FRUZSINA MÁK

VOLUME RISK
IN THE POWER MARKET

Department of Statistics

Supervisors:

Beatrix Oravecz, Senior lecturer, Ph.D.

András Sugár, Associate professor, Head of Department of Statistics, Ph.D.

Corvinus University of Budapest
Doctoral Programme of Management and Business Administration

Volume risk in the power market

Load profiling considering uncertainty

Ph.D. Dissertation

Fruzsina Mák

Budapest, 2017

TABLE OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

Acknowledgement

I would like to express my gratitude to my supervisors, Beatrix Oravecz and András Sugár, who have been helping me tirelessly throughout the PhD programme and have contributed to this dissertation with an enormous amount of support and advice – both personally and professionally.

I would also like to thank my pre-opponents, Gergely Fazakas and Gergely Mádi-Nagy for the valuable observations and recommendations they provided in the evaluation of the draft of this paper. I am grateful to Gabriella Jenei, who has contributed to completing the final form of the dissertation with a lot of inspiration and support.

I am grateful to many of my present and former Colleagues for supplying professional help and encouragement in the process of writing up this dissertation.

I am especially grateful to my Family for the incredible patience and mental support with which they have been helping my work for several years.

Other thanks are due to my Friends, who have contributed to the preparation of this dissertation with their encouraging, supportive words.

INTRODUCTION

Energy market participants face several risks in making operational and strategic decisions in the short or longer term. The handling and measurement of the majority of these risks have developed simultaneously with techniques commonly used in financial markets or as an extension of these methods adapted to the peculiarities of the energy market.

In parallel with the progress of **liberalisation,** EU objectives are bringing into prominence the necessity of realising successful **energy efficiency**, **energy saving** and the **reduction of consumption**. At the same time, those basic conditions that permit periodic checks of energy consumptions are gradually initiated with the spread of **smart metering** which often allows *online* tracking. Besides these basically micro level tendencies (interpreted at the level of the consumer) there are system level tendencies that manifest themselves, for example, in the handling of system level balancing problems or in the effort to decrease system level loss.

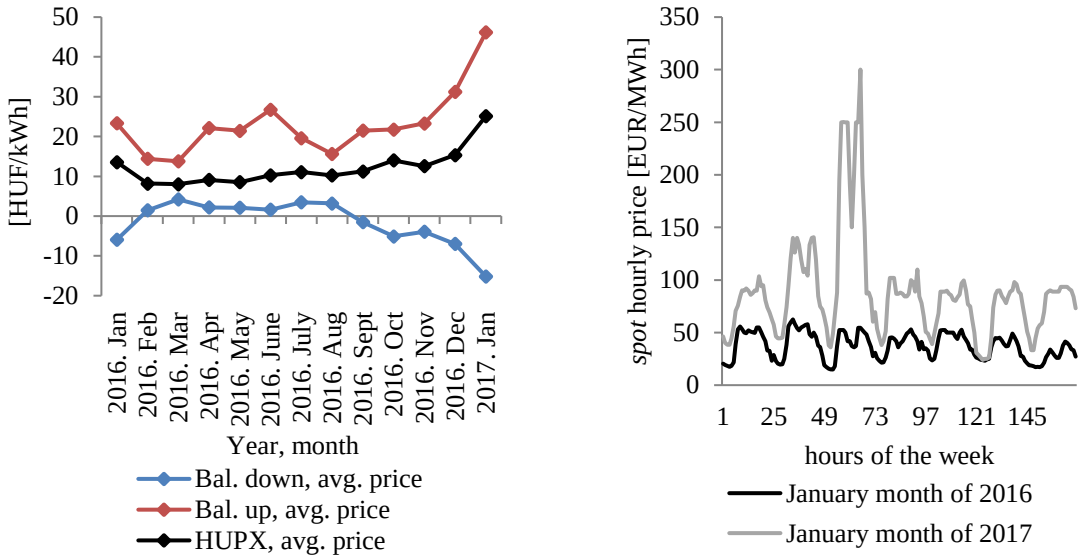
Although the source of the highest potential risks on the energy market is basically price, as a result of the above, consumer level behaviour becomes increasingly important besides the portfolio level and has growing business value from the points of view of not only energy companies and consumers, but also from the perspective of system operators.[1] On the **electricity market**, which is related to the topic of this paper – but also on other markets – there are more and more applications where it is not enough to be aware of the **(expected) consumption** but its **uncertainty** also needs to be considered, and the resulting risk needs to be dealt with.

Such a field is, for example, determining the portfolio level electricity demand (scheduling), hedging the portfolio in the long term, or the calculation of tariffs in relation to individual consumers. Certainly the above listed examples are interrelated on the one hand, cross-sectionally (the portfolio level curve is the sum of consumer curves) and on the other hand regarding time series (ex-post energy costs that occur as a result of forecasting errors while scheduling is added to the portfolio during the financial year).

[1] The proportion of the effect of consumer behaviour depends on the current energy market circumstances, such as energy market regulations and political decisions, as it is difficult to promote and encourage consumer saving with a downward pressure on prices.

Highlighting the importance of the topic from a **financial** perspective, let us examine the *dayahead* hourly *spot* and balancing energy price[2] trends in Hungary since the January of 2016.

Figure 1: The historical variation of balancing energy and *dayahead* hourly prices



Source: author's (partly) own calculations based on HUPX Ltd.[3] and MAVIR Ltd.[4]data and author's own figure.

Highlighting only a few weeks' *dayahead* hourly *spot* prices of the January months of 2016 and 2017 it can be seen that the supply side shocks (such as power plant outages, the lack of low-priced import power, etc.) can make *off-peak*, but especially *peak* consumption more expensive. Besides, observing the monthly average values, it is obvious what position the purchasing of positive (upward) balancing energy or the sell of negative (downward) balancing energy mean (compared to *spot* prices) for the ex-post trading and settlement of actual deviations from the schedule resulting from over- or underconsumption (see the left side of Figure 1); and how important the evaluation consumption-related volume uncertainty is.

Among the supply shocks it is definitely worth mentioning renewable energy producers (whose making headway is not at all unrelated to political fights). They increase

[2] For the exact definitions and explanations of terms used in the introduction see the subsequents chapters of the dissertation.

[3] Data source: www.hupx.hu, the company operating the Hungarian power exchange market: Hungarian Power Exchange (HUPX) Ltd.

[4] Data source: www.mavir.hu, the Hungarian power transmission system operator company: MAVIR Hungarian Independent Transmission Operator Company Ltd.

the volatility of prices because of the weather-dependent uncertainty of supply, so they mean risk to the power system imbalance. High power price – thinking about the relatively low price-inflexibility of consumption – may also be caused by demand side shocks[5], though they are *ceteris paribus* less likely to result in truly extreme phenomena in terms of prices or volumes.

It is often difficult and/or expensive to ensure the **supply-demand balance** of the power system at all times by the controlling of the supply (power plant) side. For this reason, it is not only the possible role of consumer habits but also their uncertainty in realising supply-demand balance that come to the fore. This is due to the fact that consumer habits are somewhat more adaptable, manageable.

The need for the quantification of the latter has greater emphasis on more developed markets. As an example, *demand side management* activities may be mentioned. On these markets it is a priority to achieve consumption reduction by tariff schemes in the short term (which means the involvement of the consumer in the balancing procedure, among others) or to guarantee longer term energy saving and to investigate related investment decisions. In these areas, the explicit consideration of risk cannot be avoided in any way, as these questions relate to establishing new pricing logic as well.

Obviously, regarding the practical tasks above, it would be way beyond the scope of this research to provide comprehensive answers. The aim is much rather to **contribute** to overcoming the above listed challenges **by the methodologically well-grounded consideration of consumption risks**.

In connection with the magnitude of the uncertainty of consumption, let us look at the following very simple statistical measures regarding the Hungarian system load, and an individual consumption curve (see Table 1).

The role of uncertainty is naturally much greater in individual consumption curves. Therefore, dispersion measures are typically higher (e.g.: standard deviation – often used to measure risk –, or the range obtained as a result of the subtraction of minimal from maximal values), though the degree differs by consumer. Such results are of course rough estimations, and do not have any fundamental explanation.

[5] In extreme summer heat, the more intensive use of air conditioners may cause extreme consumption peaks. For example, year 2015 was the first year where the yearly summer *peak* load exceeded the yearly winter *peak* load (6 456 vs. 6 447 [MW]).

Table 1: Descriptive statistics of the Hungarian system load and a consumer curve

Time series	Mean	Standard deviation	Minimum	Maximum
System load $[MW]$	4933.7	682.7	3047.7	6486.6
measure expressed in the ratio of the mean [%]		*13.8*	*61.8*	*131.5*
Individual curve $[kW]$	8.0	5.3	2.9	25.8
measure expressed in the ratio of the mean [%]		*65.5*	*35.8*	*321.4*

Source: author's own calculations and table.

As it is conventionally done in the literature dealing with a variety of risks, measuring risks is mostly expressed in money terms. This is also convenient for the investigation of energy markets (as financial losses may occur whether it is a wholesale transaction or a retail level consumer contract). In connection with the topic of this dissertation, however, apart from the financial processes the physical processes can also be interpreted and evaluated in themselves. For this reason, the analyses in this study use primarily consumption (load) time series. In international, especially English language literature, the use of the term *load curve* is more widespread than *consumption curve*. The difference between the two is, *inter alia*, in measurement unit; in this dissertation load is usually measured in $[kW]$ and consumption in $[kWh]$, nevertheless, the results and conclusions are independent from the terms used or measurement units.[6,7]

The **purpose** of this study is on the one hand, the appropriate modelling of consumption; and on the other, to develop a method for the appropriate measurement of consumption risks that is methodologically well-grounded and can also be easily interpreted, applied and used in practice. For the latter, only *ad-hoc* measures and rules of thumb exist; and classical techniques provide few opportunities, especially for modelling uncertainty.

There is a wide range of (far from uniform) literature on consumer profiles and their applications. These profiles are achieved as a result of using basically quantitative methods, and they describe how consumption is dependent on various seasonal, calendar or other effects[8].

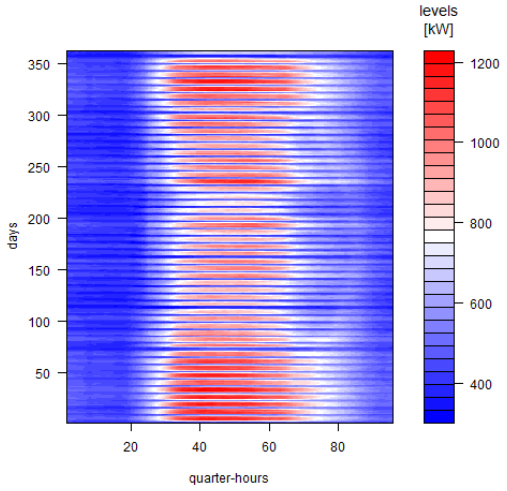
[6] In a quarter-hourly time series the average 1 $[kW]$ load realised in a given quarter-hour equals 1 $[kW] \cdot {}^{1}/_{4}\,[h] = {}^{1}/_{4}\,[kWh]$ consumption.

[7] A similar difference in measuring units exists in natural gas industry where consumption is measured either in $[m^3]$ or $[TJ]$, but since the gas year of 2015/2016 nomination (the submission of gas demand, the natural gas counterpart of scheduling) is done in $[kWh]$ here as well.

[8] Based on the literature, there is no single, general-scope definition.

The following *contour plot* (which will be described in more detail in this paper) demonstrates how much consumption can vary in time: it gives a compact picture of the yearly (y axis) and daily (x axis) behaviour by indicating different load levels with different colours. Throughout this dissertation, the comparison of such plots will be very suggestive of how much heterogeneous the patterns are that need to be captured during profiling, already at the level of expected values.

Figure 2: The *contour plot* of a portfolio time series



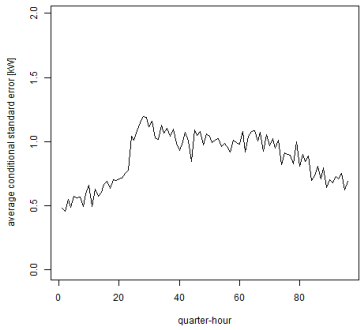
Source: author's own figure (R).

Since consumption itself is **stochastic**, its risk, uncertainty or portfolio effects also need to be taken into consideration in a similar way as it is usually done in financial time series. The fundamental difference in its handling is the result of the fact that consumption (and its uncertainty, as we shall see) is much more likely to lend itself to being modelled by various **fundamental** variables than the financial time series themselves. Hence the range of possible methods that are applicable is necessarily different, although some degree of analogy or parallelism exists. What is meant in this dissertation by modelling of 'consumption-related uncertainty', that is, modelling of 'volume risk', is the description of the behaviour of irregular component in consumption.

In the completion of the empirical results in this paper – through the relaxation of the requirement of the constant standard deviation of error terms that applies to classical regression time series – the standard errors calculated are conditional (time-dependent) with respect to given independent variables. This model-based calculation of conditional standard errors is definitely a new outcome; such empirical results are few and scarce in the

related literature. In this dissertation results seen in Figure 3 will also be computed for individual curves. The investigation will focus on when and to how extent the uncertainty of consumption is higher or lower, and what fundamental explanation can be provided.

Figure 3: The average of conditional standard errors in a consumer load curve (winter weekdays)



Source: author's own calculations (R) and figure (R).

In parallel with the logic of the previous two illustrative research results, the dissertation examines the following **fields**:

- how various consumption time series can be characterised; which so-called **stylized facts** that are otherwise described by any model of consumption need to be captured in the course of profiling as well;
- what trends can be explored in consumption **uncertainty** of various consumption time series; can (multiple) **seasonal** or any other regular pattern be observed that otherwise also characterises the consumption time series themselves;
- how all of the above can be **model**led with a special focus on, for example, the simultaneous handling of the **nonlinearity** – especially (among others) weather-dependency – and **heteroscedasticity** (non-constant standard deviation over time).

The following **hypotheses** have been formed to investigate the above questions, fields:

- H1: In electricity consumption curves the intraday seasonality is the primary source of variance in the curves.
- H2: As compared to the so-called classical methods (that rely on the typical daily profile) the extraction of the relevant individual features of the curve opens a new avenue towards the development of more realistic profiles.

- H3: Assuming constant standard deviation of the residuals results in either the under- or in certain periods the overestimation of the volume risk.
- H4: Volume risk is not constant over time, but varies depending on various exogenous variables, seasonal and calendar effects.

The **first chapter** deals with fields of application of consumer profiling. It is described what specifically is meant by profile and profile-risks, and the one (volume risk) that receives special emphasis in this paper.

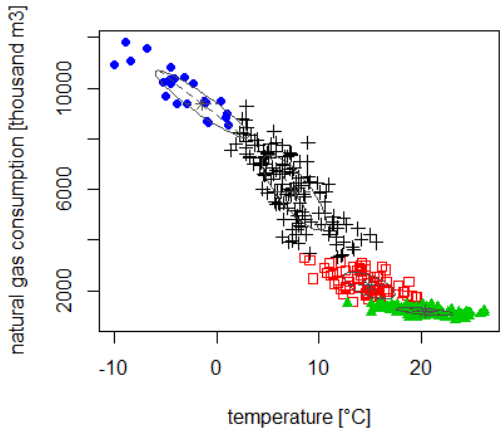
Subsequently, stylized facts will be examined for various consumption time series, with particular attention to multiple seasonality and the highly different behaviour of stochastic errors for each curve. These methods, though rarely used in practice, point to important relationships that may not be obvious based on, for example, a time series figure. The perspective taken during the analyses – in light of the results of previous studies (described later on in this paper) underlines the usefulness and justifiability of the novel viewpoint of this dissertation.

The **second chapter** is a review of the essential literature. The most commonly used methodological – basically various clustering and regression – techniques are overviewed briefly. A separate section is devoted to the most important aspects of the treatment of temperature variables. Highlighting the substantive part of a previous publication of the author of this dissertation, it is shown through the example of a natural gas consumption time series how the role of irregular effects of temperature or its magnitude can be measured and how nonlinearity can be dealt with in an elegant way. The chapter will also cover the reasons why these kinds of techniques are not so advantageous in profiling – whether it be methodological issues (e.g.: too many preadjustment steps), problems in interpretation or the applicability of the results (if, for example, some relevant features remain hidden).

The **third chapter** discusses the methods used to reach the empirical results of this study. As some of these results are common in stochastic time series analysis the emphasis is rather on outline the framework of methods used. It describes the essence of the periodic autoregressive (PAR) model[9] and the possible procedures used for analyses. This is lesser known in practice, but is essentially an extension of the classical ARMA model and its useful properties can be exploited both in integrated and stationary time series.

[9] Periodic autoregresive model.

Figure 4: Results of mixture clustering on the example of daily average temperature – natural gas consumption



Sources: author's own calculations (R) and figure (R).

The chapter comprises mostly of the discussion of the methodological (and theoretical) background of mixture models (especially the *Gaussian* mixture model). The detailed discussion is needed, *inter alia*, because this methodology has seldom been used so far in practice in the field of energy. Therefore, an empirical example will be used to illustrate (on domestic natural gas consumption data) how the method can be applied for the modelling and capturing of the relationships within a varying covariance structure of the variables. The assumption (starting point) of the mixture model has a central role both in profiling and in measuring the uncertainty of consumption, which is that the covariance structure of the variables in the sample is not constant: this is markedly important both in modelling the expected value and regarding standard error. The figure is telling of this logic: the relationship between variables differ by 'groups of dots' and it is clear that the dispersion of these 'groups of dots' are not the same either. The former phenomenon will manifest itself in capturing nonlinearity and heteroscedasticity.

The chapter contains a number of formulas and interpretations that do not appear even in foreign language literature, although they show more distinctly how mixture model based regression and classical multivariate regression are related, which is extremely useful regarding the results of this study. At the end of the chapter there will be a discussion of the use of mixture models in profiling in previous studies that have different focus from this dissertation, thereby providing the transition to the introduction of the new empirical results.

The **fourth chapter** contains the empirical research results of this study concerning the creation of consumer profiles and the measurement of their volume risk. In the sections

on both profiling and uncertainty, classical techniques appear partly to serve as a *benchmark*, to provide a basis for comparison of the new results, partly with an exploratory purpose.

It will also be examined how the framework of mixture models can add to the creation of so-called profile groups, that is, groups of consumers that can be characterised by similar consumer behaviours. To illustrate the difference, the results are compared to a classical (regression-based) technique. It is also shown what can be considered as a 'typical' consumption pattern using the logic of the new methodology.

Using the results of the methods of classical regression time series models, it was examined what – basically heteroscedastic – behaviour can be observed in residuals, the behaviour of which may also differ by consumer. The findings are then compared to a simpler, so-called heuristic indicator that is used in practice to measure volume risk. It will also be shown that by an appropriate modelling of covariance structure the variation of time-dependent standard error can be modelled well with mixture regression.

On the whole it can be said that the results of this dissertation **contribute to Hungarian and international research results and applications**, *inter alia*, the following:

- In the field of consumer profiling this thesis examines an area which takes into account uncertainty in defining consumer profile, which is more suited to the requirements continuously emerging in practice.
- The methodology used here – which has seldom been used in profiling – is such that it can model both consumption and its uncertainty as a function of various exogenous variables that mean basically seasonal or calendar effects.
- It is shown that the mixture model – which has various favourable properties even purely methodologically – can capture the fundamental reasons of consumption, to which the formation of profile groups has already given very good (indirect) evidence.
- Mixture regression is suitable for capturing the heteroscedastic behaviour of the error term, which is evidenced by investigating the consistency of errors and the calculated conditional standard deviations and the underlying factors. It is also shown that using mixture regression, it is possible – on average – to reach a much

narrower confidence interval compared to the practically almost unusably wide interval of classical techniques.

- The dissertation has a number of implications building on which either results shown here or their energy market (for instance, natural gas market) application may also open avenues for further research.

It is a pronounced and by all means highly important outcome of the dissertation that it implements the *Gaussian* mixture regression building on the *R Project* software package. Besides the completion of the listed tasks and the evaluation of the results, this is an important milestone of this research.

1. APPLICATION EXAMPLES AND TERMINOLOGY OF CONSUMER PROFILING

Both the literature and applications are rather heterogeneous regarding the definitions of (consumer) **profiling** and (consumer) **profile**. This may be due to fact that the meaning of the term 'profile' is often dependent on its field of application, and accordingly, the methods used may be diverse.

The term 'curve' is commonly used in business applications and theory (methodology) as well. On the energy market what are meant by curve may be load curves representing load, or price curves representing price time series. In different statistical or data mining fields there is an increasing interest in theoretical and empirical studies related to processes where the studied phenomena under examination are *functionals, curves* or *time series.*[10]

What definitely needs to be established is that consumer profiling means a basically methodological, modelling procedure that describes the temporal variation of the consumption or load curve for each individual consumer (or portfolio), and how it is dependent on seasonal and calendar effects. A reasonable requirement that such a task should meet is that profiles should be able to capture the so-called **stylized facts** that characterise curves in a general sense.

The dissertation focuses mainly on **consumer** profiling and on creating consumption or load curve profiles; therefore, besides the most important applications and terminological issues, the empirical study of stylised facts of consumption time series is the other main topic of this chapter. The two terms used (consumption and load curve) mainly reflect the differences between technical and business-economic aspects. This paper uses the two terms simultaneously (on the level of terminology, the term 'consumption curve' is preferred, but calculations are performed on 'load curves'), the results and conclusions of this study are independent from this.

1.1. Price- and volume uncertainty on the energy market

On the energy market the supply and demand need to be in balance at all times. The demand may be hedged by so-called **standard products** (for example yearly, quarterly,

[10] From the perspective of the dissertation the following terms may be relevant: *functional data analysis*, *functional clustering*, *curve clustering*, *time series clustering*. See more on this in Chapter 2.

monthly, etc. *base* load and *peak* load products[11]), then on the day before delivery the open position can be hedged *ex-ante* with hourly products on the *dayahead* market (or with quarter-hourly products in the *intraday* market). Deviations from the schedule are settled *ex-post* on the **balancing** energy market. The time-dependency of trading various electricity products is shown in Figure 5.

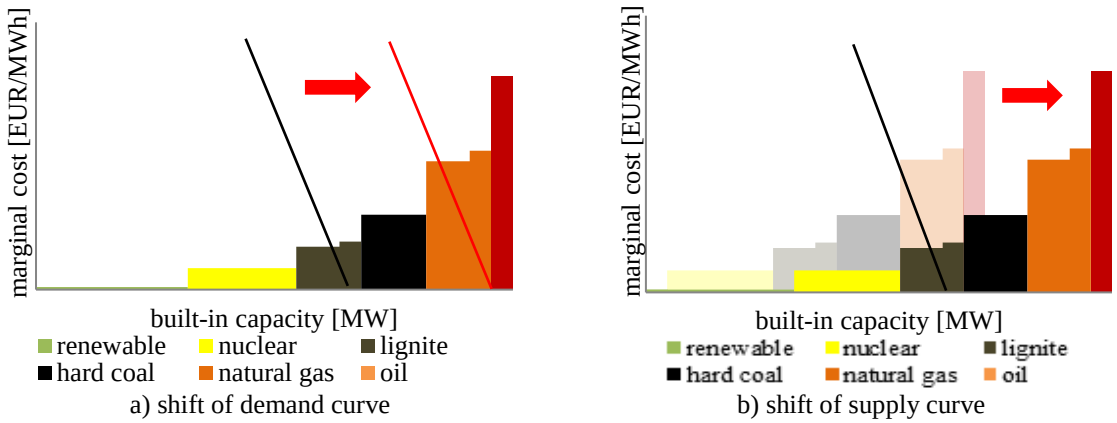
Figure 5: Wholesale products in the electricity market

Ex-ante		Ex-post
standard product[12] (*forward* products, *FM*, *FW* … *WE, DA* for *base* and *peak* periods, etc.)	*dayahead* (*spot*), *intraday*, hourly, quarter-hourly products	balancing energy

Source: author's own diagram.

The prices of various electricity products are determined by the *merit order* curve as the supply curve, given the current demand conditions. Figure 6 is an example of a *merit order* curve which shows the marginal costs of power plants as a function of (built-in) capacity.

Figure 6: Power plant merit order curve



Source: author's own figure (Excel).

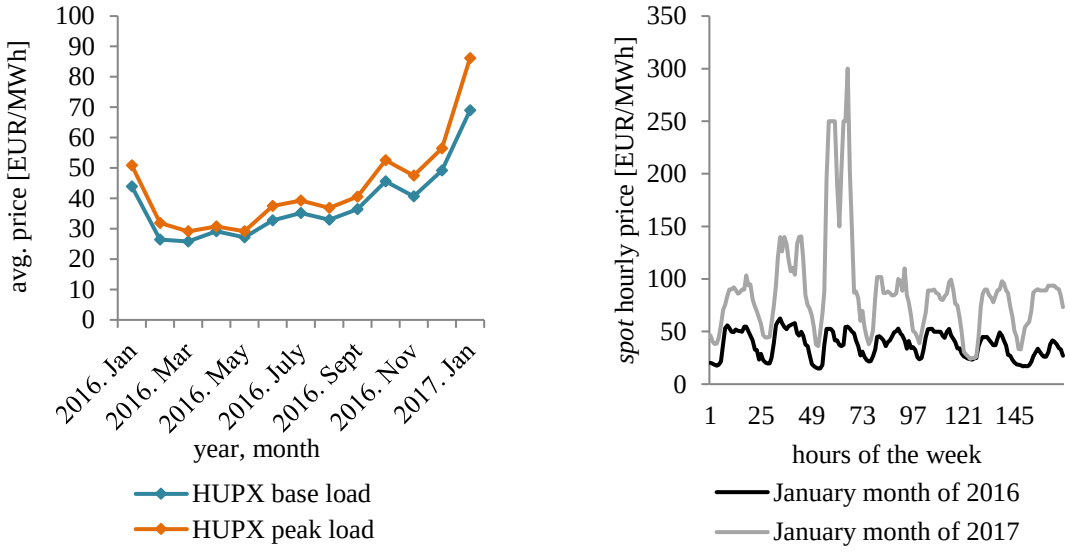
Renewables with a zero marginal cost are on the left side of the *merit order* curve, followed by nuclear plants, coal and gas-fired (OCGT and CCGT[13]) plants. The oil-fired power plants operating at a high marginal cost are on the right side of the *merit order* curve.

[11] The *base* load product is available on every day of the week in every hour. The *peak* load product is only available on weekdays between 8:00 and 20:00.
[12] The *forwards* may apply to further expiry dates. The closest expiry dates are the *front month*, *front week*, or even the next day (*dayahead*), but this category contains products that are available in the subsequent few weekends or days as well.
[13] The common short form OCGT refers to *open cycle gas turbine*, and CCGT refers to *combined cycle gas turbine* power plants, the latter represents higher efficiency.

As the figure shows, the determination of the market clearing price may be influenced by the relative position of the demand curve and the supply curve.[14]

Hence, price determination may be partly demand-driven: the higher electricity demand (the demand curve with a right shift on the left figure) of weekday and *peak* periods, or temperature-dependent periods result *ceteris paribus* in higher electricity prices. As electricity as a product cannot really be substituted, cannot be stored (or is extremely difficult to), the demand is relatively price-inflexible, the electricity prices are highly volatile; there may easily be increased prices or *peak*s (even *spike*s).

Figure 7: The historical trend of the Hungarian *dayahead* electricity market prices



Source: author's own figures (Excel) based on HUPX Ltd. data.

In recent times the supply-driven nature of prices is gaining more attention (often spiced with political fights). The outages of (conventional) power plants may result in higher electricity prices. A very good example for this is the January(-February) period in 2017, when record-high prices occurred on HUPX[15] on the *dayahead spot* market[16] partly as a result of power plants outages caused by the extreme cold, partly due to planned maintenance; in addition, to the lack of often available low-priced import electricity from southern countries.

[14] In the figure on the right a supply shock is shown, where the growing renewable capacity shifts the supply curve right. See the later part of the section on this.
[15] See the so-called inside information website of HUPX (the short name of the Hungarian power exchange) https://www.insideinformation.hu/hu/pubpages/newslistmain.aspx.
[16] There was a trading day in the first half of January when several hourly prices exceeded the historical maximum.

Figure 7 supports claims in previous sections about the average monthly *base* load and *peak* load prices on HUPX. On the right side, a week was chosen from the January months of 2016 and 2017 to illustrate the magnitude of the ominous price trend.

Of course, volatility of prices is not only related to such classical *outage* risks of conventional power plants. The production of those renewables (wind- and solar power plants) that are on the left side of the *merit order* is highly weather-dependent, which introduces a producer with a new kind of risk on the left side of the supply curve. Risk here is not only caused by the exceptional variability of their production, but also forecasting and consequently the outstanding uncertainty of the *dayahead* production (schedule) submitted. As an illustration of this, let us examine some measures of the Hungarian wind power generation trend since the beginning of 2016 (see Table 2).

Table 2: Descriptive statistics of the Hungarian wind power generation in 2016

Wind power generation	**Mean** [MW]	**Standard deviation** [MW]	**Range**[17] [MW]
Actual	71	77	311
Planned	76	70	298
Actual-planned difference	- 4	31	344

Source: author's own table based on MAVIR Ltd. data.

The total installed capacity of Hungarian wind turbines (approx. 300 MW) is by an order of magnitude lower than the average values of the Hungarian system load (which, depending on the season means a daily mean load between 4000 and 5500 MW, see Table 24 in the Appendix). In any event, it is notable that the range of differences between actual-planned productions is fairly high. That is, there is a risk not only in whether the renewable capacity between 0-300 MW is actually there and available on the left side of the *merit order*; it is also a question to what level of certainty we can determine the value of production. Due to weather-dependency, the latter cannot be planned in a way that it is usually done for conventional power plants.

This dissertation does not deal with supply side uncertainty and its measurement, although further consideration and application of the results from the supply side perspective is definitely worthwhile and opens up new avenues for further research. The above short example referring to supply side uncertainty draws attention primarily to the fact that management of the demand side uncertainty – which can be dealt with to a certain

[17] Range is the difference of the maximal and minimal generation values.

extent by demand side management – may have an important role (this issue will be discussed later).

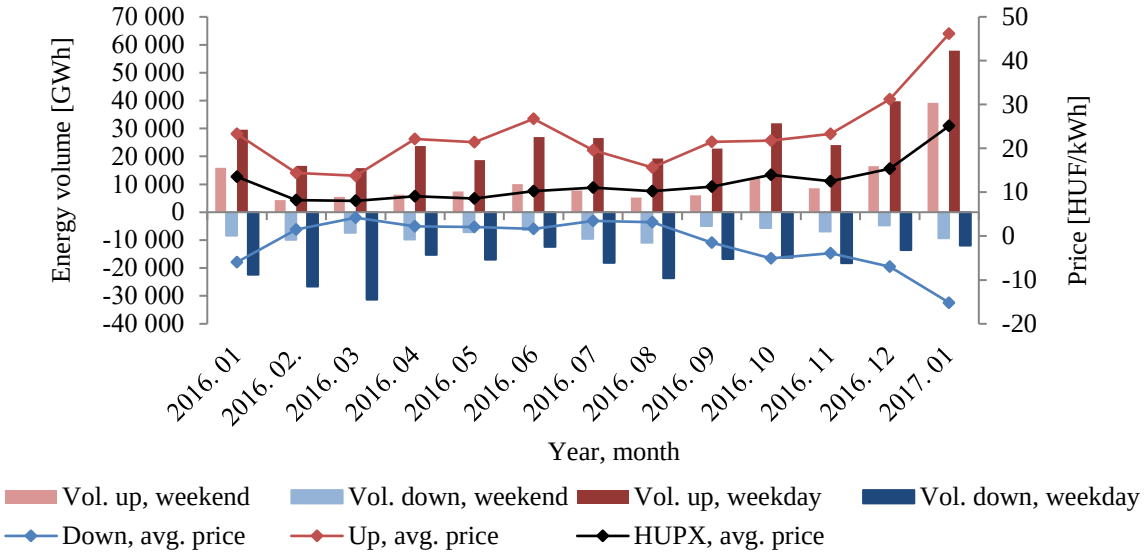
Returning to the prices of various products on the electricity market, Figure 8 shows trends in hourly *dayahead* prices, balancing energy prices and system level balancing energy volumes since January of 2016. The *dayahead* prices are determined through auctioning. The decision about which power plants will supply the balancing energy that is required because of deviations from the schedule is governed by the transmission systems operator, essentially on *merit order* basis.

Examining the figure it can be stated that the average price of the upward balancing energy exceeds by orders of magnitude the *dayahead* hourly prices, and at the same time moves closely together with the upward balancing energy amount as opposed to the downward direction. The monthly averaging on prices obviously hides short term (e.g. intraday) tendencies, but it is clear that balancing volumes are higher on weekdays, apart from the season. Behind this, on the one hand, is the higher **consumption uncertainty** of weekdays, and on the other, there are **structural** reasons. The morning ramps of weekdays are higher in magnitude and speed than those of weekends; and here, power plants on the balancing energy market have a great role in adjusting supply to the demand, as the schedule only provides a quarter-hourly step-by-step constant load[18]. Nevertheless, the supply-demand balance needs to be maintained at all times. Besides demand side explanations, a huge amount of system level balancing energy may be due to power plant capacity **failures** (that is, supply side causes), which is particularly outstanding in the 2017 January period.

Apart from the above, system level results suggest that there may be some regular pattern in demand side uncertainty (maybe even seasonally?), which is realised in system level deviations of actual versus planned consumption; in addition the realisations on the level of smaller portfolios or individual consumers may be very similar in a qualitative sense. A favourable outcome of having a system level example here is that it draws attention to both the macro level and supply side related aspects of the problem.

[18] In practise, this rather means hourly steps, which induces the use of even more balancing energy. That is, even if quarter-hourly load is estimated correctly, balancing energy will still be needed.

Figure 8: The historical trends of the hourly *dayahead* and balancing energy prices[19]



Source: author's own figure (party own calculations) based on MAVIR Ltd. data.

One of the central topics of this dissertation is that **consumption** is basically a **stochastic** process, and the role of **random shocks** may be smaller or greater depending on the given consumption time series. Applications in practice pose more and more requirements where besides consumer behaviour, its uncertainty also needs to be known.

In part due to reducing energy costs and also because of global trends (environmental protection, saving energy resources, etc.) the active participation of consumers on the energy market is increasing. These are more widespread on developed markets, and on less developed markets appear in large consumers. The essence of *demand side management* (DSM) is that the continuous balance of supply and demand at all times cannot only be maintained by the management and controlling of power plants. Demand side management may have various technical or economical (micro- or macro level) goals, including the following:

- smoothing the consumption curve by decreasing *peak* consumption and increasing *off-peak* or *super off-peak* consumption, which may help decrease system loss, or delay of big investment decisions;

[19] Based on MAVIR publications average prices are the following: the average balancing energy price is the average of the positive and negative balancing energy prices of the quarter-hourly settlements; the HUPX price is the average of the HUPX hourly prices exchanged on the official exchange (FX) rate valid on the given day of the settlement period.

- decreasing electricity procurement costs by shifting *peak* consumption to a period where energy price is lower;
- decreasing balancing power costs by shifting consumption in the case of a major deviation from the schedule;
- mitigation of uncertainty resulting from the changeable and hardly predictable (possibly decentralised) renewable energy production.

Realisation (even partially) of these poses various future challenges to energy market operators. If, for example, shift in energy usage becomes habitual behaviour that would mean a structural reshaping of profiles and this in turn would have consequences for the balancing energy market. It would be challenging from the aspects of both system balancing and infrastructure for system operators.[20] The various long and short term aims of shift in energy usage involve many pricing challenges and tasks. Due to the fact that **deviation** from the schedule is a recurring issue here, together with the **statistically grounded** evaluation of **significant** changes in *baseline* consumption scenario, the handling and modelling of consumer related uncertainty will increasingly become a central theme in the future.

In the introduction so far the emphasis has been more on system level questions. In what follows, the focus will be mainly on profiling and its applications in practice, but as we shall see, the two are not completely independent from each other.

[20] Although smoothing daily consumption is among the main goals of demand side management, what is usually mentioned here is reducing extreme consumption – there is a wider range of available tools, though. Without aiming to give a comprehensive list, the handling of daily consumption may involve some of the following solutions: *peak* clipping, valley filling, strategic conservation, strategic load growth, load shifting; see more on this in: *Macedo et al.* [2015].

1.2. Some application examples on profiling

Regarding the application of consumer or demand side profiling, the following can be identified - without aiming to give an extensive list:

- scheduling,
- (short and long term) *hedge* of consumer portfolio,
- pricing of the so-called conventional contracts[21],
- demand side management,
- creating portfolios, balancing groups.

As this dissertation places great emphasis on the uncertainty of consumption, the possible fields of application will be discussed with regard to what extent the involvement of this uncertainty can modify or complement the regular decision-making process which only examines the **expected** value of consumption, and is less concerned with **uncertainty**. At the same time, this approach highlights the fact that it is becoming an increasingly realistic requirement from profiling that it deals with uncertainty appropriately; which results in drawing important conclusions from both theoretical (methodological) and practical aspect. From a methodological perspective this is off the beaten track, it is not really examined; and its application in practice has not yet taken place even for classical techniques.

It is worth mentioning that profiling as a term usually appears in the examination of patterns in consumption curves; however, the term is often used in other instances, such as the analysis of portfolio level curves.

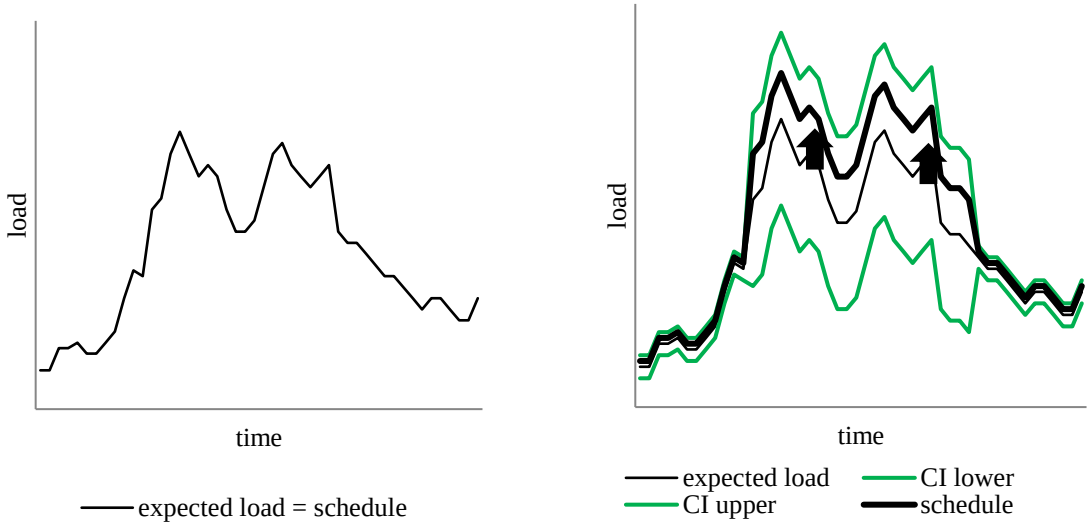
1.2.1. Short and long term hedging and pricing

Referring back to the relative price evolution of various electricity market products, given the risk-avoiding trader behaviour, the fear from high *dayahead* or balancing energy prices may result in **overhedging**. In the following, it will be shown through examples what this exactly means.

[21] The so-called full supply and schedule contracts are such. In a full supply contract, the trader ensures supplying the demand of the customer, and the customer assures the acceptance of the energy supplied. In a schedule contract the customer and the trader give an undertaking as to the supply and acceptance of a specified quantity.

In Hungary, the *dayahead* **schedule** needs to be created in quarter-hourly resolution, which is one single curve consisting of 96 quarter-hours. As a first thought, it is best to set the quarter-hourly forecast of the *dayahead* load as the *dayahead* schedule.[22]

Figure 9: The schematic representation of determining the schedule[23]



a) without considering uncertainty

b) with the consideration of uncertainty

Source: author's own diagram (Excel).

The positive balancing energy price (when consumption is higher than the schedule) is generally much higher and the negative balancing energy price (when consumption is lower than the schedule) is generally much cheaper[24]. For this reason, it might happen that traders deviate upwards from the forecast[25]. The uncertainty of load time series is not necessarily constant in time, and due to this, the schedule needs to be set for more uncertain quarter-hours at a much higher value than the expected load (in such cases, the balancing energy price is usually higher) than otherwise. In financial terms, due to risk avoiding behaviour, the extent of overhedge is higher in more uncertain periods and lower in less

[22] The *dayahead* schedule may of course be still changed within the day by modifying the schedule if there is enough liquidity on the *intraday* market.

[23] **Schedule** means a higher (upward shifted) schedule compared to the forecast load due to risk avoidance. **CI lower** means the lower limit of the confidence interval for the load given a chosen level of confidence, **CI upper** denotes the upper limit here and in Figures 10-11-12.

[24] The practice is actually much more complicated, as the price of balancing energy is not only a function of the balancing group but also of the system imbalance.

