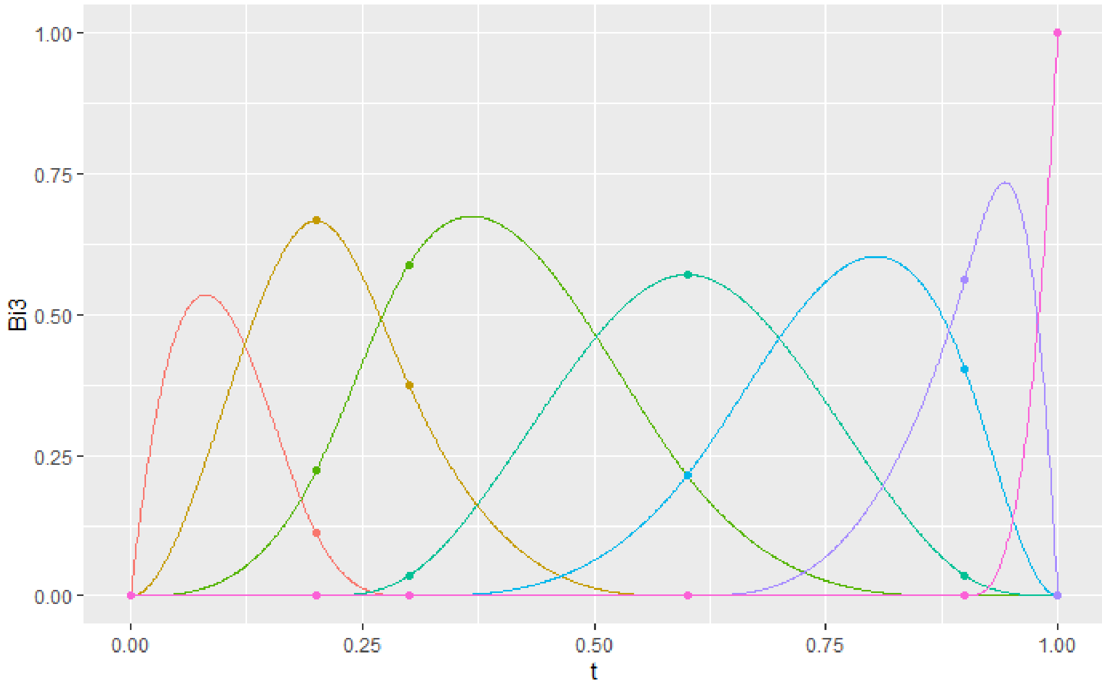

Figure 3.2.1: All possible Example of cubic B-splines (order=4) for t_i=(0, 0, 0, 0.2, 0.3, 0.6, 0.9, 1, 1, 1, 1).

If $t_i \neq t_{i+k-1}$ and $t_{i+1} \neq t_{i+k}$, once can rewrite (3.2.2) and (3.2.3) as

$$B_{ik}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(t). \tag{3.2.6}$$

B_{ik} vanishes outside the interval $[t_i, t_{i+k}]$. B-splines of order k are piecewise polynomial of degree $<$ k. Moreover they form a partition of unity :

$$\sum_{i=0}^{n-1} B_{ik}(t) = 1, \text{ for } t_{k-1} \leq t < t_n \tag{3.2.7}$$

A B-spline of order k is a B-spline of degree m, with $k = m + 1$. The basis functions are strictly local: ecah basis function is only non-zero over the intervals between k+1 adjacent knots.

In practice, cubic splines are the most common type of spline. It is claimed that they are the lowest order spline for wich the discontinuity at the knots cannot be noticed by the human eye. [2]

A spline of order k with knot sequence t can be written as a linear combination of B-splines: s=$\sum_i B_{ik} a_i$. The collection of all such spline can be written as

$$S_{k,t} := \left\{ \sum_i B_{ik} a_i : a_i \in \mathbb{R} \right\}.$$

3.2.1.2 P-Splines

"B-splines were developed as a very stable basis for large scale spline interpolation, but for most statistical work with low rank[1] penalized regression splines, you would have to be using very poor numerical methods before the enhanced stability of the basis became noticeable. The real statistical interest in B-splines has resulted from the work of Eilers and Marx [3] in using them to develop what they term P-splines." [8]

P-splines (Eilers and Marx)[3] are combination of two ideas for curve fitting: **regression on a basis of B-splines and a difference penalty on the regression coefficients**. The penalty are based on finite differences $\Delta^d \alpha$ where

$$\Delta \alpha_j = \alpha_j - \alpha_{j-1},$$

$$\Delta^2 \alpha_j = \Delta(\Delta \alpha_j) = \alpha_j - \alpha_{j-1} - (\alpha_{j-1} - \alpha_{j-2}) = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$$

$$\Delta^3 \alpha_j = \alpha_j - 3\alpha_{j-1} + 3\alpha_{j-2} + \alpha_{j-3}$$

and so on for higher d.

Let the data be n pairs (x_i, y_i) and B, computed on x and using **equally-spaced knots.** The idea is to minimize the following objective function:

$$Q_B = \|y - B_\alpha\|^2 + \lambda \|D_d \alpha\|^2 \tag{3.2.8}$$

with

$$D_d \alpha \equiv \Delta^d \alpha.$$

Minimisation of (3.2.8) leads to the following system:

$$\hat{\alpha} = (B^T B + \lambda D_d^T D_d)^{-1} B^T y.$$

A important component of P-splines is number and position of knots.

[1] There are far fewer coefficients than there are data to smooth.

3.2.1.3 Penalized regression splines

A good solution is a middle ground between fitting the data and having a relatively smooth form. Let's introduce the function $\hat{f}$, minimizing the following function:

$$\sum_i (y_i - f(x_i))^2 + \lambda P(f) \tag{3.2.9}$$

with λ a parameter of smoothing.

This function is called a penalized spline because of $P(f)$, which represents the wiggliness. The first term of (3.2.9) leads to a good fitting and the second term leads to a good smoothing. The larger λ is, this smoother the solution will be.

If splines have one basis function per data point, this is computationally wasteful. Indeed, the effective degrees of freedom, controlled by λ, will be much smaller than this. A solution to decrease the number of basis functions is penalized regression splines. This involves constructing a spline basis (and associated penalties) for a much smaller data-set than the one to be analyzed and then using this to model the original data set. There are 2 types of approach:

- Knot based: Choose a representative subset of the data: the knots. Then create the spline basis as if smoothing only those data.
- Eigen based: Choose how many basis function are to be used and then solve the problem of finding the set of this many basis functions that will optimally approximate a full spline

3.2.1.4 Cubic regression splines

An exemple of knot based approach of penalized regression spline is the cubic regression splines.

The wiggliness for a cubic spline can be represented by:

$$P(f) = \int f''(x)^2 dx. \tag{3.2.10}$$

The idea is to parametrize the spline terms of its value at the knots $x_1, ...x_k$. Let $\beta_j = f(x_j)$ and $\delta_j = f''(x_j)$. The cubic spline can be written as

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1} \text{ if } x_j \leq x \leq x_{j+1} \tag{3.2.11}$$

with

- $a_j^-(x) = \frac{x_{j+1}-x}{h_j}$
- $a_j^+(x) = \frac{x-x_j}{h_j}$
- $c_j^-(x) = \frac{1}{6}[\frac{(x_{j+1}-x)^3}{h_j} - h_j(x_{j+1} - x)]$
- $c_j^+(x) = \frac{1}{6}[\frac{(x-x_j)^3}{h_j} - h_j(x - x_j)]$

- $h_j = x_{j+1} - x_j$

Let's $\delta_1 = \delta_k = 0$ and $\delta^- = (\delta_2, ..., \delta_{k-1})^T$.

The condition that the spline must be continuous to second derivative at the x_j is obvious since $f_L^{''}(x_j) = \delta_j = f_R^{''}(x_j)$.

The condition that the spline must be continuous to first derivative can be written $f_L^{'}(x_j) = f_R^{'}(x_j)$, for $j = (2, ...k-1)$.

Since x_j is the first bound of the interval $[x_j, x_{j+1}]$,

$$f_R^{'}(x_j) = \beta_j \frac{-1}{h_j} + \beta_{j+1} \frac{1}{h_j} + \delta_j \frac{-h_j}{3} + \delta_{j+1} \frac{-h_j}{6}. \tag{3.2.12}$$

for j=(1,...,k).
Since x_j is the first bound of the interval $[x_{j-1}, x_j]$,

$$f_L^{'}(x_j) = \beta_{j-1} \frac{-1}{h_{j-1}} + \beta_j \frac{1}{h_{j-1}} + \delta_{j-1} \frac{h_{j-1}}{6} + \delta_j \frac{h_{j-1}}{3}. \tag{3.2.13}$$

for j=(2,...,k-1).

One can rewrite the first derivative condition with (3.2.13) and (3.2.12) :

$$\delta_{j-1} \frac{h_{j-1}}{6} + \delta_j \frac{h_j + h_{j-1}}{3} + \delta_{j+1} \frac{h_j}{6} = \frac{-\beta_j}{h_j} + \frac{\beta_{j+1}}{h_j} + \frac{\beta_{j-1}}{h_{j-1}} - \frac{\beta_j}{h_{j-1}}. \tag{3.2.14}$$

The equation above could be rewritten as

$$B\delta^- = D\beta \tag{3.2.15}$$

with B a $[(k-2) \times (k-2)]$ matrix and D a $[(k-2) \times (k)]$ matrix where B can b written as:

$$\begin{pmatrix}
\frac{1}{3}(h_2 + h_1) & \frac{1}{6}(h_2) & 0 & & & \dots & & 0 \\
\frac{1}{6}(h_2) & \frac{1}{3}(h_3 + h_2) & \frac{1}{6}(h_3) & 0 & \dots & & & 0 \\
0 & & & & & & & \dots \\
\dots & & & & & & & 0 \\
0 & & \dots & & 0 & \frac{1}{6}(h_{k-3}) & \frac{1}{3}(h_{k-2} - h_{k-3}) & \frac{1}{6}(h_{k-2}) \\
0 & & \dots & & & 0 & \frac{1}{6}(h_{k-2}) & \frac{1}{3}(h_{k-1} - h_{k-2})
\end{pmatrix}$$

and D can be written as:

$$\begin{pmatrix} \frac{1}{h_1} & -(\frac{1}{h_1}+\frac{1}{h_2}) & \frac{1}{h_2} & 0 & & & \dots & & 0 \\ 0 & \frac{1}{h_2} & -(\frac{1}{h_2}+\frac{1}{h_3}) & \frac{1}{h_3} & 0 & & & & \\ \dots & & & & & & & & \\ & & & & 0 & \frac{1}{h_{k-3}} & -(\frac{1}{h_{k-3}}+\frac{1}{h_{k-2}}) & \frac{1}{h_{k-2}} & 0 \\ 0 & & \dots & & & 0 & \frac{1}{h_{k-2}} & -(\frac{1}{h_{k-2}}+\frac{1}{h_{k-1}}) & \frac{1}{h_{k-1}} \end{pmatrix}.$$

In order to rewrite the spline in terms of β, one can define

$$F = \begin{bmatrix} 0 \\ B^{-1}D \\ 0 \end{bmatrix}$$

where 0 is a row of zeros. Since $\delta = F\beta$ the result is :

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)F_j\beta + c_j^+(x)F_{j+1}\beta \text{ if } x_j \leq x \leq x_{j+1} \tag{3.2.16}$$

which is in the form

$$f(x) = \sum_{i=1}^{k} b_i(x)\beta_i.$$

Hence, there is an implicit definition of new basis functions $b_i(x)$.

It can be shown [8] that

$$\int_{x_1}^{x_k} f''(x)^2 dx = \beta^T D^T B^{-1} D\beta. \tag{3.2.17}$$

To get $P(f)$ in a convenient form, one can define matrix $S \equiv D^T B^{-1} D$. Then

$$P(f) = \beta^T S\beta.$$

3.2.1.5 Cyclic cubic regression splines

A cyclic model has the same value at its upper and lower boundaries. This may be relevant for a multi-year model with seasonality as it is the case for our data. The construction of a cyclic cubic spline is based on the same principle as the cubic spline, but now $\beta_1 = \beta_k$ and $\delta_1 = \delta_k$.

Let's $\beta = (\beta_1, ..., \beta_{k-1})^T$ and $\delta = (\delta_1, ..., \delta_{k-1})^T$.

The condition that the spline must be continuous to second derivative at the x_j is obvious since $f_L''(x_j) = \delta_j = f_R''(x_j)$.

The condition that the spline must be continuous to first derivative $f_L'(x_j) = f_R'(x_j)$ at the x_j for $j = (2, ..., k_1)$ remain the same and (3.2.14) is still valid.

But now, a new condition is:

$$f_L'(x_k) = f_R'(x_1). \tag{3.2.18}$$

With (3.2.13) (3.2.12), one can rewrite (3.2.18) as (it is recalled that β_k=β_1 and $\delta_k = \delta_1$)

$$\beta_1 \frac{-1}{h_1} + \beta_2 \frac{1}{h_1} + \beta_{k-1} \frac{1}{h_{k-1}} + \beta_1 \frac{-1}{h_{k-1}} = \delta_1 \frac{h_1}{3} + \delta_2 \frac{h_1}{6} + \delta_{k-1} \frac{h_{k-1}}{6} + \delta_1 \frac{h_{k-1}}{3}. \tag{3.2.19}$$

All this leads to

$$\tilde{B}\delta = \tilde{D}\beta \tag{3.2.20}$$

with $\tilde{B}$ and $\tilde{D}$ some $[(k-1) \times (k-1)]$ matrix where $\tilde{B}$ can be written as

$$\begin{pmatrix}
\frac{h_{k-1}+h_1}{3} & \frac{h_1}{6} & 0 & & \ldots & & 0 & \frac{h_{k-1}}{6} \\
\frac{h_1}{6} & \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & & \ldots & & 0 \\
0 & & & & & & & \\
\ldots & & & & & & & 0 \\
0 & & \ldots & & 0 & \frac{h_{k-3}}{6} & \frac{h_{k-3}+h_{k-2}}{3} & \frac{h_{k-2}}{6} \\
\frac{h_{k-1}}{6} & 0 & \ldots & & & 0 & \frac{h_{k-2}}{6} & \frac{h_{k-2}+h_{k-1}}{3}
\end{pmatrix}$$

and $\tilde{D}$ can be written as

$$\begin{pmatrix}
-\frac{1}{h_1} - \frac{1}{h_{k-1}} & \frac{1}{h_1} & 0 & & \ldots & & 0 & \frac{1}{h_{k-1}} \\
\frac{1}{h_1} & -\frac{1}{h_2} - \frac{1}{h_1} & \frac{1}{h_2} & 0 & & \ldots & & 0 \\
0 & & & & & & & \ldots \\
\ldots & & & & & & & 0 \\
0 & & \ldots & & 0 & \frac{1}{h_{k-3}} & -\frac{1}{h_{k-2}} - \frac{1}{h_{k-3}} & \frac{1}{h_{k-2}} \\
\frac{1}{h_{k-1}} & 0 & & \ldots & & 0 & \frac{1}{h_{k-2}} & -\frac{1}{h_{k-1}} - \frac{1}{h_{k-2}}
\end{pmatrix}.$$

Similar reasoning to that employed for the cubic spline leads to

$$f(x) = \sum_{i=1}^{k-1} \tilde{b}_i(x)\beta_i. \tag{3.2.21}$$

Hence, there is an implicit definition of new basis functions $\tilde{b}_i(x)$. And the wiggliness for a cubic cyclic spline can be write:

$$P(f) = \beta^T \tilde{S} \beta \tag{3.2.22}$$

with $\tilde{S} \equiv \tilde{D}^T \tilde{B}^{-1} \tilde{D}$.

3.2.2 Multiple predictor variables

3.2.2.1 Thin plate splines

In order to estimate a smooth function of multiple predictor variables, let's introduce thin-plate spline (TPS) [10]. The idea is to find the function $\hat{f}$ minimizing:

$$||y - f||^2 + \lambda J_{md}(f) \tag{3.2.23}$$

where y is the vector of y_i data and $f = (f(x_1), f(x_2), ..., f(x_n))^T$. $J_{md}(f)$ is a penalty functional measuring the wiggliness of f.

$J_{md}(f)$ is the equivalent of $P(f)$ in the univariate case (formula (3.2.9)).

The wiggliness penalty is defined as

$$J_{md}(f) = \int ... \int_{\mathbb{R}^d} \sum_{\nu_1+...+\nu_d=m} \frac{m!}{\nu_1!...\nu_d!} \left(\frac{\partial^m f}{\partial x_1^{\nu_1} ... \partial x_d^{\nu_d}} \right)^2 dx_1...dx_d. \tag{3.2.24}$$

For exemple, in 2 dimensions and for $m = 2$, it gives:

$$J_{22} = \int \int \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 dx_1 dx_2. \tag{3.2.25}$$

And in 1 dimension, when $m = 2$, one returns to the penalty expressed above (3.2.10) :

$$J_{21} = \int \left(\frac{\partial^2 f}{\partial x^2} \right)^2 dx. \tag{3.2.26}$$

For "visually smooth" results it is preferable that $2m > d + 1$. [8]

Thin plate splines are **isotropic smoothers**. They are good for smooth interactions of quantities measured in same units, such as spatial co-ordinates, where isotropy is appropriate.