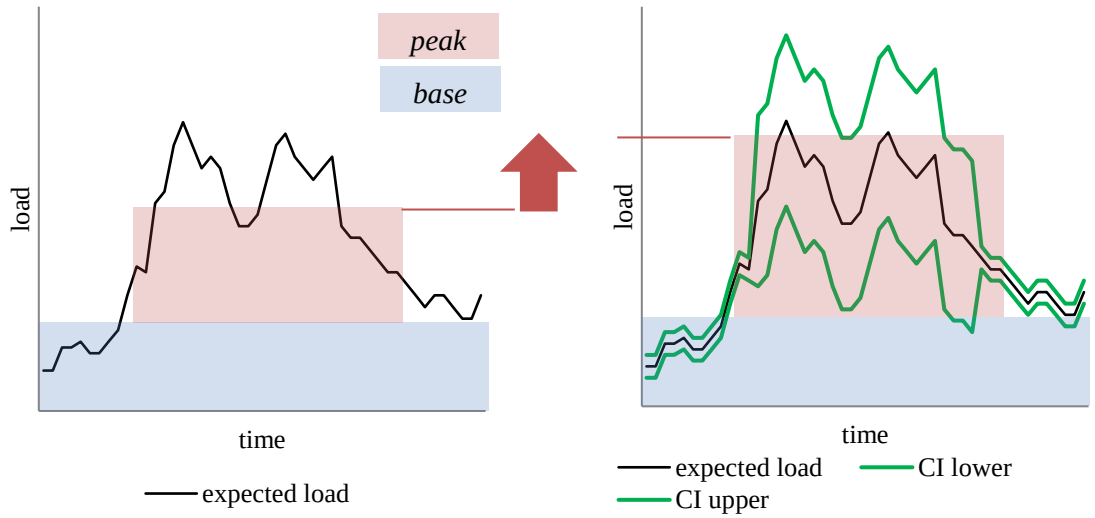
[25] This may be due to risk avoidance mentioned above, but also occur as a result of speculations.

uncertain periods. The additional cost that incurs from overhedging can be understood as a risk premium for the avoidance of high balancing energy prices.

Figure 9 is a schematic representation of this problem. Of course, in the example for the identification of the appropriate schedule it is necessary to consider not only the uncertainty of load, but the prices as well. This question is dealt with in *Lo and Wu*'s [2003] study.

Of course, the risk avoiding behaviour presented here may also be valid for the **long term** (actually, logically it precedes the scheduling example in time, though in this way, it is simpler to introduce the problem). In this case, overhedging compared to what may be explained by the expected load can be optimal, which is likely to mean overhedging at a larger amount in *peak* periods. This strategy is illustrated in Figure 10.

Figure 10: Incomplete hedging of consumer demand with *base* load and *peak* load products



a) without considering uncertainty

b) with the consideration of uncertainty

Source: author's own figure (Excel).

Actually, this logic also appears in the **pricing** of (mainly larger) **consumers**. With the progress of liberalisation, consumers entering the free market can choose which trader they buy energy from, and traders can decide for themselves what tariff they offer to each customer. Knowledge of consumer habits, and thus of consumer profiles, is especially

needed for full supply contracts[26], as in *fair* pricing every single consumer should obtain a price appropriately tailored to their **consumption habits**. A single consumption curve may be regarded a consumer demand that can be hedged with the purchase (or sale) of wholesale products. The consumer that expectedly consumes more in *peak* period pays a higher **price for the energy than whose *peak* period consumption is lower**. Taking one step further in the consideration of consumption uncertainty, if the consumption is more uncertain in *peak* periods, then as a result of possibly higher than expected consumption levels of uncertain hours a higher price is given by allocating more *peak* products.[27]

This dissertation does not intend to develop such a forecasting or optimisation (operational research) logic, because it is beyond the scope of this paper. The chapter on research results, though, provides an excellent basis for such studies.

1.2.2. Demand side management

Familiarity with consumer habits is not negligible here either – it may actually have a much more central role, and of course, various consequences on pricing. Such tariff scheme is for example the so-called *Time of Use* (ToU) tariff, when prices differ according to time zones (for example *peak* or *off-peak*); or another one is the so-called *Real Time Pricing* (RTP), where the price risk of procurement is passed on to the consumer; but other techniques may exist when only deviations from the profile are priced real time.

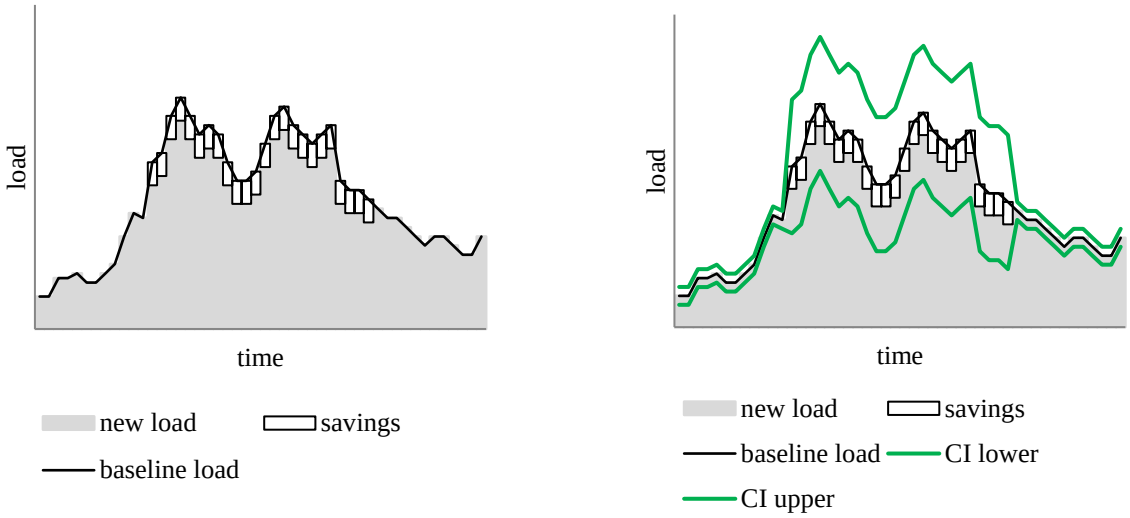
Demand side management can be interpreted not only in the short term but also with a long term perspective. For example in the case of various **energy saving** projects it needs to be considered – for example in an office building – that the *peak* period consumption is typically more uncertain than *off-peak* period; therefore, the uncertainty of the amount of expected savings based on calculations of past load data (that is, the so-called *benchmark* profile) is season-dependent. Companies specialised in energy saving (the so-called ESCO-s, *energy saving companies* or *energy service companies*) that sign contracts with costumers to realise a given amount of saving for them in a given period of time should

[26] In schedule contracts, the quantity of energy is given in detailed resolution; in this case, under- and over-procurement may only occur due to technical failures and may pose a risk, but these are usually settled contractually.

[27] Of course, there are consumers **not metered** regularly, where reading of meters only takes place in fixed intervals (see e.g. submitting meter reading), and the pricing logic shown here cannot be applied. Still, settlement is based on profiles. The basis of profile grouping may be based on a smaller sample of consumers – whose time series data is available. Then, having created profile groups, some classification methodology is used to investigate which non-consumption properties (e.g. scope of business activity) explain the profile groups (as classes) based on metering data, and on these grounds not metered consumers can also be assigned to a profile group.

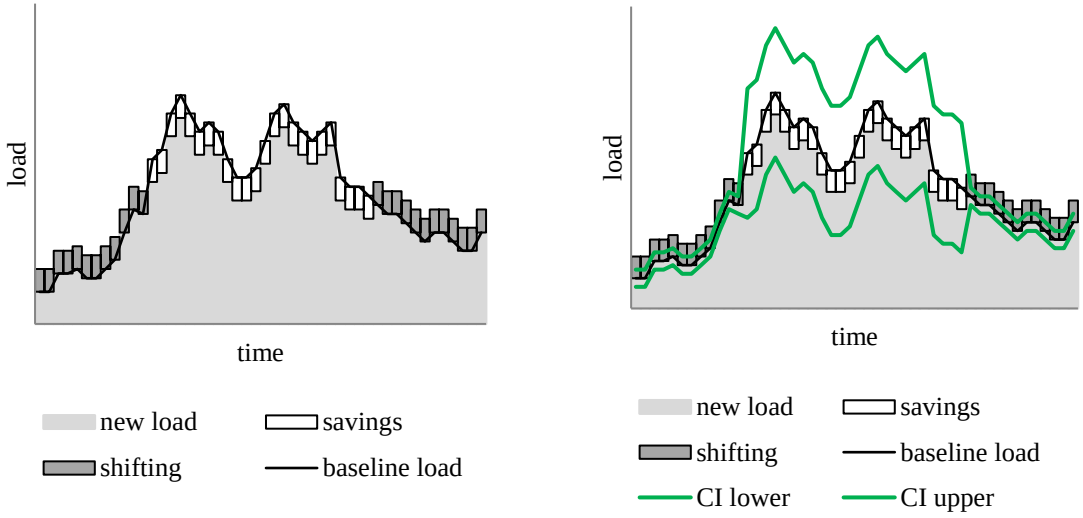
definitely take this uncertainty into consideration (see for example: *Srivastav et al.* [2013], *Heo et al.* [2012], *Manfren et al.* [2013]).

Figure 11: The evaluation of the significance of saving in energy use



Source: author's own figure (Excel).

Figure 12: The evaluation of the significance of shifting energy use[28]



Source: author's own figure (Excel).

Figures 11-12 demonstrate that deciding about whether there has been a significant amount of saving or shift in energy use in the case of a certain consumer is a question that cannot be answered without explicitly dealing with uncertainty.

[28] Notations of the figure indicate that the energy saved in the *peak* period is shifted to the *off-peak* period.

This kind of long term, or even sustainable consumer activity appears not only in the motivation to achieve savings, but also in the spread of the so-called **microgeneration**, which may also regarded as a demand side investment decision. Appropriate knowledge of the consumer profile is also important here, as it is advantageous if the microgeneration production profile can match the consumer profile of the household (see for example: *Hino et al.* [2013]).[29]

Obviously, the above fields of application require the technical capacity for computerised recording, which enables analysis of high frequency consumption measurements. In relation to demand side management activities it cannot be avoided to obtain the most detailed information of the consumption habits of the consumers.

1.2.3. Building portfolios and creating balancing groups

Although it was not mentioned in previous sections, it should be noted that consumers typically consume not in themselves, but as part of a portfolio. This portfolio may be a trading portfolio or a balancing group through which deviations from the schedule are settled through the allocation of balancing energy.

Portfolio effect actually often means that taking together a mainly *peak* period consumer with a mainly *non-peak* period consumer the outcome is a two-consumer portfolio with an almost *base* load profile, that is, primarily the contribution to the portfolio in terms of the expected value.

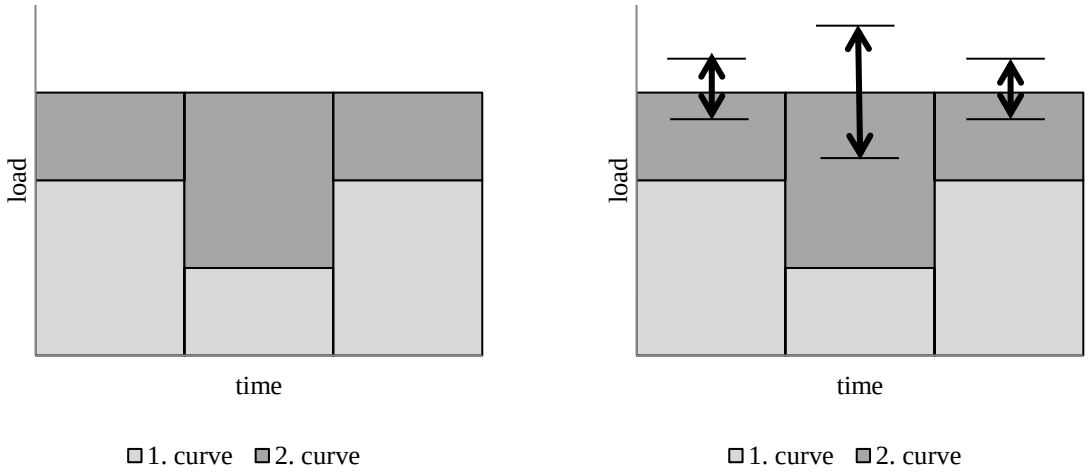
The schematic representation of this can be seen on the left side of Figure 13. Although there is uncertainty in consumption for every consumer; these may be interrelated (may correlate) to different extents. The resultant uncertainty on the portfolio level is represented on the figure by arrows and horizontal lines. If the value of the correlation coefficient of the uncertainty of two consumption curves is lower than +1, the so-called **diversification effect** will set in. This means that such uncertainty related to consumption can be diversified in a similar way as the often investigated market risk (see for example: *Brealey-Myers* [2005]) or liquidity risk (see for example: *Váradi* [2012]) in finance.[30] This

[29] It often happens, for example, that the batteries are used to handle the limited storability of the energy produced.

[30] In a simpler form, formally written, it is as follows: Let us assume that for one consumption time series the standard deviation of the error term is s_1, and for the other s_2. If the sum of the two consumption time series is examined, the standard deviation of the sum is – from the sum of the variance and twice the covariance – $s_{12} = \sqrt{s_1^2 + s_2^2 + 2 \cdot \rho \cdot s_1 \cdot s_2}$, where s_{12} is the standard deviation of the sum of the two consumption time series, and ρ is the linear correlation coefficient of the error terms of the two consumption time series. It is

diversification effect is time-dependent, as it cannot be stated unequivocally either for standard deviation or for the correlation coefficient that they are constant.

Figure 13: The schematic representation of the portfolio effect



a) without considering uncertainty b) with the consideration of uncertainty

Source: author's own figure (Excel).

The importance of what is discussed here is in that for any **pricing** task related to an individual consumer, the price should contain only non-diversifiable risk, as this is what cannot be eliminated by the portfolio effect. Returning to Figure 13, the uncertainty related to the value of the expected load is just such a non-diversifiable risk, which cannot be eliminated even by joining an ascending number of curves in a portfolio.

A similar application that touches upon the relationship between uncertainty and the portfolio effect can be found, for example in *Levy*'s [2013] study, which deals with a complex pricing problem in a market environment where demand side management is applied, however, the handling of consumer related uncertainty is relatively simple, less grounded in empirical data.

Obviously, the study of portfolio effect is much more complex: not only consumption and pricing risks, but the availability of wholesale products also need to be taken into account. This means that if a typically *peak* period consumer is put into the same portfolio with a typically *off-peak* period consumer, the portfolio effect can prevail if the portfolio

clear that the standard deviation of the sum is only $s_1 + s_2$, that is, the sum of the two standard deviations, if the correlation coefficient between the error terms equals exactly +1.

level demand curve can be hedged with a *base* load product. *Base* load products show more **liquidity**, which also contributes to the portfolio effect. In fact, disregarding uncertainty in consumption, this is what is also often meant by portfolio effect in actual practice.

1.3. Profile and profile-related risks

After a review of application examples, this section provides a more detailed definition of the term profile: its various uses in the related literature, as well as what is meant by profile in this paper. A separate section is devoted to risks related to the variation of consumption. The literature review focusing on mainly methodological issues has a place in another chapter; here the point is rather to see what previous studies focussed on in terms of terminology.

1.3.1. Definition of consumer profile

The term profile appears in various ways even in the literature on energy.[31] On the whole it is true that profiling usually covers some typing, seeking and creating typical patterns, which may appear on the level of a single consumer or group of consumers.[32]

Using Barnaby Pitt's definition (*Pitt* [2000]) by profiling (*load profiling*) we mean the modelling of how the daily *load shape*, that is, the daily *load profile*, is related to such factors as temporal variables, weather or other features that characterise consumers. The relationship between these factors and consumption is often nonlinear, and there are lots of interaction effects (see more on this in Chapter 2). The weather-dependent part is usually separated with some method from the consumption time series (for example, using some regression technique), and the time series after the removal of this effect is used from then on.

Of course, profiling makes sense for individual consumption curves, but in practice the aim is very often to group consumers with similar profiles, that is, similar consumer behaviour, to make consumer profile groups. This task is basically a methodological problem, as it is essentially about clustering time series.

[31] The aim of profiling itself is usually similar everywhere. There are characteristics, specialities that apply in various energy sectors – regarding regulation, the progress of liberalisation, but also the physical processes in the background. As the empirical part of this paper works with electricity consumption curves, most of the time we deal with electricity specific terms and terms that apply to all energy sectors.

[32] Otherwise, instead of the term **typical** the expression **average** often appears concerning profiles. Even the **mode** as a typical value is often replaced with values that result from **averaging** – specifically in connection with results are deliberately not mentioned here.

It is important to state that the majority of methods even to this day still builds on the formation of a **daily representative load curve** by consumer (*representative load curve*, RLC, see for example: *Tsekouras et al.* [2008]) which are often thought to represent typical daily load. Although these are easy to interpret and their use is practical from the perspective of creating groups, in most cases what are produced are constructed, derived, not actually realised values. In the majority of methods, it is possible to produce **daily profiles conditional on given circumstances** (such as summer, winter, transition period, or applying to different days of the week, etc.).

By profile (*Typical Daily Profile*, TDP), *Espinoza et al.* [2005] mean a daily shape where the effect of every exogenous variable (seasonal and calendar effect variable) is removed, but with the help of regression equations it is possible to produce daily profiles given a number of different conditions supplied by independent variables. The profile definition (*daily representative load profile*, RLP) provided by *Carpaneto et al.* [2003] is identical with the profile definition used in many different countries. In their study, the derivation of individual consumer profiles is completed under the assumption of various exogenous variable values (e.g.: weather-dependency, activity, available electrical devices, etc.) and loading conditions (e.g.: winter/summer, weekday/weekend, etc.). Of course, in such cases, there may be more daily load profiles according to different loading conditions. *Chicco* [2012], for example creates daily profile (*representative load pattern*) according to the above definition with the averaging of the measured values of a few collected days with the given loading conditions. In the studies by *Tsekouras et al.* [2007] and *Tsekouras et al.* [2008] the consumer profile is produced by the clustering of the daily load curves. The most obvious choice is to choose the profile from the cluster (group) with the highest frequency.

If the focus is not grouping those that share similar profiles (this is a common goal of studies in profiling) but how an individual consumer contributes to the load profile of a portfolio, then it is not only the shape but also the level of the curve that has a role. As a consequence, **in this dissertation, profiling is meant as modelling how (daily) load shape and (daily) load level are dependent on various factors, such as temporal variables, weather, or features that characterise different consumers**. That is, the definition provided by Pitt (*Pitt* [2000]) is extended: **the load profile involves the level besides the shape**, because using any type of normalisation the information concerning the level of consumption is lost.

It is difficult to find similar, publicly available results related to profiling in the Hungarian literature. According to the Electricity Act LXXXVI, the definition of profile is the following: a normalised annual consumer electricity demand curve created by statistical analysis expressed in 1000 $[kWh]$ per year (see: Act LXXXVI of 2007 on Electricity). This profile is also a composite of typical daily profiles, and this normalised curve can be rescaled from 1000 $[kWh]$ according to the annual consumption of the given consumer.

The discussion in the previous paragraphs centred on a given consumption curve. It may also happen that the definition of profile(s) is not based on an individual curve, but on the sum of curves in a profile group. This is especially useful if the consumer is not one that is metered regularly, or if the individual curves are too noisy and it is more sensible to draw conclusions about typical tendencies on the basis of the sum, where random effects – even if not in the most elegant way – are removed.

1.3.2. Profile-related risks

It is especially true for individual consumers that consumption-related risk occurs not only due to the irregular effect, but there are other risks that are typically less describable in a model-based manner.

Figure 14 depicts the various risks in a schematic form,[33] which are obviously shown (and discussed) separately, but of course, in practice they occur combined (mixed), often blurring each others' effects.

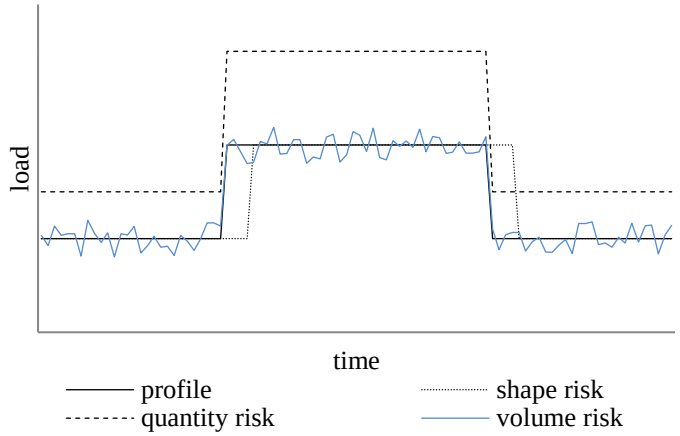
1.3.2.1. Profile risk: shape and quantity

There are basically two types of risks understood as types of profile risk (see: *Junghans* [2015]): shape risk and quantity risk.

Shape risk occurs when the shape of the consumer profile changes (for example due to changes in the daily schedule of a factory). What is meant by quantity risk is when the shape of the profile does not change, but the level of the total consumption does (for example in the period of economic boom). These are factors whose effects – especially on shorter consumption curves – are difficult to detect using statistical or similar quantitative tools; hence they are dealt with using simpler assumptions.

[33] As shape and quantity risk can often be handled only manually, in an *ad-hoc* manner, they are less likely to appear in academic work as opposed to volume risk. There are ambiguities in terms and notions used, which will be avoided here by using the previously fixed definition (which otherwise coincides with applications in practice).

Figure 14: The schematic representation of profile-related risks



Source: author's own figure (Excel).

Therefore, what is meant by profile risk is primarily when the consumer profile itself, its shape and/or level (that is, its quantity) undergoes *structural* change due to various external factors. Profile risk involves the effect of temperature variables as well, as profile (shape, level) is affected by them, and pose a risk. Being stochastic variables, irregular effect also plays an important role here (see more in this in Chapter 2). However, the latter is a field which requires more serious methodological techniques than *ad-hoc*, rule-of-thumb-like approaches.

1.3.2.2. Volume risk

Even if there is no change in the profile, consumption obviously does not always follow the path defined by the profile; smaller or larger deviations may still occur. There are various terms to define this phenomenon, among others, it may be called volume risk (e.g. *Junghans* [2015]), forecasting risk, or reliability risk (e.g. *Srivastav et al.* [2013]). The source of volume risk is basically the random or irregular factor (interpreted in the classical sense) and can rather be modelled by statistical or other quantitative methods. The emphasis is then on the appropriate representation of the random, unsystematic behaviour.

1.4. The empirical examination of stylized facts in consumption time series

In the same way as financial markets (see: *Cont* [2001] or electricity prices (see: *Marossy* [2010]), or even the relationships or interaction between energy markets and financial markets (see for example: *Leng et al.* [2014]), it is likewise possible to formulate so-called stylised facts for consumption time series. These are basically qualitative attributes that can be regarded as true and valid for the majority of consumption time series. It can be required from any model of consumption time series that it captures these stylised facts as accurately as possible. These stylised facts for consumption time series are the following – among others:

- high time-dependency,
- lack of stationarity in the strict sense,[34]
- multiple (yearly, weekly, intraday) seasonality,
- weather-dependency,
- nonlinearity and the presence of interaction effects,
- heteroscedasticity, that is, time-dependent dispersion.

Instead of the detailed study of the above with traditional methods[35] it is intended to show a few relatively simpler, but in practice, less used figures or relationships that work well to represent certain characteristics whose presence or absence is often difficult to confirm. The stylized facts can obviously be captured in the results and figures. The approach of the discussion serves as a foundation for the presentation of the author's own empirical research results in a way that it is shown how previous research results are satisfactory but at the same time, their shortcomings are also revealed.

The figures in this section are organised around the logic of the **consumption shape and level – consumption risk – temperature-dependency** triplet. This is an intuitively good way to demonstrate the contribution of the dissertation by the irregular treatment of the last three elements of the list above compared to classical solutions – keeping simplicity in mind.

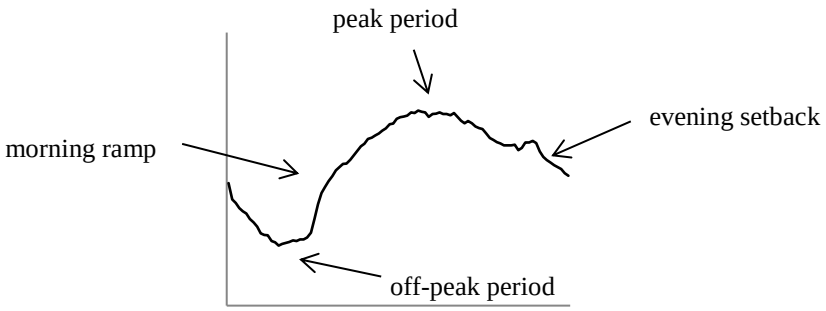
[34] See the chapter on methodology for further information.
[35] For example, calculations of autocorrelation coefficients, tests for stationarity, etc. See more on these in for example: *Hamilton* [1994], *Maddala* [2004], *Ramanathan* [2003].

In answering the research questions the data used here and in Chapter 4 are the individual load curve data of a Hungarian trading company in quarter-hourly resolution, and the Hungarian system load data publicly available from the MAVIR Ltd. website.

1.4.1. Load shape and level

Regarding the daily shape of curves, many studies on profiling distinguish between the features in Figure 15 or their transformation of some kind. Especially concerning individual consumption curves, it can readily be conceived that the transformation of these features in the form of simple or more complicated measures is often not enough for an appropriate description. In this section, *contour plot*s will be used to show what features the curve shape and level, or its variability may have in various curves.

Figure 15: Features that describe the shapes of load curves



Source: author's own figure.

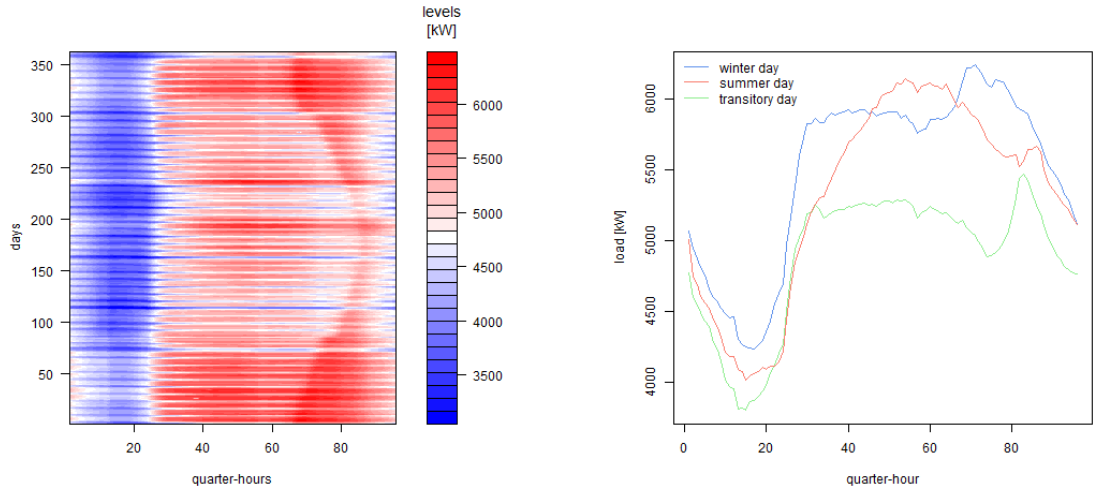
The essence of the *contour plot* is that it transforms the variation of a variable (let $z = f(x, y)$ be this variable) into a two dimensional figure as a function of two variables (let them be x and y) in a way that identical z values are connected with a line. These lines are the contour lines. It has increasingly become common with the more extensive use of graphical tools and representation that different colours are used for different levels. This way, the spatial representation of the third dimension (z) can be practically replaced by using colours even with two-dimensional display options.[36] In the figures here variables x

[36] Based on the short description above, it can be seen that *contour plot*s may be familiar from many different fields. They are used in **cartography** to draw the identical heigths above/below the sea levels around given meridian and parallel arcs, or to represent atmospheric pressure in meteorology, etc. At the same time, the **isoquant curves** used in microeconomy can be regarded such *contour plot* figures, alongside with **indifference curves** to whose curve dots identical output levels and identical demand side utility can be attached (see for example: *Varian* [2004]).

and y denote quarter-hours of a day and days of a year, while variable z – as their function – denotes the load values at a given time.[37]

Based on Figure 16 about the Hungarian system load time series it can be seen that the lowest load values apply in the morning hours up to 5:30-6:00, then the daily ramp happens at the same time throughout almost the whole year (between 6:00 and 7:00, the white line after the 20th quarter-hour of the day). The red lines after this indicate *peak* periods. The horizontal white lines among them represent the lower loads of weekend days.

Figure 16: The *contour plot* of the Hungarian system load and the time series figure of some chosen days



Source: author's own figure (R).

The second *peak* at the end of the day is displayed clearly on the figure. Its course is mainly connected with the time of the **sunset**, especially during weekdays. As in winter periods the sun sets between 4-5 pm, the second *peak* within the day starts at this time, and lasts much longer than in the summer, when the second *peak* is smaller both in magnitude and length. To support this, the right side of Figure 16 shows the development of the daily load in winter, summer and transition periods.

In addition, the effect of the two hottest periods of the summer of 2011 can clearly be seen (around the 200th and 240th days) where the *peak* during the day is much darker – this period gets a colour similar to the dark red colour of winter periods.

[37] The range of load values were divided – with 30 class boundaries – into 29 equidistant intervals, and each load value was assigned to the appropriate interval. The values belonging to the medium (15th) interval are white in the *contour plot*. The higher the load of the interval that the values belong to, the darker red colour they get, while the lower the load of the interval, the darker blue they get in the figure. With this number of intervals, the figures show a nice colour gradient effect.

Besides all this, the figure nicely shows the effect of daylight saving **time change** (in 2011 it happened on 27th March and 30th October).

The narrower, typically blue lines can be found around the 75th, 120th, 160th, 300th days and at the top of the figures mark successively greater, longer **holiday periods** (not falling only on weekends). The positions of the lines are easy to identify based on the table below (the table contains those holidays that fall not only on weekends):[38]

Holiday	Day of the year
1st January (Saturday)	1st
15th March (Tuesday, a four-day long weekend)	74th
24th-25th April (Easter)	114-115th
1st May (Sunday)	121st
12th-13th June (Pentecost)	163-164th
20th August (Saturday)	232nd
23rd October (Sunday)	296th
1st November (Tuesday, a four-day long weekend)	305th
24th -26th December (Saturday-Monday, Christmas)	358-360th
31st December-1st January (New Year's Eve)	365th and 1st

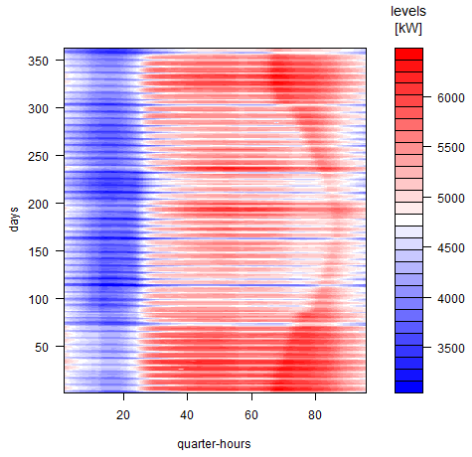
Source: author's own table.

Figure 17 shows the Hungarian system load with a portfolio and the *contour plot*s of some chosen individual curves.

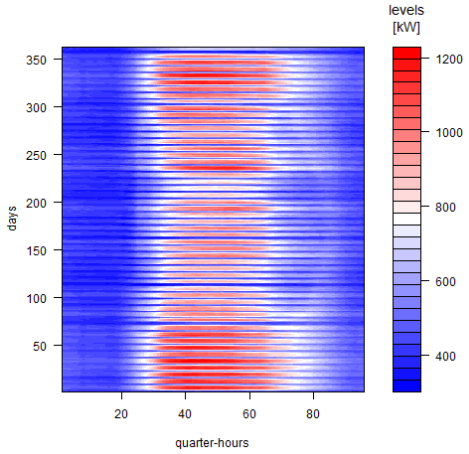
The so-called **sunset effect** (that is, when the sun sets earlier, therefore we need to turn on the lights earlier) also appears on the portfolio *contour plot* (Figure 17). As this is a portfolio that includes business consumers, the *peak* period typically ends at around 18:00 in the afternoon, and for this reason the effect is slightly smaller. Even in the portfolio the effect of **temperature** in winter and summer periods is apparent, including those two summer periods that were incredibly hot.[39] The big summer holiday period around 20th August (which is approximately between the 200th and 240th days) can be clearly seen here, as the morning ramp occurs a bit later and the consumption level is also a bit lower.

[38] August 20th was a weekend day, but the typical summer holiday period is shown on the figure around it (it is similar to the so-called 'between Christmas and New Year period').
[39] Both were out of the big summer holiday period around the 20th August.

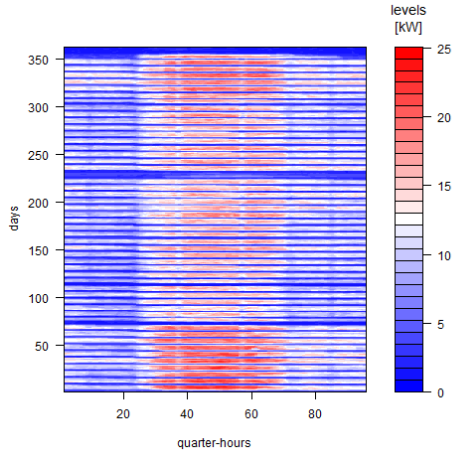
Figure 17: *Contour plots* of portfolios and individual consumer load curves



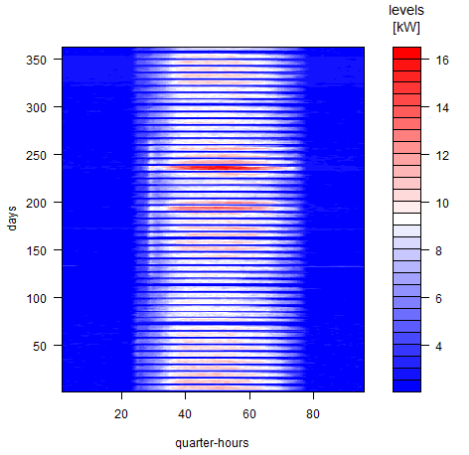
a) Hungarian system load



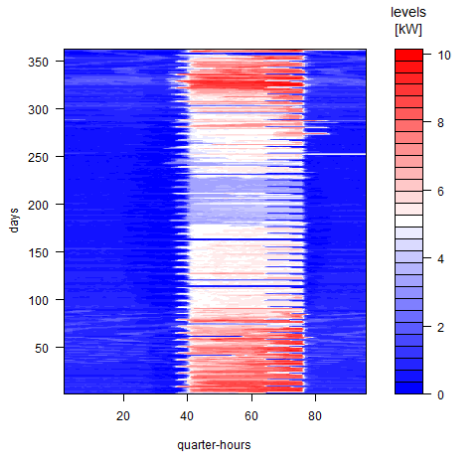
b) Portfolio



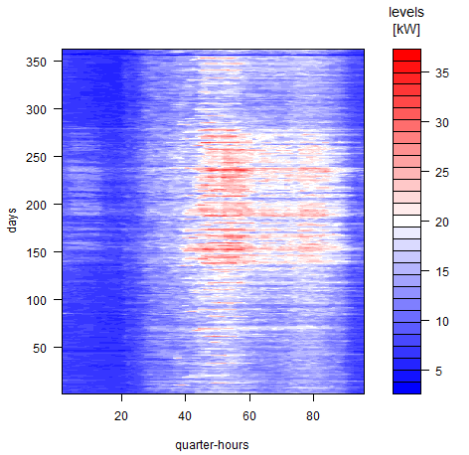
c) Curve C25



d) Curve C66



e) Curve C96



f) Curve C109

Source: author's own figure (R).

The above can be complemented mainly with the consumer-specific properties of *contour plot*s of individual curves. It can be clearly seen in the case of company V25 that the winter temperature effect is stronger and that there is practically a complete stop in holiday periods. As for company V66, it has highly regular load characteristics and – a bit interestingly – the summer **temperature** effect is only prevalent in extremely hot periods, otherwise it is not observable. In V109, summer temperature-dependency is obvious, while in V96, winter is apparent. In the case of the latter, there is a longer level shift around the 200th day for about one and a half - two months, which is probably due to some **structural** reason (e.g.: it may be a factory with some facilities shut down).

1.4.2. The intraday distribution of loads

The intraday distribution deserves a separate section as it is very telling of the shape of the daily profile, and behind the **changes** in the intraday distribution, there are often identifiable fundamental factors (e.g. often temperature). The advantage of this kind of investigation is that in the literature, profiles are usually made on the basis of daily shapes. The study of the intraday distribution is telling of the consumption-related uncertainty, which is given priority in this dissertation. Exactly for the purpose of the methodological underpinning of distribution-centeredness, this section may seem more technical compared to the others. Here, measures that characterise **distribution** (quartiles, range, minimum, maximum, etc.), their variation, (un)stability are at least as important as the **fundamental** reasons behind them.

The figures in this section show the time series of the selected (of course, not representative) weeks below (and the corresponding *boxplot*s[40]):

[40] Elements of a ***boxplot*** are defined by three values: the lower quartile, the median and the upper quartile. This way half of a day's quarter-hourly load values appear in a *boxplot*. Outside the box the two upward and downward reaching lines stretch between the lowest and highest load values. If a value is higher than the upper quartile by 1.5 times of the interquartile range (the so-called inner fence) or is smaller than the lower quartile by at least 1.5 times of the interquartile range, the value is regarded an *outlier* and is marked with a red dot on the figures (and in this case, the lines stretching upwards and downwards reach only until the above mentioned 1.5 times distance). Of course, the choice of the inner fence at 1.5 times the interquartile range is a subjective decision, other multipliers may also be chosen, which will obviously have an effect on the figure (with a smaller multiplier, more values will be regarded *outlier*s).
Note. Quartiles are measures used in descriptive statistics to help describe the distribution of some variable. The lower quartile is the value compared to which a quarter of the observations is lower and a three-fourths is higher. The middle quartile (median) is where half of the observations has a lower and the other half has a higher value. The upper quartile is the value compared to which three-fourths of the observations is lower and a quarter is higher.

Period (Season)	Start	End
Winter	17-01-2011 00:00 CET[41]	23-01-2011 23:45 CET
Summer	18-07-2011 00:00 CEST[42]	24-07-2011 23:45 CEST
Transition (spring and autumn)	18-04-2011 00:00 CEST	24-04-2011 23:45 CEST

Source: the author's own table

The left side of the a) part of Figure 18 shows the quarter-hourly historical loads of the chosen weeks and the right side shows *boxplot*s representing the Hungarian system load in daily resolution.

It is possible to draw conclusions based on the figure as those that were seen in relation to *contour plot*s (for example in winter, the level of consumption is higher due – directly or indirectly – to the *heating* and the *illumination effects*, besides, in summer the cooling effect resulting from the increased use of air conditioners increases the *peak* period consumption levels).

It is also shown that the so-called winter *off-peak* period consumption level is higher (obviously due to the heating effect), and by season, the different nature of the morning ramp and the position of *peak*s and their levels within the day (for example, in summer the cooling effect results in a higher afternoon *peak* than is produced by the evening illumination effect when the cooling effect is not so influential).[43]

The intraday distribution of loads is **asymmetric** independently of the season; it is leant to the right (that is, it is left-skewed), as the range of values above the median is much narrower than of the values under the median within a day. This asymmetry is slightly weaker at weekends, as **weekend** load in daytime is lower compared to **weekdays**. Independently from seasonal periods the range of Mondays is wider, which can be explained by the fact that the 'ramp period' needs to start from a lower level, as a result of the weekend before them.

It appears that the other main source of asymmetry within a day is **temperature**, as the distribution is much more asymmetrical on winter days and warmer summer days than otherwise[44]. This is due to the fact that in these periods, the temperature effect (whether cooling or heating effect) causes higher *peak* period load levels. These effects influencing

[41] CET: Central European Time, (UTC + 1 hour), UTC: Coordinated Universal Time
[42] CEST: Central European Summer Time, this applies in the summer time period instead of CET (UTC + 2 hours)
[43] In the evenings, household cooling use only applies in extreme heat, but it is by magnitude smaller than the peak period (workplace) use.
[44] The second week of July and the first few days of the first week of July were the hottest days in this summer.

the daily distribution may of course be combined. In the table in Appendix C) the above statements can be followed numerically. What is most important are, of course, not the exact values of the numbers but it is worth following through and unravelling the tendencies in the relative standard deviation or the variation of the skewness measure.

Based on the *boxplot*s, it can also be said that there are hardly any ***outlier*s** – based on the definition of a *boxplot* – and those few that appear in transition and *off-peak* period hours. One should be careful with the handling of *outlier*s, though, as the *boxplot*s were made in daily resolution; therefore, *outlier*s are evaluated as such – based on the position of the box belonging to a given day and as a function of the corresponding interquartile range. Neither time-dependence, nor fundamental causes are considered here.

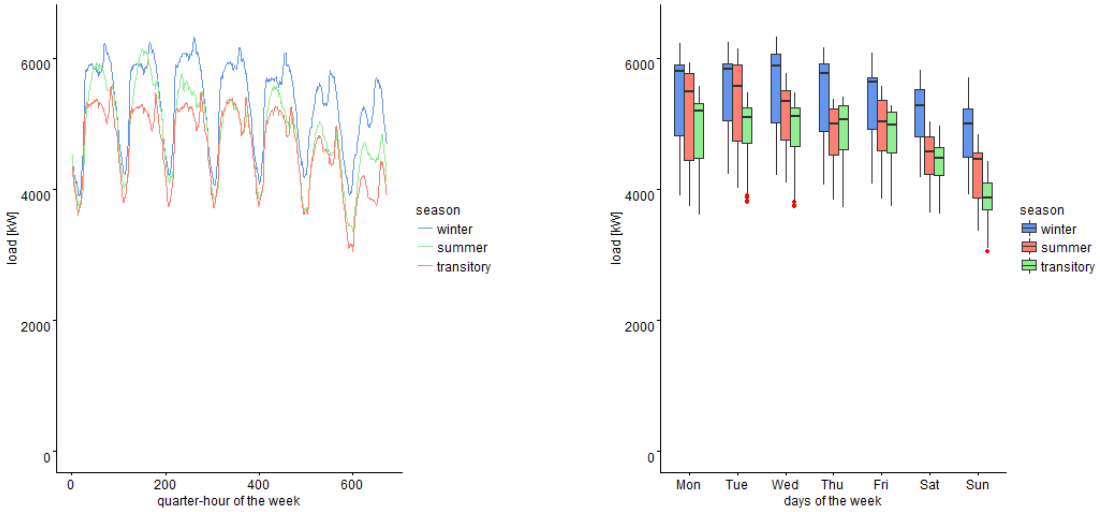
Figure 18 shows the figures of the Hungarian system load, the portfolio and individual curves, this way both similarities and differences are very easy to notice.