It can be show [10] [6] that the function minimizing (3.2.23) has the form:

$$f(x) = \sum_{i=1}^{n} \delta_i \eta_{md}(||x - x_i||) + \sum_{j=1}^{M} \alpha_j \phi_j(x) \tag{3.2.27}$$

where δ and α are unknown parameter vectors of respective dimensions $[n \times 1]$ and $[M \times 1]$ with $M = \binom{m+d-1}{d}$. δ is subject to the linear constraints that $T^T \delta = 0$ with T defining by $T_{ij} \equiv \phi_j(x_i)$ of dimension $[n \times M]$.

The null space of the penalty functional J_{md} is the M-dimensional space spanned by the polynomials in d variables of total degree $\leq m-1$. For example, for $d=2$ and $m=2$, $M=3$ and the null space is spanned by ϕ_1, ϕ_2 and ϕ_3 given by

$$\phi_1(x) = 1, \phi_2(x) = x_1, \phi_3(x) = x_2.$$

In general, $\phi_1, ...\phi_M$ denote the M monomials of total degree $<$ m.

Let's define the matrix E of dimension $[n \times n]$ is defining by $E_{ij} \equiv \eta_{md}(||x_i - x_j||)$ with

$$\eta_{md}(r) = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-\frac{d}{2})!}|r|^{2m-d} ln(|r|) & \text{d even,} \\ \\ \frac{\Gamma(\frac{d}{2}-m)}{2^{2m}\pi^{d/2}(m-1)!)}|r|^{2m-d} & \text{d odd.} \end{cases} \tag{3.2.28}$$

Hence the spline fitting problem can be reduced to find α and δ minimizing

$$||y - E\delta - T\alpha||^2 + \lambda\delta^T E\delta \text{ subject to } T^T\delta = 0. \tag{3.2.29}$$

"The thin plate spline, $\hat{f}$, is something of an ideal smoother: it has been constructed by defining exactly what is meant by smoothness, exactly how much weight to give to the conflicting goals of matching the data and making $\hat{f}$ smooth, and finding the function that best satisfies the resulting smoothing objective." [8]

3.2.2.2 Thin plate regression splines

The problem with thin plate splines is the computational cost. As said above, this cost his wasteful. This is why thin plate regression splines (TPRS) are introduced. They are an eigen based approximation of TPS : one chooses how many basis functions k are to be used and then one solves the problem of finding the set of this many basis functions that will optimally approximate a full spline.

The principle is to start with the basis and penalty for TPS and then truncating this basis to obtain a **low rank isotropic smoother**. There are far fewer coefficients than there are data to smooth. The basis truncation does not change the meaning of the thin plate spline penalty in the sense that it penalizes exactly what it would have penalized for a full thin plate spline. [9]

More precisely, the idea is truncating the space of the wiggly components (with parameters δ) by making an eigen decomposition of E in UDU^T [2] with D diagonal matrix arranged so that $|D_{i,i}| \geq |D_{i-1,i-1}|$.

Then, let U_k be the matrix of dimension $[n \times k]$ corresponding to the first columns of U and D_k be the matrix of dimension $[k \times k]$ corresponding to the top left submatrix of D. (3.2.29) becomes

$$||y - U_k D_k U_k^T \delta - T\alpha||^2 + \lambda\delta^T U_k D_k U_k^T \delta \text{ subject to } T^T\delta = 0. \tag{3.2.30}$$

[2]E is symmetric so $U^T = U^{-1}$

Let's δ_k be $U_k^T\delta_k$. (3.2.30) can be rewrite and finding $\hat{f}$ means find α and δ_k minimizing:

$$||y - U_k D_k \delta_k - T\alpha||^2 + \lambda \delta_k^T D_k \delta_k \text{ subject to } T^T U_k \delta_k = 0 \tag{3.2.31}$$

with δ_k a $[k \times 1]$ vector instead of δ a $[n \times 1]$ vector.

"Thin plate regression splines are probably the best that can be hoped for in terms of approximating the behaviour of a thin plate spline using a basis of any given low rank. They have the nice property of avoiding having to choose 'knot locations', and are reasonably computationally efficient". [8]

3.2.2.3 Tensor products

TPS and TPRS smooths treat the wiggliness equally in all directions. These approaches can be poor if there is no isotropy in the data. This is why tensor product smooths are introduced.

The idea is to start by assuming that one has known low rank bases available for representing smooth functions f_x, f_y and f_z of the covariates x, y and z. These functions are written as:[3]

$$f_x(x) = \sum_{i=I}^{I} \alpha_i a_i(x), \;\; f_y(y) = \sum_{j=1}^{J} \beta_j b_j(y), \text{ and } f_z(z) = \sum_{k=1}^{K} \gamma_k c_k(z) \tag{3.2.32}$$

Then, one converts f_x into a smooth function of x and y. This can be achieved by allowing parameters α_i to vary smoothly with y:

$$\alpha_i(y) = \sum_{j=1}^{J} \beta_{ij} b_j(y) \tag{3.2.33}$$

which gives

$$f_{xy}(x,y) = \sum_{i=1}^{I}\sum_{j=1}^{J} \beta_{ij} b_j(y) a_i(x). \tag{3.2.34}$$

Following the same argument, one gets

$$f_{xyz}(x,y,z) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \gamma_{ijk} c_k(z) b_j(y) a_i(x). \tag{3.2.35}$$

For any particular set observation of x,y and z, let's $\mathbf{X}$ be the model matrix evaluating the tensor product smooth at these observation and let's $\mathbf{X}_x$, $\mathbf{X}_y$, and $\mathbf{X}_z$ be the model matrices that would evaluate the marginal smooth at the same observations. If γ_{ijk} is ordered appropriately into a vector γ, then one can write the i^{th} row of $\mathbf{X}$ as:

$$\mathbf{X_i} = \mathbf{X}_{xi} \otimes \mathbf{X}_{yi} \otimes \mathbf{X}_{zi}.$$ [4]

This give a "tensor product" basis for representing smooth functions.

[3] The methods works for any number of covariates but the development is done for 3 covariates for a good understanding.

[4] $\otimes$ is the Kronecker product

Now the idea is having a way of measuring wiggliness. Supposing that each marginal smooth has an associated functional that measures function wiggliness expressible as a quadtratic form:

$$J_x(f_x) = \alpha^T S_x \alpha,\ J_y(f_y) = \beta^T S_y \beta \text{ and } J_z = (f_z) = \gamma^T S_z \gamma, \tag{3.2.36}$$

a way of measuring wiggliness of f_{xyz} is [8]:

$$J(f_{xyz}) = \lambda_x \int_{y,z} J_x(f_{x|yz})dydz + \lambda_y \int_{x,z} J_y(f_{y|xz})dxdz + \lambda_z \int_{x,y} J_z(f_{z|xy})dxdy \tag{3.2.37}$$

with $f_{x|yz}$ (resp. $f_{y|xz}$ and $f_{z|xy}$) considered as a function of x only (resp. of y only and of z only), with other covariates held constant.

For example, with the penalties

$$J_x(f_x) = \int \left(\frac{\partial^2 f_x}{\partial x^2}\right)^2 dx,\ J_y(f_y) = \int \left(\frac{\partial^2 f_y}{\partial y^2}\right)^2 dy, \text{ and } J_z(f_z) = \int \left(\frac{\partial^2 f_z}{\partial z^2}\right)^2 dz \tag{3.2.38}$$

the wiggliness of f_{xyz} is:

$$J(f_{xyz}) = \int_{x,y,z} \lambda_x \left(\frac{\partial^2 f_x}{\partial x^2}\right)^2 + \lambda_y \left(\frac{\partial^2 f_y}{\partial y^2}\right)^2 + \lambda_z \left(\frac{\partial^2 f_z}{\partial z^2}\right)^2 dxdydz. \tag{3.2.39}$$

Tensors smooths are invariant to rescaling of the covariates.

3.2.3 Summary

The aim is fitting data (x_i, y_i) with a smooth function. For this purpose, splines are introduced. They are smooth functions composed of piecewise polynomials. The fitting function has to be a good balance between fitting the data and be smooth enough (avoid overfitting). For this purpose, the notion of penalty is introduced, weighted by a smoothing parameter λ.

In one dimension, penalized spline are the minimizer of $\sum_i (y_i - f(x_i))^2 + \lambda P(f)$ with P(f) the representation of the wiggliness. The cubic regression splines are an example of knot based derivative approach of penalized spline. They have directly interpretable parameters but require the choice of knots. A cyclic cubic spline is similar but starts and ends at the same point.

P-splines are based on B-splines with equidistant knot. Their principle is to to penalize the coefficients of the spline directly, rather than penalizing something like $P(f)$.

Thin plate smooth are generalization of penalized spline in several dimension. Thin plate regression splines are an eigen-based law rank version of TPS. They do not need any knot selection. They can smooth with respect to any number of covariates and any penalty order. They are more relevant in case of isotropy in the data (e.g. when manipulating geographic coordinates).

In several dimensions, an alternative to TPRS is using tensor product . They are a best to smooth interactions of quantities measured in different units or where different degrees of smoothness are appropriate relative to different covariates.

The MGCV package uses splines to fit data in the form of a generalised additive model. Let's look at the different mgcv options in practice on a sample of our data.

Chapter 4

GAM in practice: gam() from MGCV

In the MGCV package, the gam() function aims to approximate data depending on a GAM model. MGCV stands for "Mixed GAM Computation Vehicle".

Recall that GAM can be estimated as a sum of smooth functions f_j (splines):

$$g(\mu_i) = A_i^\theta + \sum_j f_j(x_{ji}) \tag{4.0.1}$$

where g is a smooth monotonic function called link function,
$\mu_i \equiv \mathbb{E}(Y_i)$,
A_i is a row of the model matrix for any strictly parametric model components,
θ a vector of unknown parameters,
f_j are unknown smooth functions of (possibly vector) covariate x_{k_j} ,
and

$$Y_i \sim \text{ some exponential family distribution.}$$

And recall that splines can be rewritten with simpler functions called basis functions, wich can be expressed as follow:

$$f(x) = \sum_{k=1}^{K} \beta_k b_k$$

for b_k a basis function and β_k the associated coefficient.

GAM models are a middle ground between flexibility and interpretability. Indeed, some machine learning models like neural networks or regression trees are very flexible and good at making predictions. This being said, they are difficult to interpret and this is why they are sometimes called black boxes. On the other hand, linear models are easy to understand but fail to represent some complex phenomena.

In order to explore gam() and its functionalities, let's focus on an example: the data of gas consumption for years 2010, 2011 and 2012 in May. Let's begin with smoothing of a single covariate:

$$\text{GWhGas}_i = f(\text{tempEq}_i) + \epsilon_i \text{ where } \epsilon_i \sim \mathbb{N}(0, \sigma)$$

The default functions in gam() are the following:

- the method used to choose λ is by cross validation: **"GCV"**
- the number k of basis is 10
- the basis spline is the **"tp"** (for thin plate regression spline basis)
- the family distribution is **gaussian**

4.1 How to choose λ?

λ controls the wiggliness of the fitted function. Too big λ yield excessively smooth models and their associated loss of information from the data.Too small λ yield overfitting. Figure 4.1.1 shows model fits for different λ. As can be observed, $\lambda = 10$ and $\lambda = 1$ seem to be too large whereas $\lambda = 10^{-7}$ appears to overfit the data. $\lambda = 10^{-2}$ seems to be correct.

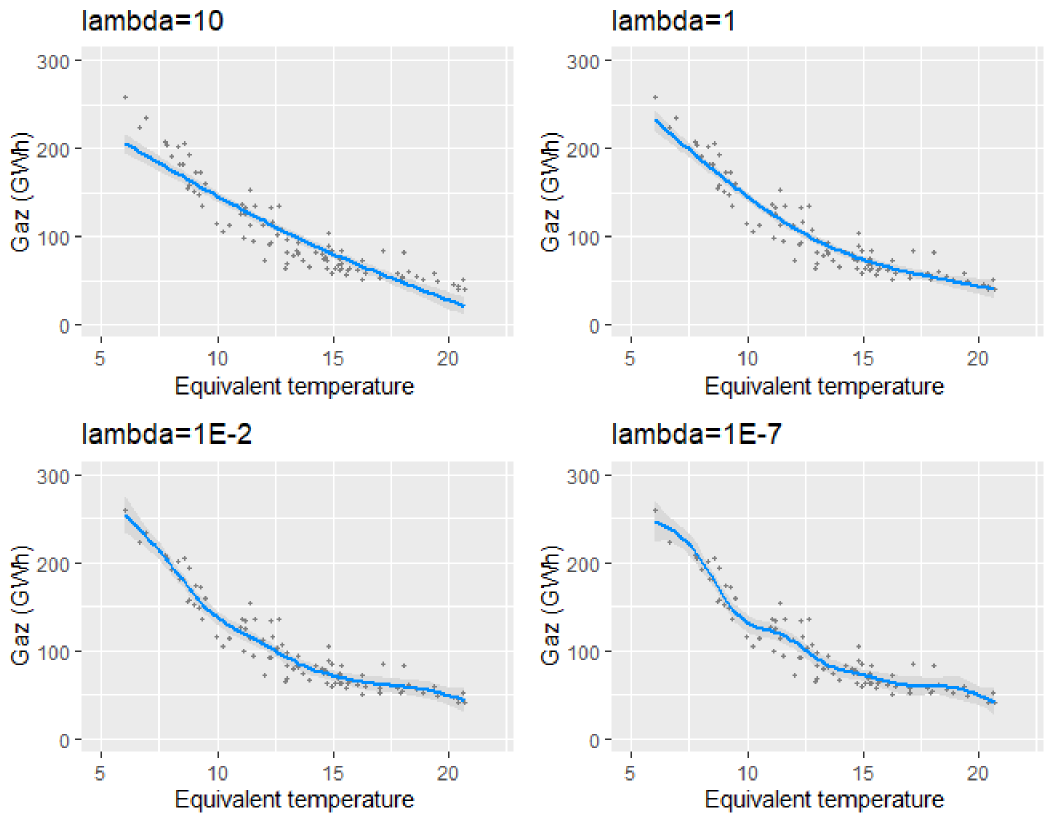

Figure 4.1.1: Models fitted with different λ. Data of Gas consumption, from May, 2010-2012

If one wants to avoid an arbitrary choice, there are methods to help select a relevant lambda. There are two ways of choosing: **predictive methods** or **bayesian methods**.

"GCV" and **"GACV"** are predictive methods. GCV stands for generalized cross validation while GACV stands for generalized approximate cross validation. Predictive techniques are based on minimising out-of-sample prediction errors. As they do not have an independent sample, they leave out each observation in turn. This is the principle of cross validation. Generalized cross validation uses a matrix representation of the problem to avoid high computational cost. GACV is is the version of GCV extended to non-Gaussian models.

REML stands for "restricted maximum likelihood" while **ML** stands for "maximum likelihood". They can also be considered as Bayesian methods. In this context, they postulate that the reason for smoothing data is because the true shape of the curve is rather smooth than wiggly. In this logic, the smoothing penalty corresponds to a Gaussian prior on the model coefficients. Then, one has to find the λ which maximize **the bayesian marginal likelihood** [11]:

$$\int f(y|\beta)f(\beta)d\beta$$

with $f(y|\beta)f(\beta)$ the joint density of the data and coefficents β.

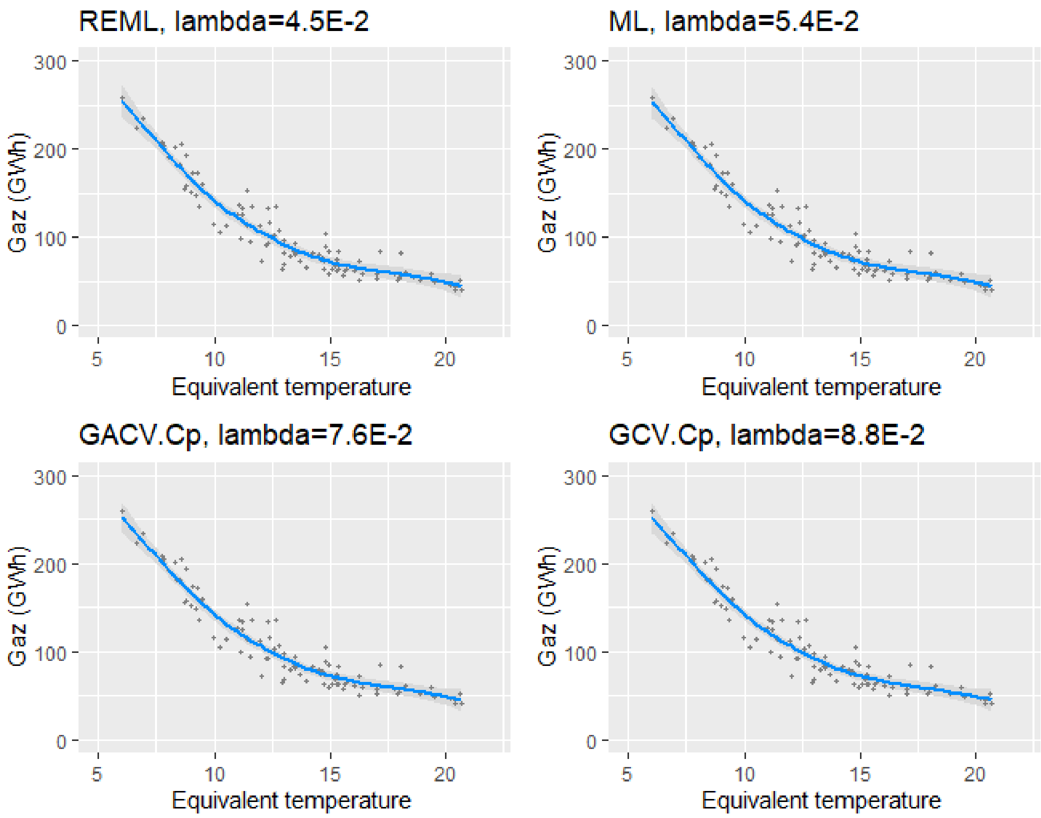

Figure 4.1.2: Models fitted with λ from different methods. Data of Gas consumption, from May, 2010-2012

In our case, the results are globally equivalent as shown in figure 4.1.2. However, the parameters from the Bayesian methods produce slightly smoother functions as the example in the appendix shows (see figure 9.3.1 with another subset of our data from the winters of 2010 to 2012.).