The role of winter heating and summer cooling effects appear similarly in the portfolio compared to the Hungarian system load. The weekend load **level**, however, is much lower than on weekdays, and in this period, the heating and cooling effects have little influence. It is important to state that – as opposed to the Hungarian system load – the **ranges** of weekday daily loads and the position of the median are much more **stable**.

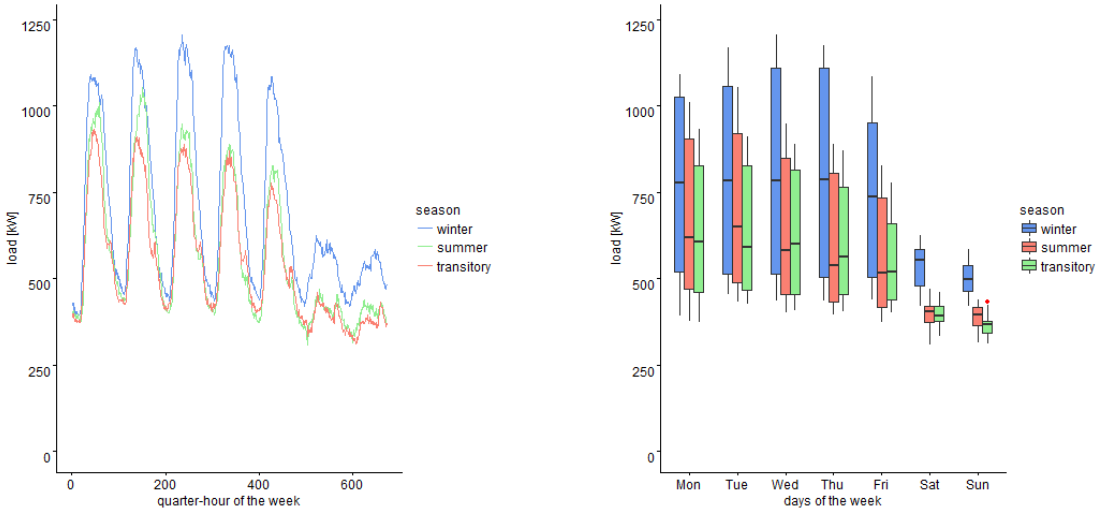
The evolution of Friday afternoons and the weekend setback are much more considerable and clear: the time series figures show that the Friday *peak* period is much shorter and based on the *boxplot*s, *ceteris paribus* the range is smaller than the daily maximum.

As regards the individual consumer load curves, in curve C25 the differences are mainly in *peak* period loads between seasons. Both figures suggest that it is mainly the winter temperature effect that needs to be considered. What make the load curve special are rather the quick Monday morning ramp and the slow setback (lasting until the middle of Saturday), otherwise the weekdays can be regarded identical. Therefore, the **variation of the curve** can best be **characterised** by a ramp at the beginning of the week, a weekend setback period, the weekday periods (shifted by the temperature in winter) and there is a weekend period, with basically constant load.

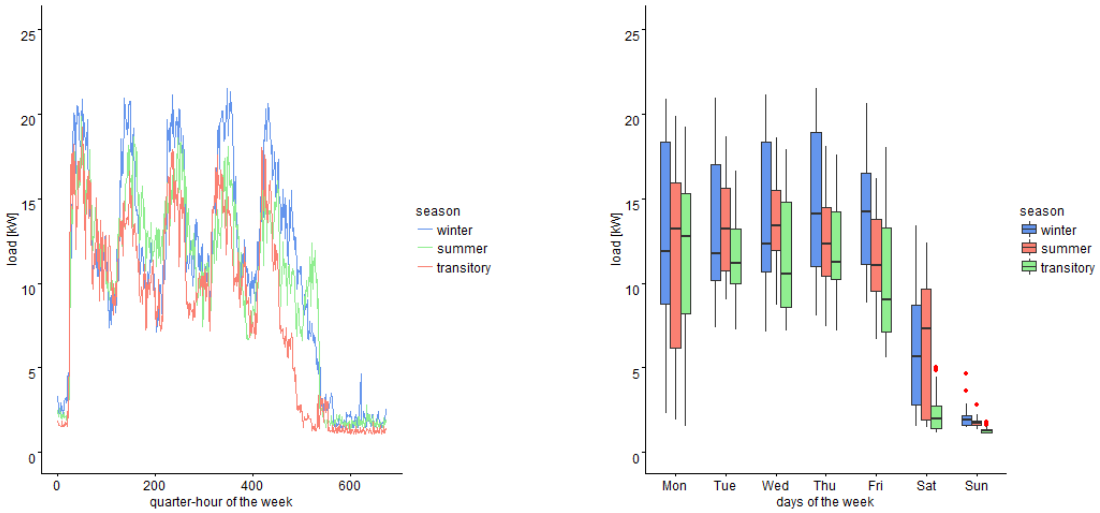
Figure 18: Weekly load time series and the related *boxplot*s of portfolios and individual consumer load curves



a) Hungarian system load



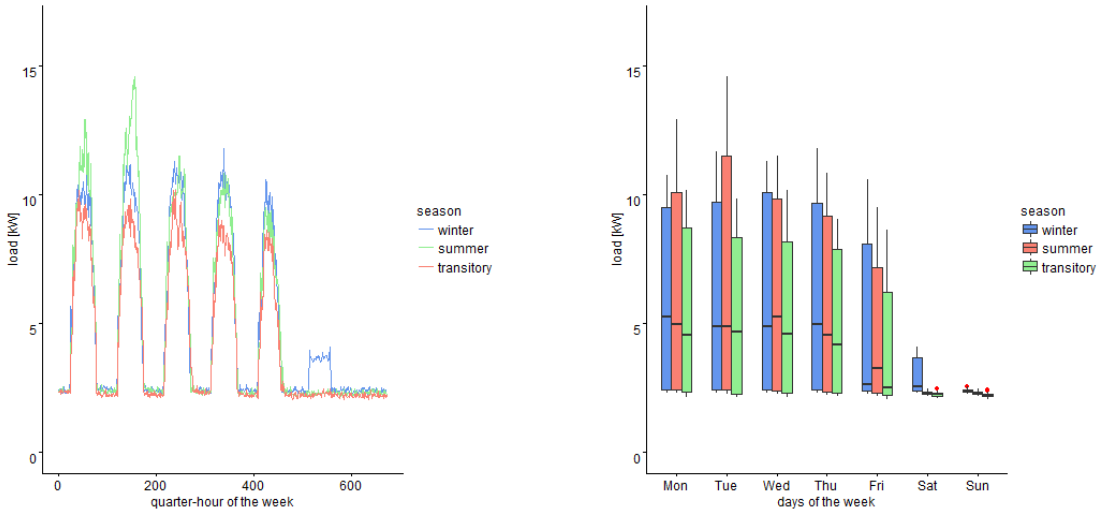
b) Portfolio



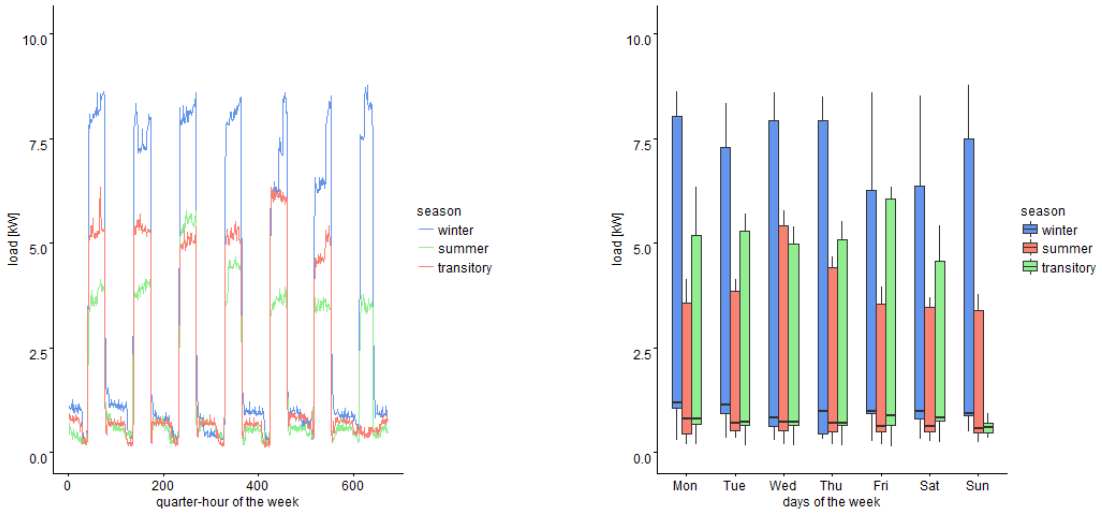
c) Curve C25

Table is continued on the next page.

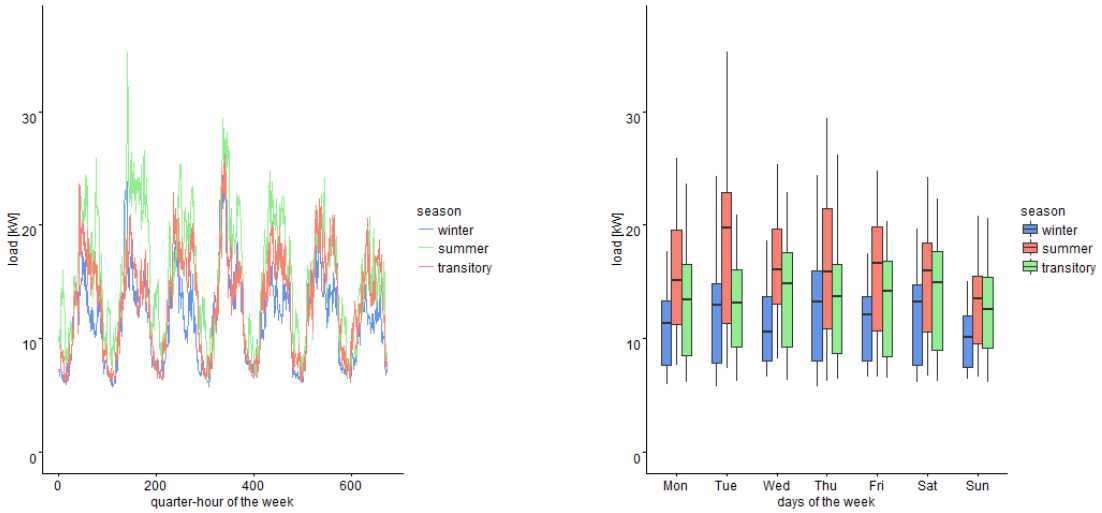
Table continued from the previous page.



d) Curve C66



e) Curve C96



f) Curve C109

Source: author's own calculations (R) and figures (R).

Figures of company C66 suggest that the **distribution** of weekdays is extremely **stable**, and the load of weekdays and weekends in *off-peak* periods is practically constant. Temperature effect only occurs in the summer (on Monday, Tuesday, that is, on the previously mentioned warmer days the *peak* period load is higher), and the winter *peak* period load level is somewhat higher than in transition periods.

In the curve of company C96 only the *off-peak* approximately 1 [kW] and the *peak* period 5-8 [kW] **regimes** alternate quite **regularly**, weekends and weekdays do not differ either. As a reflection of this, daily distributions are also very stable with a right-skewed asymmetry all the way, as the ratio of *peak* period values during the day is much lower. Regarding the effect of temperature, only winter heating effect can be considered.

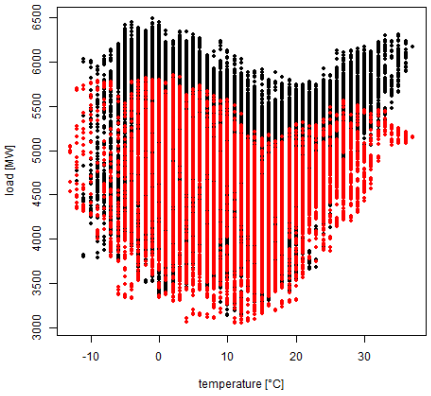
Curve C109 – as opposed to the previous one – is characterized by summer temperature effect, but this is only characteristic in *peak* periods, otherwise, the daily minimum level remains constant independently from the season (disregarding a few summer days).

1.4.3. Temperature-dependency

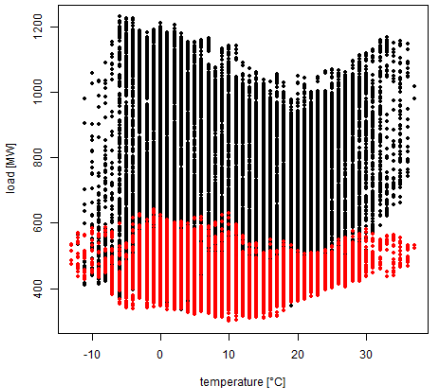
Relying on results in the literature (see Chapter 2) it can be stated that out of the weather factors that have an effect on electricity consumption, it is temperature that has the greatest effect. Based on Figure 19 it can be seen very well that this effect varies by curve. In the sample period, concerning the Hungarian system load, the heating threshold value (under which the so-called heating effect applies) starts at around 12-13 $°C$, while the cooling threshold value (above which the so-called heating effect applies) is at around 21-22 $°C$. Between these two values, the temperature effect does not apply. In a consumer portfolio, it is basically these two threshold values that seem valid, though it is worth noting that in actual practice, these are considered for the calculation of the so-called *heating degree-days* (HDD) and *cooling degree-days* (CDD) (see more on these in Chapters 2 and 4 as well).

Regarding the Hungarian system load +1 $°C$ seems to have a somewhat stronger influence regarding the cooling effect than the heating effect. It is also clear that this only applies for weekdays. At weekends the behaviour of the heating effect is more or less similar to weekdays, the behaviour of the heating effect is a bit milder, and cannot be characterised by a very clear tendency.

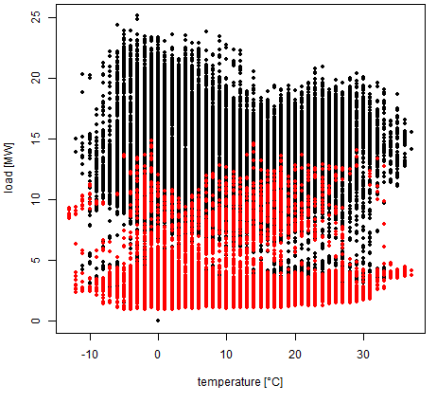
Figure 19: Portfolios and individual consumer loads as a function of temperature



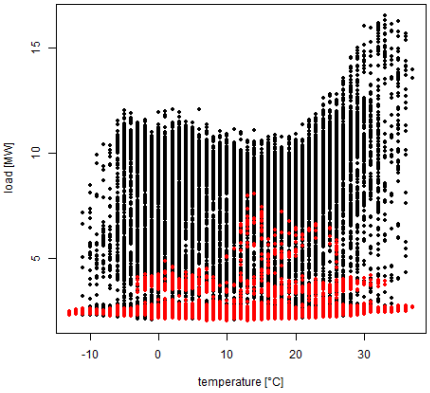
a) Hungarian system load
(weekdays (black) – weekends (red))



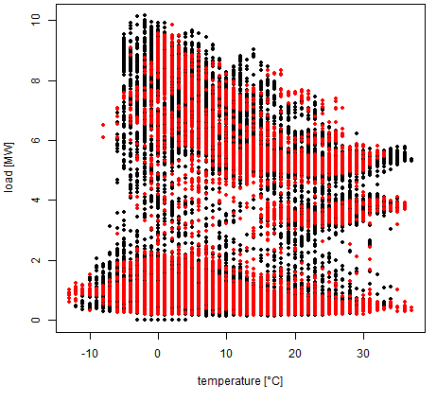
b) Portfolio
(weekdays (black) – weekends (red))



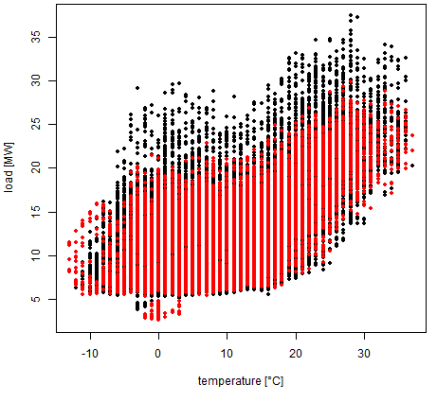
c) Curve C25
(weekdays (black) – weekends (red))



d) Curve C66
(weekdays (black) – weekends (red))



e) Curve C96
(weekdays (black) – weekends (red))



f) Curve C109
(weekdays (black) – weekends (red))

Source: author's own figures (R).

In the portfolio the weekday heating and cooling effects are almost the same, at weekends both are weaker.

Obviously, in individual curves the situation is more complicated in analyses of this; in addition, the random component has a much greater effect. It is similarly outlined how the loads are **concentrated, grouped** depending on the temperature on weekdays and at weekends, when the dispersion is smaller or greater.[45] This latter perspective will be important in the rest of this dissertation.

1.4.4. Conclusions

When formulating the conclusions, it is worth returning to the statement that the results shown have aimed at drawing attention to phenomena important from the perspective of this dissertation focussing on **the consumption shape and level – consumption risk – temperature-dependency** triplet.

The *contour plot*s give a compact picture of the level and shape of the consumption – including effects that apply in different extents in various curves. These effects include calendar effects (detecting the effects of weekdays, weekends, holidays), temperature-dependency, illumination (or sunset) effect and the effects of other **structural** changes.

Most academic work dealing with profiling approaches it from creating or forming daily profiles, therefore it was considered appropriate to study the distribution of load values for various seasons within a day. Fundamentally, similar statements were formulated regarding *contour plot*s, but here it is much easier to check to what extent the **distribution** of one day's load values can be viewed as **stable** or **unstable**, and what factors (e.g. temperature) may have an effect on the distribution of the load values of each day. This way, information is gained about not only consumption level and shape, but also about their risk and uncertainty. Obviously, the various types of risks (such as profile- or volume risk) cannot be quantified using these methods, only assumptions can be formulated regarding their behaviour.

Complementing the above, the approach of temperature-load *scatter plots* appears later in the section of the dissertation on mixture models. As a preliminary to the results there, this section has offered an opportunity to get an insight into how quarter-hourly load curves are **concentrated and grouped in the temperature-load dimension**.

[45] In C25, for example, the load is constant at weekend days. Figure 19 a) is a bit disturbing, because the weekend setback lasts until the middle of Saturday; however, in this period temperature – understandable – does not really have a significant effect on the load.

Although it is true that **the formation of daily profiles is appropriate in a certain sense – and above all, is easy to interpret – but at the same time, it is not always efficient, as typical daily profiles are not necessarily formed along daily shapes**. If the formation of profiles is performed by daily discretisation, it might mean the unnecessary estimation of many parameters, but may also result in drawing misleading conclusions. In other words, the grouping of consumption values is not necessarily most efficiently modelled along daily profile curves, though examining *contour plot*s vertically this appears to be a good approach. A line (or some lines) of the *contour plot* that is in some form typical, is (more or less) similar to the other lines, while there are ramps, *peak*s and setbacks with different degrees of differences. The results here only have informational purposes; this is what is pointed out for example by the difficult and misleading identification of the *outlier* values with the method shown here from daily discretisation.

The above analyses, though from a different perspective, support the hypothesis H1, which proposes that in electricity consumption curves the main source of variance in curves is intrday seasonality – that is, this hypotheses cannot be rejected. This can be seen from the daily variation of *boxplot*s and the organisation of the colour scale of the *contour plot*s. This result could have been very easily reached by variance analysis (see Appendix C)), but the **prior aim** of this chapter was to find out and **reveal the mainsprings and major relationships behind the variation of time series**, especially to provide a picture of the **heterogeneity of individual consumption behaviour** highlighting the complexity of goals formulated by dissertation.

2. PREVIOUS RESEARCH RESULTS ON CONSUMER PROFILING

Identification of consumer profiles goes hand in hand with creating consumer profile groups that include those that share similar profiles. For this reason, this chapter reviews research results that have been reached so far related to time series clustering, and in a narrower sense, energy (consumption) time series clustering. In connection with this, the so-called curve feature is essentially the relevant information extracted about typical consumption patterns in a specific instance of practical usage.

These tasks are closely related to the appropriate handling of weather induced effects; therefore, this is the topic of the second part of the chapter. Building on a previous publication of the author it is shown what consequences the most commonly used solutions in the handling of weather induced effects have for both profiling, modelling and evaluation of the related uncertainty.

2.1. The two-step consumption time series clustering

Based on the literature on profiling, this section gives an overview of the general framework used for creating consumer profiles and consumer profile groups. Practically, this means the adaptation of *two-step* or *two-stage clustering* to consumer profiles. Here, besides the applied clustering techniques (2nd step) what is more stressed and industry-specific is that curve features are produced (as a 1st step) to describe consumption curves to represent compressed information.[46] The overview starts with a short introduction to time series clustering.

2.1.1. Time series clustering in general

As it is commonly known, the aim of clustering is essentially grouping observations in a way that those within the group are as similar as possible, while compared to each other, the groups differ as much as possible. Algorithms used in *time series clustering* and *curve clustering* are often based on similar techniques – using some time series characteristics – such as clustering techniques in general (see, for example: *Liao* [2015]).

[46] As in consumption time series there are typically huge amounts of datum per time series (e.g. the consumption curve for a year – in a non-leap year – consists of $365 \cdot 24 = 8760$ hours or $365 \cdot 96 = 35040$ quarterhours), as a first step, it is always necessary to carry out some form of information extraction.

Related to this topic, Table 3 shows two possible classifications from two important studies that synthesise the applied methods (*Liao* [2005] and *Jacques-Preda* [2013]).[47]

Table 3: Possible classifications of time series clustering algorithms

Classification by Liao	Classification by Jacques and Preda
1. raw data clustering 2. feature-based clustering 3. model-based clustering	1. raw data clustering 2. two-stage clustering 3. nonparametric clustering 4. model-based clustering

Source: table edited by the author.

The first techniques are in both cases clustering techniques that work with raw data (the original time series), interpreting them either in time or frequency domain. It is essential in feature-based or two-step clustering that the first step is to extract some relevant feature from the raw data, and then clustering takes place in this domain. Among nonparametric techniques there are techniques that perform clustering by using various distance or similarity measures. The advantage of model-based clustering is that clustering and dimension reduction happen in one step by the estimation of the model that best suits the data. In a slightly misleading way, Liao lists here the type of clustering where (similarly to feature-based clustering) some parameters of the models applied to the time series (as some results obtained from time series models) serve as a basis for clustering.

2.1.2. The general framework for profiling

The framework used by Chicco in Figure 20 (see: *Chicco* [2012]), is the application of two-step or two-stage clustering on consumer profiling.

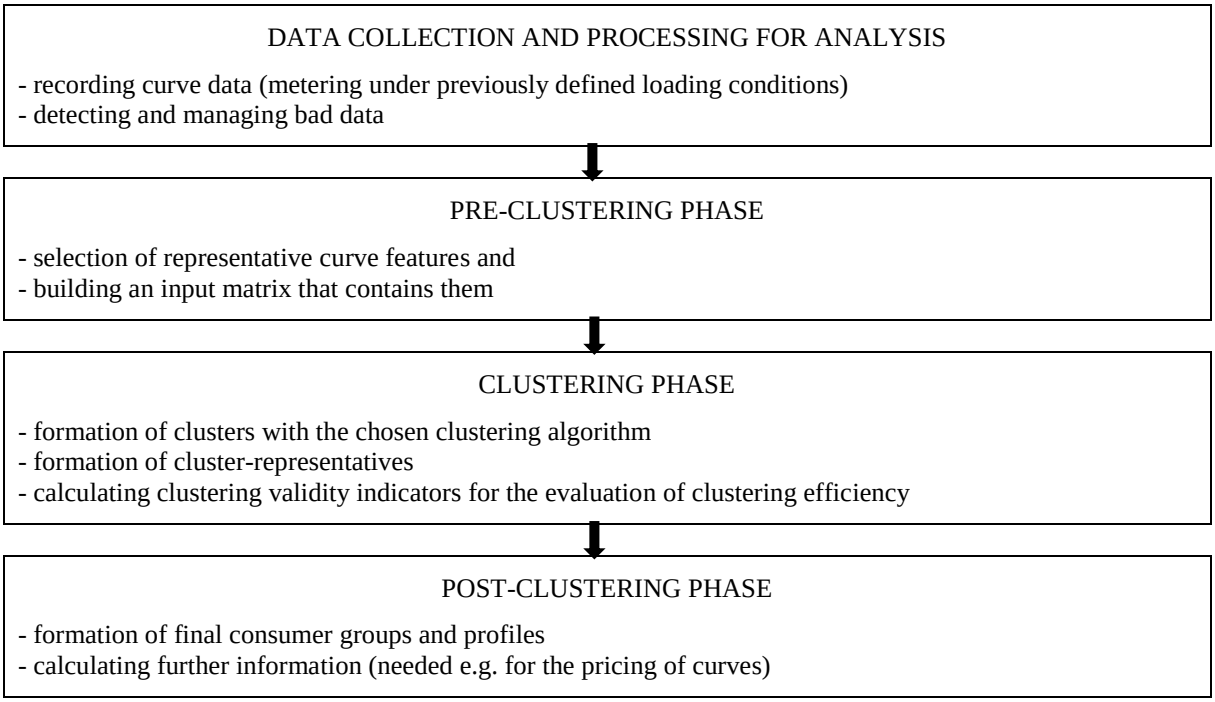
Pre-clustering is a phase in clustering where the so-called curve features are produced. These are the features that presumably represent the variation of consumption well.[48] This is followed by the compilation of *input* data.

Clustering itself is not only creating clusters with some chosen clustering technique, but it also means the formation of cluster representatives and evaluating the appropriateness of the clustering results by clustering validity indicators.

[47] In their article Jacques and Preda (*Jacques-Preda* [2013]) summarise clustering techniques for functions (*functional data clustering*), but they note that in most cases the data to be analysed manifest themselves as a function of time as a continuous variable. This field is becoming increasingly popular as a subfield of functional data analysis (FDA).

[48] Curve features are understood in the majority of the applications as features that characterise the whole curve compressed in one single day's profile, as it was mentioned in the definition of profile before.

Figure 20: The procedure of profiling consumption curves



Source: author's own figure based on Chicco [2012].

The last, so-called post-clustering phase means creating consumer groups and final consumer profiles[49]. The final number of consumer groups does not need to be identical with the number of clusters resulting from the clustering process[50], because in actual practice a market participant may not be able to manage many clusters in a transparent way, thus in such cases some form of aggregation is necessary.

Obviously, our goal is not to explain all the above steps in detail. However, from the perspective of this paper it is highly important to see how the curve features to be used in clustering are produced. This also includes the pre-adjustment of the time series, which in most cases means the removal of or dealing with the effects of temperature (see more about this in the second part of this chapter).

[49] These final profiles are basically those daily profiles produced in various conditions that have also been mentioned in connection with the definition of profiles (such as profiles that apply in summer, winter, transition periods, different days of the week, or even a combination of these, etc.). For a very neat classification of these see: *Pitt* [2000].

[50] Various clustering validity indicators exist for the methodologically appropriate decision about the number of clusters; see more on this in *Chicco* [2012].

2.1.3. Producing curve characteristics to be used in profiling

The curve features that describe each consumption curve in clustering can be produced in many ways. A possible classification of these can be found in Table 4, which was compiled on the basis of how they have been produced so far according to the literature.

Table 4: A possible classification of curve features to be used in profiling

Daily representative load curves (RLC)				Features produced with other methods
shape parameter-based features produced	**time domain-based features produced**	**frequency domain-based features produced**	**model-based features produced**	
Chicco et al. [2005] *Mathieu et al.* [2011]	*Chicco* [2012] *Li et al.* [2010] *Macedo et al.* [2015] *Panapakidis et al.* [2012] *Panapakidis et al.* [2014] *Tsekouras et al.* [2007] *Tsekouras et al.* [2008]	*Carpaneto et al.* [2003] *Carpaneto et al.* [2006] *Chicco et al.* [2005] *Panapakidis et al.* [2014]	*Espinoza et al.* [2005] *Hino et al.* [2013] *McKenna et al.* [2014]	*Räsänen et al.* [2010] *Srivastav et al.* [2013] *Verdú et al.* [2006]

Source: author's own compilation and table.

It is clear that in most cases features that characterise consumption curves are produced on the basis of some daily representative load curve. Such is the previously mentioned *representative load curve* (RLC) whose production is summarised in *Tsekouras et al.* [2008], for example. Chicco (see: *Chicco* [2012]) assumes that a one-day load curve (consisting of 96 quarter-hours) is already available, which means using the mean of quarter-hourly load values of 'representative' weekdays from the transition period of the total load curve.[51] A study which essentially combines the two solutions, the RLC-based (two-step) consumer segmentation, is the article by *Tsekouras et al.* [2007]. These methods assume that consumption curve clustering is produced from daily hourly or quarter-hourly representative load curves, that is, from characteristics produced based on **time domain** features.

It is a possible (and simpler) method, though, to extract further relevant information from the available daily RLC, which in the Table 3 were labelled **shape parameters**. In

[51] As this concerns the transition period, the removal of the effect of temperature is not dealt with in the study.

Chicco et al.'s [2005] work, these are the so-called *shape factors*.[52] In *Mathieu et al.*'s [2011] study, they are called *load shape parameters*.[53] The drawback of these methods is that the measures cannot always be determined definitely, for example for load curves where there are two *peak* periods within a day (as it very commonly happens in household consumption).

Chicco et al. [2005] use a **frequency domain** based approach (Fourier transform) for the representation of daily load curves instead of the time-domain method mentioned in the previous paragraphs. A solution very similar to this is used in the studies by *Carpaneto et al.* [2003], *Carpaneto et al.* [2006] and *Panapakidis et al.* [2014], among others. In all these studies the main idea is to transform the daily RLC into frequency-domain, and perform clustering on the results thus derived.

In the methods reviewed so far, this RLC often means a daily load curve from the transition period which is either the mean of the load values of one or more days, or the result of daily load curve clustering (see: *Chicco* [2012], *Tsekouras et al.* [2007]).

It is worth using a different category for the so-called **model-based** solutions, where the daily representative load curve is produced with the assumption of estimating an underlying model. Examples to this are the studies by *Hino et al.* [2013] and *McKenna et al.* [2014], where the daily load curve is modelled as a mixture of normal density functions, and the classification of daily curves is done on the basis of results thus derived. *Espinoza et al.* [2005] use a completely different solution based on a time series regression technique, where the effects of all exogenous variables (such as temperature, days of the week, etc.) are removed.

Besides, there are other solutions where daily RLCs are not produced (or at least, not with the aim of clustering consumption curves). In *Räsänen et al.*'s [2010] study the

[52] These shape parametres are the following:
- average daily load / maximal daily load,
- average daytime load / maximal daytime load,
- minimal daily load / average daily load,
- average evening, night load / average daily load,
- average load measured in dinnertime / average daytime load,
- minimal daytime load / average daytime load.

[53] These load parametres are the following:
- *Near-Base Load* [NBL, kW]: the 2.5^{th} percentile of daily loads,
- *Near-Peak Load* [NPL, kW]: the 97.5^{th} percentile of daily loads,
- *High-Load Duration* [HLD, hour]: the length of the time period when the value of load is closer to NPL than it is to NBL,
- *Rise Time* [hour]: the length of the time period until load reaches the beginning of HLD from NBL,
- *Fall Time* [hour]: the length of the time period until load reaches NBL from the end of HLD.

method uses dimension reduction (*self-organizing maps* (SOM)) on the basis of times chosen randomly (but identically) from time series, then groups the curves with similar characteristics by K-Means clustering or hierarchical clustering. *Srivastav et al.* [2013] use mixture models for estimating typical consumption patterns. This will be described in detail in a subsequent chapter. *Verdú et al.* [2006] perform dimension reduction on weekly load time series using the SOM method as well.

2.1.4. Clustering algorithms used in profiling

Table 5 contains, without aiming to give an exhaustive list, the most commonly used clustering algorithms for creating profile groups, and some applications of these.

The application of the methods is extremely wide ranging even beyond their use in energy; therefore, the advantages and drawbacks that are generally known also apply here. In selecting the best method, the right question to ask is rather how strongly the advantages and drawbacks affect the results of profiling in particular cases.

Although iterative K-Means clustering and its variants (K-Means++, fuzzy K-Means) are fairly simple, there is a drawback to all of them: they require the *a priori* knowledge of the number of clusters.

The advantage of the various hierarchical clustering techniques is that many distance and similarity measures can be used, whereas in an agglomerative case, for example, the various techniques for merging can produce highly different results, and – normally – it is impossible to reassign a load curve into another cluster during the agglomeration steps.

K-Medoid clustering also allows the use of various distance measures, but compared to K-Means clustering it is a more robust solution. The so-called medoid of the resulting clusters will be the observation compared to which the other cluster members are closest, that is, it can also be regarded as a representative observation characterising the cluster (as opposed to the centroid which is sensitive to averaging).

It is clear that the application of various artificial neural network (ANN) solutions is on the rise in the field of clustering tasks as well, and besides dimension reduction, they are used for creating groups. Often, they do not appear alone, as the so-called self-organising maps (SOM) map the examined input variables – similarly to multidimensional scaling – into a lower, usually two-dimensional space, and clustering itself is performed by the already listed clustering techniques.

Table 5: Clustering algorithms used in profiling

Clustering algorithm	Application
K-Means clustering	*Chicco et al.* [2005] *Chicco* [2012] *Espinoza et al.* [2005] *McKenna et al.* [2014] *Panapakidis et al.* [2014] *Räsänen et al.* [2010] *Tsekouras et al.* [2007] *Tsekouras et al.* [2008]
K-Means++ clustering	*Panapakidis et al.* [2014]
Fuzzy K-Means clustering	*Chicco et al.* [2005] *Chicco* [2012] *Tsekouras et al.* [2007] *Tsekouras et al.* [2008]
Hierarchical clustering	*Chicco et al.* [2005] *Chicco* [2012] *Hino et al.* [2013] *Panapakidis et al.* [2014] *Räsänen et al.* [2010] *Tsekouras et al.* [2007] *Tsekouras et al.* [2008]
K-Medoid clustering	*Panapakidis et al.* [2012]
Follow-the-leader (FDL) and its variations	*Carpaneto et al.* [2003] *Carpaneto et al.* [2006] *Chicco et al.* [2005] *Chicco* [2012]
Artificial Neural Networks (ANN)	*Macedo et al.* [2015]
Self-organizing Maps (SOM)	*Chicco et al.* [2005] *Panapakidis et al.* [2014] *Verdú et al.* [2006] *Tsekouras et al.* [2008]
Canonical Variate Analysis (CVA) and discriminant analysis	*Li et al.* [2010]

Source: author's own compilation and table.

In a similar way, canonical variate analysis (CVA) is not capable of clustering in itself. Here, clustering is based on canonical variables derived from between and within group covariance matrices (for example, in *Li et al.*'s [2010] study, using linear discriminant analysis).

Numerous other solutions exist; further information about them can be found, for example, in the literature listed in the Table above.

2.2. Capturing the effect of weather variables in energy time series

A fundamental question in relation to profiling consumption curves is dealing with the effects of various exogenous variables, such as variables describing weather, besides deciding about which effects of these variables are or are not regarded as **part of the profile**. As variables describing weather are basically stochastic in the same way as consumption is, deviations from the typical (or expected) value and their effects on consumption are highly relevant from the perspective of profiling (i.e. seeking typical consumption patterns). [54]

2.2.1. The relationship between weather variables and consumption

The quantification of weather variable effects is an extremely complex task. This does not only relate to exploring the **existence** of a relationship, but also **how** the effect of each variable can be captured most effectively. This involves the decisions about sampling (observation) frequency, using various structures for lagged variables, or other transformations performed on weather variables (such as the consideration of daily average-, or daily *peak* periods, etc.).

Regarding weather variables, the following are most often used: temperature, humidity, wind, the ratio of cloud coverage and precipitation (see: *Pitt* [2000]). Beyond the fact that the relationship between them is basically nonlinear, there are also interaction effects.

Nonlinearity that characterises **temperature** is usually managed by using *heating* and *cooling degree-day* (HDD and CDD, see: *Sugár* [2011] or *Mák* [2015]). *Espinoza et al.* [2005] likewise define the energy demand that arises as a result of the heating or cooling effect (HR and CR, that is, *heating* and *cooling requirement*).

Inclusion of interacting variables is supported if there are certain variables that reinforce or weaken each others' effects. Out of the other weather variables the effect of **humidity** is primarily relevant in the summer, when due to the sultry weather the use of air conditioners increases (*cooling effect* or often referred to as *discomfort effect*). The role of the **wind** is more significant in winter, especially when it is very cold, because then the co-occurrence of cold and wind greatly increases our sensation of coldness (that is, the winter

[54] Obviously, besides variables that describe temperature it is usually necessary to somehow deal with seasonality, *outlier*s, etc.; the specific way in which it is done depends on the methodology used.

heating effect will be even stronger; this is called *wind-chill effect*). Effects like this or similar to this can appear in a model in many different ways.

One solution is – similar to degree-days – to use formulas arrived at through practical experience and empirical research. Meteorological institutions have many kinds of *wind-chill* formulas.[55] The so-called *humindex* (*humidity index*), for example, is used for modelling the previously mentioned *discomfort* effect.[56]

Obviously, in some instances these indicators are not the most appropriate (likewise in the case of *heating* and *cooling degree-day*s the use of 12 of 21°C as threshold values, and the calculation of upward or downward deviations from this, as suggested by *Sugár* [2011] is not always optimal). Explorative solutions may be used to remedy all of this, such as MARS (*Multiple Adaptive Regression Splines*), for example, which is capable of capturing both nonlinearity and interactions without resorting to formulas obtained by empirically researched formulas. It makes it possible to reveal both threshold values and the possible interaction effects through a basically exploratory method, by the evaluation of exact measures (such as model selection criteria).

Throughout the research it is often a difficulty that appropriate quality data are not available for the analysis (for example, OMSZ (Hungarian Meteorological Service) data are

---

[55] Some of these can be summarised as follows:

- Canada:

$$T_{wc} = 13.12 + 0.6215T_a - 11.37V^{+0.16} + 0.3965T_aV^{+0.16},$$

where:

  - T_{wc} is the so-called *wind-chill* index, measured on °C scale,
  - T_a is the temperature of the air,
  - V is the windspeed on 10 metres (km/h),

- the United States:

$$T_{wc} = 35.74 + 0.6215T_a - 35.75V^{+0.16} + 0.4275T_aV^{+0.16},$$

where:

  - T_{wc} is the so-called *wind-chill* index, measured on °F scale,
  - T_a is the temperature of the air,
  - V is the windspeed on 10 miles (mph),

The above calculations apply only below 10°C (50°F), above 4.8 km/h (3.0 mph), these are the two threshold values in excess of which the combined, interaction effect of the variables is relevant.

[56] This is calculated in the following way:

$$Humindex = T_{air} + 0.5555\left[6.11e^{5417.7530\left(\frac{1}{273.16}-\frac{1}{T_{dew}}\right)} - 10\right],$$

where:

- T_{air} is the temperature of the air in °C,
- T_{dew} is the dewpoint in K-degree.

As there might be some approximating conversion between dewpoint and humidity, *humindex* can actually be used to quantify the interaction effect of temperature and humidity.

not available for free, even for research purposes; and internet data sources – often freely accessible – are not reliable enough).

Though slightly more indirectly related to this topic, it is worth mentioning that there might be some lagging in the effect of temperature (that is, the t^{th} period consumption is not only influenced by the t^{th} period, so-called *spot* temperature). Various (usually simpler) techniques have been developed to manage this, which is summarised here briefly through the solutions most commonly used concerning temperature:

- the use of some lags, that is, inclusion of data on the temperature of previous periods (T_{t-1}, T_{t-2} …) besides the t^{th} period temperature (T_t),
- the use of exponential weighting, that is, instead of T_t, using rather:
$$T_t(\alpha) = \alpha \cdot T_t + (1-\alpha) \cdot T_{t-1}(\alpha) \qquad \text{(where } 0 < \alpha < 1\text{)},$$
- the use of the *Noon Effective Temperature* (NET) defined by *Elexon* [2013], which produces the temperature relevant for the description of consumption on the given day from the average of the previous day and the day before that by fixed weighting in the following way:
$$T_t^{NET} = 0.57 \cdot T_t + 0.28 \cdot T_{t-1} + 0.15 \cdot T_{t-2}.$$
- and another such solution on the Hungarian natural gas market is the so-called sliding weighted average temperature used to calculate profiles for consumers (whose consumption is not measured regularly), which considers the temperatures of the days preceding the t^{th} day with gradually decreasing weight.

2.2.2. Capturing the effect of weather variables in profiling

As a consequence of their stochastic nature, there are two perspectives from which the effects of weather variables can be looked at:

- directly modelling the effect of the weather variable itself, or
- directly modelling the effect of deviation from the typical (average) tendency, the naturally occurring irregular behaviour of the weather (as a stochastic) variable.