GCV was the first method available for gam() at the creation of the mgcv package. This is why its author chose to call it "Multiple GCV" or "Multiple generalized cross validation". Simon Wood later proved that Bayesian methods were more stable for certain data sets, [7]. The package had to keep the same name, that's why it is now called "Mixed GAM Computation Vehicle" [4].

4.2 How to choose k ?

k stands for the basis dimension.

Since λ controls the actual model complexity, an exact value of k is not critical. But if k is too small, there will be oversmoothing. Whereas bigger k increases the computational cost. Hence it must be large enough but not unnecessarily large.

Any default k is essentially arbitrary.

- It is set by default to 10 for the following basis: cubic regression splines, cyclic cubic regression splines and p-splines.
- It is set by default to 5^d for tensor basis, with d the number of covariate.
- It is set by default to 10 for one covariate, 30 for 2 covariates and 110 for 3 covariates or more.

As said above, some bases are knot-based approaches. However, the user has to specify a k anyway. k gives the dimension of the basis, which implies a certain number of knots. These knots are placed at the boundaries of the covariate then evenly over its range.

On figure 4.2.1 it can be seen that k= 4 is sufficient. A regrouping of curves on the same graph can be found in the appendix 9.5.1, which shows that the curves for k=5, k=6 and k=92 are almost the same.

4.2.1 gam.check()

In order to check if k is large enough, one can use the gam.check() function of the mgcv package.

It returns 4 indices:

- **k'** gives the maximum effective degree of freedom (edf) that can be achieved by the model. For the "cr", "ps" and "tp" bases it is k-1. For the "cc" basis it is k-2. For the "te" and "ti" bases, it is the product of the k' of the corresponding bases used.
- **edf** gives the effective degree of freedom of the model. This can be compared with k'. An edf very close to k' suggests that the dimension of the basis is too small.

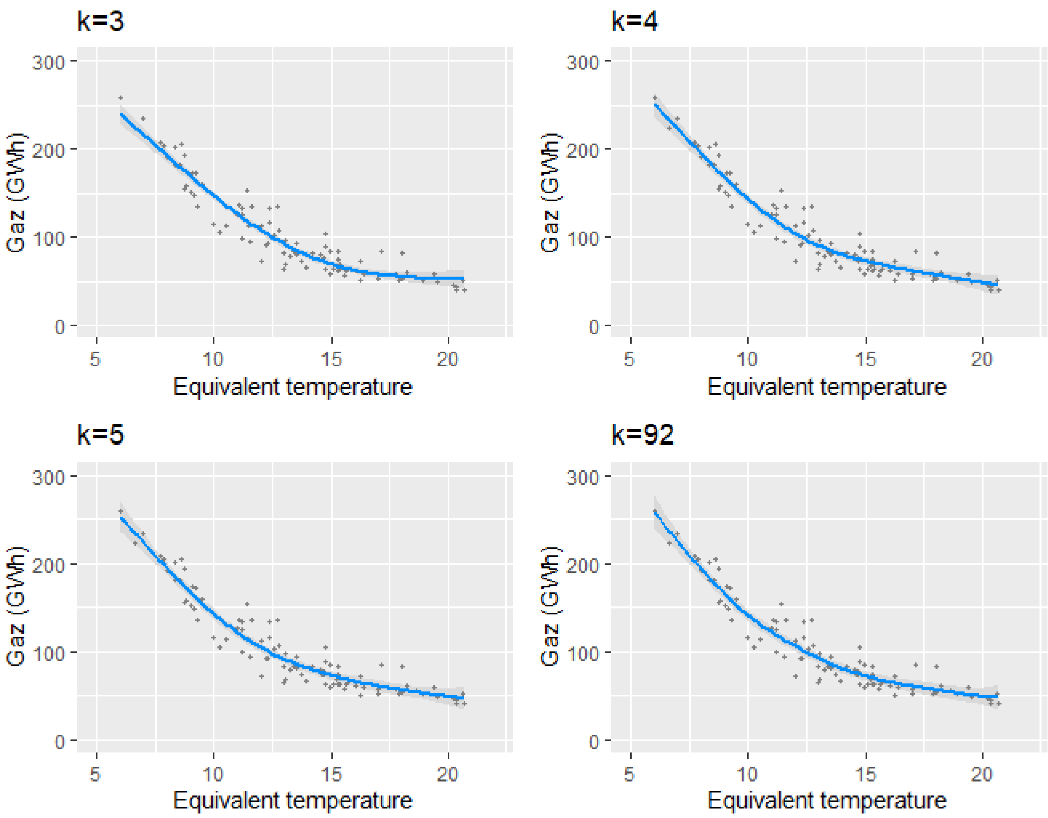

Figure 4.2.1: Smoothing with different k. Data of gas consumption, from May, 2010-2012

- **k-index** is a ratio between Δ and the variance of the residuals. Δ is calculated as the average of the differences of the residuals with their nearest neighbours. A k-index far from 1 means that Δ is abnormally small. One can suppose that the residuals are positively correlated with their neighbours, suggesting that there is missed pattern left in the residuals. This could possibly be relieved by increasing k.

- **p-value** is for the residual randomization test: Δ is calculated n times for re-shuffled residuals. It gives the null distribution of Δ. The test examines whether the original Δ look like an ordinary observation from this distribution. If the p-value is low, one can suppose that the residuals are positively correlated with their neighbours, suggesting that there is missed pattern left in the residuals. This could possibly be relieved by increasing k.

It also returns 4 plots to check the residuals.

4.3 How to choose a basis?

The default method used in the function gam() is the "tp" basis for thin plate regression spline basis. One can choose m, the order of the derivatives in the thin plate spline penalty (by default m=2).

For smoothing w.r.t one covariate, others options are available:

- The "cr" basis for cubic splines.
- The "cc" basis for cyclic cubic splines
- The "ps" basis for P-splines. The order of the basis and the order of the difference penalty can be choosen with the option m=(a,b) with a= order-2 and b=d (for d like in (3.2.8)). For example a=2 is for a cubic spline and b=3 is for a thrid order diffrence penalty. The default is m=c(2,2). products.

Let's compare the different basis fitting our data on figure 4.3.1. A regrouping of curves on the same graph can be found in the appendix 9.4.1. It can be seen that the curves for bases "tp","cr" and "ps" are almost the same. As expected, the cyclic cubic spline doesn't fit well because the data is not suitable for cyclic regression.

4.3.1 TPS vs tensors

Recall that the goal is to examine the relationship between consumption and temperature and particularly its behaviour depending on the time of year and its evolution over time.

To this end, two new variables are created:

A variable named **"FracOfYear"** represents the fraction of the year: a value ranging from 0 to 364/365 (or 0 to 365/366 for leap years).

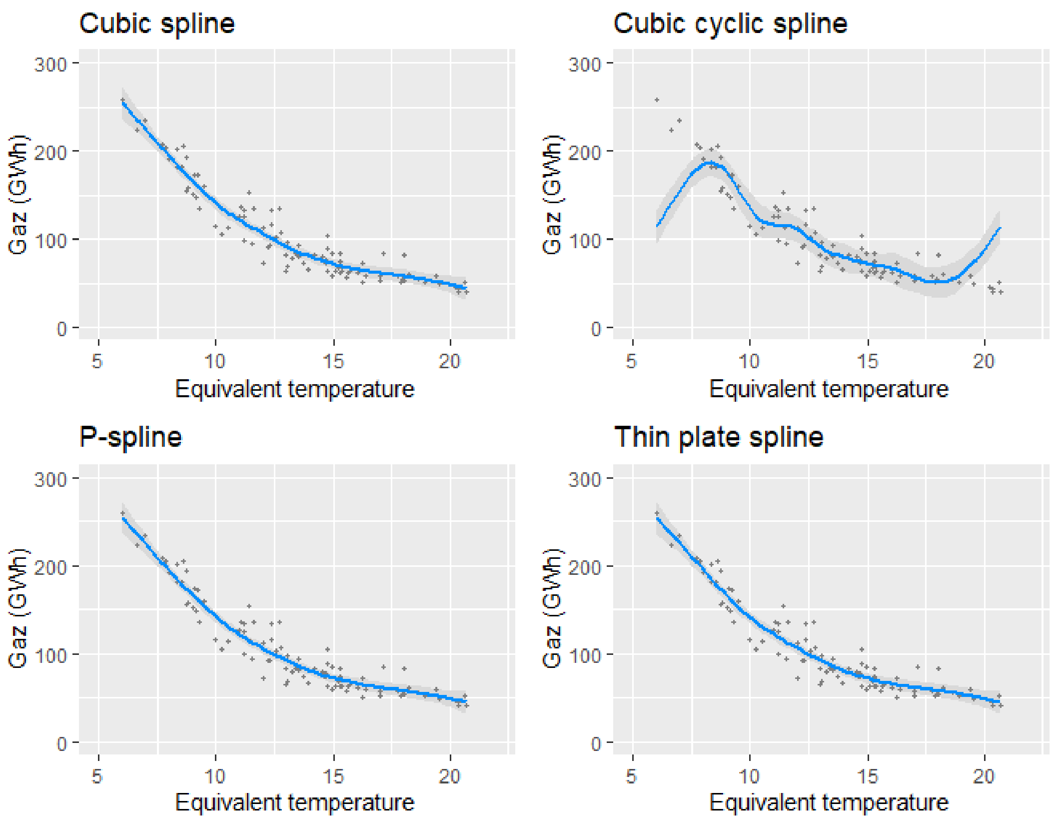

Figure 4.3.1: Models fitted with different basis. Data of Gas consumption, from May, 2010-2012

A variable named **"trend"** represents the total sum of days since January 2010 (01/01/2010 has trend 1, 02/01/2010 has trend 2, and so on)

Let's take our sample again but now add a second covariate: FracOfYear.

A model with several covariate can be fitted with the "tp" basis or with a tensor basis like "te". "te" takes into argument the argument for each of the covariate. By default, "te" use a 'cr' basis with k=5. For example, let's use a p-spline with m=c(2,2) to smooth the equivalent temperature variable "tempEq" and take a cyclic cubic regression spline for "FracOfYear".

One can compare this model with a "tp" basis model. Recall that TPRS are relevant in the case of isotropy which is not the case in our data. To remedy this, one creates a new variable, with range=1 by dividing a covariate by its range. The variable "normTemp" is created. (This is not necessary for "FracOfYear" because its range is already 1).

4.3.1.1 Comparison

The model with 'te' has a higher AIC but the modele with tprs has a higher BIC. Both have the same value for adjusted R^2.

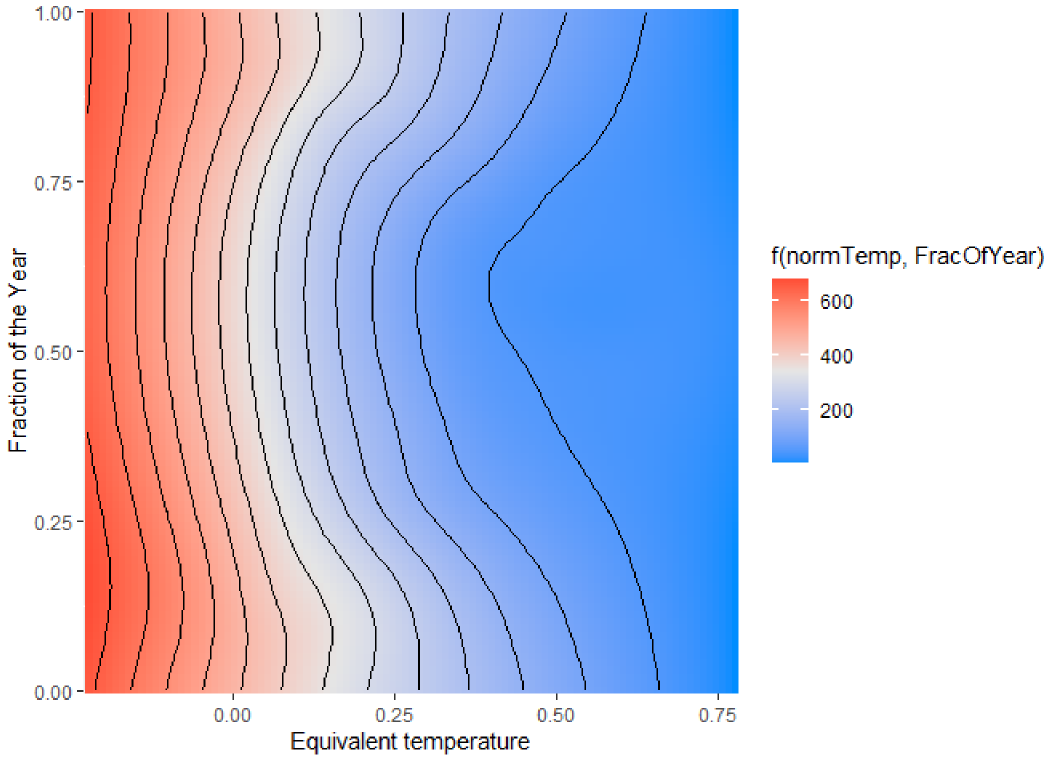

Figure 4.3.2: Model fitted with "te" basis. Data of Gas consumption, from May, 2010-2012

	edf	AIC	BIC	R^2 adj
te	33.4	33973.9	34184.6	0.987
tprs	48.2	33919.3	342222.7	0.987

Even after increasing k to 50, the tprs model still has poor index in gam.check().

The model with tprs is not better than te and te is more relevant in our non-isotropic case.

4.3.2 Difference between ti and te

ti() is also in option to smooth tensor products. te produces a full tensor product smooth, while ti produces a tensor product interaction. When the main effects (or any lower interactions) of the covariable in the tensor product are present, ti is more appropriate.

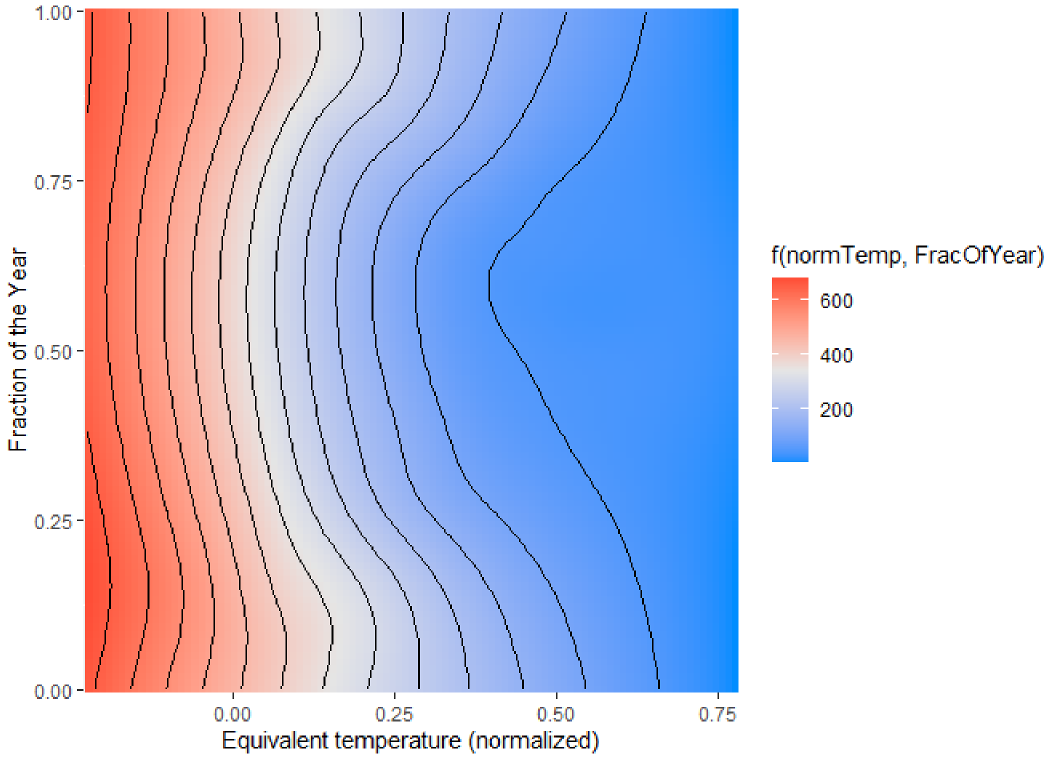

Figure 4.3.3: Model fitted with "tprs" basis. Data of Gas consumption, from May, 2010-2012

Chapter 5

Final models

5.1 Construction of models

The aim is to examine the relationship between consumption and temperature. In particular, its behaviour depending on the time of year and its evolution over time.

Since the heating of the house has a big role on the energy consumption, we use the equivalent temperature variable "tempEq". This is used by gas experts for their analyses.

The other variables on which the analysis will focus are:

- "FracOfYear" represents the fraction of the year (see 4.3.1).
- "trend" represents the trend 4.3.1).
- New factor variable:"WeekDay" represents the day of the week. (see 2.2.1).
- New factor variable: "Holiday" indicates whether the selected day is in a holiday period or not. If so, it indicates which holiday it is "SummerHoliday", "WeekHoliday", "EasterHoliday" or "WinterHoliday" (see 2.2.1). If no, the value is "No".

The focus is on the GWhElec variable.
The history of the search for the right construction criteria is given in the external annex A.

The variable "FracOfYear" is cyclic, so the base "cc" is chosen. For the others, the classic "cr" base is chosen.

To analyse the effects of several co-variables, the use of tensors is preferred since the model is not isotropic.

In this problematic, one will analyse the main effects and the interactions. For these the formula ti() is used, with either 'cr' or 'cc' basis depending on the variable.

The selection method of the lambda is "REML".

The construction methodology is as follows:

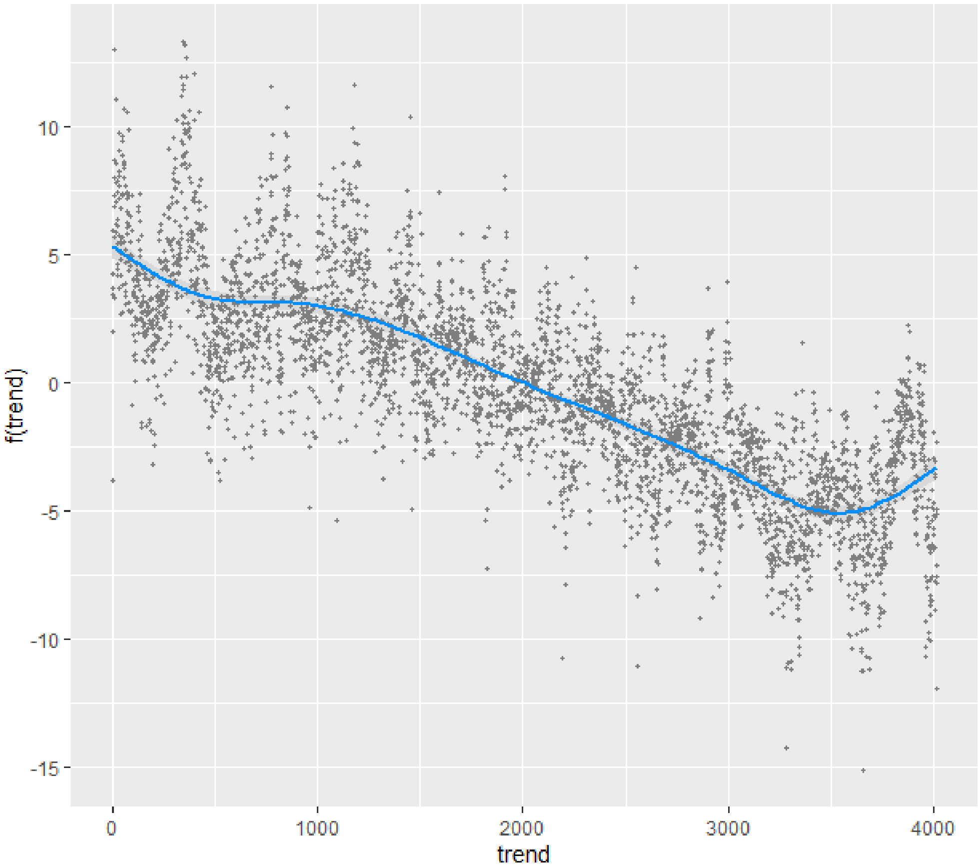

Figure 5.1.1: Main effects model. Impact of the variable trend.

First, one creates a model with only the main effects. For this one, one checks that the default k chosen is sufficient with the gam.check() option. 4.2.1. If it is not sufficient, k is increased.

The "trend" variable has the following particularity: it includes both the effects of time and the consumption pattern depending on the period of the year. This is already included in FracOfYear.

gam.check() gives poor results for the smoothing of this variable as it neglects a pattern in the data , see figure 5.1.1. The result of this test for the main effect of this variable is therefore not considered. The curve contains about 8 edf. This is sufficient for the interpretation of the time.

One tests all possible combinations of interractions. For each combination, one checks that the chosen k is sufficient with the gam.check() option. If this is not the case, a larger k is specified.

The model with the lowest AIC is the one including the 3 variables, the 3 2-to-2 interactions and the 3-variable interaction.

In addition to being difficult to interpret, the interaction of the 3 variables adds little to the model (see figure 9.6.1 in appendix). For simplicity of the model, it is not kept in the final model. This choice increases the AIC index only slightly, from 16998.68 to 17037.04.

Let's then see if the factor variables "Holiday" and "WeekDay" can bring something to the model. Since the model that takes them both into account has the lowest AIC index, they are kept.

The summary of the final models are available in sections 9.6.1 and 9.6.2.

The search for the gas model proceeded in the same way and led to the same results (see external appendix B).

5.2 Visualisation of the results

Figures in section 9.6.3 represent the impact of the covariates on the model fit. Here are some observations: For gas:

5.2.1 Gas

In figure 9.6.2 one can see the effect of the variable "tempEq" on the fit. It is a decreasing curve. It approaches a straight line with a decreasing slope for the lowest temperatures before adopting the behaviour of a second degree curve. Its impact on the final model can go from $\sim$ -150 to $\sim$ +450.

In figure 9.6.3 one can see the effect of the variable "FracOfYear" on the fit. The curve is flat at the extremes (coldest months) and forms a trough for the central months. Its impact

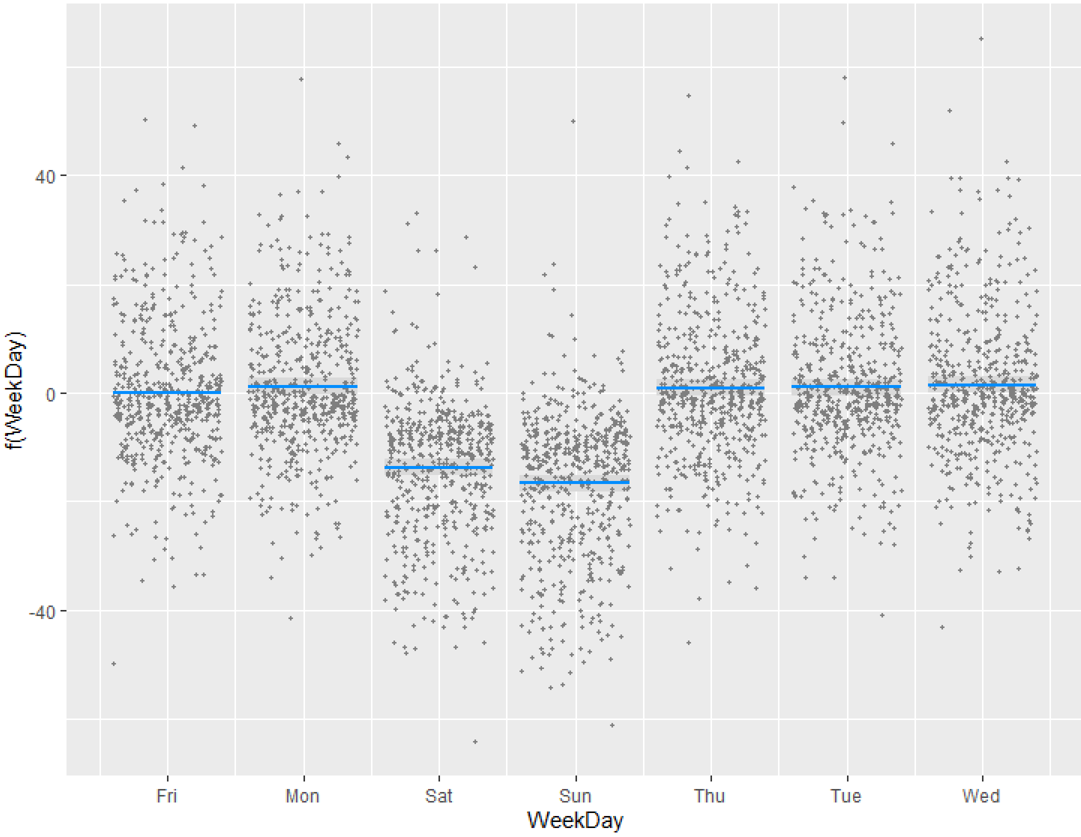

Figure 5.2.1: Effect of different days of the week on the fit for gas.

on the final model can go from ∼ -80 to ∼ +80.

On the figure 9.6.4 one can see the effect of the variable "trend" on the fit.The curve is almost flat and suggests that the variable has almost no impact.

In figure 9.6.5 one can see the effect of the interaction of the variables "FracOfYear" with "tempEq" on the fit . The yellow/lighter colours represent a positive impact on the fit value and the red colours a negative impact.

These effects "attenuate" the impact of "FracOfYear" for cold and warm temperatures: for a temperature of 20°, it will have a greater positive impact on the fit in July than in March.

However, this is reversed for central temperatures. They "exacerbate" the impact of "FracOfYear" on the fit. For a temperature of 12°, there will be a stronger positive impact on fit in March than in July.

In figure 5.2.1 one can see the effect of different days of the week on the fit. This picture suggests that less gas is consumed during the weekend.

5.2.2 Electricity

For electricity. In figure 9.6.6 one can see the effect of the variable "tempEq" on the fit. It is a decreasing curve. It approaches a straight line with a decreasing slope for the lowest temperatures before adopting the behaviour of a second degree curve and rising slightly at the end. Its impact on the final model can go from $\sim$ -10 to $\sim +30$.

On the figure 9.6.7 one can see the effect of the "FracOfyear" variable on the fit. The curve is high at the extremities (winter months) and low for the central months (summer). It goes down sharply in the first quarter of the year, then more slightly until about August. Then it goes up again until the last months of the year. It is quite wiggly. It suggests that the consumption of the year is not symmetrical. Its impact on the final model can range from $\sim$ -10 to $\sim +10$.

On the figure 9.6.8 one can see the effect of the variable "trend" on the fit. The curve is a downward slope. For the last times it goes up slightly. This observation should be taken with caution as it may be related to the change in living patterns in 2020. It will be necessary to monitor how this phenomenon evolves with the data for 2021 and subsequent years.

On the figure 9.6.9 one can see the effect of the interaction of the variables "tempEq" and "trend" on the fit. This suggests that over time the effect of warmer temperatures increases consumption more and colder temperatures increase it less.

In figure 9.6.10 one can see the effects of different days of the week on the fit. This picture suggests that the day with the highest electricity consumption is a Saturday and the day with the lowest electricity consumption is a Sunday.

Chapter 6

Prediction from final models

6.1 Construction of a second interactive dashboard for predictions

Once the final models have been found they must now be used to predict gas and electricity consumption. The objective is to extrapolate to "extreme" temperatures of -11°[1] and +40° over the next few years..

The idea is to predict the behaviour of the temperature/consumption curve with respect to the other parameters of the model:

- "FracOfYear" represents the fraction of the year.
- "trend" represents the trend.
- "WeekDay" represents the day of the week.
- "Holiday" indicates whether the selected day is in a holiday period or not, and if so, which holiday it is.

The user can choose the date with the options "Day of the month:", "Month:", "Year:". The predictions can range from 01/01/2010 to 31/12/2025[2]. The selected date is then converted to the variables "FracOfYear" and "trend".

The user can also choose the day of the week with the option "Day of the week"[3].

Finally he can choose if it is a holiday period or not with the "Holiday" option. The available choices are : "No", "Summer Holiday", "Winter Holiday", "Easter Holiday" and "Week Holiday".

The dashboard consists of two pages: one for each energy. Each page contains a graph showing the predicted consumption as a function of the equivalent temperature ("tempEq"),

[1]-11° is the temperature for extrapolation that interests the gas experts who collaborated on the study.

[2]The choice of the number of years of prediction is arbitrary.

[3]One of the advantages of the interactive dashboard is that it allows the user to play with the parameters and gauge the change in the temperature/consumption curve. In order not to interfere with this experience it has been decided not to automatically match the date with the day of the week.

for the temperature range -11° to +40°.

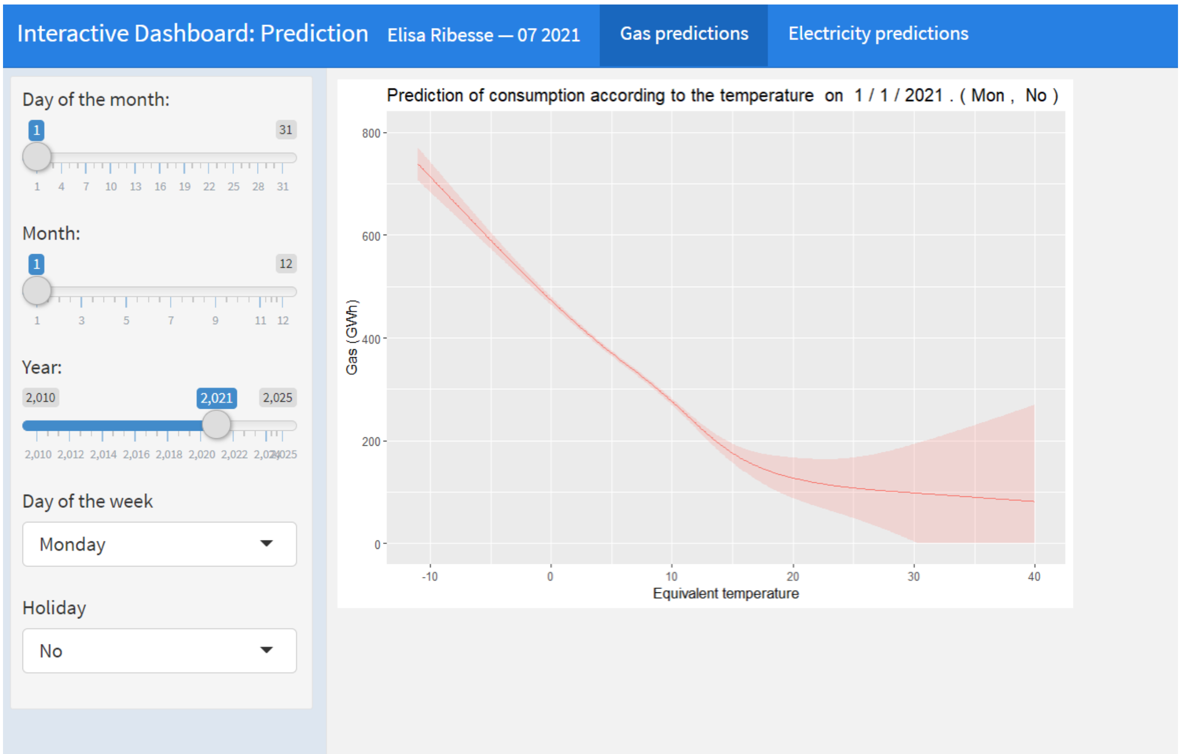

Figure 6.1.1: Overview of the interactive prediction dashboard for Gas.

The curve that appears is the fit of the corresponding model for the chosen parameters (see figure 5.1) with the help of the gam.predict() function from mgcv. This function gives also standard errors, se.fit, based on the posterior distribution of the model coefficients. The interval surrounding this curve covers 1.96·(se.fit) on each side of the curve.

With the first dashboard, the user has to make an arbitrary decision to divide the year into months, seasons etc. With the use of "FracOfYear" and "trend" in the model the predictions no longer depend on a division of the year.

6.2 Observations for gas

Let's look at the consumption prediction graph for a winter day, for example in January 2022, on figure 6.2.1. One can see a first trend: a straight line with negative slope, ∼ -21GWh/°[4]. This means that the higher the temperature, the lower the gas consumption. Hence the predictions for extremely low temperatures has the same pattern as low temperatures

While these predictions are consistent with the data provided, they may not be realistic. Indeed the model does not take into account the possible saturation of the boilers.

[4] Approximated slopes are roughly deduced by looking at the graphs.