The adaptation of these two options for the field of profiling can be summarised as follows:

- removal of the total effect of the weather variable from the profile, that is the effect of the weather is not part of the profile at all, or

- only the irregular effect of the weather variable is removed during profiling, so that only the typical (average) weather effect is included in the profile.

Typically, it is the first solution that is used in the related literature and, apart from a few exceptions, out of the variables that describe weather only temperature is taken into account.

Räsänen et al. [2010] use a simple regression technique to eliminate the effects of temperature in each of the four seasons in order to allow for **nonlinearity** (such as winter and summer temperature effects). Because of the difficulty in defining the beginnings and endings of seasons this is a disadvantageous and not too elegant solution.

Espinoza et al. [2005] define heating and cooling requirements separately (HR and CR, see further chapters about this), which is more elegant; however, the threshold values used for degree days may be problematic.

Pitt's [2000] study considers the effects of more weather variables, even profiling is performed by dividing the consumption time series into so-called *weather dependent* and *weather independent* parts. Weather variables are handled using the MARS (*Multiple Adaptive Regression Splines*) method which allows for the exploration and estimation of the best combinations of nonlinear and interaction effects to fit the time series. It is worth noting, however, that a solution such as this, which takes many weather variables into consideration, is very rare.

Mutanen et al. [2011] use a method that slightly differs from the previous ones, as correction is carried out for identical days and months, and it uses **deviations from the mean** for both temperature and consumption. The disadvantage of this method is that temperature-dependency can differ greatly within days as well, and this – while not to a great extent – influences the results, that is, it does not work so well in the case of multiple seasonality. In Section 2.2.4, dealing specifically with the extreme effect of temperature, it has been shown that a similar solution can work well with monthly data (*Mák* [2015]). In that part of the paper conclusions that apply to profiling are also described.

The phenomenon that traditionally the whole effect of the weather is removed while profiling can be explained by the fact that profiling is applied to more or less previously defined groups (e.g.: there is either winter or summer temperature-dependency for everyone, or both, or neither). This is not a problem when the 'only' aim is to create homogenous groups, and individual consumption behaviour is not relevant (neither in terms

of typical patterns nor uncertainty). Weather-dependency may be weaker or stronger for each consumer (which may be relevant in any individual treatment, for example, individual pricing), even uncertainty may be weather-dependent.

2.2.3. The extreme (irregular) effect of temperature

Sections 2.2.3. and 2.3.4. contain the slightly modified version of parts of an earlier study that are relevant for the purposes of this paper.

The more or less extreme values that occur in the variation of weather are not occasional fluctuations, but smaller or larger deviations from the mean that are constantly present; therefore, their classic modelling as *outliers* is disadvantageous, and not exactly appropriate.

The quantification of such irregular effects is important in the short and long run (this section focuses on issues related to the long term). As uncertainty (which is not like a structural break, for example) is constantly present, the management and quantification of the related risks is important and needs constant control, whether it be regulations or long term portfolio management, both on the supply (identification of necessary contracted volumes in the long run) and demand side (forecasting consumption).

This section describes, with the use of the seasonal adjustment methodology, to what extent stochastic shocks caused by temperature can affect the results of modelling (for example, the seasonal adjustment itself, or the identification of *outlier*s) and what improvements in results can be achieved by their explicit management.

The aforementioned extreme (irregular) feature can of course be dealt with using the estimation of a regression model (similar to regARIMA known in the methodology of seasonal adjustment), as the decomposition of the time series is basically independent from this, because it is used, *inter alia*, for the pre-adjustment of the time series. The methodology of seasonal adjustment – as an additional benefit – has provided an opportunity for both handling 'classic' *outliers* [57] and the validation of results (see, for example *slidings span* diagnostics). As the methodology of seasonal adjustment is not so directly related to the topic of this paper, this section focuses only on the handling of the effect of temperature, and only these parts of the earlier study (*Mák* [2015]) will be presented here.

[57] These are the following: *additive outlier* (AO), *level shift* (LS) and *transitory change* (TC).

Among the previously published results was the construction of an easily interpretable model that takes into account the nonlinear relationship of temperature and consumption from two aspects. On the one hand, it considers the effect of temperature as relevant only under a certain threshold level; this way, the use of *heating degree-day* is supported in the models. This solution is equivalent with other solutions typically used in this field. On the other hand, the model calculates with the monthly (heating) degree-day deviations from the mean, which makes it possible to consider the fact that the effect of temperature is not necessarily the same below the above mentioned threshold level.

Results shown here have numerous important consequences that are relevant from the aspect of profiling. These will be summarised at the end of the section, supplementing the study with some further model runs.

2.2.4. Seasonal adjustment and the removal of extreme (irregular) effect of temperature in the Hungarian natural gas consumption

This section presents part of an earlier study (*Mák* [2015]). Although the study is related to Hungarian natural gas consumption, it is important from the perspective of this dissertation to discuss how the effects of deviation from the mean temperature can be included in a model.

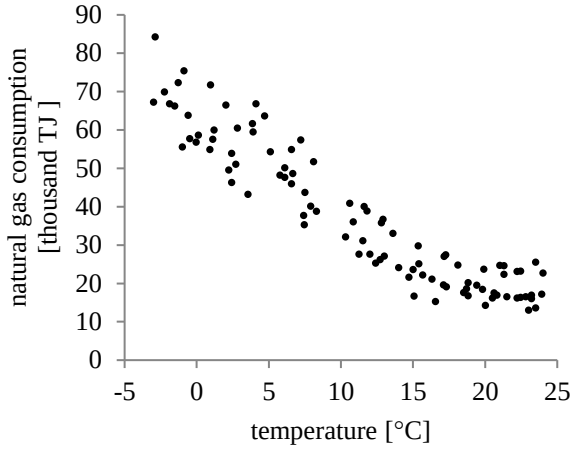
2.2.4.1. Data used in the analysis

The natural gas consumption data used in the empirical part of this study are from the Eurostat database[58] and temperature data were obtained from the Pestszentlőrinc weather station of Hungarian Meteorological Service.[59]

Figure 21 shows how temperature and natural gas consumption are related with regard to monthly average temperature and consumption data between 2006 and 2013. It is easy to see that above an average temperature of 15-16 $°C$ consumption is around a relatively stable level and as monthly average temperature drops, the value of consumption rises. Under the given threshold the relationship may even be called linear; however, disregarding the highest consumption values (which, otherwise, is January 2006, the coldest month of the observed period[60]), the linear relationship is not so evident anymore.

[58] http://ec.europa.eu/eurostat/, in thousand terajoules (TJ)
[59] http://www.varaljamet.eoldal.hu/cikkek/climate_budapest.html, in Celsius ($°C$)
[60] The only month colder than this was the February of 2012, however, because of the decreasing trend the consumption was much lower, despite the extreme cold.

Figure 21: Hungarian natural gas consumption as a function of temperature, 2006–2013



Source: author's own (Excel) figure based on Eurostat data.

The threshold value of 15-16 $°C$ can be taken into account in basically two ways. One way to calculate HDD (*heating degree-day*) is to calculate the downward deviations from the traditionally 16 $°C$-threshold (and their sum), and the other option is to only calculate with temperature values below 15 $°C$ (and only those), and take the value of their deviations from 18 $°C$ (and the sum of these values). The two methods expressed with formulas are the following:

1st method:

$HDD = \max(0, 16\ °C - temperature)$,

2nd method:

$HDD = 18\ °C - temperature$, if temperature $\leq 15\ °C$,

$HDD = 0\ °C$, otherwise.

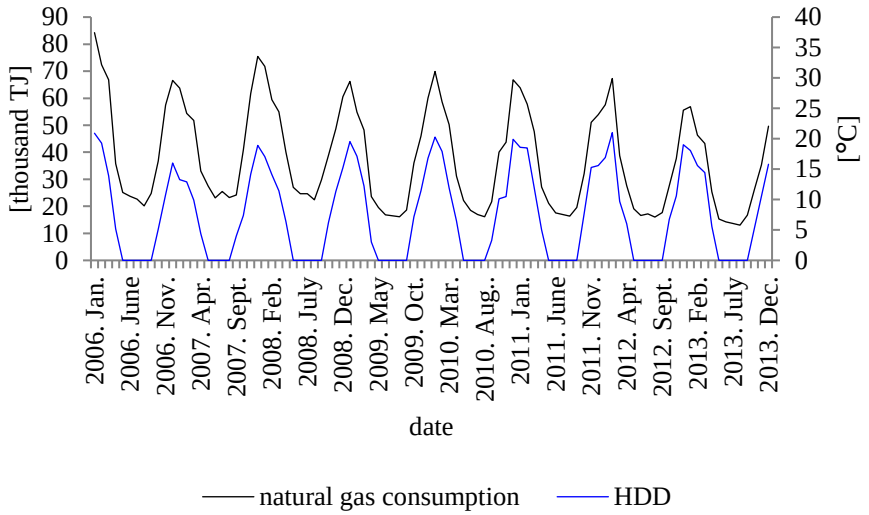
The calculation used in the second method can be done in a more sophisticated way (see for example *Howden et al.* [2001]); but in any case, the empirical studies prove that the latter method ensures a better fit, and it is also the formula used by Eurostat.

In this paper, the 2nd method was used for calculations.[61] Figure 22 shows the temporal fluctuation of HDD calculated on the bases of natural gas consumption and temperature for the period between 2006 and 2013. The downward trend in consumption is

[61] If daily data are available, aggregating the values calculated from them gives the monthly and yearly values, etc. As credible data dating back to 2006 are not available in daily resolution, and Eurostat only publishes such data until 2009, in the study – lacking daily data – results calculated from monthly means were used.

visible – which is not only due to the economic crisis of 2008 – as well as the higher consumption in colder winters, and lower in warmer winters; moreover, it shows the *outlier* effect of the cold in February 2012.

Figure 22: Hungarian natural gas consumption and HDD, 2006-2013



Source: calculations and figure by the author (Excel) based on Eurostat data.

Summarising the empirical claims made here, the HDD-method in itself is only capable of modelling the heating effect, but in observations of values under the threshold level, it only describes the relationships as linear. Still, the figures suggest that the relationships are not necessarily linear even under the threshold value.

This section shows a possible solution for the modelling and quantification of the latter empirical fact. Using the methodology shown, it will also be examined how the model can be used to deal with extreme temperature effects. It seems obvious to define the calculated HDDs and the deviations from the mean HDD, and use them as variables. The deviation may thus be positive or negative, depending on whether the given month was rather cold or warm compared to the average trend. As the months of December, January (and even February) are typically colder than the other months, it is practical to calculate the deviations from the mean by month. That is, to calculate the mean HDD looking at only the months of January[62] and to examine the differences between each January value and this mean – and proceed in the same way for all the other 11 months. This way, it is true for

[62] Mean here refers the mean in the sample period.

every month that the mean deviation from the mean HDD as a reference level is exactly zero (as the mean deviation from the mean is zero).

2.2.4.2. Estimation results

Results shown in this section were reached using the X-13ARIMA and X-13ARIMA-SEATS methods, and the Excel programme. The detailed steps will not be described, only the main results, and the difficulties and decisions that were made in the modelling process. HDD-deviations as exogenous variables (using the terms used by the seasonal adjustment software: *user-defined variables*) will be investigated for their influence on the quality of seasonal adjustment. This needs to be highlighted here, as seasonality in natural gas consumption is primarily influenced by the temperature, where the role of the stochastic nature is extremely great, therefore, it may have a considerable effect on the stability of seasonal adjustment.[63]

It may appear as if seasonality were a field rather distant from the topic of this paper. Its presence here, though, is supported by the fact that it may be one of the aims of profiling to calculate a typical (an average) trend, to arrive at a time series that is free from all sorts of irregular effects. Compared to the methodology shown here, the task is somewhat more complicated in profiling. For the methodology of seasonal adjustment in Hungarian see *Sugár* [1999a] and *Sugár* [1999b].

Concentrating primarily on the final results, this section shows how they were calculated with and without HDD-deviations. A lesson learned here is that the explicit inclusion of extreme temperature effects produces a better established model, for example, concerning the automatic selection of various structural breaks. The Reader may find the results of the most important tests and diagnostics in the cited paper.

Proceeding to the results of HDD-deviations (see Table 6), HDD-deviations of June, July and August have not been tested, as in these summer months the value and deviation of HDD was zero in the sample. The months of May and September each showed positive HDD values in the sample, but even here, HDD-deviations are not considered to be significant.

As the model was fitted after logarithmic transformation of the original time series, the time series can be written as an exponential function of various independent (or explanatory) variables.

[63] This will be examined using one of the well-known diagnostics of the X-13ARIMA method (*sliding spans*). The results arrived at will not be explained here, they can be found in the cited study (*Mák* [2015]).

In this case – as we know – the coefficients (let us mark β) are not interpreted directly, but as exp(β), due to the exponential function formula. That is, if the parameter for the month of January is 0.0363, then exp(0.0363) = 1.0370, that is, if the month of January is 1 $°C$ colder than the mean (that is, the value of HDD-deviation is 1 $°C$ higher), then the value of natural gas consumption is, *ceteris paribus*, on average 3.70 percent higher. The coefficients of the others months can be calculated and interpreted in a similar way.

Table 6: The main results of the regression model on the example of Hungarian natural gas consumption

Variable	Coefficient	Standard error	*t*-value
intercept	–0.0010	0.0016	–0.68
LS_2008_oct	–0.2203	0.0494	–4.46
HDD_deviation_jan	0.0363	0.0075	4.85
HDD_deviation_feb	0.0478	0.0066	7.30
HDD_deviation_mar	0.0556	0.0101	5.53
HDD_deviation_apr	0.0749	0.0145	5.17
HDD_deviation_oct	0.0439	0.0111	3.96
HDD_deviation_nov	0.0427	0.0091	4.70
HDD_deviation_dec	0.0320	0.0093	3.46
peak_deviation_summer	0.0001	0.0001	1.78

Source: author's own calculations (X-13ARIMA-SEATS) and table.

Finally, it might be worth observing some figures to see what advantages building temperature into the model has from the perspective of seasonal adjustment.

Without the inclusion of HDD-deviations the seasonally adjusted time series, that is, the original time series after the removal of the seasonal component, is rather zigzag, as the temperature induced irregular effect is still there (see Figure 23). This is particularly outstanding on the winter of 2011/2012, when February was very cold. The period around the economic crisis has a similar zigzag shape. As it was mentioned in the original study, this model did not recognise the structural breaks during the crisis either.

A similar phenomenon can be observed in the examination of SI ratios[64], where these values show great variability within a month (See Figure 24).

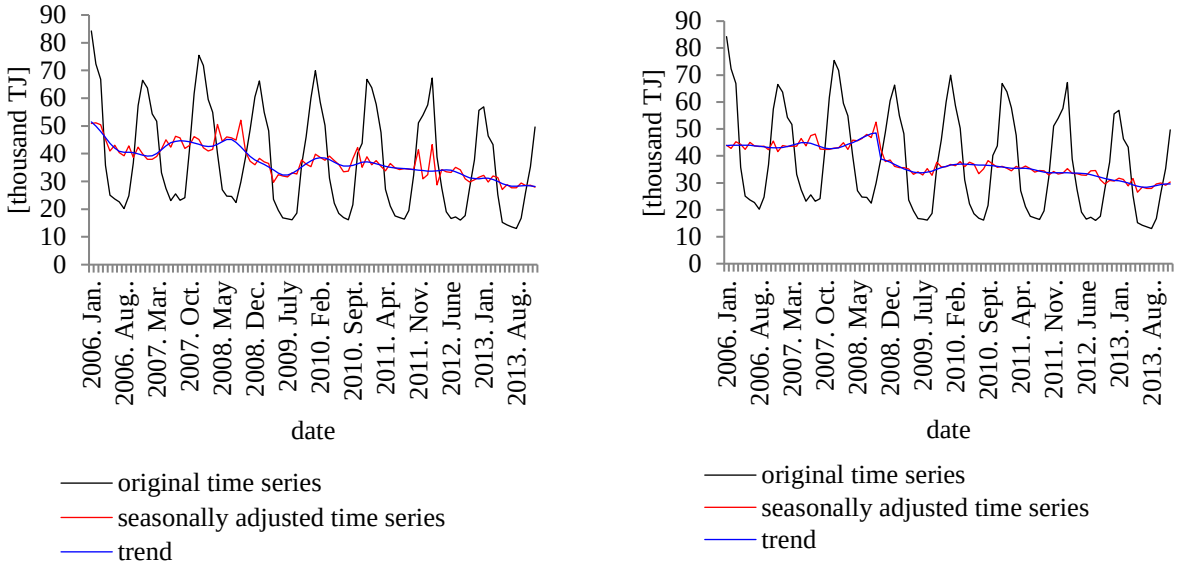
[64] SI ratio is the sum of the seasonal (free from regression effects, the latter ones are defined by the user or built-in ones) and the irregular components (see below) in the framework of seasonal adjustment. It is only briefly noted here what it means to define components and the relationships between them that provide the starting point concerning seasonal adjustment. In an additive model the time series can be written as:

$$Y = T + C + S + TD + H + O + I,$$

where the components are the following:

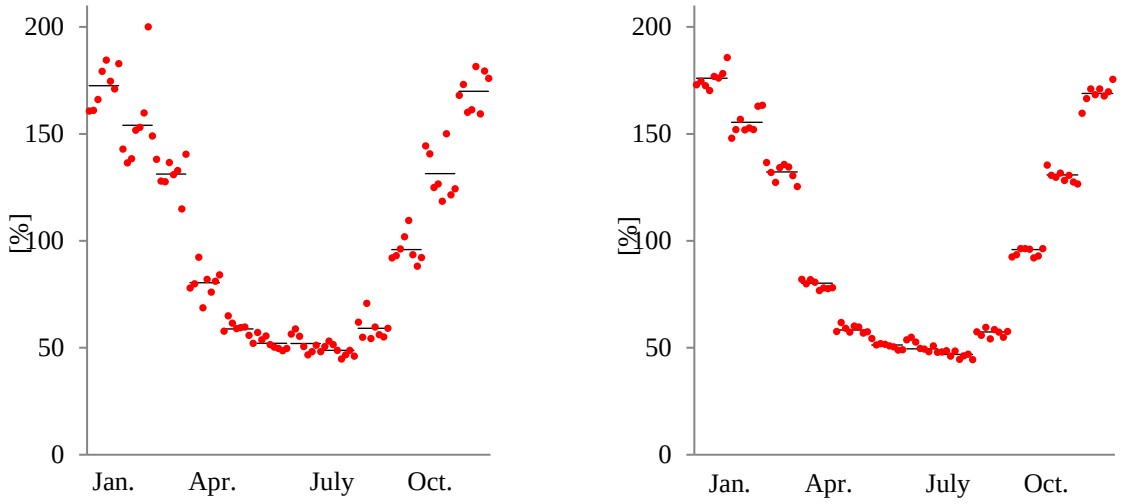
T – is long term trend,

Figure 23: Final results of seasonal adjustment without and with using HDD-deviations



Source: author's own calculations (X-13ARIMA-SEATS) and figure (Excel).

Figure 24: SI ratios without and with using HDD-deviations



Source: author's own calculations (X-13ARIMA-SEATS) and figure (Excel).

Having built HDD-deviations into a model, the outcome is a much less rugged, seasonally adjusted time series, as HDD-deviations were defined as part of the seasonal

---

C – represents the effect of the midterm cycle,
S – is the seasonality that describes regular fluctuation in a year,
TD – represents the effect of the different number of working days,
H – represents the effect of holidays,
O – represents the effects of observed *outlier*s,
I – is the error term (irregular component).

There are, among others, multiplicative and log-additive models, which involve writing the model as the product of these components or the sum of their logarithms.

component. The effect of the economic crisis is clearly recognisable both in the trend, and the seasonally adjusted time series (see Figure 23). Continuing the previous line of thought, here SI ratios fluctuate much less within each month (see Figure 24).

2.2.4.3. Removal the extreme (irregular) effect of temperature

Table 7 shows the estimated values of HDD-deviation coefficients and the results derived from them. According to this, if January is 1 $°C$ colder than the mean (the mean January value), then the consumption is *ceteris paribus* 3.70 percent higher on average (in February this value is 4.90, in March 5.72).

The results are not surprising in that December and January are typically the coldest; therefore, if in these months the temperature is 1 $°C$ lower, the consumption does not increase so much as in other months. This is partly because heating systems have their limits, and partly because if temperature decreases from –1 $°C$ to –2 $°C$, it has a greater effect than if it decreases from –5 $°C$ to –6 $°C$ (for example, due to economising). This also probably explains why the April value is the highest.

Table 7: The values of HDD-deviation parameters and their corresponding interpretations on the example of Hungarian natural gas consumption

Month	β	$\exp(\beta) - 1$ [percentage]
January	0.0363	3.70
February	0.0478	4.90
March	0.0556	5.72
April	0.0749	7.78
May	*	*
June	**	**
July	**	**
August	**	**
September	*	*
October	0.0439	4.49
November	0.0427	4.36
December	0.032	3.25

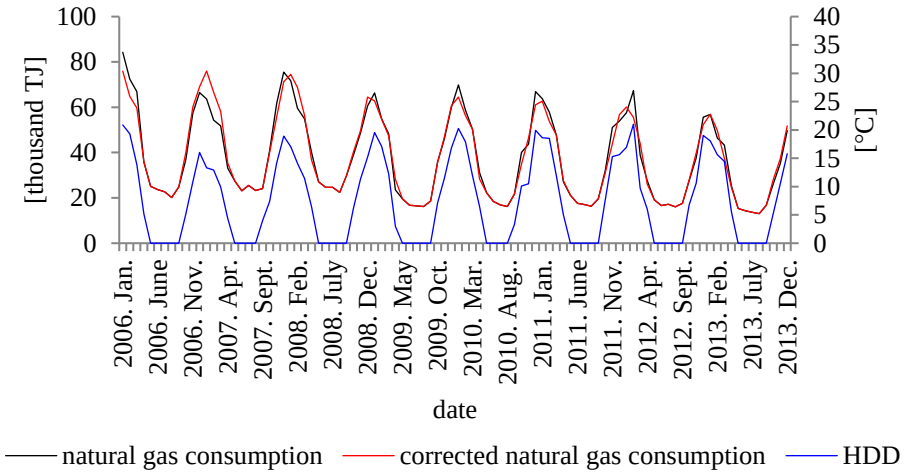
* The parameter is not significant.
** Without parameter estimation (HDD-deviation zero).

Source: author's own calculations (X-13ARIMA-SEATS) and table (Excel).

It may, of course, occur in an extremely cold May or September too (although it is rare, because the heating season typically starts in the middle of October), but this does not cause excess heating effect, hence HDD-deviation in these months is not significant.

Consequently, if there is HDD-deviation in September, correction with this is not needed either. In January months 1 $°C$ of HDD-deviation causes *ceteris paribus* 3.70 percent higher consumption. Therefore, if a January was 1 $°C$ colder than the mean, the temperature corrected consumption of this month needs to be decreased to 3.70 percent lower than the actual consumption, because the excess consumption was caused by the weather being colder than the mean; if it was 2 $°C$, then double the amount, and so on. The correction is also valid *vice versa*, that is, if a January was 1 $°C$ warmer than the mean, than the temperature corrected consumption of this month needs to be raised by 3.70 percent, as the weather being warmer than the mean was the cause of the lower consumption.

Figure 25: Temperature corrected natural gas consumption, 2006–2013



Source: author's own calculations (X-13ARIMA-SEATS) and figure (Excel).

In general, temperature correction in the above model can be performed as follows:

$$Y \cdot exp(-\beta * HDD\ deviation),$$

where

Y – contains data of the original time series,

β – denotes the coefficients for the estimated HDD-deviation,

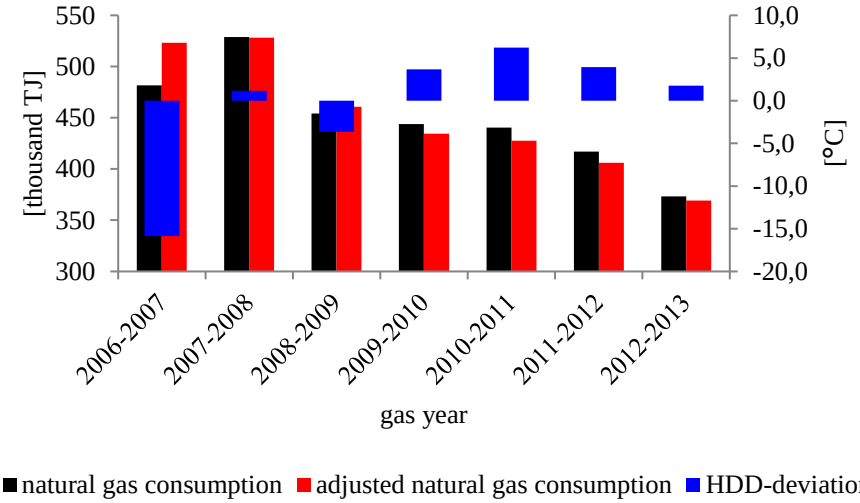
HDD-deviation – denotes the monthly time series of HDD-deviation.

Figure 25 shows the temperature corrected times series of the Hungarian gas consumption based on the formula above.

It is observable that while, for example, in 2006/2007 the winter was relatively mild, the temperature corrected quantities there are higher than the actual quantities, while in the extremely cold February of 2012, the correction was downward.

Figure 26 shows the aggregate of the previous results according to gas year. A gas year lasts from 1st July to 30th June in the next calendar year, this way, the gas year does not cut the winter in two, and the effect of relatively mild or cold winters can be evaluated more easily (see gas years 2006-2007 and 2010-2011).

Figure 26: Temperature corrected natural gas consumption according to gas year, 2006–2013



Source: author's own calculations (X-13ARIMA-SEATS) and figure (Excel).

2.2.4.4. The main conclusions from a profiling perspective

Though this study uses monthly resolution time series, at a first glance, it provides a technique that can be easily adapted for the removal of extreme weather (temperature) effects from daily, hourly or quarter-hourly time series.

Profile may as well be defined with the removal of the effect of weather (temperature) deviation from its typical (average) tendency. For this reason, the results shown here may be interpreted as a kind of profiling, as this can also be regarded as the derivation of the consumption under given typical (average) temperature circumstances. As a matter of fact, this method is also used by *Mutanen et al.* [2011]. The technique, however, is not so advantageous if applied to high-frequency time series, which can be explained by, *inter alia*, the multiple (intraday and yearly) seasonality in temperature and by phase shifts (if summer or winter starts earlier or later). It becomes more difficult to define what typical (average) temperature means in a given point in time[65]; besides, this form of discretisation may result in too high standard deviation, as a result of which certain coefficients do not

[65] *Mutanen et al.* [2011] use a technique where group means are made on the basis of differentiating months or days of the week; what is meant here by irregular temperature effect is the deviation from the partial means.

seem significant. This kind of technique is able to handle nonlinearity primarily in aggregate (for example, monthly or quarterly) time series.

Given the focus of the analysis in this section, it deals less with time-dependent dispersion, heteroscedasticity; however, some related statements can be made to supplement the discussion above.

Based on Figure 21, natural gas consumption seems to have constant standard deviation in temperature-dependent and non-temperature-dependent periods. Though it was not among the results of this section, it might provide a useful foundation for the rest of the dissertation to perform the seasonal adjustment working only with HDD instead of HDD-deviations, and without the **logarithmic transformation**. Appendix D) nicely shows that if the model here is applied on a non-logarithmised time series, there is some loss in the stability of SI ratios within months (Appendix D, Figure d)); and working with HDD instead of HDD-deviations, there is even more of such loss (Appendix D), Figures a) and c)).[66,67]

Another important experience is that the regular 'yearly shape' of SI ratios only appears if models involving HDD-deviations are used. The reason behind this may be that in such cases HDD and the seasonal components after the removal of regression effects are correlated, and if so, the removal of the HDD effect itself – from the viewpoint of the dissertation as well – is not (so) appropriate. This problem is similar to the problem of multicollinearity (the - strong - correlation of independent variables). It is like including seasonal (monthly) *dummy* variables in multivariate regression, where not only the interpretation but also the logic of regression decomposition are difficult to realise.

It should be noted that slight heteroscedasticity is present even in the example here (that is, the dispersion of temperature-dependent periods is somewhat higher). This, however, is less apparent in Figure 21 due to the higher summer consumption levels at the beginning of the time period. Since this is an aggregate and not very long (a few 100 or 1000 element) time series, heteroscedasticity – due to these properties – is much less detectable numerically (for example, using statistical tests), but is much more apparent graphically (see for example: figures in Appendix D)).

[66] Concerning the results that appear in the Appendix – for ease of comparison – no monthly varying HDD-deviation parameters are estimated.

[67] As a matter of fact, similar conclusions can be drawn based on the residuals of the SARIMA model. The illustrative opportunity for interpretation is provided by the framework of seasonal adjustment and the relatively more elegant analysis using the SI ratios.

High-frequency (hourly, quarter-hourly) time series are typically more heteroscedastic (like it was assumed while examining stylized facts). Moreover, the problem of handling the nonlinearity (or additionally, to interaction effects) arises in an even more complex way, and for this reason, the co-presence of the two phenomena and their appropriate treatment is even more critical. The similar practical transformation of the variables of a basically linear model is not necessarily plausible. In addition, the outcomes of the present technique are not so advantageous either, as both logarithmising and the use of HDD-deviations handled (basically 'eliminated', removed) heteroscedasticity.[68]

Summarising the above, if the behaviour of the irregular component of the time series is relevant from the research perspective, the removal of the total effect of temperature (HDD) is not really advantageous from neither technical nor interpretational aspects. The removal of the irregular temperature effect is likely to reduce the heteroscedasticity of consumption. Similar tendencies are also valid concerning high-frequency time series, with the presence of the difficulties mentioned at the beginning of this section.

The listed results and experiences can all be regarded important both from the aspect of choosing the **appropriate methodology** and for formulating future **research questions**. A supplement to all this is the theoretical consideration that one of the major sources regarding the variation and uncertainty of the otherwise typically price inflexible electricity consumption is weather (primarily temperature). For this reason, removing its effect and the analysis of the so-called (extreme) temperature-free time series may be a limiting factor regarding possible analyses, relevant methods and conclusions that could be drawn.

As a closure of the topic of handling temperature (weather) effects and as a preliminary to the subsequent chapter, it should be stressed in connection with the methodology applied here, that profiling itself is in many aspects similar to seasonal adjustment (for example in pre-adjustments, decomposition logic, existence of multiple solutions, etc.) and the complexity of the latter method may be revealing regarding profiling as well. However, this dissertation does not deal with the possibility of applying such a multi-stage complex methodology.

[68] Logarithmic trasformation – that results in different function form and parameter interpretation – is otherwise often used in practice, for example when the logarithm of consumption is represented as a function of temperature (cf. the *scatter plot*s in Section 1.4). This implies that the effect of $1°C$ is much more related with the relative (%) variation of consumption than with its absolute variation. The model selection criteria in this study have also validated this concept.

3. AN OVERVIEW OF METHODS USED IN THIS DISSERTATION AND THEIR APPLICATIONS IN PROFILING

The part of Chapter 1 on examining stylised facts that characterise consumption curves contained the short overview of some methods because of their simpler logic, and often because of their being less known or less widely used.

The classical (or so regarded) stochastic time series regression models serve as a *benchmark*, or reference point in Chapter 4, discussing the empirical research results of this paper, therefore, they will only be reviewed here briefly. The SARIMA model is often referred to as the basic model of stochastic time series analysis. The periodically autoregressive model – as an extension of that – and the methodological background of its practical application on stationary time series that relates to the topic of this paper will be dealt with here – drawing on a previous publication by the author on the same topic (*Mák* [2014a], *Mák* [2014b]).

The majority of the methodological review consists of mixture models and within that the overview of the *Gaussian* mixture model; including, among others, the more detailed description of the regression approach building on it (the so-called *Gaussian* mixture-regression). This is supported by the fact that this field is lesser known, especially in the Hungarian literature and practice. The chapter closes with the presentation of the previous results of mixture models from the energy (partly profiling) field.

3.1. Classical stochastic time series regression models

One of the central concepts of stochastic time series analysis is **stationarity**. Among other things, the importance of stationarity is essentially in that stationary time series can be modelled in the framework of stochastic time series analysis. If this is not met, then the time series needs to be transformed in some way to become stationary. From the perspective of this paper, it is worth mentioning that **the error term (the stochastic shock)** has a different kind of role in case of stationary and non-stationary (integrated) time series both in terms of methodology and (practical) consequences. Before **writing** of these **classical time series models,** these issues will be discussed less formally. Further detail can be found in the related literature (see, for example: *Hamilton* [1994], *Maddala* [2004], *Ramanathan* [2003]).

3.1.1. The definition of stationarity and testing for unit root

A time series has **weak or covariance stationarity** if its expected value and variance are constant in time. A time series is strictly stationary if for any $t_1, t_2 \ \dots \ t_m$ time set it is true that its joint probability distribution is identical with the joint probability distribution of any $t_{k+1}, t_{k+2} \ \dots \ t_{k+m}$ time set, where k is any integer. This means that the joint distribution of the variables only depends on the distances of $t_1, t_2 \ \dots \ t_m$ times from each other, but not on t. Weak stationarity does not imply strict stationarity, however, it is not true *vice versa*, as the higher order moments are not necessarily time-independent.

Unit root tests[69] are basic tools for identifying the **presence or absence of stationarity**. The most widespread versions that are still used to this day are the ADF-test (Augmented Dickey-Fuller-test (*Dickey-Fuller* [1979]), KPSS-test (Kwiatkowski-Phillips-Schmidt-Shin-test, *Kwiatkowski et al.* [1986]); and for time series influenced by seasonal factors, for example, HEGY-test may be used to test for the presence of seasonal unit root (*Hylleberg et al.* [1990]).

3.1.2. The role of the error term in integrated time series

It is especially true for energy time series that non-stationarity has two main sources: the presence of **trend** and **seasonality**. Concerning trend, there may be two cases: the presence of deterministic or stochastic trend (that is, it has a unit root) – or in rare instances, both. Making a parallel with seasonality, it can also be modelled in a deterministic or stochastic way. If non-stationarity is due to a deterministic origin, the role of the error term is practically the same as for stationary time series. This is described in the following section.

If the source of non-stationarity is of a stochastic kind, it can be handled by non-seasonal (first) differencing ($y_t - y_{t-1}$) or seasonal differencing (in general $y_t - y_{t-s}$[70]); application of the latter, though, has to fulfil two extremely marked and serious assumptions: on the one hand, the presence of all unit roots (one non-seasonal and $(s-1)$ seasonal; due to which the problem of overdifferencing may occur) and on the other hand, the independence of the non-seasonal and the corresponding seasonal components. In the author's studies that have been cited earlier (*Mák* [2014a] and *Mák* [2014b]) the requirements for the relaxing of this assumption of independence regarding the model

[69] *Unit root* is an expression related to the formal definition of the SARMA model in Section 3.1.4. A process is stationary if the roots of the difference equation are outside the unit circle, and a unit root process is one where there roots are on the unit circle.

[70] s refers to the periodicity of the time series.

components are discussed in detail together with its implications for testing for stationarity.[71]

For the purposes of this paper it would rather be worth highlighting that the framework introduced in these studies is capable of examining the **fluctuation of shocks over time**, that is, which seasons are the ones that have the largest **long-run effect** and are more likely to be incorporated into the value of the time series, or which ones undergo the heaviest impact from the **accumulation** of shocks. As the long term incorporation of shocks and their appearance in the time series differ seasonally, the model can be viewed as the coherent modelling of stochastic trend and seasonality (that are not independent from each other). In other words, it means that the risks are not identical across seasons, and this solution can handle this **season-dependent risk** in case of integrated time series.

3.1.3. The role of the error term in stationary time series

In stationary time series – as opposed to integrated time series – stochastic shocks do not have a permanent effect on the expected value of the time series on the long term.[72] Based on the assumptions of classical multivariate regression model the error term is *White Noise* (*WN*), that is:

- it has zero mean,
- its standard deviation is constant in time (that is, it is homoscedastic),
- its values are independent from each other in time (that is, it is not autocorrelated).

In practice, however, it is not true in most cases that the error term is perfectly random, that is, an independent, identically distributed (**IID**) random variable.

In financial time series, one of the most conspicuous characteristics is the phenomenon of time-dependent dispersion, or heteroscedasticity. According to the empirical experience in the financial area (the so-called *volatility clustering* (see *Cont* [2005])) in certain periods prices are more hectic, more widely dispersed than they are in other periods, which is one of the risks posed by stationary time series.

---

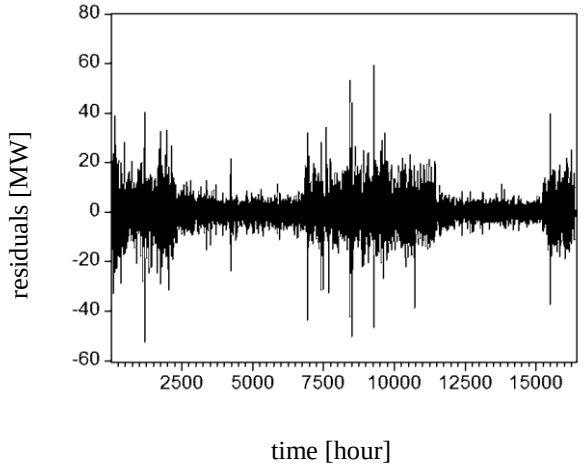
[71] The studies show with the use of the periodic autoregressive model structure how the presence or absence of stationarity can be tested for if the stochastic trend and seasonality are not independent from each other. Relaxing of the assumption of independence, that is, the assumption that the value of the seasonal component is dependent on the trend (or *vice versa*), is usually identified in practice as the multiplicative relationship of the components. Therefore, the technique demonstrated there can also be practically seen as testing for unit root in a multiplicative model environment. Otherwise, there are techniques that help the selection between additive and multiplicative models, but the discussion of this would exceed the scope of this paper (see more on this in *Sugár* [1999a], [1999b]).

[72] This means that those time series are also regarded stationary which are only stationary with the inclusion of various deterministic variables (like e.g. analytical trend, *dummy* variables).

This heteroscedasticity on the energy market appears due to **fundamental** reasons (e.g.: daily seasonality, seasons, temperature) more often than in financial markets, but at the same time, it is also a much less mapped and modelled field.

Figure 27 shows the fluctuation of the SARMA model residuals fitted on a heavily winter weather-dependent time series (heating/warming energy use, heat consumption). It is a slightly more than two-year-long hourly resolution time series where the standard deviation of the residuals is higher in winter periods and lower in summer periods.[73] The tendency that characterises the standard deviation can be explained by the (low) temperature and human activity, and the resulting 'increased' random behaviour.

Figure 27: The fluctuation of residuals in a SARIMA model



Source: author's own figure (EViews) based on the author's calculations (EViews).

In financial markets the fundamental stock price models are less likely to be short term, besides for the short term the use of (G)ARCH-type models is more widespread ((G)ARCH, *(generalized) autoregressive conditional heteroscedasticity*, see: *Bollerslev* [1986], *Engle* [1982]). In the variance equation written here (which is the explicit realisation of modelling standard deviation) their use with additional fundamental variables is not so advantageous anymore.[74]

[73] On the horizontal axis 'unfortunately' only times appear, not dates.

[74] In (G)ARCH models as compared to the SARIMA specification (see Section 3.1.4) the ε_t residual component is not assumed to be *White Noise*, its **standard deviations is the function of time**, as follows:

$$\varepsilon_t = \sigma_t \cdot u_t,$$

where the time series of the so-called underlying residuals is a *White Noise* process, that is $u_t \sim WN(0, \sigma^2)$. Beyond the mean equation the **variance equation** needs to be estimated to model the non-constant dispersion of residuals over time. A general GARCH(q,p) model can be written as:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2.$$

It is worth mentioning that heteroscedasticity may have various sources. In energy time series the appropriate quantification of **nonlinearity** is highly relevant. There are many options for its modelling (the degree-day method is such, or in an exploratory view, the use of MARS[75] type models), therefore it is worth examining if heteroscedasticity is caused by the inappropriate quantification of linearity or is there because of the omission of some other variable.

Time-varying dispersion or heteroscedasticity does not have a negative effect on estimated parameters; they are unbiased, which means that such models work very well, for example, in forecasting. The standard errors of parameters, however, are biased and inconsistent.

3.1.4. Seasonal autoregressive moving average (SARMA) model

The basic tools for the analysis of stationary time series are the so-called *autoregressive moving average* (ARMA) models, written:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where ε_t is the error term, which is assumed to be *White Noise* ($\varepsilon_t \sim WN(0, \sigma^2)$) and $t = 1, 2, \ldots T$, where T is the length of the observed time series, p is the number of AR (autoregressive), and q is the number of MA (moving averages) terms. Resulting from the presence of autoregressive terms the value of the time series at a given time depends on the past time values of itself, that is, its **past realisations.** The moving average terms mean that the value of the time series at a given time is dependent on the **past errors**. The extension of ARMA models with seasonal lags is found, for example, in *Box* and *Jenkins* [1976]. **Seasonal lags** are built into the SARMA (*seasonal autoregressive moving average*) model in a similar way as non-seasonal lags, where in the notation P usually refers to the number of seasonal AR (autoregressive) and Q to the number of seasonal MA (moving average) terms.[76]

---

where q is the order of ARCH terms, p is the order of GARCH terms. Accordingly, the conditional variance of a given period is dependent on the squared residuals of the previous period and the conditional variance of the previous period, in addition, and the so-called unconditional or long term variance can be easily derived. Definition of conditional variance in the framework of another model (mixture regression) will surface later in the section.

[75] MARS, *multiple adaptive regression splines* (see more on this in: *Friedman* [1991]).

[76] It is only for clarification of terminology that in models where the original time series was made stationary after d non-seasonal or D seasonal differentiation are generally denoted by $SARIMA(p,d,q)(P,D,Q)_S$ (the notation p and q for the number of AR and MA lags, and P and Q are the number of seasonal AR and seasonal MA lags is already known).

Estimation of SARMA model parameters is normally performed by the *Maximum Likelihood* (ML) procedure.[77] It is worth noting that besides autoregressive and moving average terms, the models may include exogenous variables (that is how SARMAX models are produced), which in case of energy time series typically means *dummy* variables that represent seasonality or variables that describe weather. There is an abundance of literature in Hungarian as well about multivariate regression analysis (see for example: *Hunyadi-Vita* [2003], *Kerékgyártó et al.* [2008]) besides other previously mentioned studies that discuss the topic primarily from the perspective of econometric applications.

3.1.5. Periodic autoregressive (PAR) model

This section contains highlights from two more detailed studies (see *Mák* [2014a] and *Mák* [2014b]) that concentrate primarily on integrated time series from practical aspects as well.[78] Here, the emphasis is on stationary processes regarding possible analyses relevant for the discussion in this study.

3.1.5.1. Definition of the model

Starting out from the classical autoregressive model of order *p*:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t,$$

where ε_t error term is *White Noise* ($\varepsilon_t \sim WN(0, \sigma^2)$); its extension to a periodic autoregressive model of order p is the following (for the sake of simplicity, quarterly time series is assumed):

$$y_{t,s} = \phi_{1,s} y_{t-1} + \phi_{2,s} y_{t-2} + \ldots + \phi_{p,s} y_{t-p} + \varepsilon_t,$$

where according to periodicity in quarterly time series $s = 1, 2, 3, 4$. The formula shows that $\phi_{p,s}$ parameters that pertain to lags of order p are seasonally or periodically varying. The alternating of autoregressive coefficients from period to period is supported empirically by the fact that in different time series the value of the first quarter does not depend in the same way on the fourth quarter of the previous year as the second quarter on

[77] Choosing the right number of lags is usually performed using correlogram and/or model selection criteria. The most commonly used **model selection criteria** in practice are *Maximum Likelihood*-based model selection criteria, Akaike information criterion ($AIC = -2(ln\ L) + 2m$) and Schwarz information criterion ($BIC = -2(ln\ L) + m(ln\ n)$). In the above m denotes the number of independent variables and L is the value of the optimalised likelihood function. The criterion minimum marks the model to be chosen. The error term of a well-specified SARMA model is *White Noise*, besides the correlogram there are other methods for **testing for *White Noise***, such as Ljung-Box Q-test (see: *Ljung-Box* [1978]), or Breusch-Godfrey LM-test (see: *Breusch* [1978], *Godfrey* [1978]).

[78] This is the so-called *Periodically Integrated Autoregressive* (PIAR) model.

the first, etc. As different autoregressive coefficients are estimated for each period, the model is also suitable for the estimation of periodically varying autocovariance.

It is also worth mentioning how the model is written as a system of equations (as this approach is the most plausible for the empirical part of the paper), where the number of equations is – obviously – identical with the number of seasons, that is:

$$\Phi_0 Y_{T,s} = \Phi_1 Y_{T-1,s} + \Phi_2 Y_{T-2,s} + \ldots + \Phi_p Y_{T-p,s} + E_T,$$

where the error terms of vector $E_T = \left[\varepsilon_{T,1}\ \varepsilon_{T,2}\ \varepsilon_{T,3}\ \varepsilon_{T,4}\ \right]^T$ are *White Noise* ($\varepsilon_{T,s} \sim WN(0, \sigma^2)$) and $s = 1, 2, 3, 4$.

The variables in the system of equations are $Y_{T,s} = \left[y_{T,1}\ y_{T,2}\ y_{T,3}\ y_{T,4}\ \right]$ and $Y_{T-1,s} = \left[y_{T-1,1}\ y_{T-1,2}\ y_{T-1,3}\ y_{T-1,4}\ \right]$; these yearly vectors contain quarters of the years T and $(T-1)$. It can be seen that the indices of the y_t variables observed quarterly are changed. The time variable t recorded quarterly is replaced by T, s variables, which also record quarterly data, but at the same time show which year and which quarter it is.

In matrices that contain parameters the first index refers to the order of lag, the second shows to which period's equation the given time lag order applies; thus simplifying the above to a model with four lags, the parameter matrices are the following:

$$\Phi_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\phi_{1,2} & 1 & 0 & 0 \\ -\phi_{2,3} & -\phi_{1,3} & 1 & 0 \\ -\phi_{3,4} & -\phi_{2,4} & -\phi_{1,4} & 1 \end{bmatrix}, \text{ and } \Phi_1 = \begin{bmatrix} \phi_{4,1} & \phi_{3,1} & \phi_{2,1} & \phi_{1,1} \\ 0 & \phi_{4,2} & \phi_{3,2} & \phi_{2,2} \\ 0 & 0 & \phi_{4,3} & \phi_{3,3} \\ 0 & 0 & 0 & \phi_{4,4} \end{bmatrix}$$

For example, the first row of the Φ_1 matrix shows how the first quarter of year T depends on the first, second, third and fourth quarters of year $(T-1)$. Obviously, the order of the arrangement in the matrix is the reverse of what the order of the time lag would suggest, as the first quarter of year T is preceded by the fourth quarter of year $(T-1)$, whose coefficient for this reason is $\phi_{1,1}$.

Based on the above, the system of equations can be reconstructed in the following way (after rearrangement – see previous paragraph):

$$y_{T,1} = \phi_{1,1} y_{T-1,4} + \phi_{2,1} y_{T-1,3} + \phi_{3,1} y_{T-1,2} + \phi_{4,1} y_{T-1,1} + \varepsilon_{T,1}$$

$$y_{T,2} = \phi_{1,2} y_{T,1} + \phi_{2,2} y_{T-1,4} + \phi_{3,2} y_{T-1,3} + \phi_{4,2} y_{T-1,2} + \varepsilon_{T,2}$$

$$y_{T,3} = \phi_{1,3} y_{T,2} + \phi_{2,3} y_{T,1} + \phi_{3,3} y_{T-1,4} + \phi_{4,3} y_{T-1,3} + \varepsilon_{T,3}$$

$$y_{T,4} = \phi_{1,4} y_{T,3} + \phi_{2,4} y_{T,2} + \phi_{3,4} y_{T,1} + \phi_{4,4} y_{T-1,4} + \varepsilon_{T,4}$$

It is noteworthy that matrices have practical content, as Φ_0 contains parameters that pertain to the quarters of the same year, while Φ_1 contains this information for the previous year.

In summary, the PAR model reviewed here may be written in several ways. The first way presented here makes it easier to interpret the basic notion behind the model. Writing it as a system of equations may appear strange compared to the traditional one equation formula of univariate autoregressive models, still, as a result of its periodicity it is a practical representation on the one hand, and on the other, can be used to derive certain results. Options for model selection are identical with those in classical multivariate regression methodology.

3.1.5.2. Calculating the long term equilibrium (mean)[79]

In a previous section it was assumed that the time series is centred, that is, its expected value is zero; therefore, none of the model formulas contained the intercept term. There are practical applications when this solution is not appropriate. In this case the model written in a matrix form is modified according to the following:

$$\Phi_0 Y_{T,s} = C + \Phi_1 Y_{T-1,s} + \Phi_2 Y_{T-2,s} + \ldots + \Phi_p Y_{T-p,s} + E_T,$$

where C contains the periodically constant parameters (which may be identical or different across periods). If the time series is stationary, the long term average, that is, expected value of the time series can be defined:

$$E\left(Y_{T,s}\right) = \left(\Phi_0 - \Phi_1 - \Phi_2 \ldots - \Phi_p\right)^{-1} C,$$

where $E(.)$ denotes the expected value. The formula can be obtained easily, because if this long term average really exists, then $E\left(Y_{T,s}\right) = E\left(Y_{T-1,s}\right) = E\left(Y_{T-2,s}\right) = \ldots = E\left(Y_{T-p,s}\right)$, besides $E(E_T) = 0$, which easily yields the above formula.[80]

There will be applications in this dissertation were besides the intercept and autoregressive terms there will be other exogenous variables among the independent variables (e.g.: heating or *cooling degree-day* values). Following *Espinoza et al.* [2005] profile can be interpreted as a daily curve in hourly or quarter-hourly resolution, from

[79] Though the title may seem too general, the application of the result as profile makes it highly relevant from the perspective of this dissertation.

[80] The result though is nothing else, but the generalisation of long term mean formula for the well-known AR(p) model, which is the following:

$$E(y_t) = \frac{c}{1-\phi_1-\phi_2 \ldots -\phi_p},$$

where parameters ϕ_1, $\phi_2 \ldots$ ϕ_p denote the coefficients of the given autoregressive terms.

which the effects of all the other variables have been removed. This is equivalent with the notion that the values of all exogenous variables (such as degree-day values, various *dummy* variables) are made equal to zero, and the above formula is used. Thus the previously described long term equilibrium (mean) can be interpreted as a definition for profile.[81] This will be discussed further in the chapter on empirical results.

3.2. Mixture models

The application of the so-called *(Finite) Mixture Models* (that use a finite number of components) is gaining ground in an increasing number of fields, and provides a basis form many future applications in practice. *Mixture Models* (MM) have appeared in many fields including biology, agriculture, medicine, economics or signal processing (see for example: *McLachlan-Basford* [1988], *McLachlan-Peel* [2000]). The family of mixture models includes many applications. It involves the representation of the distribution of various phenomena as a mixture of known distributions (such as normal distribution), clustering, discriminant analysis or regression estimation based on mixture models.

The methodologies used in this paper are the *Gaussian Mixture Model* (GMM) and a regression approach based on it; these will be reviewed in detail. It is an advantage of the model and one of the reasons for using it in this paper, that the standard errors of point estimates of the dependent (output) variable can be written as a function of the given values of the independent variables, therefore, the model may be suitable for dealing with heteroscedasticity more fundamentally. **This section contains many formulas that cannot be found in other studies and that show – not necessarily obvious – similarity to classical multivariate regression.** Besides, the paper uses high resolution (quarter-hourly) time series, which usually requires deviation from the general estimation procedure, which is also new compared to other empirical studies (see for instance the examples of Section 3.3.3). The methodological background to this is provided by the application of mixture models that is based on random sampling, besides using clustering and discriminant analysis (see for example: *Fraley-Raftery* [2000]).

[81] The article also mentions that heteroscedasticity can be handled in the PAR model (as an extention of the basic model described here) in a way that periodically different residual standard variance is estimated (which on a quarter-hourly basis would mean 96 further estimated parameters). This, however, does not take into account the difference in heteroscedacticity between e.g. weekdays-weekends and seasons.

3.2.1. Description of the mixture model (MM) and the *Gaussian* mixture model (GMM)

Given data y with independent observations $(y_1, y_2 \ ... \ y_n)$ of number n, that is $y = (y_1, y_2 \ ... \ y_n)$, where:

- y_i is a vector of size $(m \times 1)$ containing the attributes of observation i $(i = 1, 2 \ ... \ n)$,
- n is the number of observations,
- m is the number of attributes. [82]

Let us assume that the observations are generated by a mixture distribution with *K* components whose density function can be written as:

$$f(y) = \prod_{i=1}^{n} \sum_{k=1}^{K} \tau_k \cdot f_k(y_i|\theta_k),$$

where:

- $f(.)$ is the density function of the mixture distribution,
- $f_k(.)$ is the density function of component k,
- θ_k denotes the parameters that describe component k,
- the *prior* probability τ_k is the probability of observation i belonging to component k,
- k denotes the components ($k = 1,2, \ ... \ K$), and
- K is the number of mixture components.

In most cases – as in this paper – it is assumed that the distribution of component *k* is normal, that is $f_k(.)$ denotes the multivariate normal *Gaussian* density function parameterized by mean vector μ_k and covariance matrix Σ_k parameters, so the distribution of component *k* can be written:

$$f_k(y_i|\theta_k) = \varphi(y_i|\mu_k, \Sigma_k) = \frac{1}{|2\pi\Sigma_k|^{-1/2}} exp\left[-\frac{1}{2}(y_i - \mu_k)^T \Sigma_k^{-1}(y_i - \mu_k)\right],$$

where:

- $\varphi(.)$ is the multivariate normal *Gaussian* density function,
- $|.|$ is the determinant.

[82] The term *attribute* used in international, mainly data mining literature is identical with the term *variable* in regression terminology.

Components are normally identified as clusters when applied for grouping in the framework of mixture models and clustering, so in the rest of this paper the terms *cluster* and *component* may be used interchangeably: with the former term denoting interpretation and the latter emphasising the methodological approach.

3.2.2. Expectation-Maximization (EM) estimation procedure

Estimation of mixture model parameters is carried out by the *Maximum Likelihood* (ML) method, the *Expectation-Maximization* (EM) algorithm (see for example: *Dempster et al.* [1977], *McLachlan-Krisnan* [1997]).[83] The EM algorithm consists of the successive iteration of *estimation steps* (*E-step*) and *maximization steps* (*M-step*).

The algorithm views observations as an incomplete data set (with missing, unobserved variables), which means that they are thought of as pairs of variables (y_i, z_i). Here, variable z_i is not observed, it denotes the so-called indicator variable which shows which observation belongs to which component. That is, z_{ik} equals 1, if observation i belongs to component k, otherwise it is 0.[84] In so far as these component memberships z_{ik} are missing or non-observed values, they need to be estimated when using the EM algorithm, which is realised in the form of *posterior* probabilities p_{ik}.

Let $\psi = (\tau_1, \tau_2 \dots \tau_K, \theta_1, \theta_2 \dots, \theta_K)$ denotes the parameters to be estimated, that is the *prior* probabilities of the components and the parameters of the normal distribution. The *likelihood*-function is the following:

$$L(y) = \prod_{i=1}^{n} \prod_{k=1}^{K} f_k(y_i|\theta_k)^{z_{ik}},$$

and the *loglikelihood*-function is:

$$log\big(L(y)\big) = l(y) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \cdot log(f_k(y_i|\theta_k)).$$

---

[83] It is worth noting that the structure of the covariance matrix Σ_k in the mixture model can be defined in several ways. According to the general, so-called unconstrained version (see: *Banfield-Raftery [1993]*), the Σ_k covariance matrix can be written as:

$$\Sigma_k = \lambda_k D_k A_k D_k^T,$$

where:

- D_k denotes the orthogonal matrix of eigenvectors,
- A_k is the diagonal matrix (whose diagonal elements are proportional to the eigenvalues of the matrix),
- λ_k denotes the so-called constant of proportionality.

These parameters may be different by cluster or there may be constraints for each parameter. In Fraley and Raftery's study there are examples for which specific cases are equivalent with which clustering techniques (for example: *Fraley-Raftery* [2000], *Fraley-Raftery* [2007]), as the eigenvalue decomposition of the covariance matrix Σ_k described above is only a framework. This paper will always use the general, unconstrained version, as it was found in the other reviewed literature (e.g.: *Srivatstav et al.* [2013], *Eirola-Lendasse* [2013]).

[84] Of course, this way it is also true that $\sum_{k=1}^{K} z_{ik} = 1$ for every observation i.

Given the observations $y = (y_1, y_2 \ ... \ y_n)$ of number n, iteration $(r + 1)$ means performing the following steps.

In the *E-step* on the basis of the set of parameters in iteration *r*, that is $\psi^{(r)}$, for every observation i the *posterior* probability p_{ik} of belonging to the component k is calculated:

$$p_{ik}{}^{(r+1)} = \frac{\tau_k^{(r)} \cdot f\left(y_i|\theta_k^{(r)}\right)}{\sum_{k=1}^{K} \tau_k^{(r)} \cdot f\left(y_i|\theta_k^{(r)}\right)},$$

and in addition, using this, the value of the Q function is calculated, which provides the expected value of the *loglikelihood* that applies to the whole data set given the estimated parameters and the observed values of the variables in the sample, that is:

$$Q\left(\psi|\psi^{(r)}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik}^{(r+1)} log\left(f_k(y_i|\theta_k)\right).$$

Using the calculated *posterior* probabilities $p_{ik}^{(r+1)}$ as weights in the *M-step* the values of the parameter set $\psi^{(r+1)}$ are obtained by maximising the Q function, which means optimising in the following way:

$$\psi^{(r+1)} = arg \max_{\psi} Q\left(\psi|\psi^{(r)}\right),$$

or written differently:

$$Q\left(\psi|\psi^{(r)}\right) \underset{\psi}{\rightarrow} max,^{85}$$

whose result gives the optimal solution, that is, the estimated values of the parameters in the $(r + 1)$ iteration:

$$\psi^{(r+1)} = (\tau_1^{(r+1)}, \tau_2^{(r+1)} \ ... \ \tau_K^{(r+1)}, \theta_1^{(r+1)}, \theta_2^{(r+1)} \ ... \ \theta_K^{(r+1)}).$$

As a result of the *M-step*, the *prior* probabilities of the components can also be expressed using the following formula:

$$\tau_k{}^{(r+1)} = \frac{\sum_{i=1}^{n} p_{ik}{}^{(r+1)}}{n}.$$

The *E-* and *M-steps* are iterated successively until the parameter estimations start to converge, or a maximal iteration number is reached. The choice between the models is based on model selection criteria (e.g.: AIC, BIC). Model selection criteria are helpful not only in determining the optimal number of variables but also the optimal number of

[85] The two ways to write the optimalisation task mean the same; the second may be simpler, the first is a bit more formal. The *argmax* notation is the short form of the expression *arguments of maxima* and marks those points of the domain of some function at which the function values are maximal.

components, which is one of the advantages of model-based clustering as opposed to other clustering solutions.[86]

3.2.3. An empirical example on the daily natural gas consumption data of Budapest

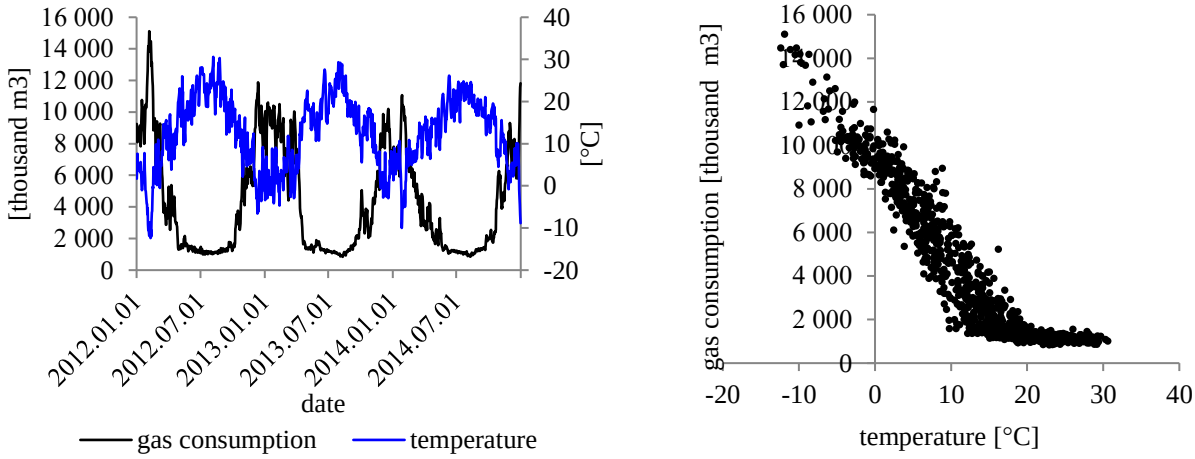
For ease of understanding the use of mixture models, model-based clustering is demonstrated on a simpler example, comparing the results to those of a traditional clustering method, K-Means clustering, which is again iterative. Some results are in Appendix B). Figure 28 shows the procedure of K-Means clustering, the so-called pseudo-code that summarises its logic.

Figure 28: The pseudo-code of K-Means clustering

1.	*(random[87]) selection of starting cluster centroids of number k*
REPEAT	
2.	*assignment of observations to clusters*
a.	calculating the distance between observations and cluster centroids
b.	assigning observations to the cluster from whose cluster centroid the distance is the shortest
3.	*recalculating cluster centroids*
a.	calculating average observations by cluster and
b.	identifying it as a clustercentroid
UNTIL	*the given convergence-criterion is met*

Source: author's own figure.

Figure 29: Hungarian daily gas consumption, daily average temperature and daily average temperature – gas consumption *scatter plot*



Source: author's own figure (Excel).

[86] The formula of the BIC criterion – with the known formula – is: $BIC = -2 \cdot l(y) + log(n) \cdot bp$, and the formula of the AIC criterion is: $AIC = -2 \cdot l(y) + 2 \cdot bp$, where bp marks the number of estimated parameters (which may differ according to the structure of the covariance matrix).

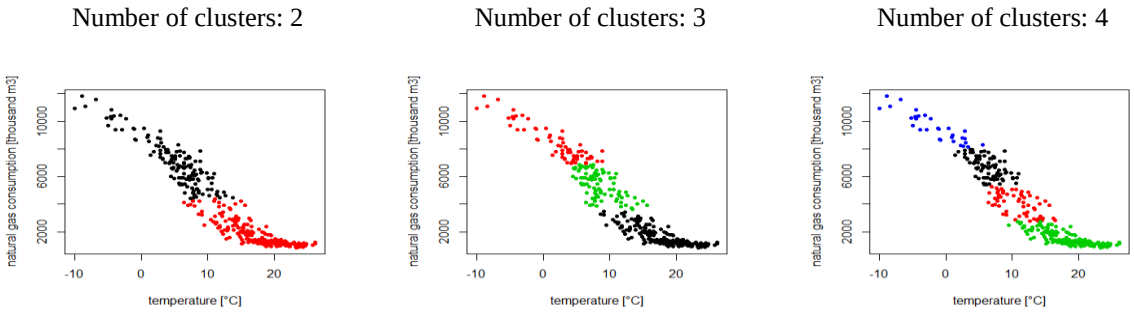
[87] The random selection of initial cluster centroids can be ensured in various ways. The conventional K-Means clustering and K-Means++differ basically in that in K-Means++ they are selected more efficiently; therefore, the running time is shorter than using the conventinal method which, in addition, is often stuck in a local optimum.

The differences between the clustering techniques are shown on the evolution of the daily natural gas consumption in Budapest[88]. The development of the natural gas and temperature time series is shown in Figure 29. Clustering will be performed in the space of these two variables or attributes (temperature and gas consumption), that is, the observations to be grouped are the days. Based on the figure on the right side, it can be assumed that at least two clusters need to be created, one for temperature-dependent and another for temperature-independent days.

Selecting only year 2014 the results of K-Means clustering are shown on Figure 30. As for K-Means clustering previous knowledge of the number of the clusters is required, the results are shown for 2, 3 and 4 clusters as well.

The bottom line of the figure shows the ratio of the so-called Between Sum of Squares (BSS) and Total Sum of Squares (TSS). In K-Means clustering this ratio can (also) be relied on in selecting the optimal number of clusters[89]. Given the selected number of clusters between 2 and 10, Table 19 in the Appendix shows these values. At the same time, besides the fact that deciding on the number of clusters in this way is fairly inconvenient (between 4-5 clusters, where there is a break in the BSS/TSS trend) the interpretation of the results is not too neat either. As Figure 30 shows, the clusters seem to be organised around the bands parallel with the temperature axis (which is especially visible when using fewer clusters).

Figure 30: The results of K-Means clustering on the example of daily average temperature – gas-consumption



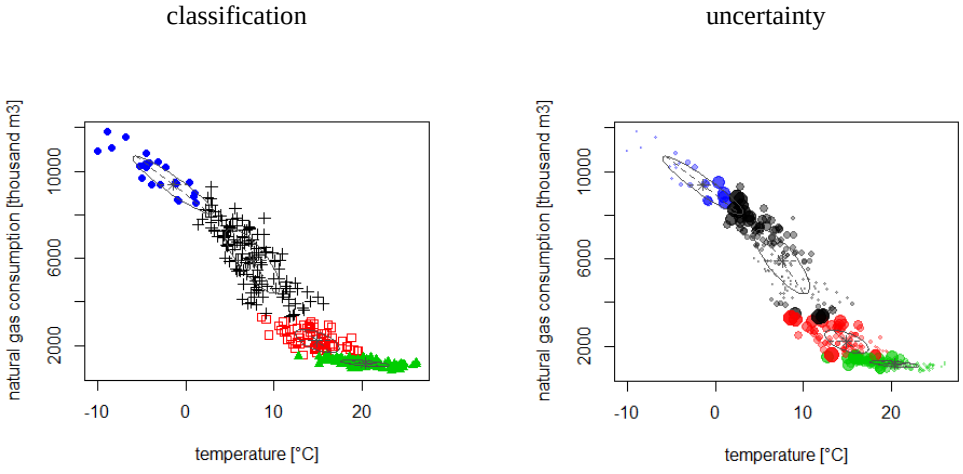
Source: author's own calcuations (R) and figure (R).

[88] Data source: FGSz Ltd. (the Hungarian natural gas transmission system operator), www.fgsz.hu
[89] Of course, there are much more complex, sophisticated indicators for the selection of the most appropriate number of clusters.

As opposed to K-Means clustering, mixture clustering has the advantage that there the selection of the number of clusters is or may be determined automatically, based on model selection criteria (BIC), therefore, it is not necessary to assume *a priori* knowledge of the number of clusters. In Table 20 in the Appendix the minimum of BIC marks the cluster to be selected, and for this reason, the mixture model selected contains four clusters (components).

As it was mentioned in the previous section, for every observation the probabilities of belonging to a cluster will be calculated, whose sum for each observation is 1 (or 100%). When wishing to assign observations clearly to a cluster, cluster membership is produced by the maximum of the conditional probabilities, that is, every observation will be classified in the cluster where it is most likely to belong. Based on this, the left side of Figure 31 shows the 4 clusters thus created. Days marked with green are not weather-dependent, and the remaining weather-dependent days can be divided into three parts. Red marks the days where the so-called heating effect starts to apply (the so-called transition periods), black marks the days which are obviously temperature-dependent, and blue marks the colder days. For each different cluster the weather-dependency of the consumption is different. We may imagine fitting regression lines on the 4 'groups of dots'[90], or could examine the direction of the axes of the ellipses[91] in the figure to support this.

Figure 31: Results of mixture clustering on the example of daily average temperature – gas consumption



Source: author's own calculations (R) and figures (R).

[90] In what follows, the regression based on mixture model assumes such logic; though not by each 'group of dots', but for the full data set using *posterior* probabilities as weights.

[91] The ellipses show contours according to the $mean \pm standard\ deviation$ value of the two-dimensional normal distribution.

Figure 31 represents the uncertainty of classification. The uncertainty of classifying observation (u_i) can be defined as follows:

$$u_i = 1 - \max_k p_{ik},$$

that is, the maximum probability of classification is subtracted from 1, or 100%. The lower the maximal *posterior* probability is for an observation, the more uncertain the classification is.

Using the graphs of the R programme the colours of the clusters are inherited from the graph that represents classification, and an observation is marked with a bigger and darker symbol if its uncertainty is higher. Without knowledge of the specific values, it is clear that uncertainty is higher on the domain where clusters slightly intersect, and that uncertainty is highest on the intersections of the coldest (blue colour) and colder (approx. below 10 $°C$) days, and in transition (red colour) days.

Obviously, cluster centroids can be calculated here as well (see Table 21): in clusters where the daily average temperature is lower the daily average consumption is higher. These can be treated as typical, representative consumption patterns resulting from normal distribution. **This statement will be further discussed in Chapter 4 on profiling, constructing typical consumption patterns.**

It is worth examining the distribution of days by months and by days of the week (between weekdays and weekends) (see Tables 22 and 23). It clearly shows the distribution of days among months resulting from their weather dependency (extremely cold, cold, transition and no heating) and the related gas consumption. In addition, it shows that the distribution of weekdays and weekends within a cluster is approximately the same (the ratio of weekends is around 2/7 = ~28.6% everywhere), which means that there is no separate weekend cluster, as the effect of heating is much more dominant than the weekly calendar effect.[92]

As this section had mainly didactic purposes, it has raised a number of questions and possible analyises that cannot be discussed here in enough detail. Only one of these will be mentioned here, which is important regarding the rest of this dissertation. The question is if observations have different dispersions by cluster (here, for example the dispersion of weather-independent days with a given temperature seems much lower, but it looks as if the same was true for extremely cold days) then to what extent can they be transformed to

[92] This statement is important because in energy time series the variance explained by different levels of seasonality is completely different.

measure uncertainty of consumption?[93] In the rest of this chapter the methodological aspect of this question will be discussed, and in the following empirical chapter, the results will be presented using quarter-hourly power consumption time series.

3.2.4. Further methodological questions related to the *Gaussian* mixture model

This section reviews some methodological questions that remain after the general discussion of the EM algorithm and the GMM.

3.2.4.1. Determining the initial cluster memberships and posterior probabilities

The disadvantage of using the conventional EM algorithm is that it may easily stuck in a local optimum or in 'itself' it does not find a solution that can be interpreted easily. As the final result depends on how initial cluster memberships or initial parameters are obtained, these can be determined manually, but usually it is advisable to run the procedure multiple times using some randomized initial cluster membership values or parameters (for more on this topic see: e.g.: *Biernacki et al.* [2003]).

It is also a possible solution to apply other, modified EM algorithms first (such as: CEM, SEM, hierarchical EM), then use their results (initial cluster memberships values or parameters) as a basis for the conventional EM algorithm.

3.2.4.2. Hierarchical model based clustering

In the package *'mclust'* of the R Project programme used in this dissertation, the initial cluster memberships are obtained by using the so-called hierarchical EM algorithm. Hierarchical model based clustering is agglomerative, that is, each observation forms a unique cluster at the start. The procedure is based on maximizing the so-called *classification loglikelihood* (*cl*) which may be written as:

$$cl\left(\psi|z_1, z_2 \dots z_n, y_1, y_2 \dots y_n\right) = \sum_{k=1}^{K} \sum_{\{i|z_i=k\}} log\left(\tau_k f_k(y_i|\,\theta_k)\right),$$

where $\{i|z_i = k\}$ is the set of observations that belongs to component k. Due to the presence of this $\{i|z_i = k\}$ condition, the previous conventional EM algorithm cannot be applied here. The step-by-step merging of clusters ensures that the value of classification likelihood increases at the highest possible rate in each step. Of course, in the last step all observations are classified as belonging to one cluster. The advantage of this solution is that

[93] Section 2.2 has already addressed a similar question in connection with the extreme (random) effect of temperature.

as a result of the hierarchical, agglomerative solution the evolution of merging steps is not dependent on the number of clusters.

3.2.4.3. Gaussian mixture model in discriminant analysis

Using the logic of mixture models it is possible to perform discriminant analysis as well, which as opposed to clustering is a so-called *supervised* procedure. Here, known classification is modelled using various (independent) variables while in clustering there is no *ex-ante* classification, but observations regarded in some way similar are classified into one cluster or group (the so-called *unsupervised* procedure).

In the framework of mixture models there are various options for this (see: *Fraley-Raftery [2000]*):

- *Eigenvalue Decomposition Discriminant Analysis* (EDDA), and
- *Mixture Discriminant Analysis.*

The former assumes a single normal component for each class, the latter allows fitting a mixture model as a density estimate for each class.

This dissertation uses the EDDA discriminant analysis method. Essentially, this is nothing else but the completion of one (discrete) *M-step* and *E-step*. While performing the discrete *M-step* the highest of the *posterior* probabilities for each observation that results from the mixture clustering is equalled to 1, and all the others to 0. An *M-step* is run on the data set, and then the observations that do not form part of the classification procedure can be assigned in the *E-step* to the cluster where the *posterior* probability is the highest.

3.2.5. The regression approach based on the *Gaussian* mixture model (GMR)

The *Gaussian* mixture regression is based on the *Gaussian* mixture model by identifying one of the variables as the dependent (output) variable, and a regression is written using the other variables as independent (input) variables.

The notations used in the previous section is slightly modified for the purposes of the regression application:

- y_i is the value of the dependent variable for observation i,
- x_i is the vector of the length $(p \times 1)$ that contains the values of the independent variables for observation i $(i = 1, 2 \ \ldots \ n)$,
- n is the number of observations,

- p is the number of independent variables.

Previously, the variables were uniformly denoted by y_i and the number of variables m. Besides, where needed, matrix representation will be used in the overview of the literature on *Gaussian* mixture regression.

3.2.5.1. Derivation of the Gaussian mixture regression

To facilitate understanding and a better parallel with the results of the literature, partitioning of the mean and the covariance matrix may be done as follows:

$$\mu^k = \begin{bmatrix} \mu_Y^k \\ \mu_X^k \end{bmatrix}, \text{ with the sizes: } \begin{bmatrix} 1 \times 1 \\ p \times 1 \end{bmatrix},$$

and

$$\Sigma^k = \begin{bmatrix} \Sigma_{YY}^k & \Sigma_{YX}^k \\ \Sigma_{XY}^k & \Sigma_{XX}^k \end{bmatrix}, \text{ with the sizes: } \begin{bmatrix} 1 \times 1 & 1 \times p \\ p \times 1 & p \times p \end{bmatrix},$$

where the partitioned means and covariance matrices are likewise weighted means and weighted covariance matrices using the *posterior* probabilities p_{ik} as weights, that is:

$$\mu_y^k = \frac{Y^T diag(W_k)}{\sum_{i=1}^n diag(W_k)}, \qquad \mu_x^k = \frac{X^T diag(W_k)}{\sum_{i=1}^n diag(W_k)},$$

$$\Sigma_{\mathrm{XX}}^{\mathrm{k}} = \frac{\mathrm{X^T W_k X}}{\sum_{i=1}^n diag(W_k)}, \qquad \Sigma_{\mathrm{XY}}^{\mathrm{k}} = \frac{\mathrm{X^T W_k Y}}{\sum_{i=1}^n diag(W_k)},$$

$$\Sigma_{\mathrm{YX}}^{\mathrm{k}} = \frac{\mathrm{Y^T W_k X}}{\sum_{i=1}^n diag(W_k)}, \qquad \Sigma_{\mathrm{YY}}^{\mathrm{k}} = \frac{\mathrm{Y^T W_k Y}}{\sum_{i=1}^n diag(W_k)}.$$

where:

- X is a matrix of size $(n \times p)$ containing the x_i independent variables ,
- W_k is a diagonal matrix of size $(n \times n)$ containing the p_{ik} weights for each component k,
- Y is a vector of size $(n \times 1)$ containing the values of the dependent variables y_i,
- $k = 1 \ldots K$ denotes the components.

The regression coefficients for each component k are obtained by the weighted least squares method on y_1, $y_2 \ldots$ y_n and x_1, $x_2 \ldots$ x_n variables with using the *posterior* probabilites p_{ik} as weights, that is:

$$\widehat{\beta_k} = (X^T W_k X)^{-1} X^T W_k Y,$$

or they can be obtained by using a different formula based on the partitioned covariance matrices:

$$\widehat{\beta_k} = \Sigma_{YX}^k \left(\Sigma_{XX}^k\right)^{-1}.$$

The variance of the error terms for each component k is:

$$\widehat{\sigma_k^2} = \frac{\sum_{i=1}^n p_{ik}\left(y_i - x_i^T \widehat{\beta_k}\right)^2}{\sum_{i=1}^n p_{ik}},$$

or with a matrix formula:

$$\widehat{\sigma_k^2}^{(r+1)} = \frac{\left(Y - X\widehat{\beta_k}\right)^T W_k \left(Y - X\widehat{\beta_k}\right)}{\sum_{i=1}^n diag(W_k)} =$$

$$= \frac{Y^T W_k Y - \left(\widehat{\beta_k}\right)^T X^T W_k Y - Y^T W_k X\widehat{\beta_k} + \left(\widehat{\beta_k}\right)^T X^T W_k X\widehat{\beta_k}}{\sum_{i=1}^n diag(W_k)} =$$

$$= \frac{Y^T W_k Y - \left(\widehat{\beta_k}\right)^T X^T W_k Y - Y^T W_k X\widehat{\beta_k} + \left(\widehat{\beta_k}\right)^T X^T W_k X (X^T W_k X)^{-1} X^T W_k Y}{\sum_{i=1}^n diag(W_k)} =$$

$$= \frac{Y^T W_k Y - \left(\widehat{\beta_k}\right)^T X^T W_k Y - Y^T W_k X\widehat{\beta_k} + \left(\widehat{\beta_k}\right)^T X^T W_k Y}{\sum_{i=1}^n diag(W_k)} =$$

$$= \frac{Y^T W_k Y - Y^T W_k X\widehat{\beta_k}}{\sum_{i=1}^n diag(W_k)} = \frac{Y^T W_k Y - Y^T W_k X (X^T W_k X)^{-1} X^T W_k Y}{\sum_{i=1}^n diag(W_k)}$$

which, written with the partitioned covariance matrices is identical with:

$$\widehat{\sigma_k^2} = \Sigma_{YY}^k - \Sigma_{YX}^k \left(\Sigma_{XX}^k\right)^{-1} \Sigma_{XY}^k.$$

Regarding $\widehat{\beta_k}$ and $\widehat{\sigma_k^2}$ the results are obviously the same as the results of other studies in the literature (see for example: *Srivastav et al.* [2013]), but the derivation of the formulas is also detailed here.[94]

3.2.5.2. Conditional mean and standard deviation of the dependent variable

The component-based calculation of the conditional mean and conditional standard error is based on conditional mean and conditional standard deviations for each component using

[94] The results obtained here may look strange, however, the differences compared to the classical multivariate regression are only do to the weighting with *posterior* probabilities. In classical multivarate regression the formula for estimated parameters written with the matrix formula used here is:

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

The regression equation is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots \hat{\beta}_p x_{ip},$$

or put differently:

$$\hat{y}_i = x_i^T \hat{\beta},$$

and the variance of the error term is:

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^n \left(y_i - x_i^T \hat{\beta}\right)^2}{n-p-1}.$$

This way, it is much easier to see the similarity between the results of classical multivariate regression and mixture regression.

posterior probabilities as weights. Assuming normal distribution for each component, the density function of the dependent variable y_i can be written – based on *Srivastav et al.* [2013] – as:

$$\Phi\big(y_i, \lambda(x_i)\big) = \sum_{k=1}^{K} p_{ik} \cdot \frac{1}{\sqrt{2\pi}s_{ik}} \cdot exp\left(-\frac{1}{2}\left(\frac{y_i - m_{ik}}{s_{ik}}\right)^2\right),$$

where $\lambda(x_i) = \{p_{ik}, m_{ik}, s_{ik}\}$. That is, as seen from the notation, the value of the parameters $\{p_{ik}, m_{ik}, s_{ik}\}$ depends on the given values of independent variables x_i.

The calculation of probabilities p_{ik} is the same as before:

$$p_{ik} = \frac{\tau_k \cdot f(y_i|\theta_k)}{\sum_{k=1}^{K} \tau_k \cdot f(y_i|\theta_k)}.$$

The values of the other parameters are:

$$m_{ik} = \mu_Y^k + \Sigma_{YX}^k \big(\Sigma_{XX}^k\big)^{-1} (x_i - \mu_X^k),$$

and

$$s_{ik}^2 = \Sigma_{YY}^k - \Sigma_{YX}^k \big(\Sigma_{XX}^k\big)^{-1} \Sigma_{XY}^k.$$

It is easy to see that the formula of m_{ik} is nothing else but a substitution in the regression equation assuming that there is no constant. So the formula is written for centred variables (that is, for variables from which the corresponding means are substracted resulting in zero-mean variables), which means the following:

$$m_{ik} - \mu_Y^k = \Sigma_{YX}^k \big(\Sigma_{XX}^k\big)^{-1} \big(x_i - \mu_X^k\big) = \widehat{\beta_k}\big(x_i - \mu_X^k\big),$$

It is familiar from general regression terminology that if regression is estimated with an intercept using the original variables, and without an intercept using centred variables, the estimated $\widehat{\beta_k}$ parameters are the same.

The formula of s_{ik}^2 is identical with the residual variance per component.

From the above it follows that the expected value and variance of the output variable (that is, the squared standard error) can be written as:

$$\hat{y}_i = \sum_{k=1}^{K} p_{ik} \cdot m_{ik},$$

and

$$var(\hat{y}_i) = \sum_{k=1}^{K} p_{ik} \cdot (s_{ik}^2 + m_{ik}^2) - \left(\sum_{k=1}^{K} p_{ik} \cdot m_{ik}\right)^2.$$

The well-known formula that variance is the difference of the squared quadratic mean and the squared arithmetic mean is used twice in deriving the variance of the output variable.

3.2.5.3. Confidence interval for the dependent variable

In the mixture regression described here there is no way to calculate confidence interval in the conventional way, as the underlying distribution is the estimated mixture distribution.[95] The integral of the conditional density function described in the previous section can also be written (based on *Srivastav et al.* [2013]) as:

$$\Phi\left(y_i, \lambda(x_i)\right) = \sum_{k=1}^{K} \frac{p_{ik}}{2}\left(1 + erf\left(\frac{y_i - m_{ik}}{\sqrt{2}s_{ik}}\right)\right),$$

where *erf(.)* is the so-called *Gaussian* error function. This is used in the theory of probability and statistics, and it is related to the distribution function of standard normal distribution in the following way:

$$\Phi(x) = \frac{1}{2} + \frac{1}{2} erf\left(\frac{x}{\sqrt{2}}\right) = \frac{1}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right),$$

from which formula its relationship with the above formula is quite obvious.