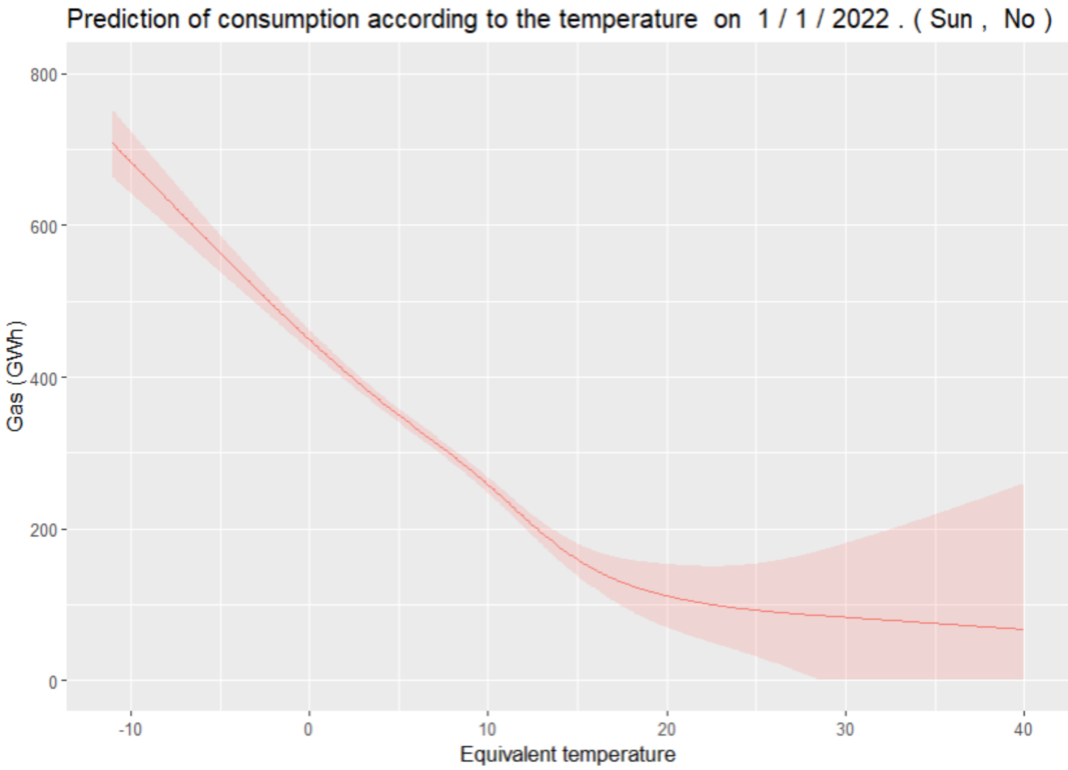

Figure 6.2.1: Extrapolation of gas consumption depending on the equivalent temperature, January 2022. Options: Sunday and not holiday day.

The interval around the warmest temperatures is very large and nothing can be derived from it. This is not impact the study because this curve is supposed to represent the consumption behaviour in January.

Now let's look at a summer month, for example July 2022, figure 6.2.2.

One can see three trends. First a negative slope for the lowest temperatures. This is of little interest since we are analysing a summer month. Then a second degree curve. Finally, from ∼17° the curve takes a linear form with a very slight negative slope. This slope is smaller than a slope of -1Gwh/°. The predictions for extremely warm temperatures thus follow the trend of warm temperatures.

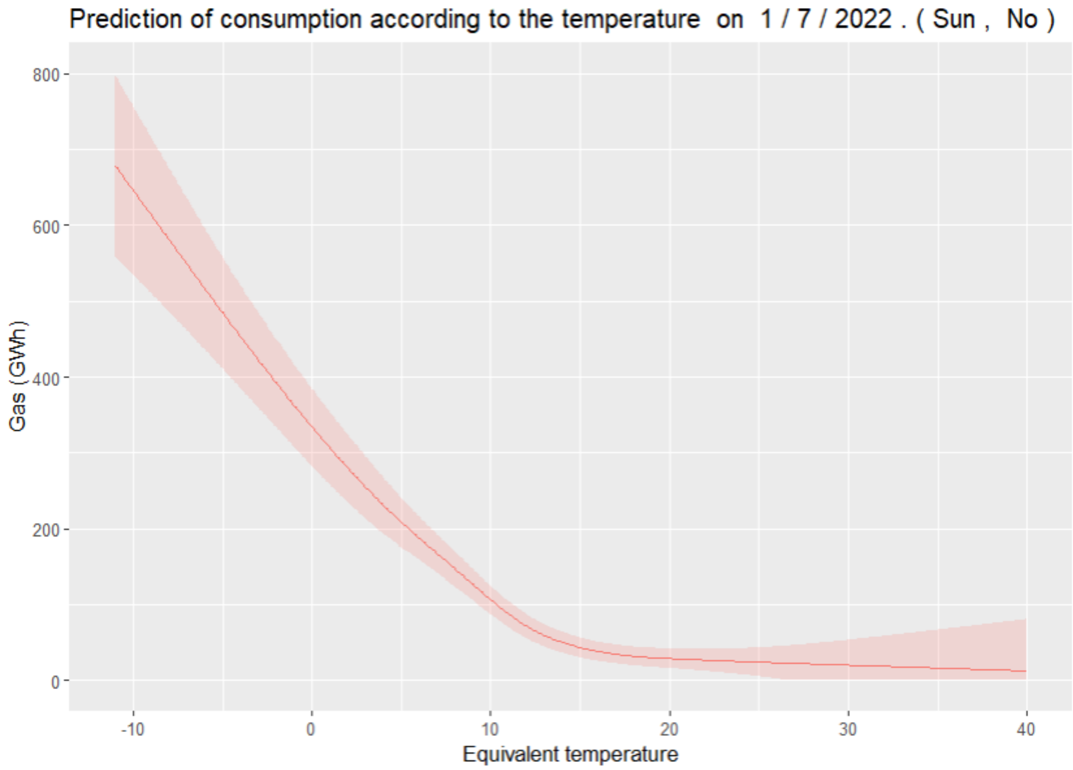

Figure 6.2.2: Extrapolation of gas consumption depending on the equivalent temperature, July 2022. Options: Sunday and not holiday day.

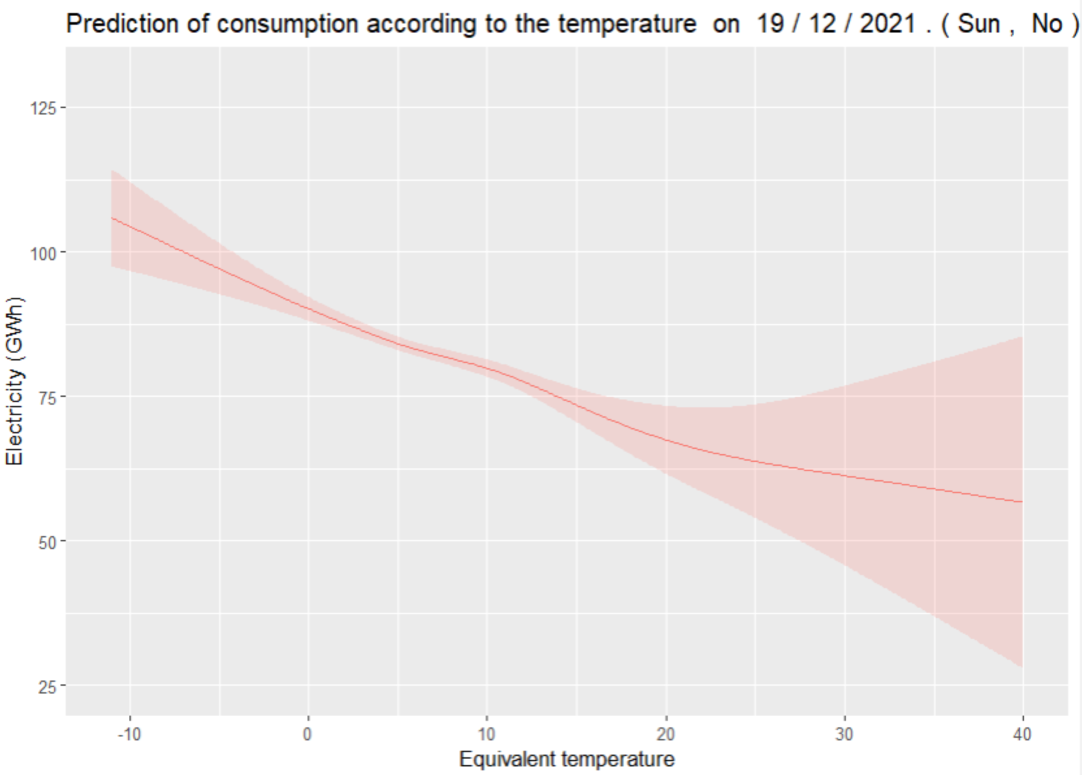

Figure 6.3.1: Extrapolation of electricity consumption depending on the equivalent temperature, December 2021. Options: Sunday and not holiday day.

6.3 Observations for electricity

Let's look at the forecast graph for a winter day, for example on 19 December 2021, see 6.3.1.

One can see a decreasing curve. This means that consumption decreases when temperatures increase. For the lowest temperatures, the curve follows a straight line of about -1.25Ghw/°.

The curve decreases even at extremely warm temperatures. Besides, the interval around the curve is extremely large for warm temperatures and nothing can be learned from it.

This is not a problem. This curve is supposed to represent the winter consumption pattern, which it does. It is unlikely that it will reach 30° in December in Belgium (for the next few years).

Now let's look at a summer day, for example 1st July 2022, see 6.3.2.

This curve is formed by three trends. Firstly, its behaviour is similar to a straight line with negative slope for the coldest temperatures. This means that as the temperature increases, the electricity consumption decreases.

Secondly, it behaves like a second degree curve with a minimum of $\sim$ 63 GWh for a

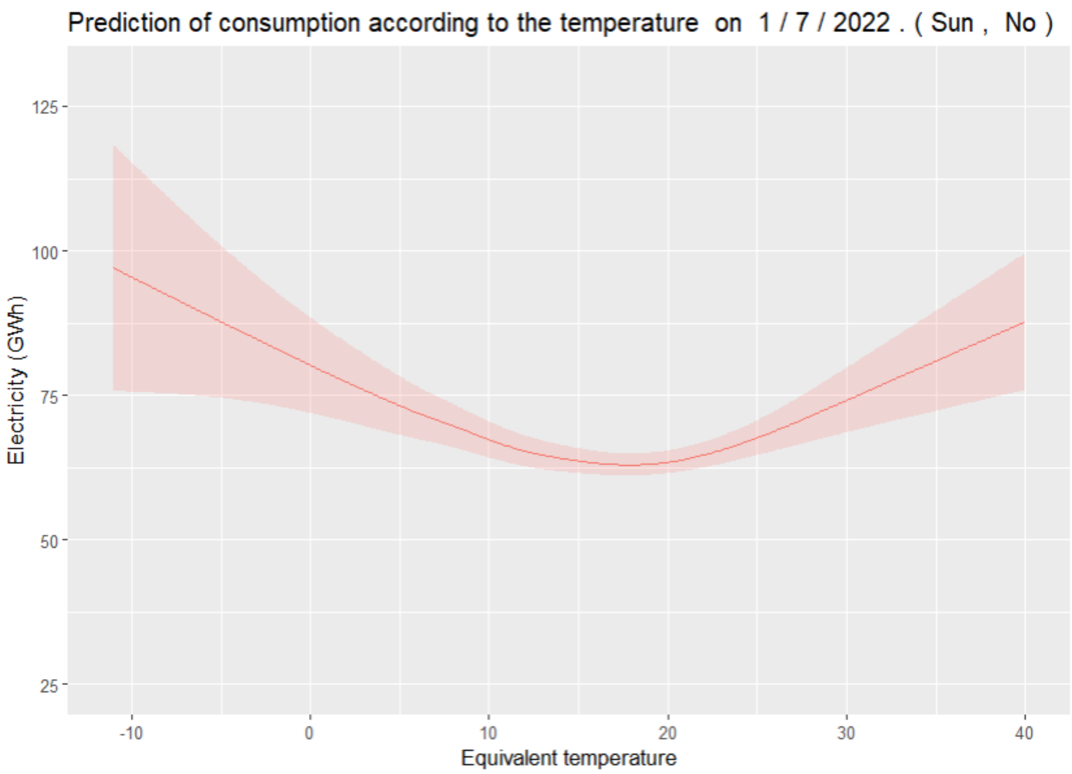

Figure 6.3.2: Extrapolation of electricity consumption depending on the equivalent temperature, July 2022. Options: Sunday and not holiday day.

temperature of ∼18°.

Finally, it behaves like a linear line with a positive slope. This means that the higher the temperature, the higher the electricity consumption.

This is a phenomenon that one could already observe on the graphs of the first dashboard. However, it becomes more pronounced over time.

Let's take the same date in 2014 as a comparison, figure 6.3.3. The same shape is noticeable. However, the slope for lower temperatures is stronger. And the slope for higher temperatures is weaker.

This illustrates the phenomenon that electricity consumption is less sensitive to very cold temperatures. But it is more sensitive to very warm temperatures.

These predictions can give indications to the electricity suppliers. Let's assume that they organise the maintenance of their lines in July. Because it is a month where temperatures are the warmest and there is less need to supply the lines than in winter. In the long term, it might be better to shift this maintenance to a slightly cooler month, like May.

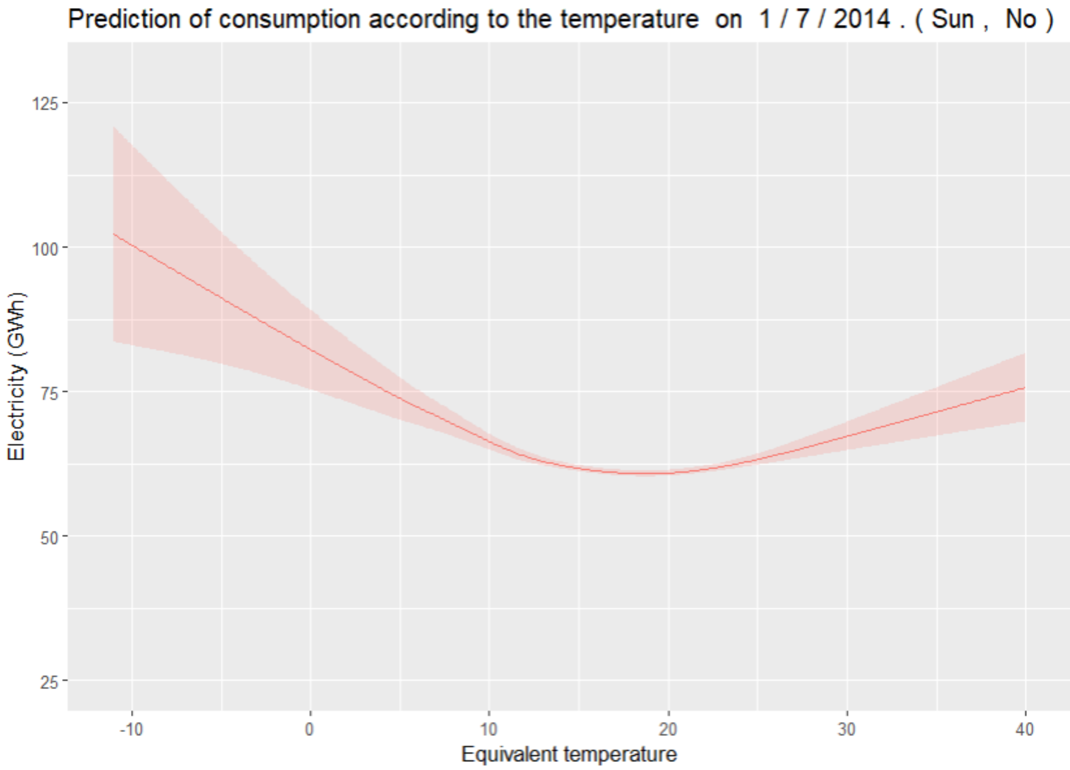

Figure 6.3.3: Extrapolation of electricity consumption depending on the equivalent temperature, July 2014. Options: Sunday and not holiday day.

6.4 Summary

A new interactive dashboard is created. It allows to visualize the extrapolation of the model at temperatures ranging from -11° to +40° for each combination of parameters.

The gas consumption for very cold temperatures in winter follows the trend of cold temperatures, with a straight line behaviour. This may not be the case in reality, with the possible saturation of boilers, not taken into account here.

The gas consumption for very hot temperatures in summer follows the hot temperatures, with a straight line behaviour with a very small negative inclination.

The general pattern of gas consumption moves very little with time.

The electricity consumption for very cold temperatures in winter follows the trend of cold temperatures, with a straight line behaviour. Predictions suggest that its inclination is lower over the years, which would indicate that consumption is less sensitive to very cold temperatures.

Electricity consumption for very hot temperatures in summer follows the trend of hot temperatures. It has a straight line behaviour, with a positive slope. Predictions suggest that its slope is steeper over the years, which would indicate that consumption is more sensitive to hot temperatures.

Chapter 7

Conclusion

7.1 Conclusion

The consumption of gas and electricity by households in Belgium is an important part of the total consumption[1]. It therefore has to be expressed by detailed models to allow accurate information for capacity planning and for the monthly settlement process. It is in the interest of the energy sector to continuously monitor and improve these models.

The main influence of energy consumption is temperature. But it does not affect it in the same way, depending on variables such as the time of the year.

In order to model these relationships, non-parametric models of daily gas and electricity consumption are created, using the fitting of GAM models with the help of mgcv package.

These models are based on the principle of a linear predictor involving a sum of smooth functions of covariates, called splines functions. Splines are piecewise polynomials. They are based on the notion of penalisation in order to get smoothness.

The constructed models are displayed in a dashboard, realised with the help of shiny package.

The observations that can be drawn from the new models are the following:

The gas consumption for very cold temperatures in winter follows the trend of cold temperatures. It increases proportionally to the decline in temperature.