For the calculation of the lower limit of the confidence interval with α confidence level it needs to be defined which value y_L, if integrated from -∞ to value y_L, gives the integral value $\frac{\alpha}{2}$, that is, the y_L value is sought where the condition below applies:

$$\frac{\alpha}{2} = \sum_{k=1}^{K} \frac{p_{ik}}{2}\left(1 + erf\left(\frac{y_L - m_{ik}}{\sqrt{2}s_{ik}}\right)\right).$$

In calculating the upper limit of the confidence interval with α confidence level it needs to be defined which y_U value, if integrated from value $y_U$to +∞, gives the integral value $\frac{\alpha}{2}$, that is, the y_U value is sought where the condition below applies:

$$1 - \frac{\alpha}{2} = \sum_{k=1}^{K} \frac{p_{ik}}{2}\left(1 + erf\left(\frac{y_U - \hat{m}_{ik}}{\sqrt{2}\hat{s}_{ik}}\right)\right).$$

The above equations can be solved using, for example, the Newton-Raphson-method (see for example: *Srivastav et al.* [2013]).

This dissertation used a different, iterative method to identify the upper limit of the confidence interval in a way that the value of the mixture distribution function was calculated for the point estimation, for a much higher value compared to point estimation, and for the mean of these two. If the value of the mean is higher than $1 - \frac{\alpha}{2}$, the upper limit of the confidence interval is between this mean and the point estimation, so the logic

[95] The formula of the confidence interval in classical multivariate time series regression – using the notations of the previous footnote – is the following symmetrical interval: $\hat{y}_t \pm t_{1-\frac{\alpha}{2}}(n - p - 1) \cdot \hat{\sigma}$, where instead of index i index t is used to empasize the time series feature.

described was successively applied until the deviation from the $1-\frac{\alpha}{2}$ probability was minimal. The solution can be used for the lower limit of the confidence interval as well. The iterative method described here definitely converges to adequate results, as the distribution function of the mixture distribution – due to the characteristics of the distribution function itself – is monotonically non-decreasing.

3.2.6. *Gaussian* mixture regression for time series

In previous sections it was mentioned that the *Gaussian* mixture model is often applied for time series, e.g.: in speech recognition or signal processing. Then, the task becomes multivariate in a way that in addition to the time series observed the other variables are obtained by producing the lags of the time series itself.[96] *Eirola and Lendasse* [2013] review the use of the regression application of the *Gaussian* mixture model for time series, including its application for forecasting and interpolating missing data.

When there is a given stationary time series $z = (z_1, z_2 \dots z_T)$ of length T, the following data matrix can be produced using the lags of the time series:

$$y = \begin{bmatrix} z_1 & z_2 & \dots & z_d \\ z_2 & z_3 & \dots & z_{d+1} \\ \vdots & \vdots & & \vdots \\ z_{n-d+1} & z_{n-d+2} & \dots & z_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-d+1} \end{bmatrix},$$

where one row of matrix y is a vector with d number of elements. In total, there is d number of variables together with the original time series, and because of the lags the total sample available becomes shorter; of length $(n-d+1)$. Having clarified this, the *Gaussian* mixture model can be written with normal density functions as:

$$f(y) = \prod_{i=1}^{n-d+1} \prod_{k=1}^{K} \tau_k \cdot \varphi(y_i|\mu_k, \Sigma_k),$$

where $\varphi(y_i|\mu_k, \Sigma_k)$ is the density function of the multivariate normal distribution, with mean vector μ_k and covariance matrix Σ_k for each component k, and mixture weights τ_k. From this point on estimation follows the procedure as described in the previous section, and the regression application is likewise valid.

Correction of the results may be necessitated by the time lags, as the variables are not independent from each other. *Eirola and Lendasse* [2013] suggest a solution that performs

[96] Regression applications of the *Gaussian* mixture model, for example, in the fields of speech recognition or signal processing, practically rely only on autoregressive variables, typically in constructions where for a given periodic sign t the other variables are created with calculating the lags of order 2, 4, 6… or 6, 12, 18…. Instead of the term autoregressive, the so-called *delay embedding* is much more widespread. See more on these in: *Shekofteh-Almasganj* [2013] and *Povinelli et al.* [2004].

an *ex-post* correction of parameters estimated by the EM algorithm. Their results suggest that correction is only needed when there are many estimated components.[97]

3.3. Mixture models and their applications on energy time series

As a preamble to the discussion of the empirical research results in this paper, some of the favourable properties of mixture models will be highlighted here, because they are exploited both in other studies and in the present dissertation. An evident field of application is the model-based representation of load curve features, where the underlying assumption is that the dispersion of load values is **multimodal**, and the load values are basically organised, concentrated (clustered) around these modes. Using the *Gaussian* mixture model a practical advantage lies in the interpretation of these. The other useful field of application is the examination of phenomena whose dispersion **cannot be described by classical distribution functions**. This may mean the presence of multiple modes or heteroscedasticity (time-dependent dispersion).

This section reviews the fields of application where mixture models have been applied in modelling energy, as the dissertation aims to contribute to such practical applications.

3.3.1. Construction of typical daily consumption curves

Due to the fact that recently more and more high resolution data are becoming available concerning individual consumers, the discovery of typical consumption patterns is an ever-present and extensively studied topic, and applications of the *Gaussian* mixture models have appeared in several studies.

The vast majority of the applications build basically on the clustering of daily consumption curves similarly to the previously reviewed methods. Some of them are relevant for the understanding of the empirical results of this dissertation; therefore, a few

[97] The reason for the necessity of correction is that the lagged variables – apart from being lagged – are identical with the original time series. Thus it is a requirement that the global mean for each variable should be equal, and the global covariance matrix should be symmetrical and Toeplitz-type (using the mixture weights for the calculations regarding the global parameters). All the decreasing diagonals of the latter are the same, that is:

$$\Sigma = \begin{bmatrix} r(0) & r(1) & \dots & r(d) \\ r(1) & r(0) & \dots & r(d-1) \\ \vdots & \vdots & & \vdots \\ r(d) & r(d-1) & \dots & r(0) \end{bmatrix},$$

where $r(k)$ denotes the autocovariance coefficient of time series z with lag k. The practical explanation of this is that, for example, the autocovariance value between the time series and its first lag, between its first and second lags, etc., should have the same value.

will be reviewed in more detail. Giving emphasis to the author's own empirical results and their novelty, some of the results will be reproduced here on the data analysed in the empirical parts of the thesis.

McKenna et al. [2014] explore typical consumption patterns on water consumption data obtained from smart metering. In this study, mixture model is used in a way that an (average) daily consumption curve is regarded as a density function, which is expressed as a mixture of normal density functions. Of course, it is required from a density function that its integral – that is, the area under the density function – equals 1 (100%). This way, the estimation of the mixture density function is capable of recording the daily shape. To arrive at the final result, it is necessary to rescale the values of the density function in a way that after rescaling, the area under the curve gives the daily total consumption. Using the mixture distribution parameters (mean, standard deviation and the ratio of mixture components) as in *McKenna et al.*'s [2014] study, the consumers themselves can be clustered; which is performed by the classical K-Means clustering here, and as a result, in a fairly disadvantageous way, because it does not make full use of the special (mixture distribution type) property of the parameters.

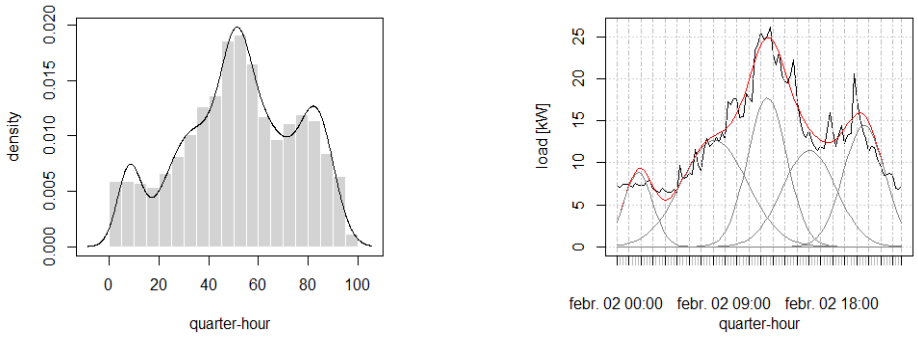
Another important statement of the study is that instead of hourly resolution time series it is possible to obtain much more stable results using half-hour interpolated data (using the so-called *Hermite* type interpolation polynomials[98]), which is definitely promising regarding high resolution metering data; though interpolation also means the smoothing of time series, which – as opposed to half-hourly measurements – might blur or smooth some features. On the whole, a disadvantage of the solution is that the mixture density function is fitted on daily averages and not data for each separate day. In the latter case, derivation of the typical daily curve by consumer would need to be solved (for example, by using some clustering technique to produce daily profiles for each consumer, then grouping these consumers using these daily profiles – in this case, however, it should be handled somehow if the number of daily profiles is different for each consumer, which suggests that the modelling on daily averages used by the authors is an elegant avoidance of this problem).

Hino et al. [2013] also performs clustering of daily consumption curves in both time series (the grouping of a consumers' daily consumption curves) and cross-sectional

[98] On *Hermite* type interpolation see more at: https://en.wikipedia.org/wiki/Hermite_interpolation.

(grouping many consumers' consumption curves on the same day) dimensions. In this study, a daily consumption curve is regarded as a density function produced by the mixture of normal distribution density functions.

Figure 32: Fitting the mixture density function on daily load curves (load curve C109)

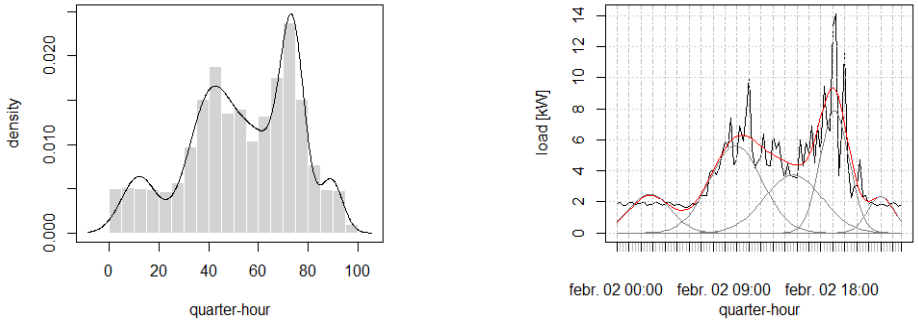


a) Histogram created on the basis of quarter-hourly load values and mixture density function.

b) Rescaled mixture density function, normal distribution components and the quarter-hourly load values.

Source: author's own results (R) and figure (R).

Figure 33: Fitting the mixture density function on daily load curves (load curve C148)



a) Histogram created on the basis of quarter-hourly load values and mixture density function.

b) Rescaled mixture density function, normal distribution components and the quarter-hourly load values.

Source: author's own results (R) and figure (R).

The clustering of days, however, has a much more sophisticated methodology here: by applying the so-called symmetrical generalised Kullback-Leibler (divergence) distance measure (which is basically used to measure the differences between distributions,

developed further by the authors) in the framework of hierarchical clustering.[99] The other great advantage of the study is that the metering results are on a per second basis, which makes fitting the density functions highly plausible (as a day consists of $24 \cdot 60 \cdot 60 = 86400$ seconds). The following figures show the mixture density functions fitted on daily curves for some curves that are used in this paper as well.

On the whole, this solution is a really good instrument for estimating where the **daily *peak*s** are, and on the long run, how they shift or change. This is not only interesting on the level of consumers, but also on an aggregate (even system) level of consumption. Examination of how spikes shift is one of the important research questions of studies focused on long term consumption saving or consumption shifting. It is a more **strategically oriented** field of research which brings us further than the scope of this dissertation.

3.3.2. Modelling the distribution of consumption using mixture density function

The basic idea in the studies by *McKenna et al.* [2014] and *Hino et al.* [2013] is that the mixture density function describes daily consumption characteristics, and does so in a way that if there is more energy consumed in a given (unit of) time, then the value of the density function is higher, while at other times it is smaller (that is, the *likelihood* of the consumption of a unit of energy is higher in some periods than in others). *Singh et al.* [2010] model the distribution of half-hourly consumption data using the estimates of mixture density functions, that is, it is not the modelling of when the consumption of a unit of energy is most likely, but the modelling of which consumption levels are the ones where half-hourly consumptions typically (are most likely to) pool, cluster.[100]

The following figures show histograms from the compiled data of some quarter-hourly load curves, and the fitted mixture density functions including the normal distribution components. For C25, the lower weekend and higher weekday consumption levels are typical, but on the whole, five components were estimated. It is similar to what can be seen regarding C109, where the two components with higher average consumption

[99] It has a relatively simple, closed formula for normal distribution, in case of mixture of normal distributions the value can be approximated (see: *Hershey-Olsen* [2007]); see also: the chapter presenting the author's own empirical research results.

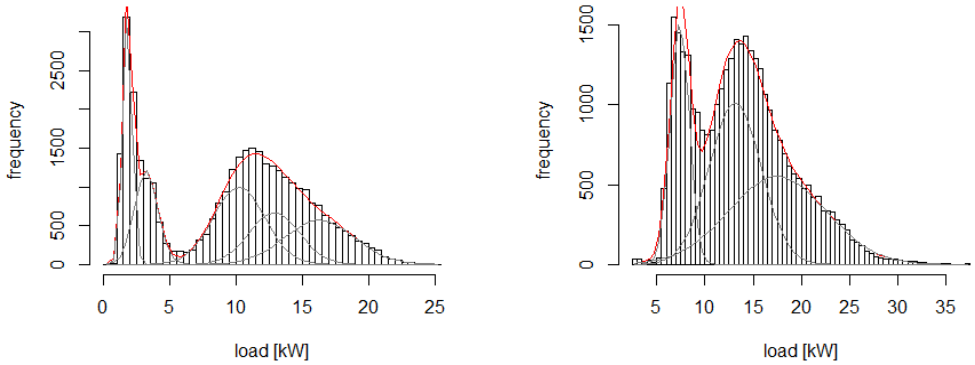
[100] In addition, in *Singh et al.*'s [2010] study there is an example using the mixture model for the joint distribution of two time series, which may provide a promising basis for examining the portfolio effect in the future.

probably indicate summer (weather-dependent) and non-summer (not weather-dependent) consumption levels.

Indeed, the multivariate extension of the analyses seems to be much more forward-looking (e.g.: taking into consideration the effect of temperature). There will be an example of this in this section, and in fact, the dissertation also follows a similar path. The need for a multivariate extension also arises in *Hino et al.*'s [2013] study (see the previous section), but there the inclusion of this extra information can be conceived in the framework of a multi-step solution.

None of the above studies have examined consumption together with the factors that influence it (such as weather or the effect of the time of day) or applied these in mixture modelling. The rest of this section shows just such a publicly available example.

Figure 34: Fitting the mixture density function on the empirical distribution of load values



a) Curve C25 b) Curve C109

Source: author's own results (R) and figure (R).

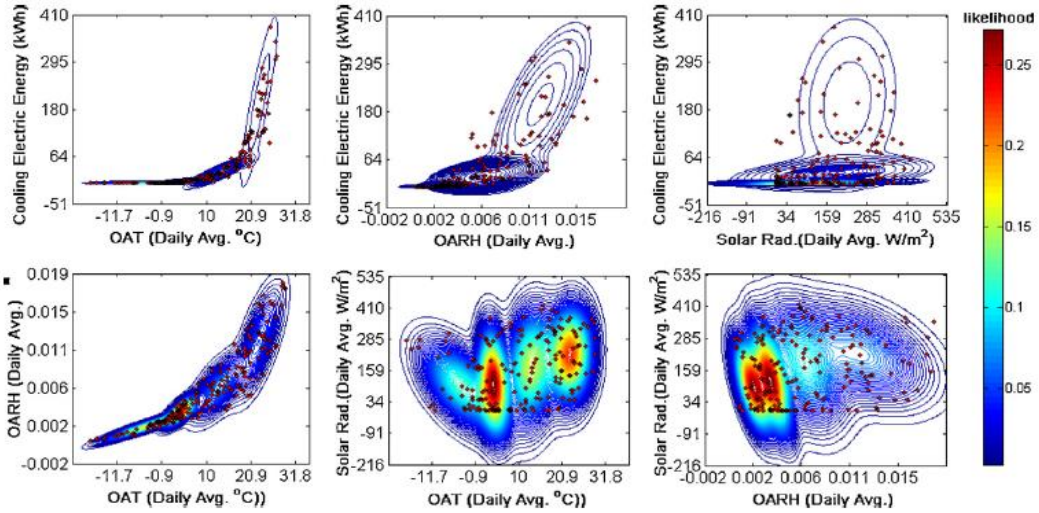
3.3.3. Modelling the distribution of consumption using mixture density function and regression

Relatively few examples exist for the application of mixture regression, but *Srivastav et al.*'s [2013] study is one of them. The study shows how cooling energy use can be modelled in the framework of a mixture model using the variables of temperature, humidity and solar radiation for both daily and hourly resolution time series.

In the article, the advantage of using the mixture model is not primarily in that the modes of the components can be regarded as typical consumptions, but that through using mixture regression it enables the calculation of such, localised confidence intervals that can be written as a function of the given independent variables. As a possible field of application the authors mark opportunities regarding electricity savings pricing, because if

the consumption risk varies seasonally, the risk of savings that can be realised from consumption also varies.

Figure 35: Estimation of the mixture model and visual representation of the fitting per variable pairs



Source: Srivastav et al. [2013]

Returning to the modes of the components, the drawback (or not necessarily an advantage) of this solution is that the typical value of a component (or cluster) – the mean vector of multidimensional normal distribution – does not give a typical daily profile, which is otherwise common in literature and applications concerning typical consumption patterns. The advantage, though, is that it is possible to directly derive confidence intervals (there are examples of indirect solutions, but based on previous experience, these are not the best, see more in this in Section 4.2 about *Subbarao et al.*'s [2011] kNN- method); in addition, it is relatively fast (compared e.g. to the *Gaussian process*-based regression).

Technically, because of the assumption of multidimensional normality of the variables the confidence interval for the dependent variable of the regression can also be derived from the mixture of normal distributions; there is no closed formula for that. The above mentioned study is thus related to such applications of mixture models where the purpose of modelling is to **relax** the assumption that the **error term** of the regression model **is an independent, identically distributed (normal)** random variable. In actual practice in energy consumption time series the time-dependent dispersion, heteroscedasticity, is the most obvious example that – together with the lack of normality (the latter is even less likely to be fulfilled in the case of power price time series) – the above mentioned IID feature does not apply.

In the authors' study, when discussing the results it is difficult to see, but the underlying assumption of using multivariate mixture models is exactly that clustering observations in the given variables' dimensions is possible. **Cluster** (or group) observations are characterised by **different covariance structures** and dispersion, whose results can be captured in the **error terms** in the course of regression fitting.

4. CONSIDERING UNCERTAINTY OF CONSUMPTION IN PROFILING – EMPIRICAL RESEARCH RESULTS

In practice it is increasingly becoming a requirement from profiling that it should properly describe not only the expected consumption characteristics, but also the uncertainty of consumption. This was kept in mind while choosing the methods used in the empirical research.

When examining the stylized facts of consumption time series, it seemed to be obvious that uncertainty is to some extent related to variables to which the characteristics of time series themselves are related (such as temperature, multiple seasonality, etc.).

This chapter shows a solution for profiling that suits the above requirements, as it deals with both the expected value of consumption and its dispersion. The chapter also shows how this solution allows the extraction of additional information following from the model structure used. The results are presented in comparison with models that may be regarded classical, where the consideration of uncertainty is not originally an aim in itself, or has a subsidiary role and is thus difficult to derive.

4.1. Creating typical consumption patterns

It has already been mentioned that literature on profiling – independently of whether it focuses on a single consumption curve in itself, or the grouping of similar curves – aims at deriving typical daily (consumption) profiles as a final output (this may be performed by using some clustering technique based on daily discretisation, or some regression methodology based on removing the effects of different variables). An indisputable advantage of this is its simplicity and easy interpretation. Emphasising the results of this particular paper, the *benchmark* will be a similar solution using regression logic.

The output of the mixture model is not directly comparable to the output of classical profiling solutions, and for this reason, the comparison can only be formulated in a qualitative sense, highlighting the advantages and disadvantages of the different methods. At the same time, this is an explanation of the title of this section (the term ‘typical consumption patterns’), because independently of the methodology **typical consumption pattern** may be some daily profile, but also the parameter set of the mixture model to be shown here (the composite of *prior* or mixture probabilities, means and covariance

matrices).[101] **Curve feature** (as extracted information that applies to the curve; also used in Section 2.1.3) is basically the statistical, methodological counterpart of typical consumption pattern, which is often – in a slightly misleading way – called profile. This may be due to the fact that the curve feature often occurs in the form of a curve (after the removal of various effects) called typical daily profile curve.

4.1.1. Using the mixture model to create typical consumption patterns

The mixture model used in profiling is shown in detail on a portfolio curve, then highlighting the most important results, it is also described for individual curves. These are preceded by the description of the methodological steps used in the calculations.

4.1.1.1. Description of the applied methodology

In the mixture model the observations to be clustered are quarter-hours, which means the treatment of 35 040 observations for a yearly curve (365 days * 96 quarter-hours/day). A drawback of using mixture models is that for a large number of observations, they work slower (because of the longer running time of the optimisation of the objective function). It is not very practical to use the hierarchical model-based clustering for the calculation of initial cluster memberships in this case either (as hierarchical clustering procedures are not used by large samples in general). For this purpose, the integrated use of mixture clustering and mixture discriminant analysis is accomplished in this paper (EDDA method, see for example: *Banfield-Raftery* [1993]). The main steps can be summarised as follows:

1. A random sample was selected from the whole 2011 year, which means 35 040 / 8 = 4 380 quarter-hours.
2. Clustering was performed on its 20%, that is, on 4 380 / 5 = 876 quarter-hours, using unrestricted covariance matrix. [102,103]
3. Building on clustering, the EDDA method was performed as described below:
   a. The probabilistic classification was discretised based on the *posterior* probabilities calculated in the 2nd step. This means that each quarter-hour was assigned to the cluster where the *posterior* probability of the component

[101] In the title of the section the term 'typical consumption patterns' appears instead of profiling, because in the majority of studies related to profiling profile appears as some typical daily curve. In the solution used in this paper the typical, characteristic values are derived directly from the results and the parameters of components – as we shall see at the end of the section – can likewise be interpreted well.
[102] This is what is typically used in the reviewed literature, see Chapter 3.
[103] The choice of both the eighth part of the year and the 20% was made via a subjective but empirical way, after having tried out many different options.

is maximal and this maximal *posterior* probability was equalled to 1 (100%), and the others were equalled to 0. The discriminant analysis was carried out by this discretised classification and by using independent variables, which means performing a single M-step and a single E-step.

b. As results of the M-step, the estimated parameters were produced (*prior*- or mixture probabilities, means and covariance matrices),

c. and in the E-step the *posterior* probabilities of cluster memberships were calculated using the estimated parameters. The M-step only uses the quarter-hours used in clustering, but the E-step uses the whole random sample (4 380 quarter-hours). Of course, even the total number (that is 35 040) quarter-hours can be used in the E-step.[104]

The main advantage of the solution is its speed and transparency, but simultaneously, it uses the favourable property of model-based clustering that it aims basically at identifying structure; therefore, the existence of only a small sample is not a problem. This is especially useful from the aspect that for each curve it is only one year's data that are available, hence there are very few extremely cold or hot days where the temperature-load relationship may be different from typical (that is, not so extreme winter or summer days). Another advantage of random sampling is that this way the quarter-hours in the analysis may be regarded independent from each other. In the estimation of the mixture model, independence is assumed, which is hardly ever true in practical applications – but with a random sample the task fits the model assumptions better.

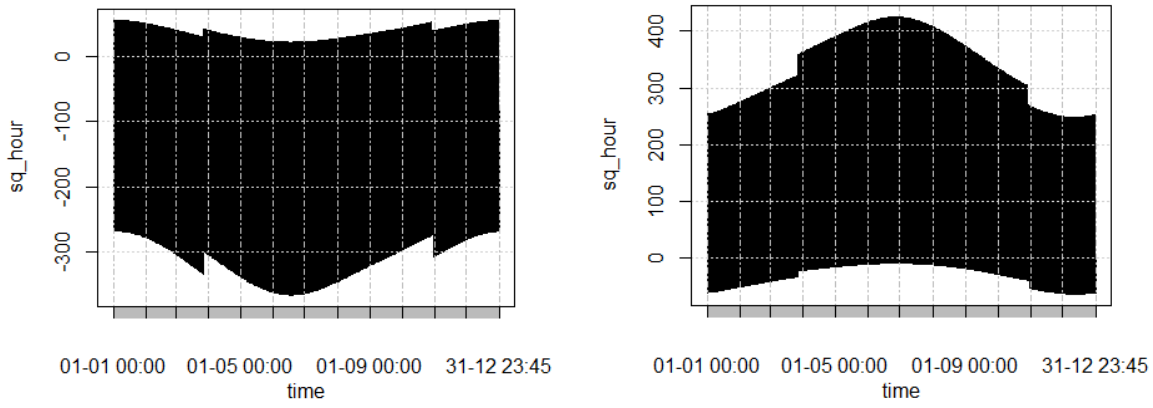
Besides the quarter-hourly time series, the following variables were used in the clustering of the quarter-hours:

1. daily average temperature data,
2. signed, squared deviations in minutes for the given quarter-hour from the time of sunrise and sunset within the day,
3. one period (that is, one quarter-hour), and one day (that is, 96 quarter-hours) lags of the time series.

[104] If not all quarter-hours are used, then the unused quarter-hours can be used for out-of-sample assessments (see the section on regression applications).

Daily average temperature is important for handling the winter heating and summer cooling effects, and at the same time, it may be regarded as a *proxy* variable for the handling of (yearly) seasonality. One day lag can be used for the treatment of weekly seasonality (thinking about the one-day (that is, 96 quarter-hours) distances between Fridays-Saturdays or Sundays-Mondays). The point of using squared distance from the time of sunrise and sunset is that they capture daily seasonality (temporal behaviour) and will support grouping in this dimension. After squaring, it is important to maintain the value as signed because of the position within the day. The yearly variation of the latter variables nicely echoes the tendency of the length of days. It is primarily not the specific values of the variables, but much rather the 'shape' that may be reminiscent of the yearly variation of the sunset-effect already seen in the *contour plots* in Section 1.4.[105] Thanks to squaring, the sunrise variable 'expands' the end of the day, while the sunset variable 'expands' the beginning of the day, which will be useful in clustering (see later).

Figure 36: Variation of the signed, squared differences of quarter-hours from the time of sunrise and sunset



Source: author's own figures (R).

The components produced while applying the mixture model do not necessarily need to form separate groups. It often happens that similar components close to each other are regarded one group. As each component is represented with a normal density function the distances between the components are practically measured with the so-called Kullback-Leibler divergence (or distance) measure, which serves to compute the distances between probability distributions. If $p_i(x)$ and $p_j(x)$ are two distribution functions, they are defined as (see for example: *Cover-Thomas* [1991]):

[105] A similar construction will appear in Section 4.1.2 at the description of classical solutions.

$$D(p_i, p_j) = \int_{-\infty}^{\infty} p_i(x) log \frac{p_i(x)}{p_j(x)} dx.$$

If the two density functions are normal distributions with means μ_i and μ_j, and with standard deviations σ_i and σ_j , the formula is simplified to the following closed formula:

$$D(p_i, p_j) = \frac{1}{2}\left[log\left(\frac{\sigma_j^2}{\sigma_i^2}\right) + \frac{\sigma_i^2}{\sigma_j^2} - 1 + \frac{(\mu_i - \mu_j)^2}{\sigma_j^2}\right].$$

The Kullback-Leibler divergence – being an entropy measure[106] – is not symmetrical[107], and for this reason usually both $D(p_i, p_j)$ and $D(p_j, p_j)$ values are computed, and their average is used as a measure of distance, that is:

$$D_{KL}(p_i, p_j) = \frac{D(p_i, p_j) + D(p_j, p_j)}{2},$$

where $D_{KL}(p_i, p_j)$ denotes the values to be used in the subsequent part of this chapter.

4.1.1.2. Results of the portfolio load curve

Results using the previously described variables are shown on *scatter plots* for the portfolio in Figure 37. The different colour groups of dots denote different components. In the centres of the ellipses are the means of the components, and the ellipses draw the line corresponding to the mean $\pm$ standard deviation regarding the pairs of variables – in the same way as it was seen in Section 3.2.2 on the example of the Budapest daily natural gas consumption.

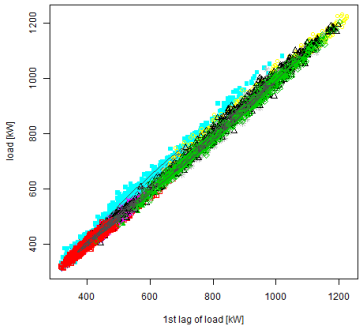
Figure 37. b) shows how, for example, Mondays and Sundays are separated, which are usually different in terms of load characteristic (they are almost parallel with the y and x axes, successively). Figures c) and d) show the clustering result as a function of the times of day. In these, the red and blue 'groups of dots' in the bottom right corner denote the early morning hours. While in Figure c) they are not so nicely separated along those two dimensions, they do in Figure d). This is why it is useful to include the variables that indicate difference from both the sunrise and sunset in clustering. Compared with Figure e) not only the markedly better separation is visible, but also that one denotes the summer, the other the winter morning quarter-hours. It nicely outlines the temperature-dependency of

[106] It is also common to use the term relative entropy for the same notion, see for example: http://mathworld.wolfram.com/RelativeEntropy.html.

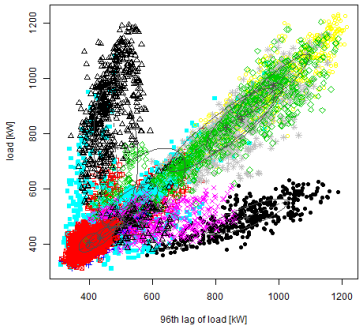
[107] In connection with similarity or distance measures the requirement that they should be symmetrical is usually formulated, that is, the similarity or distance should not depend on the order (that is, the distance of A from B should be the same as the distance of B from A). In the case of more complex measures this is often not fulfilled, in such cases it is common that similarity and distance are calculated in both directions and their averages are used.

the load as well, especially in the daytime hours. The positions of the ellipses were possible to foresee based on the *scatter plots* in Section 1.4, but there the focuses were only on temperature and load variables.

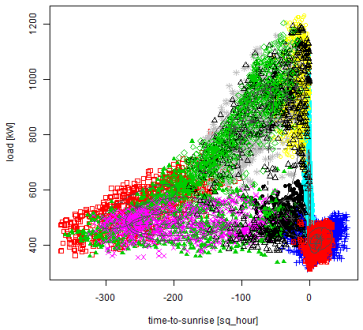
Figure 37: Estimation of the mixture model and visual representation of the goodness of fit for pairs of variables for the portfolio



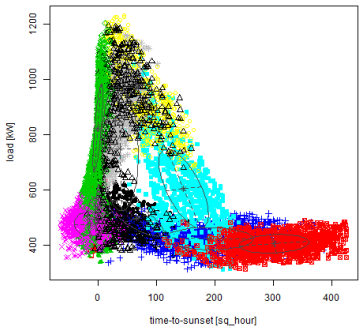
a) Load and the 1st lag of the load



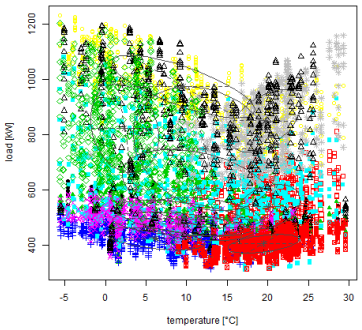
b) Load and the 96th lag of the load



c) Load and the signed, squared deviation from the time of sunrise



d) Load and the signed, squared deviation from the time of sunset



e) Load and temperature

Source: author's own calculations (R) and figures (R).

Although it may be projected from the *scatter plots*, it can also be examined which quarter-hours belong to which component. The 2nd-3rd-8th components contain mainly morning, while 4th-9th-10th-11th components contain primarily evening and the other

components contain *peak* period hours. In the same way, it can be examined which components have a larger or smaller role in the description of weekdays and weekends. Figure 38 shows the distribution of hours, months, weekdays and weekends among the components.[108]

Highlighting just a few more components, the 6th component marks the *peak* period quarter-hours of the summer weekdays, when due to the use of air conditioners the cooling effect is prevalent. The 9th and 10th components mark evening (partly setback) period quarter-hours of winter and summer months. An interesting – but absolutely logical – result is that the 3rd component which groups morning ramp periods has a much smaller ratio in weekend quarter-hours than during weekdays. Of course, the distribution of the components is not so 'clear', as a quarter-hour was assigned to a component where the *posterior* probability was highest; however, most quarter-hours can be the realisation of more components.

Closely related to this, Figure 38 shows the dendrogram made on the basis of the Kullback-Leibler divergence. The merging was performed using Ward's agglomeration logic (see later).

Based on the two dimensional figures and the distributions within the components practically every component can be regarded as a single component. The dendrogram, however, may be used to identify which components are similar, or in the case of merging of overlapping components it can also be used for the ease of interpretation.[109]

If we wish to group components, three main groups can be identified (these could be predicted on the basis of Figure 38):

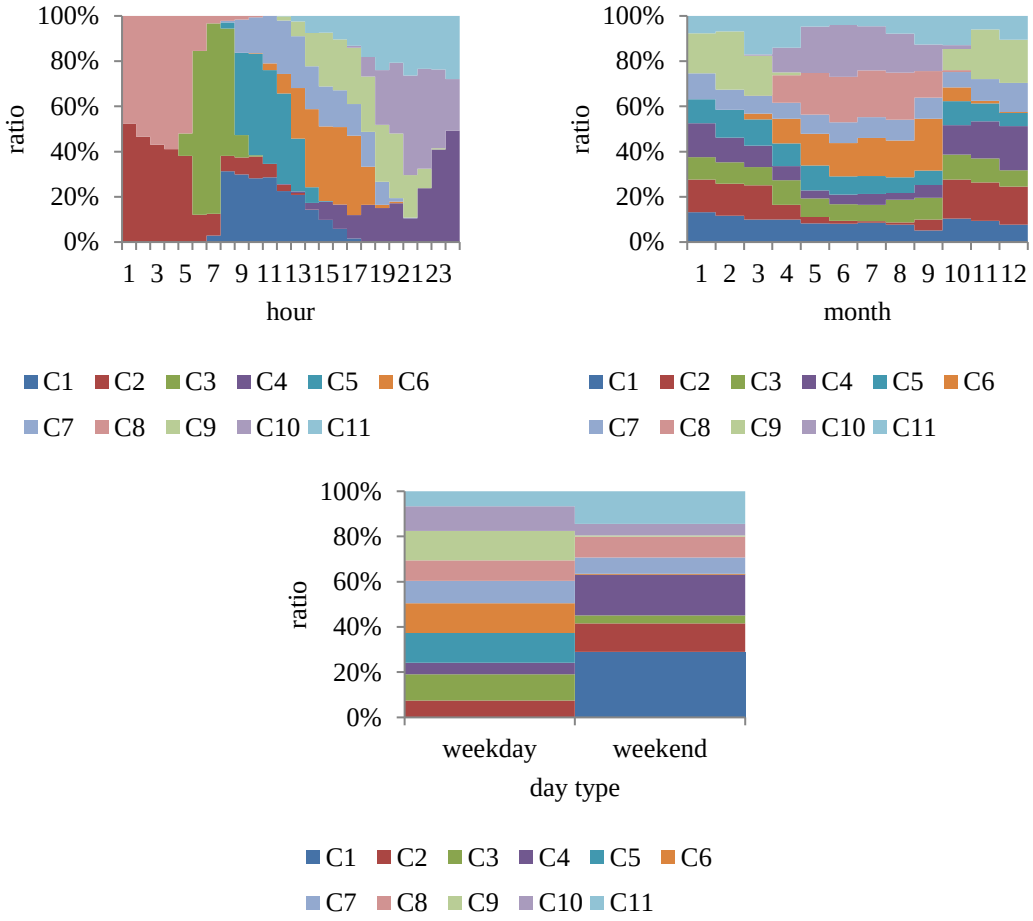
- 1st group: 2nd, 3rd and 8th components (morning quarter-hours),
- 2nd group: 4th, 9th, 10th and 11th components (evening quarter-hours),
- 3rd group: 1st, 5th, 6th and 7th components (*peak* period quarter-hours).

Seeing the merging steps it seems that the time of day is one of the primary discriminative factors. The time of day is much more of a determining factor in clustering

[108] Of course, besides the ones that appear here, it is possible to make *scatter plots* for the other pairs of variables. In the figures here one of the pairs of variables is always load, thinking about regression, the dependent variable in the regression approach based on the mixture model.
[109] A similar, likewise entropy-based logic for merging components can be found in *Baudry et al.*'s [2010] study.

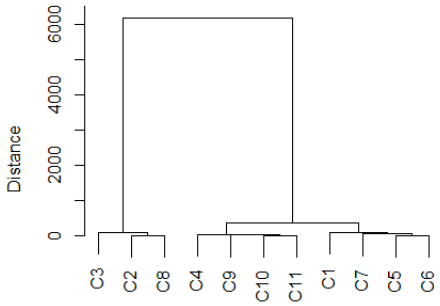
than for example, temperature, which has a smaller effect on the distribution of the load values of each day (see for example the *boxplots* in Section 1.4).

Figure 38: The composition of components in the portfolio[110]



Source: author's own calculations (Excel) and figure (Excel).

Figure 39: Dendrogram based on the distances between components in the portfolio



Source: author's own calculations (R) and figure (R).

[110] C1, C2 … C11 denote the 1st, 2nd … 11th components. The notation will be similar in the next section as well.

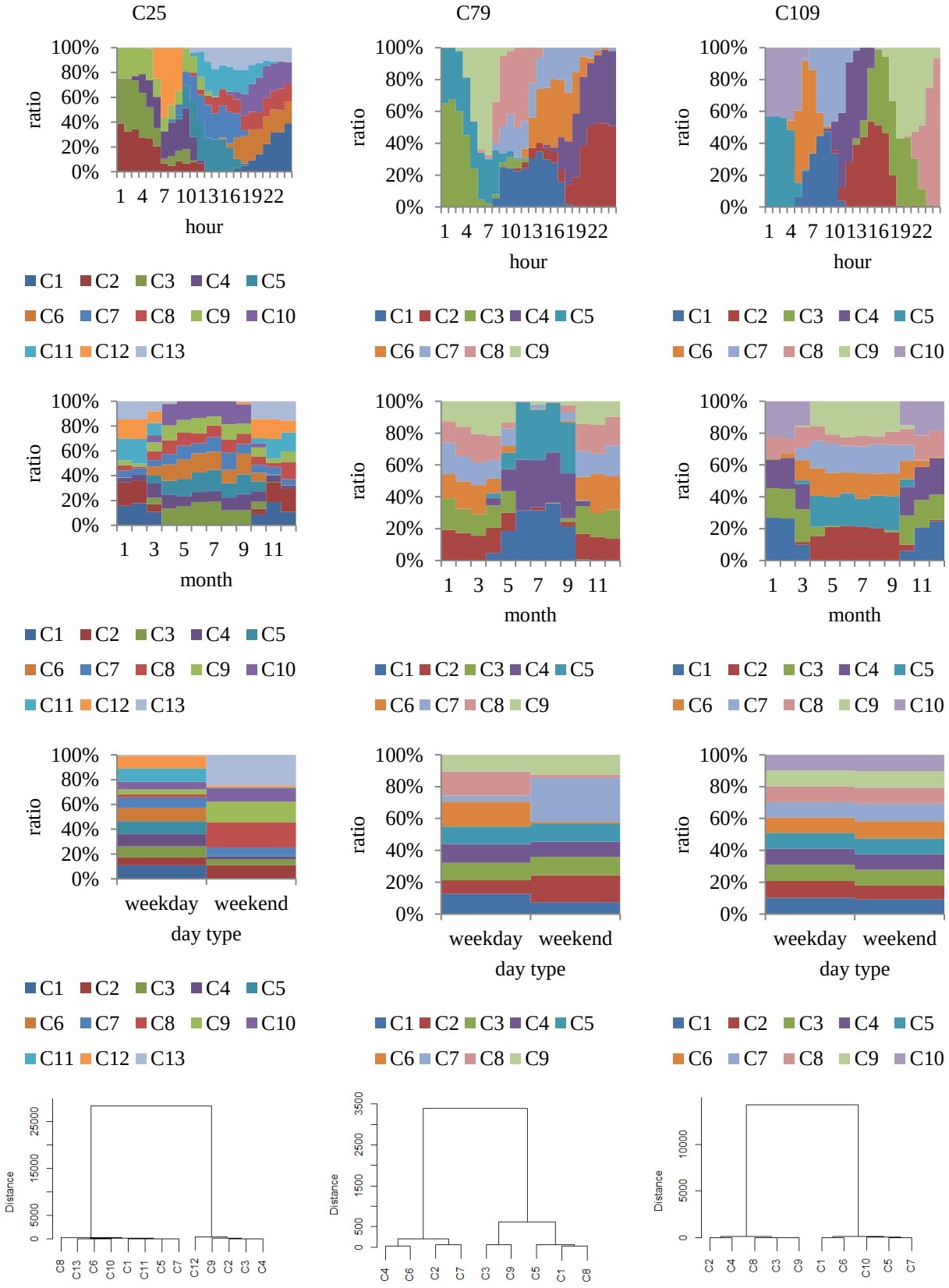
4.1.1.3. Results of individual load curves

In connection with individual curves similar conclusions can be drawn as about the portfolio. The next figure shows the dendrogram based on the Kullback-Leibler divergence of the components (familiar from the previous section), and the composition of components is shown in some individual curves.

The ratio of hours within the components compared with the dendrogram show that for all three curves it is mostly 'time of day' that determines grouping. Some results worth mentioning are highlighted here.

In C25, for example, the components that record the Saturday setback after the Friday night and the Monday morning quick ramp are the 9th and 4th components successively. As their variability does not really change during the year, clearly, these components have approximately the same ratio in every month. Quite spectacularly, in the case of C79 three components (the 1st, 4th and 5th) describe the variation of the summer load, which is primarily explained by the fact that the dispersion of the curve in this period greatly exceeds that of other months' (see Section 4.2 as well). As in C109 there is only a minor difference between weekdays and weekends, it is not surprising that the quarter-hours in weekdays and weekends are distributed among the components uniformly (that is, the 'weight' of each component in the description of weekdays and weekends is approximately the same).

Figure 40: The composition of components and dendrograms based on the distances between components in individual curves



Source: author's own calculations (R and Excel) and table (R and Excel).

4.1.2. Using classical time series regression to create typical consumption patterns

Using classical time series regression in profiling is based on the periodic autoregressive (PAR) model described in the chapter on methodology.

4.1.2.1. Description of the methods used

The regression solution aiming at the formation of profiles is based on the notion that thanks to regression logic the effect of certain variables can be removed from the time series. In short, if the value of all exogenous variables is equalled to zero, the so-called typical daily profile (TDP) is identical with the quarter-hourly long term mean (see Section 3.1.5); meanwhile, the given consumption may differ from this due to exogenous variables and the realisations of the error term.

Table 8 below contains the exogenous variables used in regression.

Table 8: Independent variables used in regression and their short description (PAR model)

dummy variables denoting quarter-hours	the value of the *dummy* variable is 1, if it is the given quarter-hour of the day, otherwise it is 0
dummy variable denoting weekend days	the value of the *dummy* variable is 1, if it is the quarter-hour of a weekend day, otherwise it is 0
interaction variables denoting quarter-hours on weekends	variables that are constructed as interaction (that is, as product) of *dummy* variables denoting quarter-hours and *dummy* variables denoting weekend days
variables denoting holidays and other special days (official non-working days and transferred days)[111]	the value of the given *dummy* variable is 1, if it is the quarter-hour of a given holiday, official non-working day or transferred working day, otherwise it is 0
the so-called sunset-effect	the signed deviation of the sunset time from 18:00, see Figure 41 below
heating degree-day (HDD)	the downward deviation of temperature from $12°C$ to capture the heating effect
cooling degree-day (CDD)	the upward deviation of temperature from $21°C$ to capture the cooling effect

Source: author's own table.

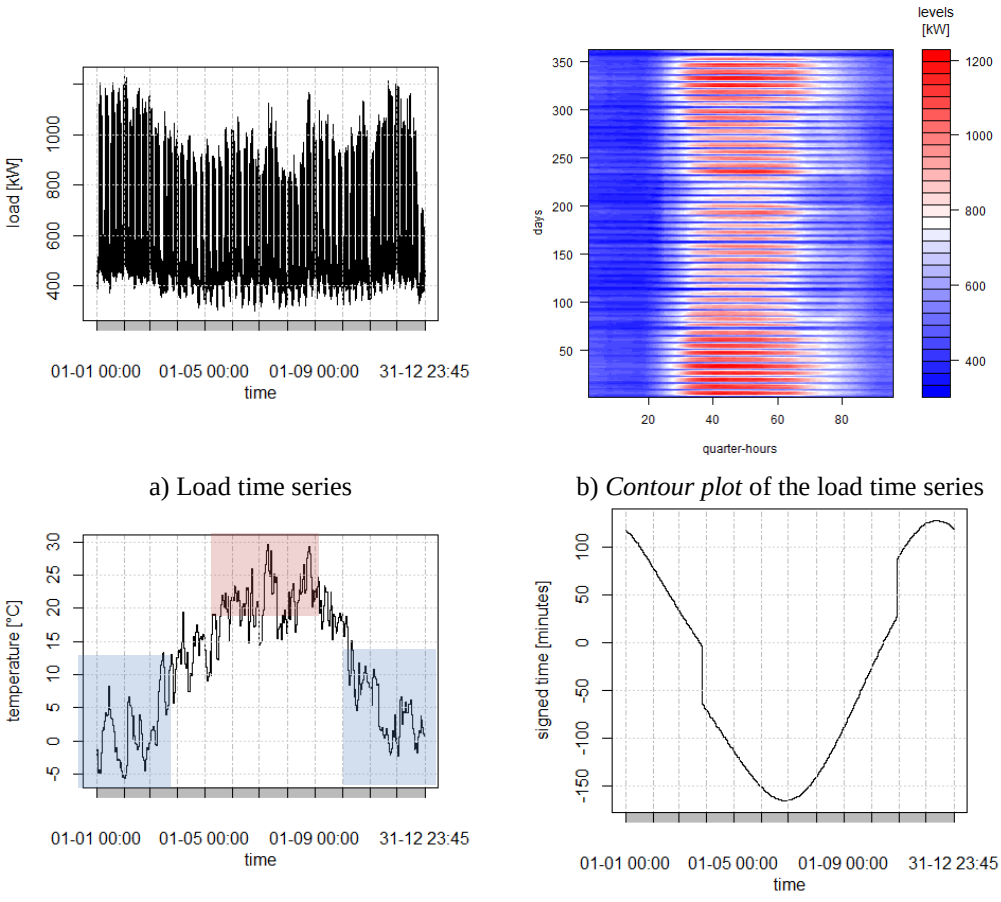
The bottom row of Figure 41 shows the temperature and sunset-effect variables also used in the regression model. To support the inclusion of these variables in the model, the upper rows of the figure show the load curves of the portfolio itself and its *contour plot* shown previously.

[111] These are the following (referring to the whole 2011 year):
- holidays: 1st January, 15th March, 24-25th April, 1st May, 12-13rd June, 20th August, 23rd October, 1st November, 24-26th December
- official non-working days: 14th March, 31st October
- transferred working days: 19th March, 5th November.

In addition, transferred working days include days that are between two winter holidays (from 27th to 31st December), because while they are weekdays, in most cases (if possible) these days are holiday periods, as in transferred working days.

Including the sunset-effect in the way it is shown in the table means applying the technique also used by the distribution systems operator in Great Britain (*Elexon* [2013]). It is very similar to using the variable with the content of the length of day (the period between sunrise and sunset in hours, see *Sugár* [2011]). The technique used here is a bit more realistic given that the shifts resulting from changing our clocks also appear in the variable, as it could be noticed in the *contour plots* in the tendencies of the time series. The fluctuation of the sunset variable used here is almost a one-to-one representation of the shape in the *contour plot* after its rotation by 90 degrees.[112]

Figure 41: The effect of temperature and sunset on the load in the portfolio



a) Load time series

b) *Contour plot* of the load time series

c) Temperature variable

d) Sunset-effect variable

Source: author's own calculations (R) and figures (R).

The use of *heating* and *cooling degree-day*s with the threshold values of 12 and 21 $°C$ may not be the most optimal solution resulting in the best fit in all the time series. However, using these will suffice here; the techniques providing the best fit (that is,

[112] This latter effect can be seen much better on the Hungarian system load time series.

producing the maximal coefficient of determination, R^2) do not have a significant effect on the results. According to *Sugár* [2011] these two threshold values provide the best fit if applied to the Hungarian system load.[113] In the figure, the areas covered by the blue and read squares are aimed at illustrating, highlighting this seasonally different effect of the temperature.

Some additional remarks need to be added primarily to the construction of *dummy* variables. Usually the regression model is written in a way that assuming one intercept term, the estimated coefficients of the *dummy* variables shift this intercept (if the values of the variable equal 1). For example, given the intercept term, 95 *dummy* variables can be defined to estimate the effects of quarter-hours, with a chosen quarter-hour as a reference category. A technique equivalent to this without an intercept is the inclusion of 96 *dummy* variables, but thinking about writing the periodic autoregressive model as a model including 96 equations, estimating 96 intercept terms the same result can be arrived at.

Briefly summarising the essence of all this, during the estimation of typical daily profiles, all of the effects of the – exogenous – variables that appear in the table are removed and the long term equilibrium of the 96 quarter-hours are computed on the basis of 96 intercept terms. These typical daily profiles will appear as compressed information (in other words: curve feature) in what follows.

4.1.2.2. Results on the portfolio and individual consumer load curves

This section only deals with results derived from the PAR model, that is, typical daily profiles (TDPs).

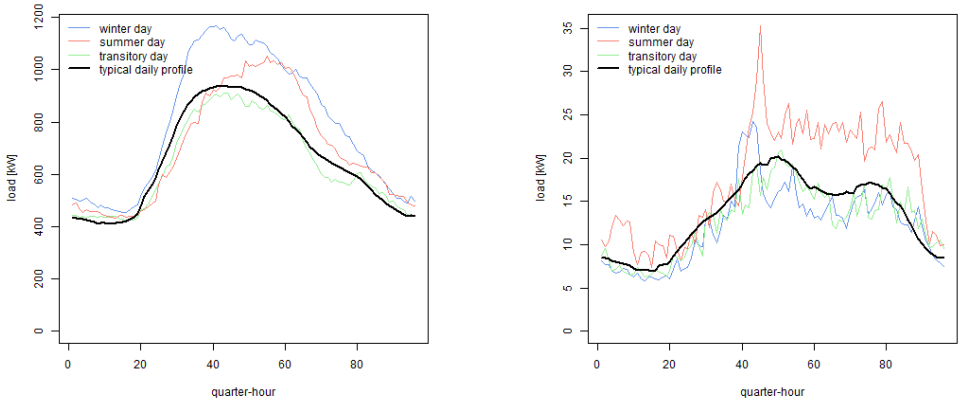
In Figure 42, what can be seen are the variations of daily load curves in a winter, summer and transitory period and the typical daily profiles (TDPs) for the portfolio and the C109 individual curve (referring to the other curves see the first figure of Appendix E)).

It can be seen very clearly that the effects of exogenous variables have been removed, thus the shape of the curves resembles the shape of the (non-temperature-dependent) transition period the most. Besides, of course, the typical daily profile is basically a smooth, noise-free curve with the removal of realisations of the error term. It is worth noting that the winter and summer temperature effects do not modify the shape of the curve in the

[113] Although it is common to use the threshold level of $16°C$ for *heating degree-day* in natural gas consumption, the value concerning electricity is lower. This can be explained by the phenomenon that the consumption of heating systems can be notable when it is much colder.

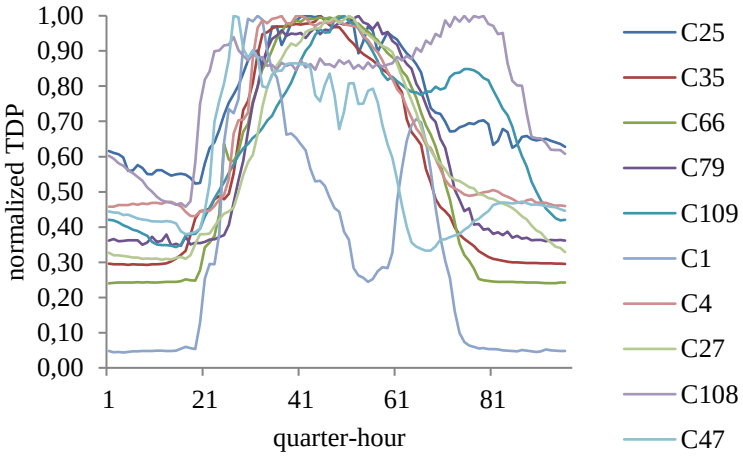
same way within a day: for the winter effect it is more likely that it occurs in the first part of the day, while in the summer period it appears in the middle of the day, in the afternoon. (It is not so evident on the C109 curve, where there is only summer temperature-dependence.)

Figure 42: The variation of daily load curves and typical daily profiles (TDPs) in the portfolio and the C109 individual curve



Source: author's own calculations (R) and figure (R).

Figure 43: Normalised typical daily profiles (TDPs) in individual curves



Source: author's own calculations (R) and figures (R).

Figure 43 shows the normalised[114] typical daily profiles calculated for different curves. It is apparent that this classical method also captures a number of features (e.g. *peak–off-peak* ratios, location of *peak* period(s) within the day). However, it is also visible

[114] Normalising supports the visualisation here. It is used in profiling so that the level of the curve does not influence the results. Normalising means using the following formula: TDP_t / TDP_{max}, where TDP_t is the quarter-hour value at time t of TDP, and TDP_{max} is the maximum value of TDP.

that not only noise, but everything that is a likely realisation of profile related risk is also removed. This, regarding its content, is not so advantageous, but the smoothness of the curve, their noise (and basically risk)-free nature are all useful in clustering with TDPs using classical clustering methods.

4.1.3. Creating profile groups

The main focus of this dissertation is rather the exploration of individual consumer behaviour; therefore, this section is meant as a supplement to the previous research results, and also as providing a foundation for future research efforts. Nevertheless, its relationship with the previous section is strong. This way, the quality of the information compression realised for each individual curve can be monitored better. Methodologically, it manifests itself in the use of the Kullback-Leibler divergence – from the aspect of mixture models.

The relationships between the curves that have different consumption patterns may be studied in two main dimensions:

1. on the one hand, based on typical consumption patterns the consumers with similar load features can be grouped into homogenous (profile) groups,
2. on the other hand, for each individual curve, it is possible to rank them according to a(n abstract) measure (the given value of distance) how much the profile of a given curve is different from the profile of the whole portfolio.

Formation of profile groups is, of course, performed in both classical regression approach and the mixture-model.

4.1.3.1. Description of the methods used

In mixture models distance is measured using the Kullback-Leibler divergence that is valid for *Gaussian* mixture distributions. Parameters of the mixture models are regarded as curve features that describe the curve well; therefore, this measure is capable of measuring their distances and differences.

Let us assume that $\tilde{p}_i(x, \theta_i)$ and $\tilde{p}_j(x, \theta_j)$ (or in a shorter form $\tilde{p}_i(x)$ and $\tilde{p}_j(x)$) are the two *Gaussian* mixture distribution functions. Then the formula of the distance measure can be approximated (based on *Hershey-Olsen* [2007]) as follows:

$$D(\tilde{p}_i, \tilde{p}_j) = \int_{-\infty}^{\infty} \tilde{p}_i(x) log \frac{\tilde{p}_i(x)}{\tilde{p}_j(x)} dx =$$

$$= \sum_{m=1}^{M} \pi_m \cdot log \frac{\sum_{m*=1}^{M} \pi_{m*} \cdot exp(-D(\Phi_m, \Phi_{m*}))}{\sum_{m**=1}^{M**} \omega_{m**} \cdot exp(-D(\Phi_m, \Phi_{m**}))},$$

where:

- $\{\pi_{m*}\}_{m*=1}^{M*}$ and $\{\omega_{m**}\}_{m**=1}^{M**}$ denote the probabilities of the *Gaussian* mixture-distributions, where the number of components is M^* and M^{**},
- Φ denotes the appropriate normal density function,
- $\theta_i = \{\pi_{m*}, \mu_{m*}, \sigma_{m*}\}_{m*=1}^{M*}$ and $\theta_j = \{\pi_{m**}, \mu_{m**}, \sigma_{m**}\}_{m**=1}^{M**}$ denote the parameters of the normal distribution components.

Of course, the distance measure is not symmetrical here either, and for this reason what is used is the mean of the $D(\tilde{p}_i, \tilde{p}_j)$ and $D(\tilde{p}_j, \tilde{p}_i)$ values:

$$D_{KL}(\tilde{p}_i, \tilde{p}_j) = \frac{D(\tilde{p}_i, \tilde{p}_j) + D(\tilde{p}_j, \tilde{p}_j)}{2},$$

where $D_{KL}(\tilde{p}_i, \tilde{p}_j)$ denotes the distance measure used to compare the curves in the rest of this section.

The advantage of the distance measure is that it uses the assumption that **each component is described by the multidimensional normal distribution**. **It is not necessary that the number of components in the two mixture distributions are equal**.

With the classical technique the distances between the curves are defined using the Euclidean distance with the following familiar formula:

$$D_{eucl}(TDP_j, TDP_k) = \sqrt{\sum_{i=1}^{96} (TDP_{ij} - TDP_{ik})^2},$$

where TDP_j and TDP_k denote the typical daily profiles of the j^{th} and k^{th} curves, TDP_{ij} and TDP_{ik} denote their values in quarter-hour i.

Distances were calculated at all times for normalised curves.

Both with classical and mixture models Ward's agglomeration method with hierarchical clustering was used. The Ward's method or Ward's minimum variance method is an algorithm that is often used in hierarchical agglomerative clustering. Its main idea is the minimisation of the variance within clusters. According to the logic of hierarchical clustering the step-by-step merging of clusters is done in a way that the variance within the clusters increases to the smallest extent as a result of the agglomeration.

4.1.3.2. Formation of profile groups of curves with similar consumption patterns[115]

The following figures show the matrices containing the distances of pairs of curves in the two solutions, and the dendrograms based on them created by hierarchical clustering – in only a few curves. In the distance matrices the highest and lowest distance values are in bold.

Based on the different merging distances reflected in the dendrograms, it can be stated that the two solutions give basically the same group assignments (classifications). The conclusion is much more interesting from the perspective of why the two techniques have lead to approximately the same results.

Figure 44: Distance matrices created by TDP-based clustering and mixture clustering

Euclidean distance matrix (TDP-based clustering)

$$d = \begin{array}{c} \begin{matrix} C25 & C35 & C66 & C79 & C109 \end{matrix} \\ \begin{bmatrix} 0 & & & & \\ 2.26 & 0 & & & \\ \mathbf{2.55} & \mathbf{0.71} & 0 & & \\ 1.87 & 0.88 & 0.93 & 0 & \\ 1.47 & 2.15 & 2.31 & 1.68 & 0 \end{bmatrix} \end{array},$$

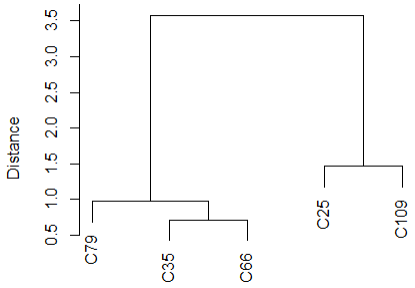
Kullback-Leibler divergence (distance) matrix (mixture clustering)

$$d = \begin{array}{c} \begin{matrix} C25 & C35 & C66 & C79 & C109 \end{matrix} \\ \begin{bmatrix} 0 & & & & \\ 32.08 & 0 & & & \\ 70.74 & 14.57 & 0 & & \\ 25.88 & 12.19 & \mathbf{12.18} & 0 & \\ 17.10 & 16.86 & \mathbf{73.84} & 17.14 & 0 \end{bmatrix} \end{array}$$

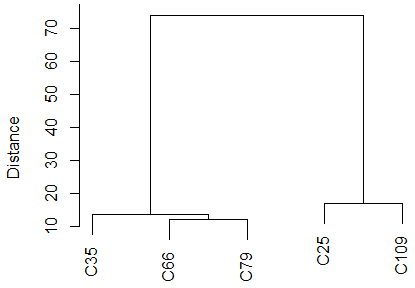
Source: author's own calculations (R) and figures (R).

Figure 45: Dendrograms based on the distances between the curves

TDP-based clustering



Mixture clustering



Source: author's own calculations (R) and figures (R).

In the TDP-based solution C35 and C66 are merged first, as both their *peak* and *off-peak* period variations are very similar. C79 differs from them basically in that the *off-peak* load is slightly higher in level compared to the other two. Compared to this, the evening load is much higher in the other curves and their shape is also very different; for example in C109 there are two intraday *peak*s (the second is slightly lower).

[115] For an easier tracing of results see the Figures in Appendix E).

According to the solution based on the mixture model the two curves that are most similar to each other are C66 and C79, which is not surprising, because in both, the weekend load is practically constant. What is especially important, the **weekday *off-peak* load level is practically identical with the *flat* level of the weekend load**. Otherwise, there is only summer temperature-dependency in both. C35 differs from them only in that it also shows winter temperature-dependency – most certainly, this is the reason why C35 joins this cluster as a third member, as the **weekday *off-peak* and weekend load level** are related to each other in the same way as those of C66 and C79. The weekend loads are not constant in C25 and C109, in C25 there is a long Saturday setback, and in C109 there are weekend variations that are identical to weekdays. The most distinctive factor in C25, however, is not the presence of non-constant weekend load' (as it is only the first part of Saturdays), but that compared to the C35-C66-C79 triplet, the level of the weekend load is significantly lower than the level of the weekday *off-peak* load. The above also provide appropriate explanation as to why C25 stands out most among the curves (this is where the means of the distances from the other curves is highest (the first column of the distance matrix), somewhat higher than in C109 (the last column of the distance matrix)).

Technically, Table 9 does not provide any additional information, but it supports what has been stated above: the distances of the curves from the portfolio can be seen as measured by the Kullback-Leibler divergence and by the Euclidean distance:

Table 9: Distances of consumer curves from the portfolio

Distance measure / Curve	**C25**	**C35**	**C66**	**C79**	**C109**
Euclidean distance	0.79	1.63	1.91	1.29	1.13
Kullback-Leibler divergence	23.68	33.36	79.19	63.01	20.41

Source: author's own calculations (R) and table.

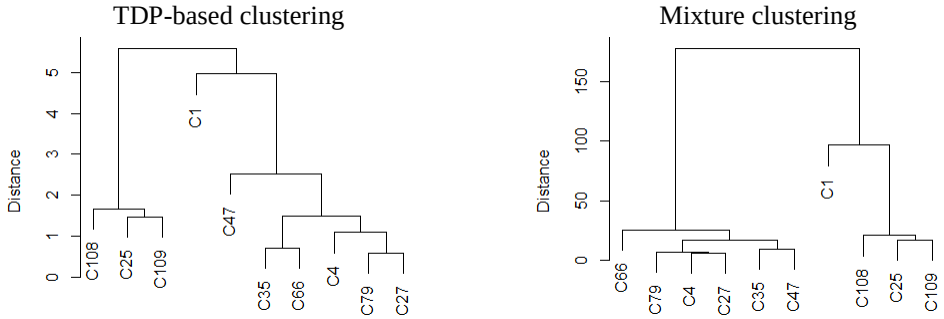
Based on the Kullback-Leibler divergence, the curves that are closest to the portfolio are C25 and C109, as their load at weekends is not baseload like in the portfolio.

The reason why the difference between the distances is relatively smaller regarding the Euclidean distance is definitely that it "does not consider" the weekends, and it can only draw conclusions from the weekday daily forms. C25 seems to be closer because regarding the morning ramp and the evening setback; this is what – on weekdays – mostly resembles the portfolio.

4.1.3.3. Formation of profile groups of curves with similar consumption patterns (extended example)

Regarding the extended example, the difference is not only in the 'suggested' number of clusters (in a TDP-based model choosing four (or even three) clusters is advised, while using the mixture model, it is better to choose three clusters,[116] see Figure 46), but also in how they are assigned to groups. It is clear that both techniques regard C1 as an *outlier* observation; it forms a one-member cluster in both types of group assignments (classifications). The *outlier* nature is caused by the fact that in the case of this consumer **the *off-peak* and weekend loads are practically zero**, while it is higher in the other curves. C47, based on the typical daily profile (see earlier in Figure 43) is somewhere between C1 and the five-member cluster. From there, the TDP-based technique can be traced easily. Figure 47 shows the same typical daily profiles as Figure 43; the identical colours indicate identical cluster memberships.

Figure 46: Dendrograms based on the distances between the curves (extended example)

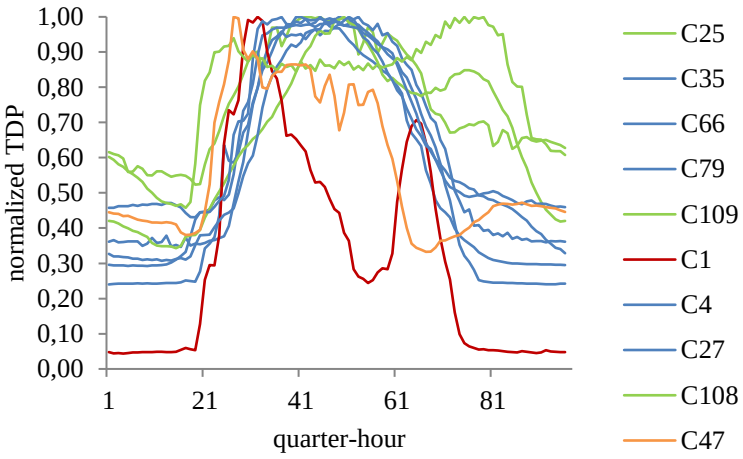


Source: author's own calculations (R) and figure (R).

C1 seems to be more similar to the C25-C108-C109 triplet's cluster using the mixture model, because these are the ones where there is a **second evening *peak*** (or a rather high evening load level – as in C25). This feature is not noticed by the TDP-based technique, and as a result, C1 is placed closer to the cluster where the *off-peak* loads are lowest (see the six-member cluster on the left side of Figure 46).

[116] If the *outlier* C1 curve is regarded as a separate cluster in both cases, see also in the main text.

Figure 47: The result of clustering normalised typical daily profiles (TDPs)



Source: author's own calculations (R) and figure (Excel).

Using the mixture model, the only real difference between the two techniques is in the six-member cluster. The level of weekend load is much smaller in every curve compared to weekdays (in C66 and C79 the expected value is practically constant). The guiding principle of the formation of clusters is probably that in C35 and C47 there is a very strong winter-dependency, while in C66 and C79 only the summer temperature-dependency is strong. Temperature modifies the ***peak–off-peak ratios*** compared to periods without a temperature-dependency, but still, the TDP-based model can only consider the latter due to the construction of the typical daily profile (that is, removing the effect of all exogenous variables).

Although the tracing and decomposition of the two types of methods was not very easy, the examples have shown the advantages of the mixture model. Here, they were most apparent in their role in the more fundamental representation of the daily profile and the role of the effect of temperature in the formation of clusters.

The two techniques shown (the TDP-based and the mixture model) are considerably different in their philosophy. While this was not discussed in this chapter in detail, it is still worth mentioning the **number of estimated parameters** used by the different techniques. In the classical technique the number of estimated parameters is **294**.[117] The number of estimated parameters used in the mixture model is provided by the following formula:

[117] Checking it on the basis of Table 8 containing the independent variables, the number of 294 is the sum of the following: 96 quarterhour *dummy*, 1 weekend *dummy*, 95 interaction *dummy*, 3 so-called special day *dummy*, 1 sunset-effect, 1 heating effect, 1 cooling effect, and 96 autoregressive coefficients.

$$(K-1)+K\left[d+\frac{d(d+1)}{2}\right],$$

where K denotes the number of components and d denotes the number of variables. The first term in the formula is the number of *prior* probabilities, the number of means and the numbers of estimated covariance matrix parameters per component are given in square brackets. The estimated parameters depending on the number of components is shown in Table 10. In this study the number of variables is always 6, that is, for 10-11 components the number of estimated parameters is more or less similar to classical techniques. In the majority of the cases, much more components have not been estimated. The differences are not really large; especially in view of the fact that in the mixture model not only the expected value, but variance is also estimated. This way, much more relevant and useful information is extracted. The number of estimated parameters for each curve and the values of the model selection criteria are in Table 11.

Table 10: The number of estimated parameters depending on the number of components using mixture models

No. of comp.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
No. of est. par.	27	55	83	111	139	167	195	223	251	**279**
No. of comp.	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
No. of est. par.	**307**	335	363	391	419	447	475	503	531	559

Source: author's own calculations and table.

Table 11: Model selection criteria and the number of estimated parameters using mixture models

Curve	**C25**	**C35**	**C66**	**C79**	**C109**
BIC	140 376	122 375	102 456	109 914	136 212
No. of comp.	13	11	13	9	10
No. of est. par.	363	307	363	251	279
Curve	**C1**	**C4**	**C27**	**C108**	**C47**
BIC	60 362	93 320	101 882	138 566	93 149
No. of comp.	13	11	7	11	11
No. of est. par.	363	307	195	307	307

Source: author's own calculations (R) and table.

4.1.4. Results, summary of conclusions

In this chapter, the description of the results for one single load curve served two main goals: to understand the logic of the mixture model and to provide a foundation for the sections on volume risk. The extent to which the parameters of components – as the statistical counterparts of typical consumption profile (or curve features) – show a more realistic picture of each curve was only truly revealed during the clustering process, through

the comparison of various curves. It was found that the mixture model can capture the structure of the whole curve much better (regarding both the daily shape, and the *peak–off-peak*, weekday/weekend periods). Therefore, **the H2 hypothesis cannot be rejected: by estimating the individually characteristic components for each curve, the mixture model supports the formation of much more realistic profiles than the classical solutions.**

The advantage of the technique based on the mixture model is obviously in that it considers the whole distribution of load variation, and works basically with **compressed information**, while using classical models, it is rather **information loss** that can be observed. This occurred here, in the present examples, during the construction of typical daily profiles (TDP), where the effects of weekend, temperature, etc. were also removed.

The mixture model used in the chapter form groups based on **covariance structure**, whose output does not directly include the typical daily profile in the classical sense. The produced typical consumption patterns, the so-called components, do not follow the logic of daily discretisation. They are not organised along days, but along the co-movement and distribution of the variables used. Of course, typical daily profiles can be produced here (due to the regression method building on the mixture model, see: next section), but a typical (or expected) daily profile is produced from the mixture of more components.

Clearly, this does not suggest that classical (daily profile based) techniques are useless, as a major part of the variance of the curve is explained by the intraday variability of the load, and these techniques can capture that. In individual curves, the weekend load is so negligible compared to weekdays that the information thus neglected does not mean much of a loss. In addition, based on for example yearly or weekly seasonal behaviour, it is possible to assign consumers to macro categories, and perform clustering within those.

Based on the second example it is much easier to see the advantage of the mixture model that it deals with the daily shape in a more fundamental manner (see for example: daily *peak* periods, temperature-dependency). The reason for this is probably that through the Kullback-Leibler divergence, it also takes standard deviation (covariance) into account. **Besides shape, risk is taken into account when forming groups; this way, it is not only the curves that have a similar profile, but also those with similar risk features that are assigned to the same group with this method.**

4.2. Modelling the uncertainty of consumption

Describing the uncertainty of consumption, that is, volume risk means the characterisation of the behaviour of the irregular component. The task assumes having some kind of a model, based on which the realisations of irregular components are generated. Of course, in practice this is not always satisfied, so there are simpler heuristic methods to measure volume risk, which are useful even if not having a specific model.

Classical time series regression models (SARMA and PAR[118]) are used to investigate to what extent the techniques that assume the constant dispersion of the irregular component cannot handle consumption uncertainty, in addition, depending on time they under- or overestimate it. The results allow us to conclude that the heteroscedastic feature of time series is basically characterised by seasonal or calendar effects, that is:

1. as multiple seasonality appears in the expected value of consumption time series, this multiple seasonality also appears in the variance of time series in the same way;
2. this way, it seems plausible to choose a technique which describes the variance of the time series as a function of the independent variables in the same way as the expected value of the time series itself.

The regression approach based on the *Gaussian* mixture model provides just such an option that fulfils all the above requirements. Using this, so-called conditional standard errors and confidence intervals are calculated, and it is shown how consistent they are with the realisations of the irregular component. The results are illustrated on various individual curves, stressing the general usability of the results, and their practical (also financially measureable) usefulness compared to classical methods.

4.2.1. Volume risk in classical time series regression models

This section describes, besides the results of using classical time series regression, a so-called heuristic method that is easy to implement in practice to measure volume risk.

4.2.1.1. The heuristic method to measure volume risk

Lo-Wu [2003] suggests the so-called **risk index** for the measurement of volume risk if not having a forecasting model. Due to the fact that forecasting errors can also be regarded as

[118] In the chapter on methodology it was described in detail that technically the difference is only in that the PAR model estimates periodically (in this case, quater-hourly) different autoregressive coefficients (which in this way means – *ceteris paribus* – 95 more estimated parameters). Due to the appropriate extension of the classical SARMA logic the PAR model is also regarded as a classical time series model.

the realisations of the differences from the forecast (expected) value, this is a suitable measure of volume risk. In Hungary the quarter-hour is the smallest interval that can be traded on the market; therefore, it is reasonable to use this resolution to calculate the risk index, so in the following $k = 1, 2 \ldots 96$ denotes the quarter-hours. Essentially, the risk index is a standard deviation-type measure and is calculated as follows:

$$R_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(L_{k_i} - \overline{L_{k_\iota}}\right)^2},$$

where:

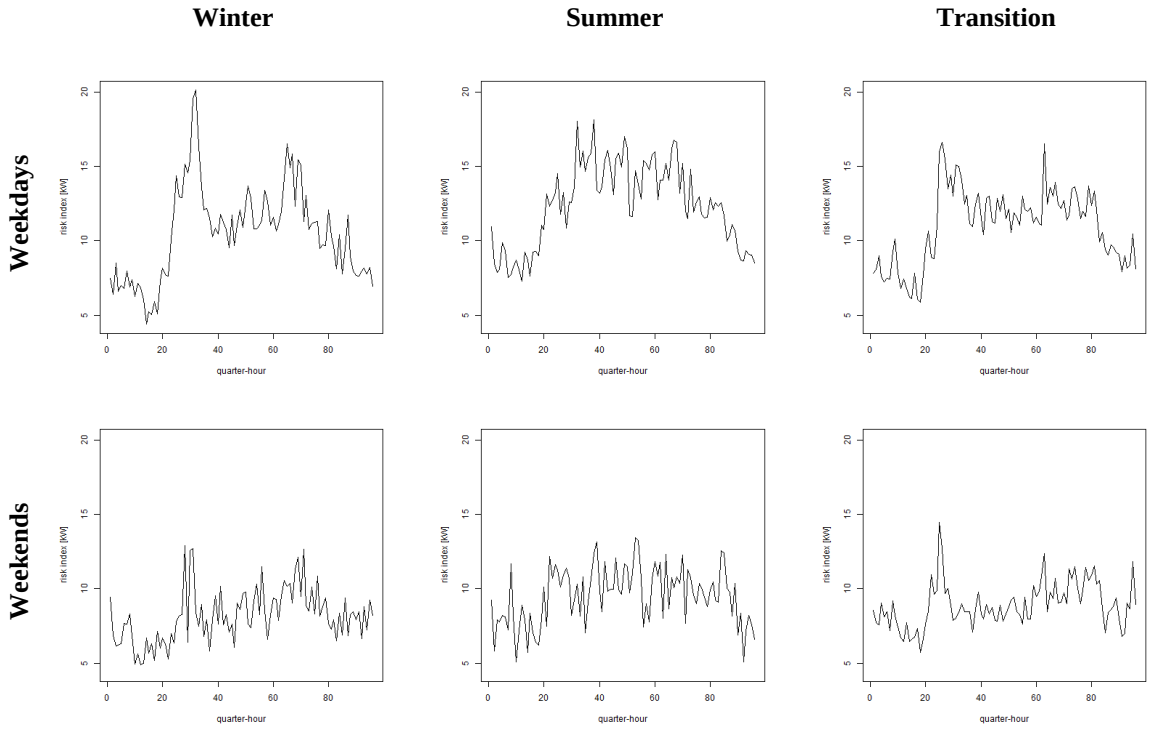
- R_k denotes the so-called risk index of the k^{th} quarter-hour,
- n is the length of the sample period (the number of periods used in the calculations),
- L_{k_i} denotes the load increment in the k^{th} quarter-hour in the sample period,
- $\overline{L_{k_\iota}}$ denotes the mean of the load increments in the k^{th} quarter-hours in the sample period.

Due to the multiple seasonality assumed based on previous results, it is worth calculating the variation of the risk index separately for each season and day type (weekdays and weekends). The results are examined concerning the portfolio. The relevant risk indices can be seen in the following figures,[119] where despite the noisy nature, the basic tendencies are shown in the sense that it can be seen where the volume risk may be higher or lower. Before summarising these, some scepticism concerning the results should be raised.

Clearly, it is an advantage of the technique described that with the choice of the appropriate sample size, uncertainty can be calculated based on the specific period (e.g. the preceding two weeks) that is regarded **manually, subjectively** the most characteristic. Another methodological advantage is that the measure is based on the standard deviations of the same hours; and **standard deviation** is otherwise a **measure of risk and uncertainty**. But at the same time, its arbitrariness is indeed a disadvantage. What it actually means is that the selected length of the sample period can only be defined based on experience. If the period is too short, the **existence of small sample size** may be problematic (the variation of the risk index will be too noisy), but if the chosen period is too long, the **'localisation'** will not be valid.

[119] The figures were calculated for a quater-hourly resolution, using the whole-year time series.

Figure 48: The variation of risk indices in the portfolio



Source: author's own calculations (R) and figures (R).

Neither is there any fundamental justification to these calculations, as the volume risk of a given day is based on the preceding days, which is a very simple, 'naive' method. On the other hand – even assuming a large sample size – these risk indices remain quite hectic, because they are not so smooth and even at the very best, only uncertain tendencies can be gathered, which makes their application unfeasible in practice. With such a simple calculation logic, it is difficult to decide if the peaks are real or not, or if they actually reflect higher risk in a given period. This unquestionably points toward the need to use some model-based approach.[120]

4.2.1.2. The model-based exploration of the characteristics of volume risk

The modelling of curves with classical time series techniques was performed using SARMA and PAR regressions. As only one year of data is available for individual curves,

[120] As an alternative, the technique that is often used is the examination of the standard deviations derived by computing the increments on the same hours of the previous week. This is rather a logically similar pair of the method where in practice the load of a given day is forecast using the load of the same day from the previous week – this, so-called weekly risk index, is much more widespread in practice. Though the conclusions using this indicator are qualitatively similar, the risk index studied here is the one that can be practically compared to model-based results. Otherwise, an application (software) implemented by the Enoro Smart Energy Management measures the uncertainty or variability of a consumer curve in such a way (as well) – based on weekly increments, see: http://www.enoro.com/. Of course, in such applications calculations disregard holidays, transferred working days, and days with 23 or 25 hours due to clock changes.

the estimation was made for the whole sample. As a consequence, the out-of-sample performance cannot be evaluated, but by all means, the examination of the latter – that is, making forecasts – needs to be treated as a separate issue.

The aim here is primarily the exploratory description (or verification) of the problem concerning the extent to which classical methods over- and underestimate volume risk, and in which period this is so. The discrete steps are described in more detail about the portfolio; a separate section contains the comparative study of individual curves.

Table 12: Independent variables used in regression and their short description

dummy variables denoting hours[121]	the value of the *dummy* variable is 1, if it is a quarter-hour of the given hour of the day, otherwise it is 0
dummy variable denoting weekend days[122]	the value of the *dummy* variable is 1, if it is the quarter-hour of a weekend day, otherwise it is 0
interaction variables denoting quarter-hours on weekends	variables that are constructed as interaction (that is, as product) of *dummy* variables denoting hours and *dummy* variables denoting weekend days
variables denoting holidays and other special days (official non-working days and transferred days)	the value of the given *dummy* variable is 1, if it is the quarter-hour of a given holiday, official non-working day or transferred working day, otherwise it is 0
the so-called sunset effect	the signed deviation of the sunset time from 18:00, see also: Figure 46 below
heating degree-day (HDD)	the downward deviation of temperature from $12°C$ to capture the heating effect
cooling degree-day (CDD)	the upward deviation of temperature from $21°C$ to capture the cooling effect

Source: author's own table.

The independent variables used in regression estimation are in Table 12. Apart form the different treatment of the effects of quarter-hours (different or equal in an hour) the variables are the same as the ones that were used in creating typical consumption patterns in Section 4.1. At this point, it needs to be added to the above that a **major advantage of the mixture regression used in this chapter is that variable transformations**, such as the ones that were used for calculating degree-days from temperature **need not be done, because the model deals with nonlinearity automatically**. Here, the stress is not only on

[121] It would also be possible to use *dummy* variables denoting quater-hours (in the same way as it was done in the chapter on profiling), but this would mean having to use 95 variables instead of 23; and including interaction variables would mean having to work with even more estimated parameters. The choice between the two solutions can be tested for using model selection criteria, but this will be not introduced here.

[122] Of course, it would be possible to use as many *dummy* variables as needed to differentiate the effects of all the days of the week; that is, to use 6 *dummy* variables instead of 1; however, being *parsimonious* is also regarded more important here.

nonlinearity, but also on automatism, as the choice of threshold levels in individual curves – keeping at degree-days – can be difficult and inefficient to accomplish.

The choice among classical regression models is based on the Schwarz-Bayes informational (BIC) criteria (see Table 13).

Table 13: Measures describing the goodness of fit in classical regression models[123] for the portfolio

Model	BIC	RMSE [kW]	MAPE [%]	MAE [kW]
ARMA(1, 0)	272 989.81	11.81	1.56	9.05
ARMA(2, 0)	272 980.76	11.81	1.56	9.05
ARMA(0, 1)	358 614.89	40.04	4.90	29.59
ARMA(0, 2)	332 575.50	27.61	3.50	20.77
ARMA(1, 1)	273 009.57	11.80	1.56	9.05
SARMA(1, 0)(1, 0)$_{96}$	**270 413.01**	11.51	1.52	8.83
SARMA(1, 0)(0, 1)$_{96}$	271 077.58	11.48	1.52	8.84
SARMA(0, 1)(0, 1)$_{96}$	348 205.91	34.50	4.32	25.79
PAR(1)	**270 483.40**	11.28	1.49	8.64
PAR(2)	**270 212.60**	11.03	1.46	8.47
PAR(3)	270 506.20	10.92	1.44	8.36

Source: author's own calculations (R) and table.