This is also true for electricity consumption. Predictions suggest that its inclination is lower over the years, which would indicate that consumption is less sensitive to very cold temperatures.

Electricity consumption for very hot temperatures in summer follows the trend of hot temperatures. It increases proportionally to the increase in temperature.

[1] consumption which is not continuously monitored (annual metering only)

Electricity consumption decreases globally over the years and is less and less sensitive to cold temperatures. On the other hand it is more and more sensitive to very hot temperature, probably due to air conditioning expansion.

These predictions can give indications to the electricity suppliers. Let's assume that they organise the maintenance of their lines in July. Because it is a month where temperatures are the warmest and there is less need to supply the lines than in winter. In the long term, it might be better to shift this maintenance to a slightly cooler month, like May.

7.2 To go further

Here are a few ideas for further research:

One could analyse the effect of other weather criteria such as cloudiness or precipitation and see if they can improve the model significantly.

Figure 9.6.8suggests that in recent times the trend has a positive impact on electricity consumption. One could monitor how this phenomenon evolves with the data for 2021 and subsequent years.

One could improve the gas consumption prediction model to take into account the possible saturation of the boilers.

Chapter 8

Bibliography

[1] *Analyse de la consommation énergétique des ménages en Belgique*. SPF Economie, P.M.E., Classes moyennes et Energie, 2019.

[2] Patrick Breheny. *STA 621 : Nonparametric Statistics*. 2018-2019. URL: https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fmyweb.uiowa.edu%2Fpbreheny%2Fuk%2Fteaching%2F621%2Fnotes.html&data=04%7C01%7Celisa.ribesse%40student.uclouvain.be%7Cfe0b593a19e74379c67b08d965ffc35c%7C7ab090d4fa2e4ecfbc7
7C0%7C0%7C637652971436984787%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luM
3D%7C1000&sdata=5JxehgPH08ALcoPKnEy85ph%2Fb%2BU%2B5NMk%2BwGId9wMoy0%3D&reserved=0.

[3] Paul H. C. Eilers and Brian D. Marx. "Splines, knots, and penalties". In: *WIREs Computational Statistics* 2.6 (2010), pp. 637–653. DOI: https://doi.org/10.1002/wics.125.

[4] *Frequently Asked Questions for package mgcv*. URL: https://stat.ethz.ch/R-manual/R-patched/library/mgcv/html/mgcv-FAQ.html.

[5] *Market size of heating, ventilation, and air conditioning (HVAC) in selected countries in Europe in 2018 and 2019*. URL: https://www.statista.com/statistics/1192124/hvac-market-size-europe/.

[6] Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990. DOI: 10.1137/1.9781611970128.

[7] Simon Wood. "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.1 (2011), pp. 3–36. DOI: https://doi.org/10.1111/j.1467-9868.2010.00749.x.

[8] Simon Wood. *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman and Hall/CRC, 2006.

[9] Simon Wood. *Package 'mgcv'*. URL: https://cran.r-project.org/web/packages/mgcv/mgcv.pdf.

[10] Simon Wood. "Thin Plate Regression Splines". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 65.1 (2003), pp. 95–114. ISSN: 13697412, 14679868.

[11] Simon N. Wood. *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC, 2017.

Chapter 9

Appendices

9.1 Graphs relating to section 2.4

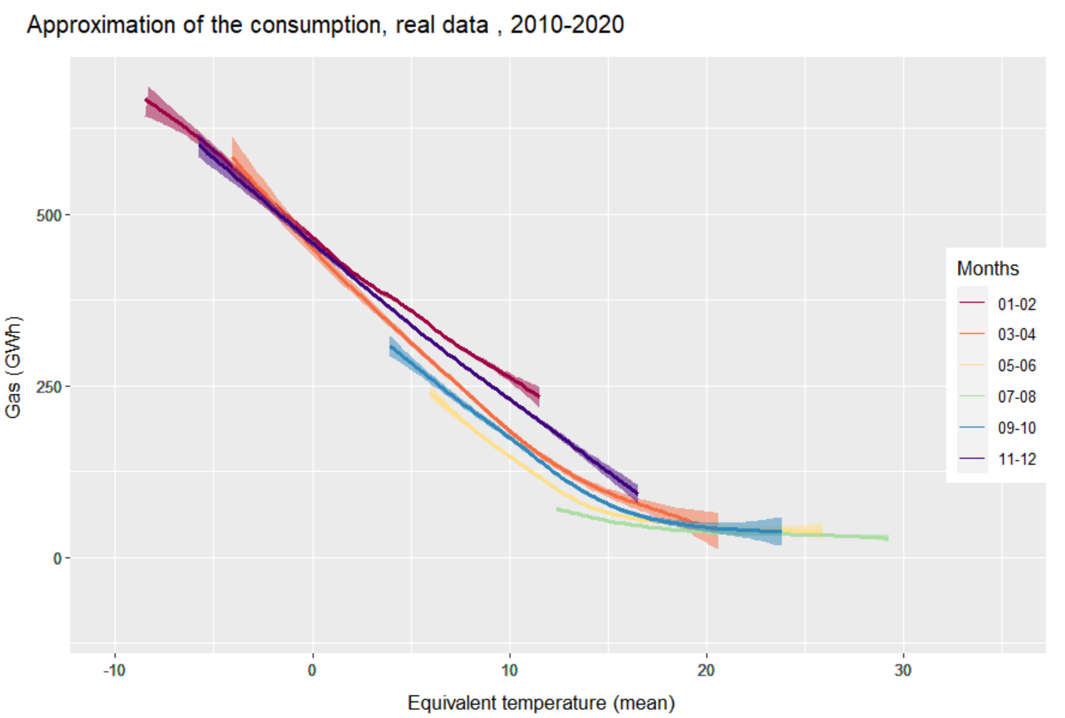

Figure 9.1.1: Approximation of Belgian gas consumption as a function of temperature (calculated based on the equivalent temperature formula "tempEq"). The year is split into periods of 2 months.

9.2 Graphs relating to section 2.5

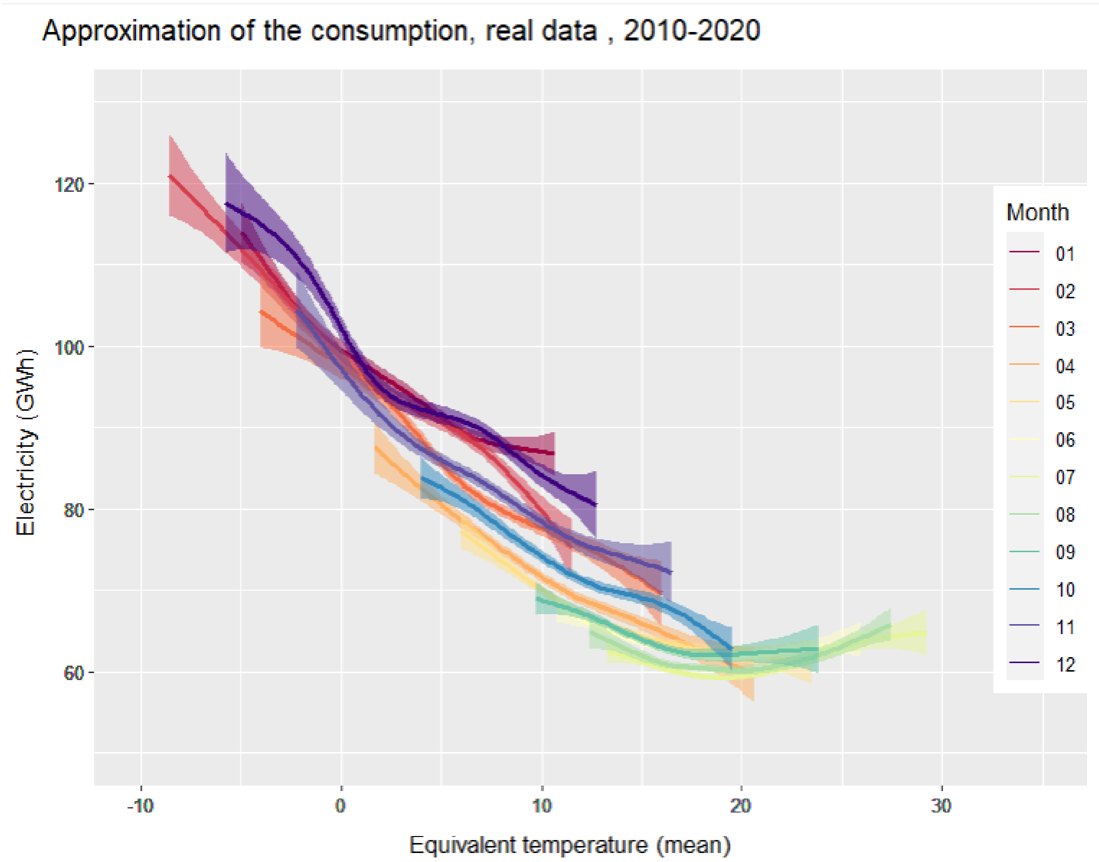

Figure 9.2.1: Approximation of Belgian electricity consumption as a function of temperature (calculated based on the equivalent temperature formula "tempEq"). The year is split into periods of 1 month.

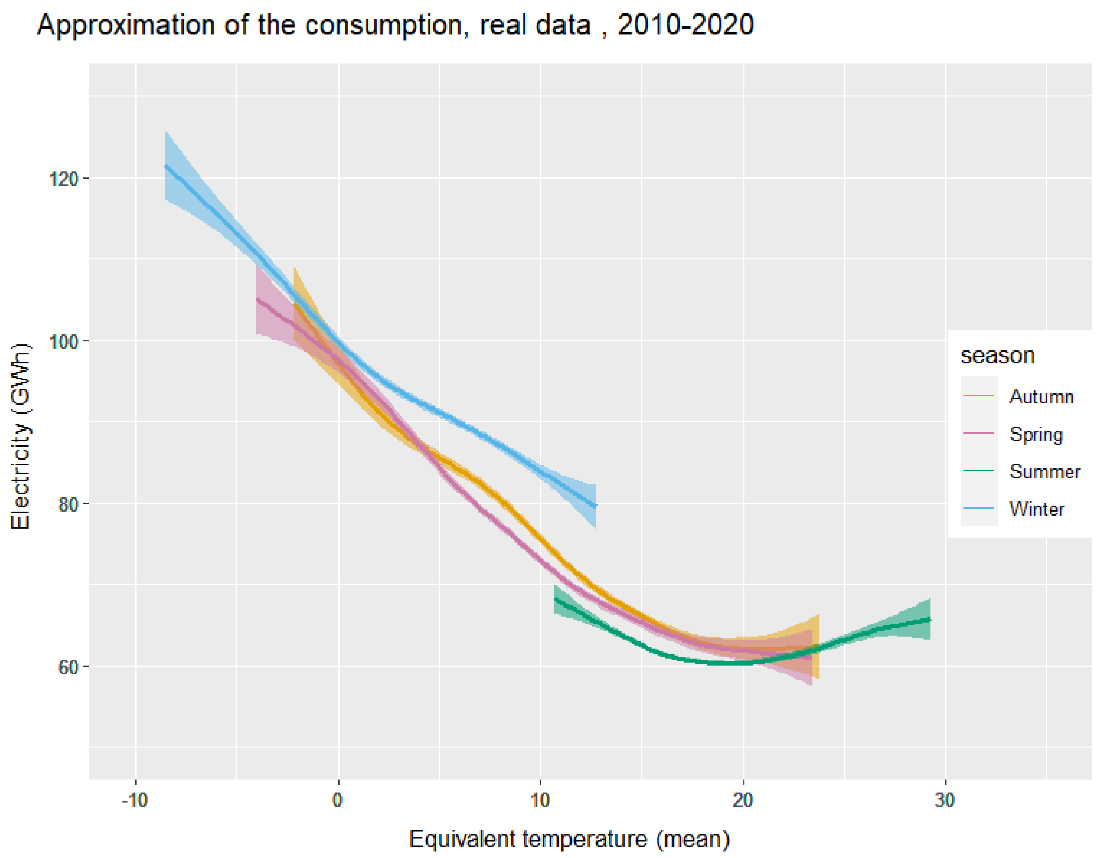

Figure 9.2.2: Approximation of Belgian electricity consumption as a function of temperature (calculated based on the equivalent temperature formula "tempEq"). The year is split by season with default months.

9.3 Graphs relating to section 4.1

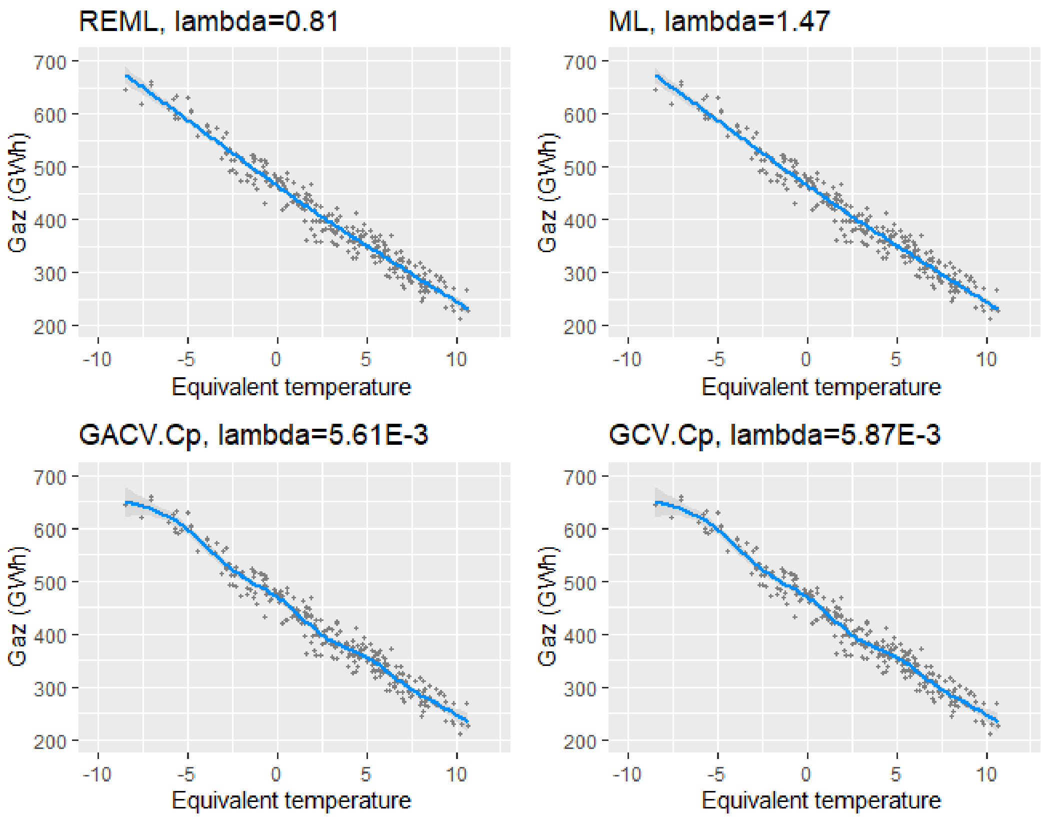

Figure 9.3.1: Models fitted with λ from different methods. Data of Gas consumption, from winter, 2010-2012

9.4 Graphs relating to section 4.3

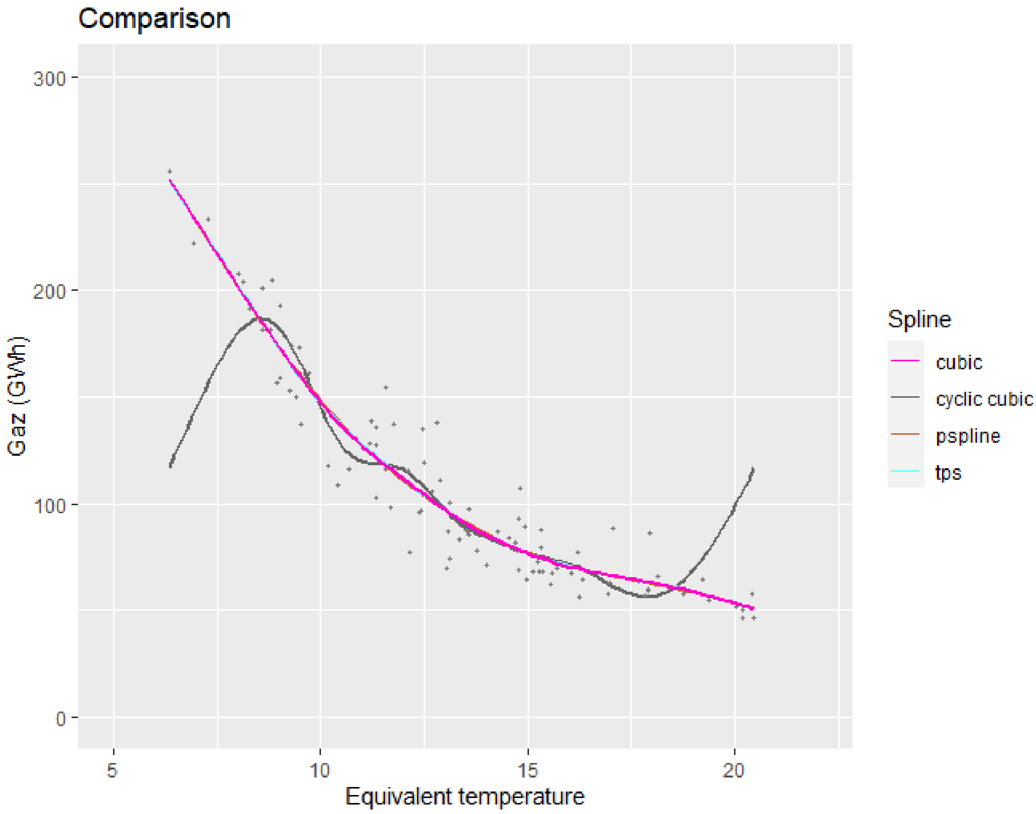

Figure 9.4.1: Models fitted different basis. Data of Gas consumption, from May , 2010-2012

9.5 Graphs relating to section 4.2

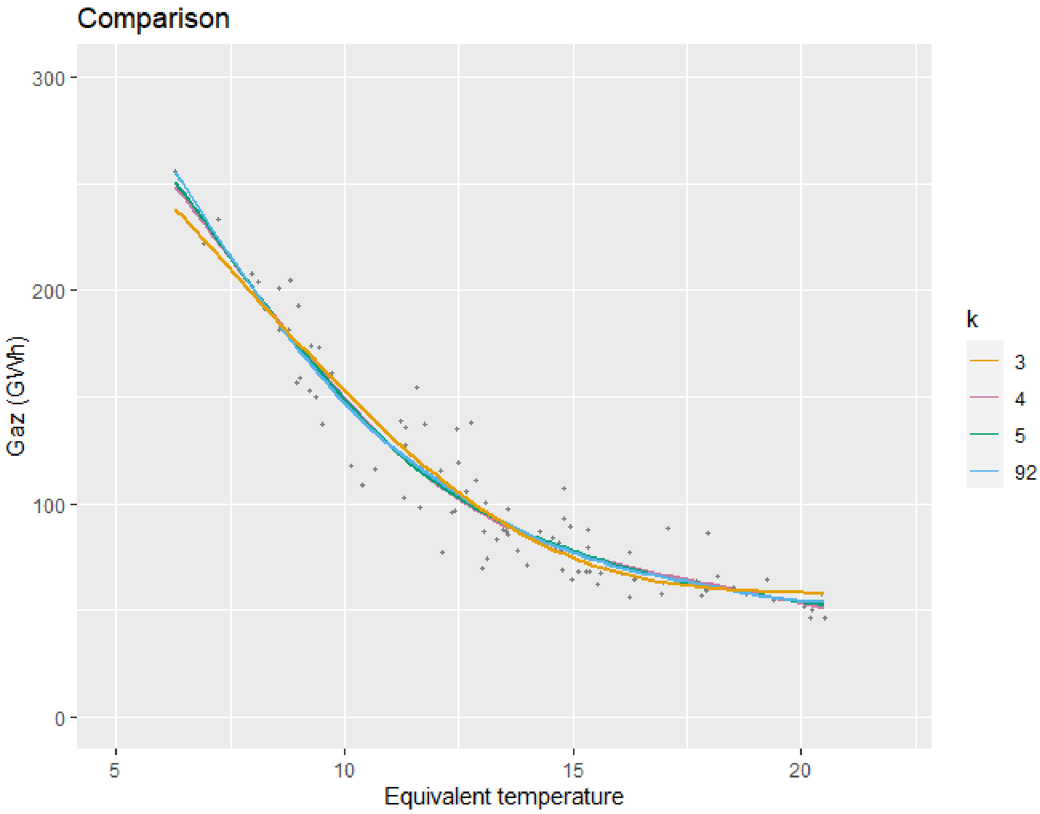

Figure 9.5.1: Smoothing with different k. Data of gas consumption, from May, 2010-2012

9.6 Section 5.1: results and graphs

9.6.1 Summary of final model for electricity displayed by Rstudio

```
Family: gaussian
Link function: identity

Formula:
GWhElec ~ s(tempEq, bs = "cr") + s(FracOfYear, bs = "cc",
    k = 15) + s(trend, bs = "cr", k = 7) + ti(tempEq, FracOfYear,
    bs = c("cr", "cc"), k = c(7, 7)) + ti(trend,
    FracOfYear, bs = c("cr", "cc"), k = c(7, 12)) +
    ti(tempEq, trend, bs = c("cr", "cr")) + as.factor(WeekDay) +
    as.factor(Holiday)

Parametric coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     74.14683    0.34691 213.735  < 2e-16 ***
as.factor(WeekDay)Mon           -0.54422    0.10912  -4.987 6.39e-07 ***
as.factor(WeekDay)Sat            1.19009    0.10937  10.881  < 2e-16 ***
as.factor(WeekDay)Sun           -1.13359    0.10950 -10.353  < 2e-16 ***
as.factor(WeekDay)Thu            0.05818    0.10896   0.534 0.593424    
as.factor(WeekDay)Tue           -0.34304    0.10907  -3.145 0.001673 ** 
as.factor(WeekDay)Wed            0.29632    0.10903   2.718 0.006600 ** 
as.factor(Holiday)No             0.76220    0.20938   3.640 0.000276 ***
as.factor(Holiday)SummerHoliday  0.46533    0.32446   1.434 0.151598    
as.factor(Holiday)WeekHoliday   -1.68879    0.34432  -4.905 9.74e-07 ***
as.factor(Holiday)WinterHoliday -0.35371    0.31348  -1.128 0.259255    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                         edf Ref.df        F  p-value    
s(tempEq)              7.185  8.030   72.675  < 2e-16 ***
s(FracOfYear)         12.707 13.000  104.072  < 2e-16 ***
s(trend)               5.929  5.996 1848.665  < 2e-16 ***
ti(tempEq,FracOfYear) 14.835 30.000    6.848  < 2e-16 ***
ti(trend,FracOfYear)  45.607 60.000    7.081  < 2e-16 ***
ti(tempEq,trend)       9.555 11.767    5.577 2.33e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.982   Deviance explained = 98.3%
-REML = 8325.8  Scale est. = 3.3972    n = 4017
```

9.6.2 Summary of final model for gas displayed by Rstudio

```
Family: gaussian
Link function: identity

Formula:
GWhGaz ~ s(tempEq, bs = "cr") + s(FracOfYear, bs = "cc",
    k = 15) + s(trend, bs = "cr", k = 7) + ti(tempEq, FracOfYear,
    bs = c("cr", "cc"), k = c(7, 7)) + ti(trend,
    FracOfYear, bs = c("cr", "cc"), k = c(7, 12)) +
    ti(tempEq, trend, bs = c("cr", "cr")) + as.factor(WeekDay) +
    as.factor(Holiday)

Parametric coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     187.1508     2.5191  74.293  < 2e-16 ***
as.factor(WeekDay)Mon             1.1042     0.7739   1.427 0.153725    
as.factor(WeekDay)Sat           -13.7563     0.7759 -17.731  < 2e-16 ***
as.factor(WeekDay)Sun           -16.6210     0.7767 -21.401  < 2e-16 ***
as.factor(WeekDay)Thu             0.8486     0.7729   1.098 0.272302    
as.factor(WeekDay)Tue             0.9731     0.7734   1.258 0.208408    
as.factor(WeekDay)Wed             1.3674     0.7728   1.769 0.076902 .  
as.factor(Holiday)No              3.1417     1.4768   2.127 0.033448 *  
as.factor(Holiday)SummerHoliday   1.5114     2.2831   0.662 0.508009    
as.factor(Holiday)WeekHoliday    -7.5582     2.4375  -3.101 0.001944 ** 
as.factor(Holiday)WinterHoliday  -8.3719     2.2135  -3.782 0.000158 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                          edf Ref.df       F  p-value    
s(tempEq)               7.965  8.530 247.114  < 2e-16 ***
s(FracOfYear)          12.557 13.000  83.935  < 2e-16 ***
s(trend)                5.792  5.977  58.595  < 2e-16 ***
ti(tempEq,FracOfYear)  17.167 30.000   9.904  < 2e-16 ***
ti(trend,FracOfYear)   43.542 60.000   5.534  < 2e-16 ***
ti(tempEq,trend)        4.178  4.787   4.837 0.000278 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.992   Deviance explained = 99.2%
-REML =  16163  Scale est. = 170.93    n = 4018
```

9.6.3 Graphs

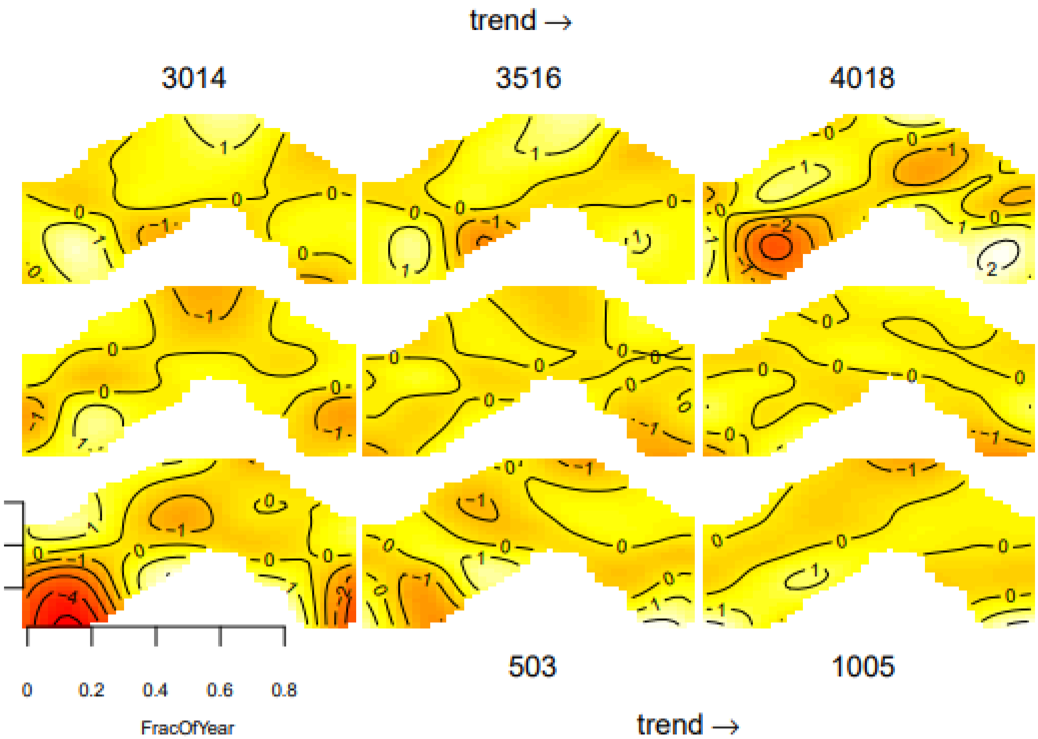

Figure 9.6.1: Impact of the interaction between "tempEq", "FracOfYear" and "trend" on fitting model for electricity. "tempEq" is represented on the y axis, "FracOfYear" on the x axis and "trend" troughout the different cases. The value of the contour lines is between -1 and 1 almost everywhere.

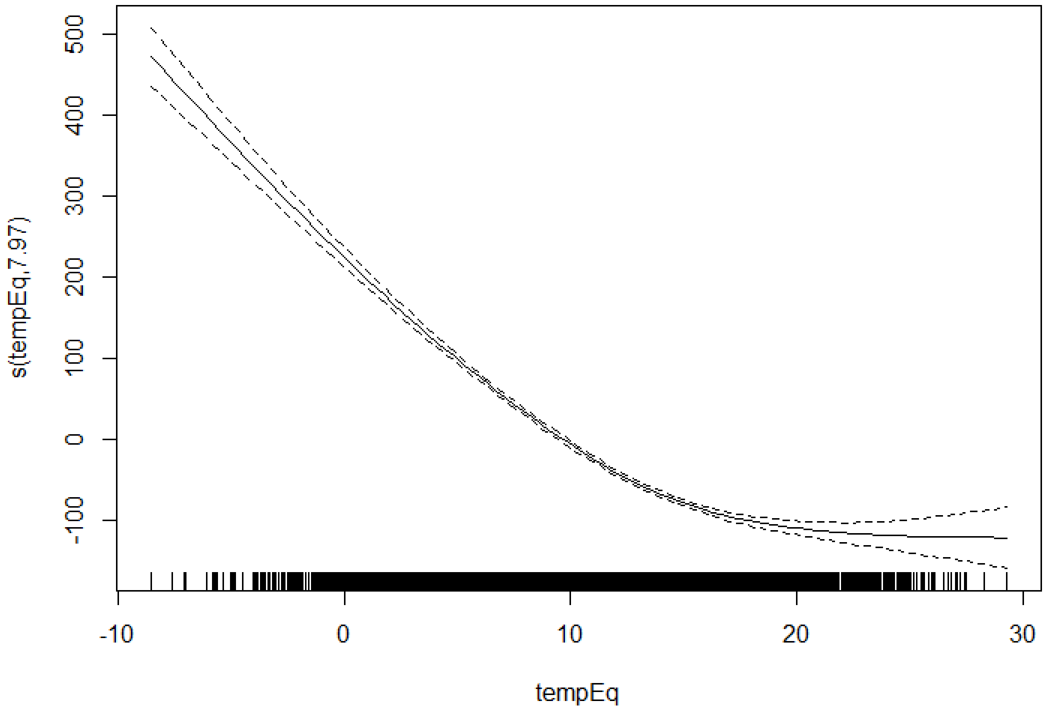

Figure 9.6.2: Effect of the variable "tempEq" on the fit for gas.

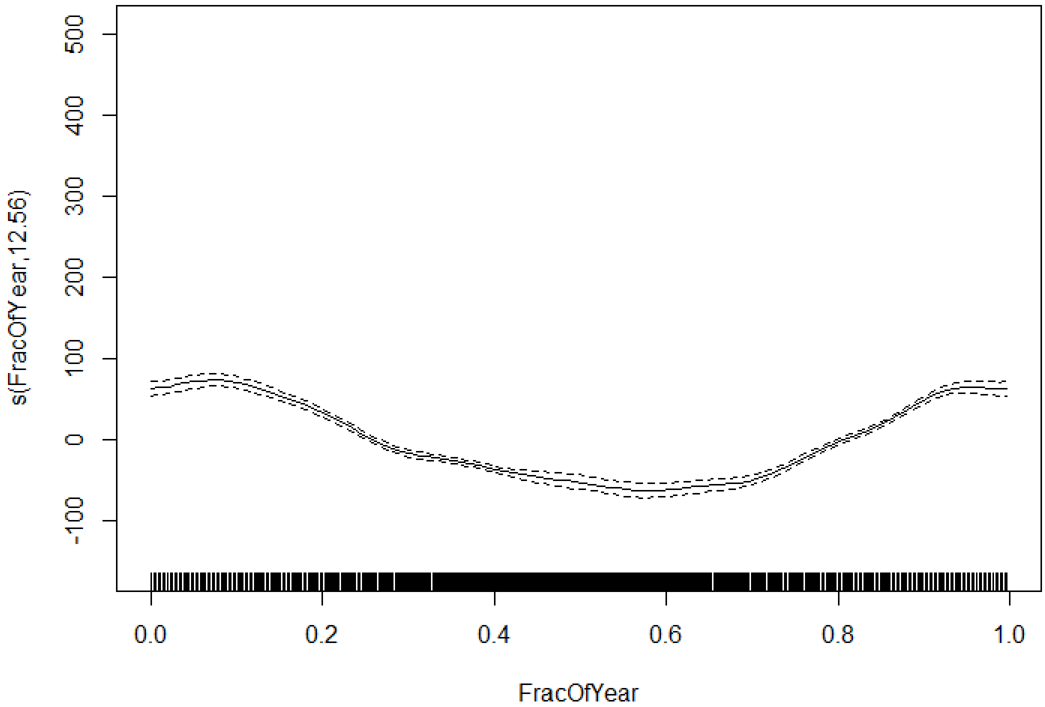

Figure 9.6.3: Effect of the variable "FracOfYear" on the fit for gas.

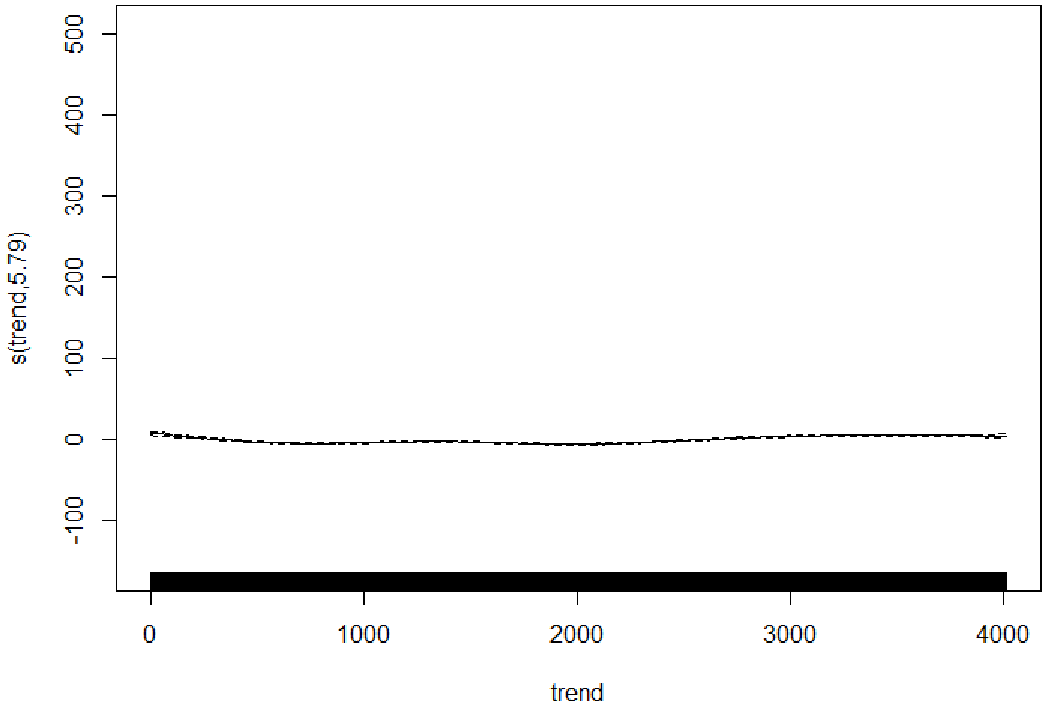

Figure 9.6.4: Effect of the variable "trend" on the fit for gas.

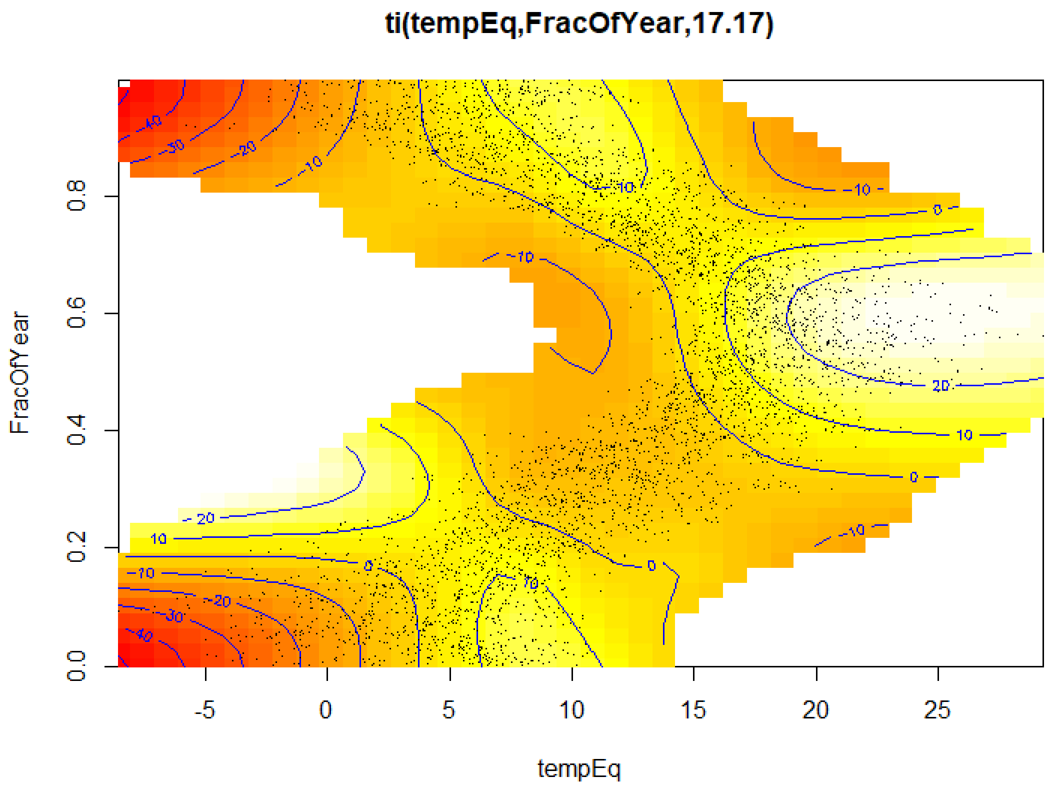

Figure 9.6.5: Effect of the interaction of the variables "FracOfYear" with "tempEq" on the fit for gas.

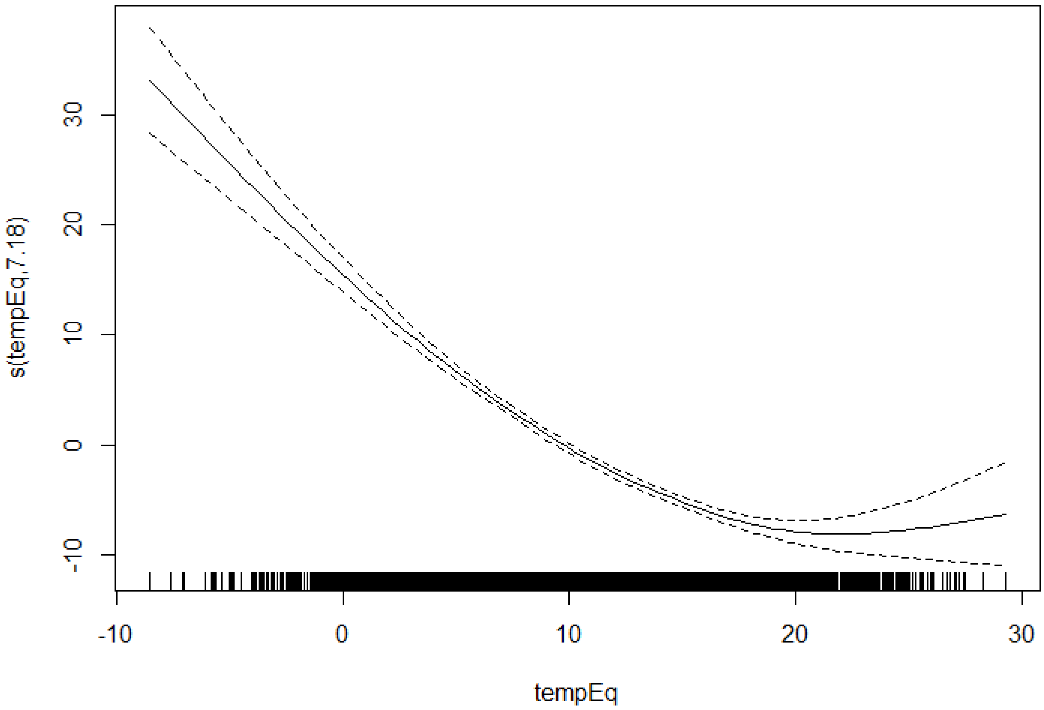

Figure 9.6.6: Effect of the variable "tempEq" on the fit for electricity

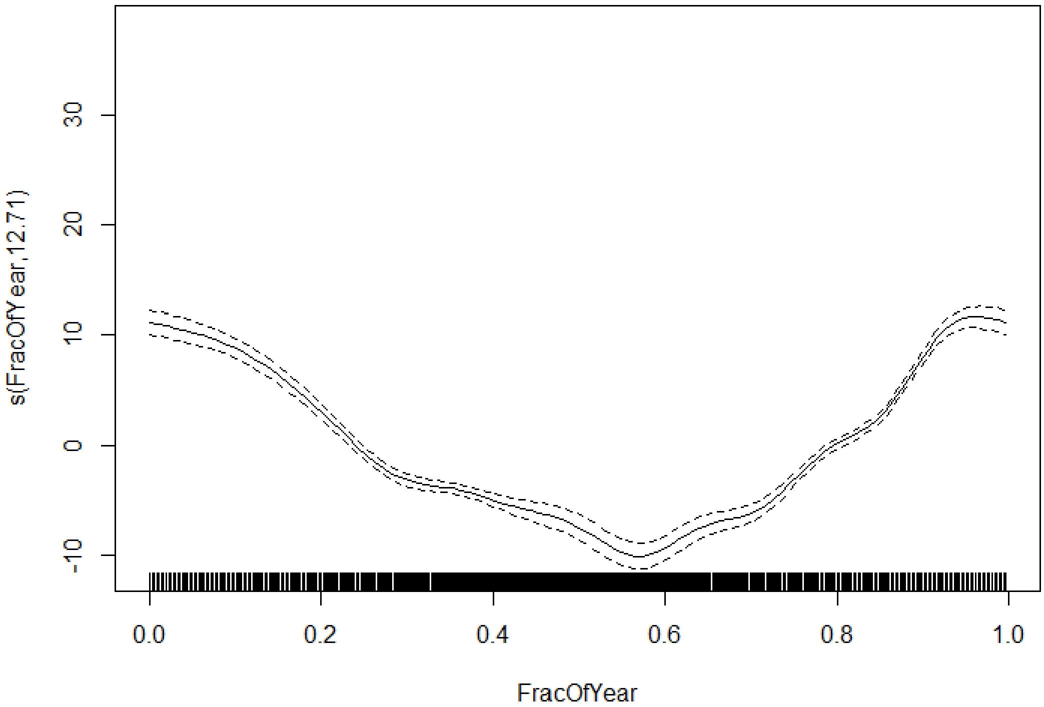

Figure 9.6.7: Effect of the variable "FracOfyear" on the fit for electricity.

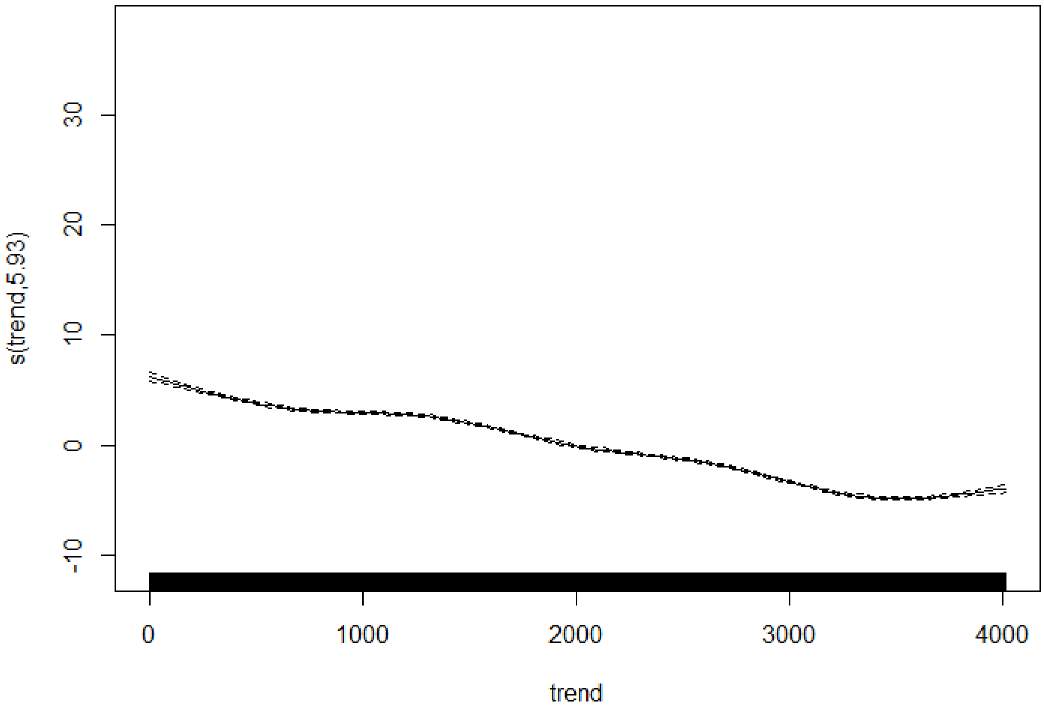

Figure 9.6.8: Effect of the "FracOfyear" variable on the fit for electricity.

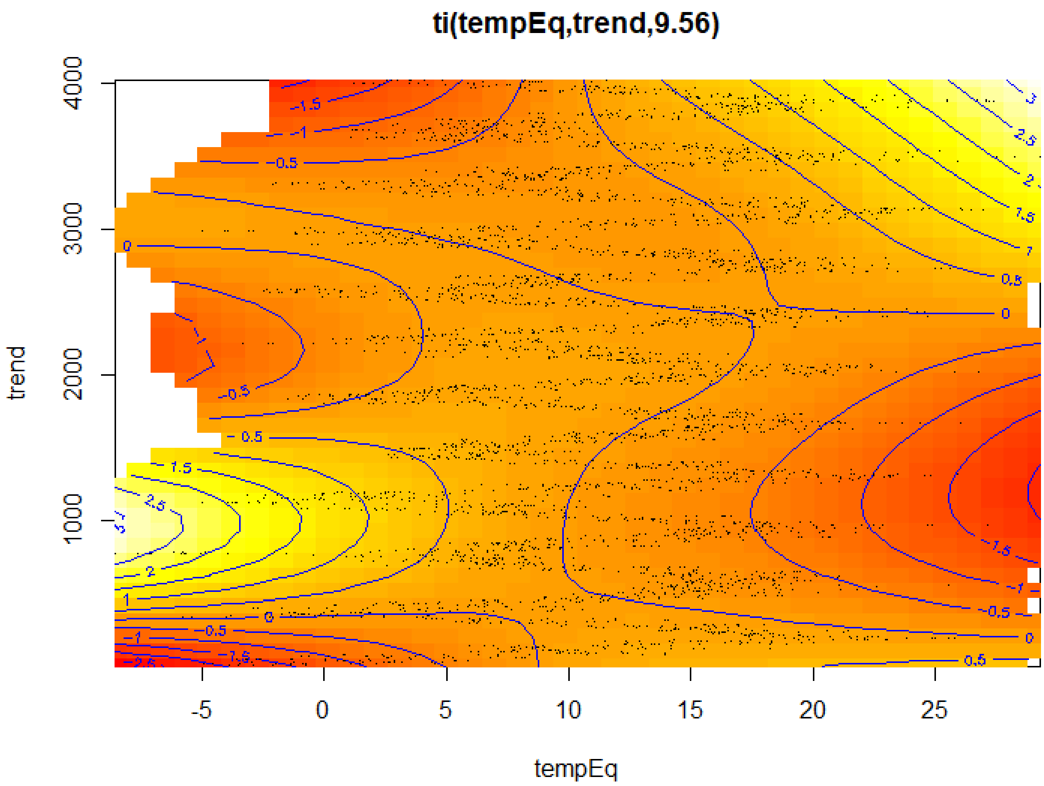

Figure 9.6.9: Effect of the interaction of the variables "tempEq" and "trend" on the fit for electricity.

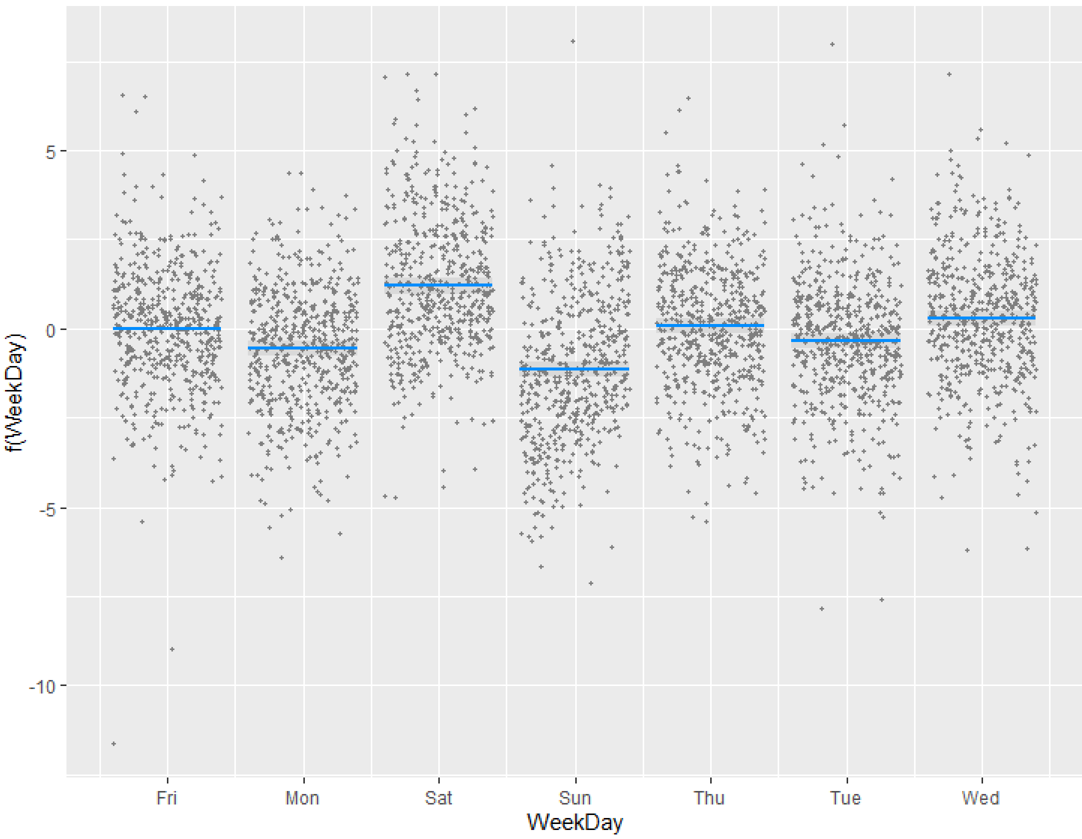

Figure 9.6.10: Effects of different days of the week on the fit for electricity.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**