Regarding strictly only the minimum of the **BIC** criterion, PAR(2) should be chosen, but considering the 96 more estimated parameters, the decrease achieved on the model selection criterion is not so considerable, therefore SARMA(1, 0)(1, 0) $_{96}$ and PAR(1) could

[123] The measures that appear in the table below are used in evaluating the goodness of fit of the model and the quality of forecasts (y_i denotes the realised values of the time series, $\hat{y}_i$ denotes the values estimated based on the model, and h is the length of the time series in the examination of in-sample fit, or the length of the horizon in forecasting):

Name of the measure	Formula
Mean Squared Error	$MSE = \frac{\sum_{i=1}^{h}(y_i - \hat{y}_i)^2}{h}$
Root Mean Squared Error	$RMSE = \sqrt{\frac{\sum_{i=1}^{h}(y_i - \hat{y}_i)^2}{h}}$
Absolute Percentage Error	$APE_i = \frac{\lvert y_i - \hat{y}_i \rvert}{y_i}$
Mean Absolute Percentage Error	$MAPE = \frac{\sum_{i=1}^{h} APE_i}{h}$
Mean Absolute Error	$MAE = \frac{\sum_{i=1}^{h} \lvert y_i - \hat{y}_i \rvert}{h}$

RMSE considers greater errors with larger weight (that is why it is better for evaluating in-sample fit), while $MAPE$ can be used to express mean error in percentages, hence it is a measure independent from the measurment unit, which is a useful characteristic in making comparisons.

be chosen. The results also show that the models that deal with time dependency only in the moving average term (that is, dealing with it in the error term) have a much worse fit.[124,125]

Although among the measures in the table (for exact definitions, see footnote) for in-sample evaluations usually **RMSE** is used, it is worth examining the others to derive some additional information. The others might be more useful in forecasting, or for out-of-sample evaluations. The majority of studies on assessing fit and forecast prefers using relative (%) errors, and **MAPE** primarily. The disadvantage of this is that errors are characteristically higher in *peak* periods, when the load values are – usually – higher. This way, a greater error in absolute value may seem relatively smaller. Likewise, an error made when the load is lower may seem much higher considering the level of load, thus increasing the value of MAPE in an unreasonable extent. Consequently, the values of MAPE are decreased by the errors in the more expensive (*peak*) periods, and increased by the errors in the cheaper (*off-peak*) periods. Considering all of these, the **MAE** indicator seems to be the best choice for energy time series[126] as differences resulting from the errors made will need to be handled in volume, that is, in $[kWh]$.

Proceeding to the investigation of errors, Figures 49-50 show the residuals of the SARMA(1, 0) (1, 0) $_{96}$ model with the confidence band at a 95% confidence level for the whole year and the three chosen weeks previously investigated in Section 1.4.

Of course, the confidence band remains constant, and on the whole, the majority of residuals are within the confidence band. The fitted model is regarded suitable if 5% of the residuals are outside the confidence interval at the 95% confidence level (or in general, at a

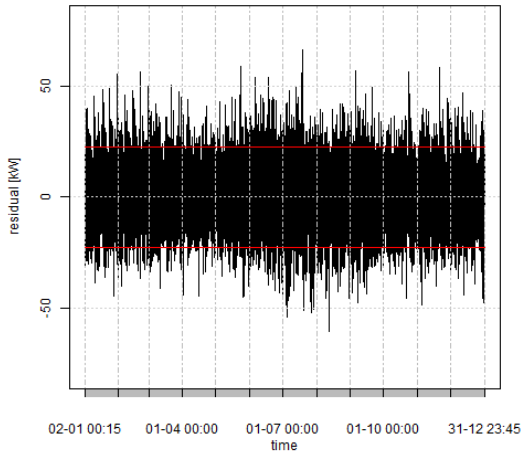
[124] Considering, however, that the peak value of loads in the examined time series is around 1000-1200 $[kW]$, any model fit can be regarded appropriate, but if the emphasis is on better fit (and more accurate forecasts), it is worth choosing some autoregressive model. But if the emphasis is on interpreting parameters (betas), due to their *ceteris paribus*-type interpretation, it is only possible in models where there is no autoregressive term. In this case, it is only possible to choose models that include moving average terms.

[125] The interpretation of the parameters is in itself useful, as it supports the decomposition of the time series along the independent variables, but its practical importance is somewhat smaller. As the aim here is not primarily the evaluation of accuracy but that of volume risk, this focus supports the idea of choosing the model with the best fit (the ones that contain an autoregressive term), and examine the behaviour of residuals in this relation.
In subsequent chapters we shall see that in individual consumption curves it is usually not the moving average but the autoregressive term that gives a more accurate capture of temporal behaviour. This probably assumes that there is some slight trend in the time series. This phenomenon can be captured in practice on the out-of-sample forecasts. Even in the case of very complicated and complex models, it may happen that forecasts with multiple periods ahead are under- or overestimated. This may be explained either by the slight trend, or the – often occurring but hardly noticeable – level shifts, that might be captured by more complicated techniques. This, *inter alia*, is the reason why out-of-sample performance should be treated as a separate field.

[126] Not mentioning the power price rate time series, where there may also be negative prices; APE and MAPE would take negative values in these hours, therefore, its application is not advantageous in any way.

confidence level of (1–α) % its α %). Its value in this case is 5.61 % (see Table 14), which means that the model can be regarded appropriate on the whole, or **globally**, because the difference from 5% is negligible.

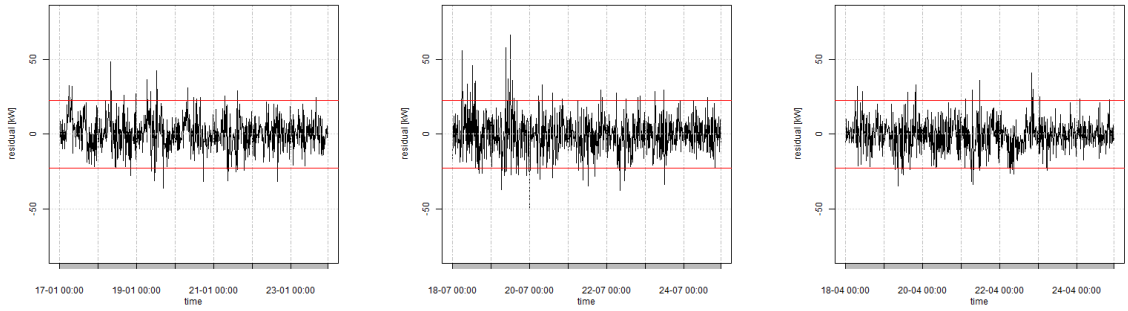
Figure 49: The variation of the SARMA(1, 0)(1, 0)$_{96}$ model residuals in the portfolio



Source: author's own calculations (R) and figure (R).

Figure 50: The variation of SARMA(1, 0)(1, 0)$_{96}$ model residuals in the portfolio for some weeks

Winter **Summer** **Transition**



Source: author's own calculations (R) and figure (R).

The reason why the proportion of observations outside the confidence interval is close to what was expected can probably be explained by the calculation logic of the regression model. The principle of the procedure is that the squared difference of the actual and the model-based prediction value of the dependent variable should be minimal. The **standard error** (as a standard deviation-type measure) of the model due to the **greater differences** will be very **high**. For this reason, regarding all the residuals, the ratio of observations outside the confidence band – exactly because of the width of the confidence

band – are approximately correct (even though based on the figures – from the other aspect – in certain periods the confidence band might seem unnecessarily wide).

The previous figures may be created for models with similarly good fit, but the differences are negligible (this can be inferred from the similar results of Table 14 and Figure 51 on PAR(1) and ARMA (1, 0) models.).[127]

Table 14: The ratio of observations outside the confidence interval (CI95)[128] and the average size of the confidence interval for the portfolio (classical regression models)

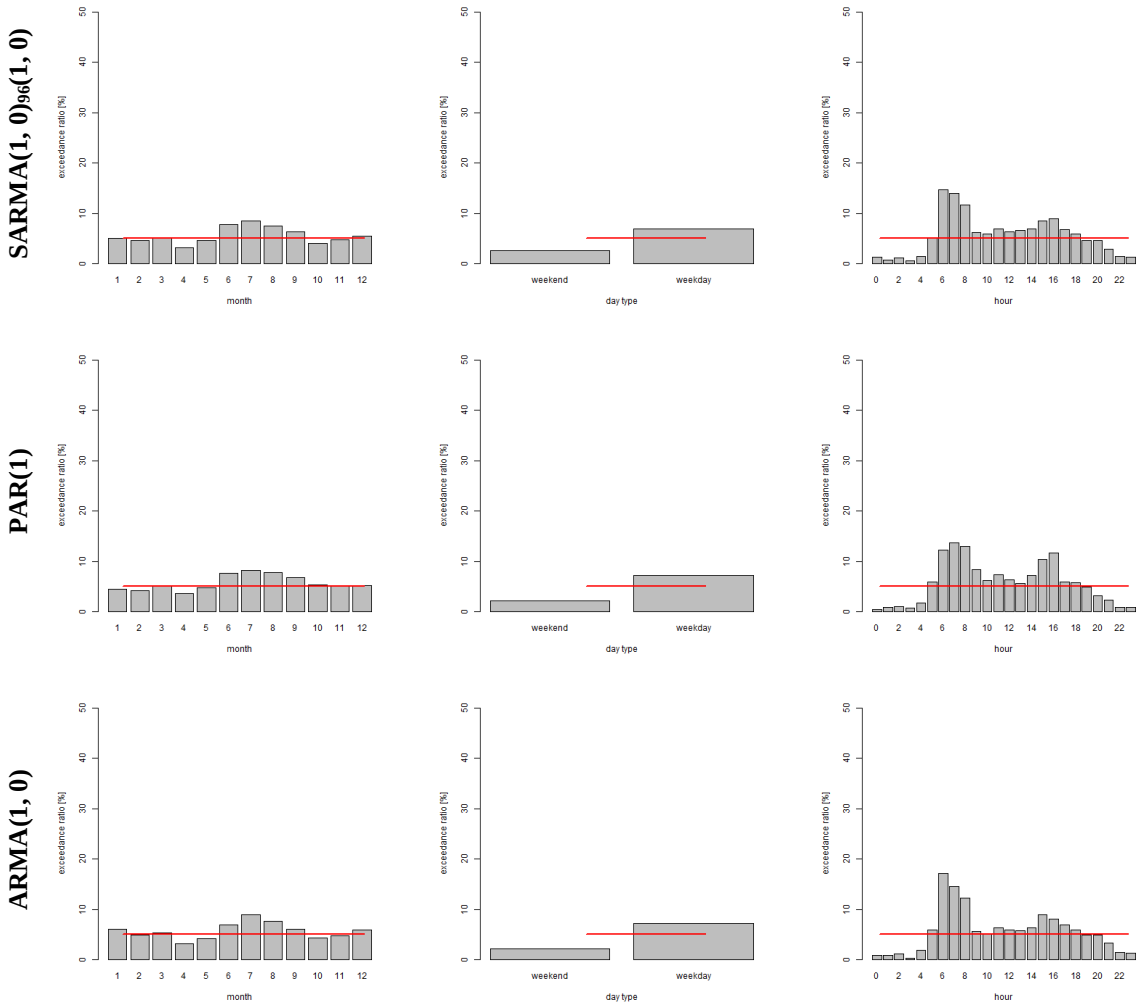
Model	Ratio [%]	Average CI95 [kW]
SARMA(1, 0)(1, 0)$_{96}$	5.61	45.12
PAR(1)	5.75	44.01
ARMA(1, 0)	5.66	46.30

Source: author's own calculations (R) and table.

For a more detailed insight into the assumed seasonal heteroscedastic behaviour of errors it is worth looking into the variation of the ratio of the observations outside the confidence interval (see Figure 51) in various periods (for example: seasons, days of the week, hours in the day). If the hypothesis of constant standard deviation is correct, there should only be random deviations from – the expected – 5% level (marked by a red line in the figure) regarding the ratios of observations outside the confidence interval in various periods. In winter and summer months, when the consumption is temperature-dependent, the ratio of observations outside the confidence interval is somewhat higher. At weekends, the ratio of observations outside the confidence interval is well below 5%[129], while on weekdays it considerably exceeds it, just as in morning or *peak* period hours.

[127] The lower standard error due to the higher number of estimated parameters is reflected in the average sizes of the confidence intervals.
[128] Here, and in subsequent tables, CI95 denotes the confidence intervals at a confidence level of 95%.
[129] In Figures 51-52 this is what was most spectacular: the confidence band, which remains constant in the observed period is much wider in weekend periods and weekday off-peak hours as what seems to be explained by the fluctuation (dispersion) of the residuals.

Figure 51: The ratio of observations outside the confidence interval (CI95) for the portfolio (classical regression models)



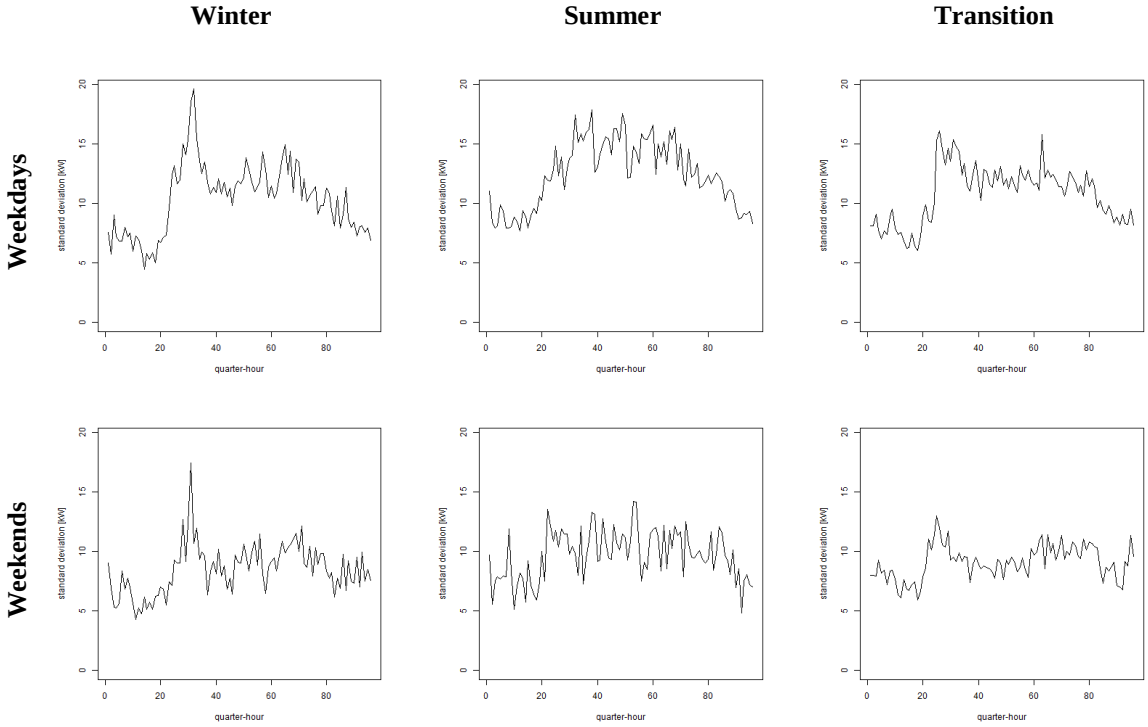
Source: author's own calculations (R) and figures (R).

This heteroscedastic behaviour can obviously be tested for using any heteroscedasticity test. The added value of the results here is primarily in that it was also revealed what the essential **characteristics of heteroscedasticity** are. This is a rather **financial logic**, where it is also important to **measure the risks of** more volatile and less volatile **periods in an accurate and consistent way**. However, there, heteroscedasticity is of a completely different nature, which supports the use of different kinds of methods.

Regarding the classical models, the conclusion mainly concerns the shortcoming that the distribution of errors outside the confidence interval is not uniform. This means that volume risk is under- or overestimated by an order of magnitude depending on the time period. This way, the results lead us to conclude that it is necessary to calculate some time-dependent, conditional standard error (higher in some periods, lower in others). The

calculation of the standard deviation of the residuals fitted to the grouping shown in Figure 51 is definitely a sound basis concerning the appropriate time-dependent magnitude. Figure 52 shows the quarter-hourly standard deviations of the residuals with the grouping used in calculating the risk index.

Figure 52: The standard deviation of residuals in the SARMA(1, 0)(1, 0)$_{96}$ model for the portfolio



Source: author's own calculations (R) and figures (R).

The figures are practically identical with the related figures of the risk index, therefore, **in summary, the findings are the following**:

- the volume risk is lower at weekends and higher on weekdays,
- the volume risk is characteristically the highest during morning ramps and the evening setbacks (around 18:00-20:00),
- the volume risk of the daytime (around 10:00-18:00) 'real' *peak* period is lower than the risk of the ramp and setback periods, but higher than those of night-time and early morning hours.

The behaviour of morning ramp and setback periods can basically be explained by **human activity**, which is negligible in *off-peak* periods, and is most uncertain in morning ramp and setback periods (which have the highest gradient). Only summer is an exception to this, where *peak* periods have the highest risk – this is mainly due to the cooling use of

electricity. The most important findings and conclusions are summarised in the next section.

4.2.1.3. Summary of results and conclusions

As both risk index and the standard deviation of residuals calculated in the appropriate resolution are standard deviation-type measures, they are both equally suitable for measuring risk. Nevertheless, the residuals calculated from the model in themselves – beyond being model-based – have seemingly little added value compared to the risk index, as the variation of the residuals contradicts the – constant – confidence interval resulting from the model.

The magnitude of the coincidence of the risk index and the standard deviation of residuals is otherwise present as a result of the assumption of **linearity**. The fact that the risk index was calculated from the standard deviations of the load increments, the mean of these increments is what a linear model can capture. As the standard deviation appears in the errors that remain after using a linear model, consequently, the two solutions lead to similar results.[130]

It might appear at first sight that there is something referring to the inappropriately handled sunrise and sunset effect (which changes throughout the day) or some other seasonality behind the higher dispersion of morning ramp and evening setback periods. However, in the case of SARMA(1, 0)(1, 0)$_{96}$ due to using the 96th lag, this effect is dealt with very well (the improvement of indicators measuring goodness of fit compared to the ARMA(1, 0) model is largely due to this), as a result, it is presumably **not** some **omitted, inappropriately handled seasonality** that explains the phenomenon. Compared to the risk index, the **model-based** approach is a more elegant solution, because the omission of a relevant variable may be the cause of the heteroscedastic behaviour. All of this, however, cannot be tested for if not having a model.

This section has concentrated so far on a better fit. This way, it has been an abandoned feature that the periodic autoregressive model – by estimating different autoregressive coefficients for each period (that is, quarter-hour) assumes that the **autocovariance function** – of course, with the condition of given exogenous variables – is

[130] Similarly to the risk index, the absolute (not relative, interpreted in %) increments are calculated if it can be assumed about the variation of the phenomenon that the absolute periodic change can be regarded constant, that is, the phenomenon can be characterised by a linear tendency. This is exactly the linearity assumption that lies behind the SARMA model.

not constant, but only **periodically constant**.[131] In theory, the possibility is given that the estimation of periodically different autoregressive coefficients slightly changes the behaviour of the residuals, since the quarter-hourly different autoregressive coefficients actually handle some quarter-hourly different seasonal effect. Based on the results of Table 14 and Figure 51 the patterns of the ratios of the residuals outside the confidence band have not changed considerably due to the inclusion of the periodically changing autoregressive coefficient.

It is definitely worth mentioning here that *Subbarao et al.*'s [2011] so-called kNN-method follows a logic for calculating the confidence interval which uses the residuals of the model that describes consumption – basically, with a historical (empirical) method; with the addition that for a given point in time it uses a sample of the errors of some past time periods that can be regarded similar based on a well-defined distance measure.[132] The interval limits (or bounds) are derived from this sample by calculating percentile values. The time-dependence of interval bounds is provided by the distance measure. However, as the standard deviations of errors are fairly high in themselves (see for example Figure 52, but the adverse effect of *outlier*s or multicollinearity may also be thought of), this noisiness also appears in confidence intervals, which is not so advantageous.

4.2.2. Modelling volume risk with mixture regression

The model used for the quantification of volume risk is identical with the one used in an earlier section (Section 4.1) on the formation of typical consumption patterns. There, only part of the total year available was used for the estimation, which provides an opportunity to perform the out-of-sample evaluation on the remaining period.

The aim of this section is to examine whether – using the logic described in the previous section – it is possible to calculate conditional standard error using regression

[131] Of course, it is possible to test for whether the estimation of a periodically autoregressive coefficient (that is, the periodically different autocovariance) is necessary. This can be performed by an F-test, using the model selection criteria in Table 13, or the *likelihood*-based test building on it.

[132] The logic of the distance measure lies in the logic of the Euclidean distance. It weighs the distances among the values of the independent variables of times by the variables' partial effect on the dependent variable (assuming for example multivariate linear regression, the weights would be the estimated beta parameters). In the calculation of the confidence interval, the number of k closest neighbouring errors are considered. The extent to which the confidence interval becomes dependent on the specific values of the independent variable is influenced by the value of k, which definitely means some form of abritrariness; however, neither a too high nor a too low value is really suitable (it is blurring the dependence of the given specific value of the independent variable in case of large sample size or results in the hectic behaviour of the confidence interval in case of small sample size).

based on a mixture model, and whether it is possible to derive confidence interval on the dependent variable (load) which consistently reflects the volume risk at the given time.

The **comparison** between classical time series regression and mixture regression results in this section – similarly to the section on typical consumption patterns – can only be **qualitative**, as due to the different model constructions the independent variables are not the same. The results are detailed here regarding the portfolio; and it is primarily the most important conclusions that are mentioned with the comparison to the classical solutions concerning individual curves.

4.2.2.1. Capturing the characteristics of volume risk

Table 15 shows the in-sample and out-of-sample goodness of fit measures of the mixture model. The aim is not primarily to improve the model fit or forecast. It is much more stressed that the performance of mixture regression is approximately the same as that of classical time series regression models.

Table 15: The variation of goodness of fit measures of models for the portfolio

Model	Prediction	RMSE [kW]	MAPE [%]	MAE [kW]
SARMA(1, 0)(1, 0)$_{96}$	In-sample	11.51	1.52	8.83
GMR	In-sample	12.23	1.60	9.27
	Out-of-sample	12.28	1.62	9.38

Source: author's own calculations (R) and table.

The ratio of observations outside the 95% confidence interval (see Table 16) slightly exceeds 5% (as in classical models)[133], but a more relevant statement is formed on the basis of the subsequent Figure 53.

Table 16: The ratio of observations outside the confidence interval (CI95) and the average size of the confidence interval for the portfolio

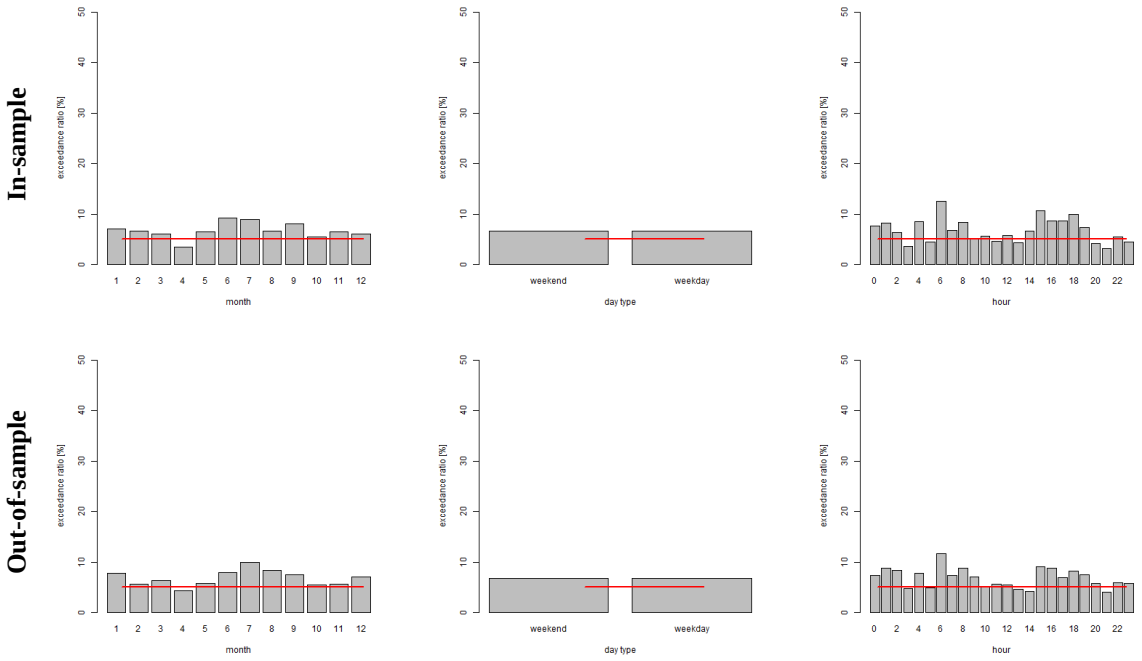
Model	Prediction	Ratio [%]	Mean CI95 [kW]
SARMA(1, 0)(1, 0)$_{96}$	In-sample	5.61	45.12
GMR	In-sample	6.69	43.89
	Out-of-sample	6.82	43.94

Source: author's own calculations (R) and table.

[133] There still might be many reasons behind the differences in exceeding the 5% such as the lack of the relevant independent variable, the inappropriate handling of the independent variable, or the wrong assumption about the distribution of the error term, etc.

The distribution of the observations outside the confidence interval is much more uniform than they were using the classical models; that is, the ratio of observations outside the confidence interval is approximately the same for every month, weekday and weekend, and hour. The relevant statement here is the more **uniform** distribution during the day, and the smoothing of the periodically (regularly, seasonally) occurring inequalities.

Figure 53: The ratio of observations outside the confidence interval (CI95) for the portfolio (mixture regression)



Source: author's own calculations (R) and figure (R).

In the following, it can be examined numerically if the **standard errors** estimated for the various quarter-hours are consistent with the **standard deviations of the residuals** of the various quarter-hours. The calculations of the conditional standard errors (or conditional standard deviations) were of course according to what was described in Section 3.2.5 on the methods used. The conditional standard deviation of the dependent variable, similarly to its conditional expected value (using the notation of the referred chapter, it is the $\hat{y}_i$, and square root of $var(\hat{y}_i)$, that is, meant as standard deviation) is to be understood with the condition of the given independent variable values. In light of the results here this is what the methodological realisation of calculating time-dependent standard errors means.

As conditional standard deviations and of course, residuals, may differ from quarter-hour to quarter-hour, in order to get a compact picture of their variation, it is worth

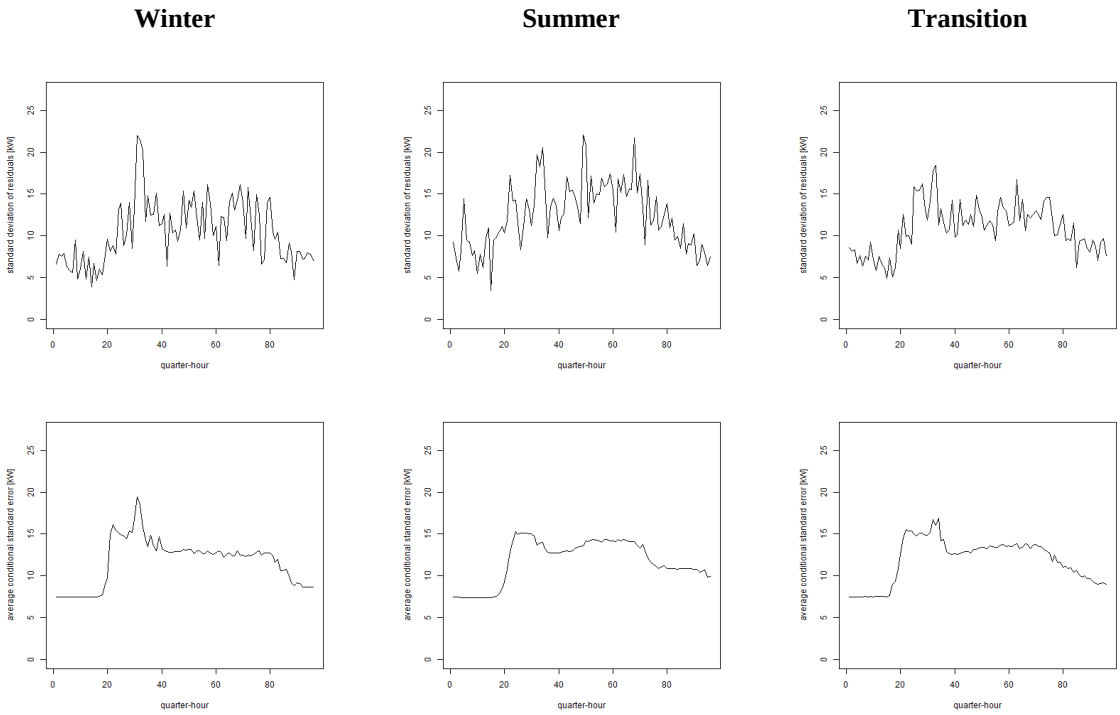
examining the standard deviations of residuals and the mean of the conditional standard deviations estimated on the basis of the mixture model for every quarter-hour according to seasons (winter, summer, transition) and day types (weekdays, weekends) (see Figures 54-55). This kind of discretisation has already appeared at the heuristic measures and the classical time series models. Here, it mainly serves the rough validation of the results. It should be stressed that regarding conditional standard deviations 'means' are presented here; but they may be different – depending on the independent variables – in different quarter-hours.

Again, it can be stated according to the uncertainty measures (that is, standard errors) calculated on the basis of the model that the uncertainties of morning and early morning hours are the smallest, while the risks of the morning ramps and evening setbacks are relatively high. In addition, the risk of daytime at weekends (in the hours that correspond to *peak* period hours) is smaller than on weekdays. The figures here reflect a shape similar to what has been shown in an earlier section concerning the standard deviation of residuals. Some differences may occur, which is due to the fact that – in contrast with techniques used in previous sections – the mixture model is capable of identifying **nonlinear** relationships in an automatic, exploratory way (as opposed to the previous classical models assuming linearity).

A last note to the conclusion is that it is true that the results in this section do not contradict the hypotheses or the results in previous sections in the variation of volume risk. What is more important is that the model that was estimated calculates standard errors that are consistent with the standard deviation of the residuals. In the figures, the aspect from which the standard deviations of residuals and the means of standard errors can be seen is the similarity of the shapes of the curves (the range of the y axes are the same, which definitely makes the comparison easier). The use of mixture regression is, *inter alia*, in this **consistence** and the underlying or supplementary model-based approach.[134]

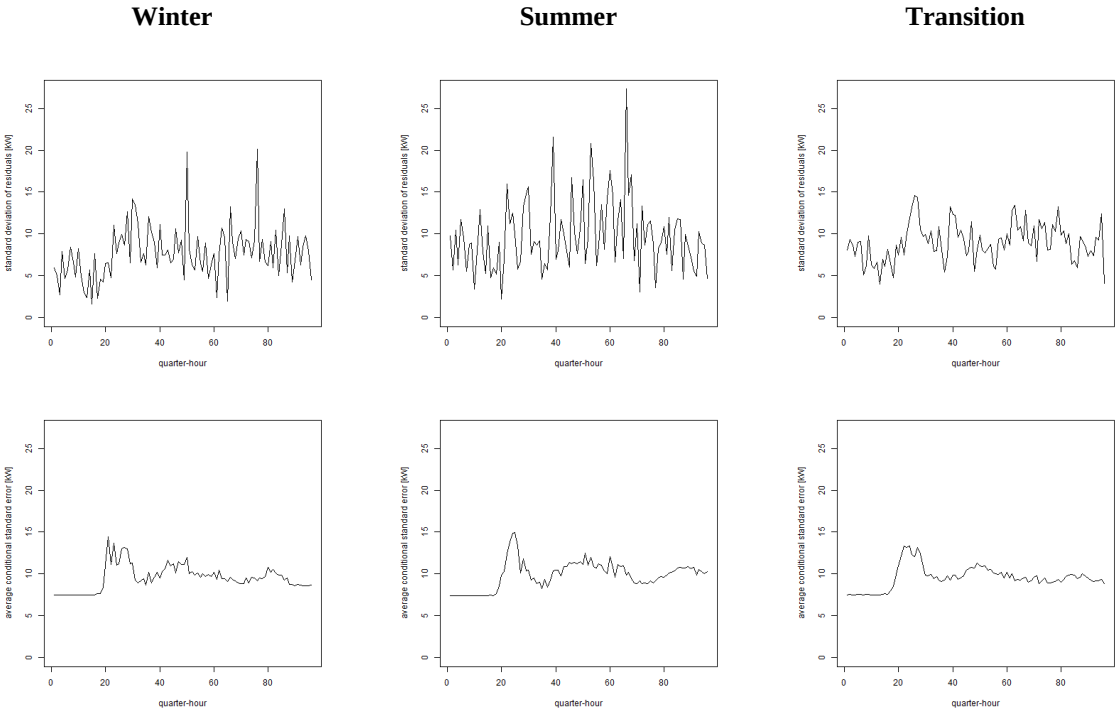
[134] Figures 54-55 include the results inside the sample; as practically, there are no differences between the in-sample and out-of-sample results.

Figure 54: The standard deviation of mixture regression residuals and the average conditional standard deviations for the portfolio (in-sample results, weekdays)



Source: author's own calculations (R) and figure (R).

Figure 55: The standard deviation of mixture regression residuals and the average conditional standard deviations for the portfolio (in-sample results, weekends)



Source: author's own calculations (R) and figure (R).

4.2.2.2. Volume risk of individual curves and their comparative analysis

Table 17 shows some goodness of fit measures in some individual consumer curves. No major differences have been found between the results of classical regression and mixture models regarding goodness of fit, indeed, mixture regression often outperforms the other.

Table 17: The variation of goodness of fit measures of models for individual curves

Curve	Model	Prediction	RMSE [kW]	MAPE [%]	MAE [kW]
C25	ARMA(1, 0)	In-sample	0.93	9.11	0.65
	GMR	In-sample	0.93	8.40	0.65
		Out-of-sample	1.01	8.51	0.69
C35	ARMA(1, 0)	In-sample	0.66	4.62	0.36
	GMR	In-sample	0.62	4.40	0.36
		Out-of-sample	0.59	4.35	0.34
C66	ARMA(1, 0)	In-sample	0.40	5.06	0.24
	GMR	In-sample	0.36	4.43	0.22
		Out-of-sample	0.38	4.48	0.22
C79	ARMA(1, 0)	In-sample	0.39	19.90	0.48
	GMR	In-sample	0.64	16.58	0.41
		Out-of-sample	0.64	16.63	0.41
C109	ARMA(1, 0)	In-sample	1.64	9.22	1.23
	GMR	In-sample	1.59	8.60	1.17
		Out-of-sample	1.64	8.95	1.21
C1	ARMA(1, 0)	In-sample	0.10	36.55	0.05
	GMR	In-sample	0.10	34.77	0.05
		Out-of-sample	0.10	33.78	0.05
C4	ARMA(1, 0)	In-sample	0.17	11.15	0.11
	GMR	In-sample	0.18	11.55	0.11
		Out-of-sample	0.17	11.48	0.11
C27	ARMA(1, 0)	In-sample	0.25	11.36	0.18
	GMR	In-sample	0.25	11.18	0.18
		Out-of-sample	0.25	11.32	0.18
C108	ARMA(1, 0)	In-sample	1.62	11.09	1.23
	GMR	In-sample	1.54	10.25	1.17
		Out-of-sample	1.59	10.32	1.20
C47	ARMA(1, 0)	In-sample	0.18	10.32	0.11
	GMR	In-sample	0.19	10.13	0.11
		Out-of-sample	0.19	10.84	0.11

Source: author's own calculations (R) and table.

The conclusion that can be drawn from Table 18 is much more important. The ratio of observations outside the confidence interval is also somewhat different from 5%, similarly to what was calculated with SARMA models. However, in subsequent figures, it

will be visible that occurrences outside the confidence-interval are much more uniformly distributed – similarly to what was observed in the portfolio.

Table 18: The ratio of observations outside the confidence interval (CI95) and the average size of the confidence interval for individual curves

Curve	Model	Prediction	Ratio [%]	Mean CI95 [kW]
C25	ARMA(1, 0)	In-sample	6.29	3.64
	GMR	In-sample	7.96	3.13
		Out-of-sample	7.87	3.32
C35	ARMA(1, 0)	In-sample	4.41	2.59
	GMR	In-sample	7.12	1.78
		Out-of-sample	7.09	1.67
C66	ARMA(1, 0)	In-sample	7.89	1.55
	GMR	In-sample	7.23	0.99
		Out-of-sample	7.72	1.00
C79	ARMA(1, 0)	In-sample	7.90	2.62
	GMR	In-sample	6.35	1.92
		Out-of-sample	7.07	1.87
C109	ARMA(1, 0)	In-sample	5.72	6.45
	GMR	In-sample	6.33	5.61
		Out-of-sample	7.04	5.64
C1	ARMA(1, 0)	In-sample	6.99	0.38
	GMR	In-sample	6.70	0.23
		Out-of-sample	6.69	0.24
C4	ARMA(1, 0)	In-sample	5.79	0.68
	GMR	In-sample	7.91	0.56
		Out-of-sample	7.66	0.56
C27	ARMA(1, 0)	In-sample	5.95	0.97
	GMR	In-sample	6.77	0.87
		Out-of-sample	7.04	0.85
C108	ARMA(1, 0)	In-sample	5.28	6.35
	GMR	In-sample	7.50	5.46
		Out-of-sample	7.90	5.48
C47	ARMA(1, 0)	In-sample	6.11	0.72
	GMR	In-sample	7.20	0.55
		Out-of-sample	7.44	0.55

Source: author's own calculations (R) and table.

From a practical perspective it is a much more relevant finding that the average size of the confidence interval is much smaller in mixture regression. The last column of Table 18 shows these results. It was earlier indicated at the portfolio, but its discussion here is more sensible. In SARMA models, the confidence interval at 95% confidence level is produced in the following – familiar – way:

$$\hat{y} \pm 1.96 \cdot s_{\hat{y}},$$

where $\hat{y}$ is the value of point estimation, $s_{\hat{y}}$ denotes standard error, and $t_{1-\frac{\alpha}{2}}(\infty) = 1.96$ is the value of the confidence multiplier at a high (marked by ∞) sample size. The width of the confidence interval in this case, given its symmetry, is $2 \cdot 1.96 \cdot s_{\hat{y}}$, and will obviously be the average width of the confidence interval everywhere else due to the assumption of constant dispersion.

Using a mixture model, it is not directly the standard error that is used (as the error does not necessary have normal distribution and the confidence interval is not necessarily symmetrical at all times, so the mixture feature is inherited by the error term). With the method of calculation described in the chapter on methodology (Section 3.2.5) the confidence interval is calculated as the difference of the interval limits in the following way:

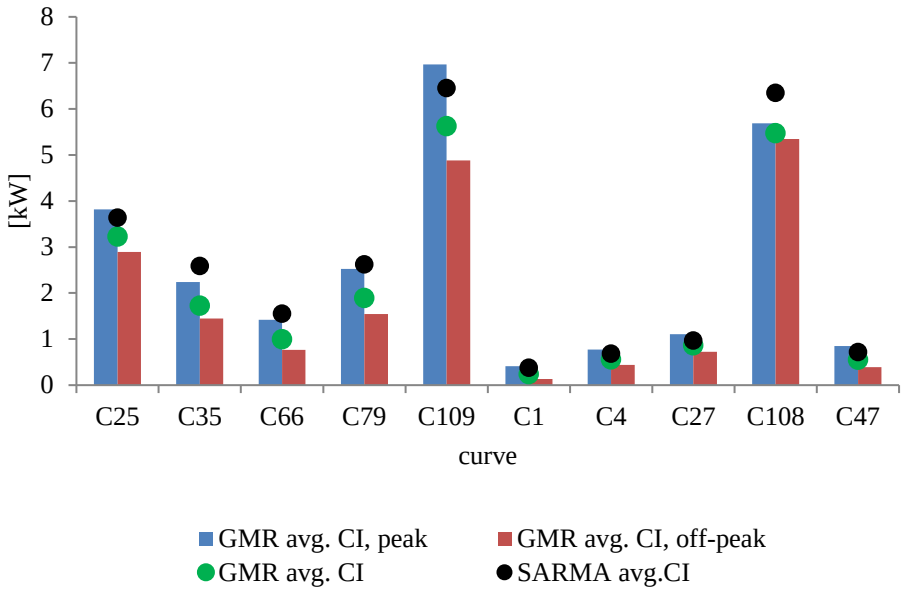
$$\hat{y}_u - \hat{y}_L,$$

where $\hat{y}_u$ denotes the upper limit of the confidence interval and $\hat{y}_L$ is the lower limit of the confidence interval. These differences may vary for any given time, and their averages for each curve give the results appearing in the last column of the table using a mixture model.

In some cases (C35, C66, C79, C1 and C47), the average width of the confidence interval decreases by more than 25%, but a 10% decrease can be found in other cases as well compared to the classical time series regression. It is sensible to show the **width of the confidence-interval in mean value**, but its practical use for a variation of a curve examined in a given interval is relevant as a **sum**. This means that it is possible to give a much smaller interval concerning the daily, weekly, monthly, etc. variation of a curve, at which intervals the actual values occur with a certain probability.

Figure 56 illustrates graphically the variation of the average confidence intervals. As in trading practice it is more sensible to distinguish between *peak* and *off-peak* periods than weekdays and weekends, the columns show the average interval sizes of *peak* and *off-peak* periods using mixture regression. (The width of the confidence interval in case of classical regression is obviously always the same independently of the time period.) As these results (e.g. the risk of *off-peak* may be even only half of the *peak* period risk) are also verifiable from the former figures of quarter-hourly resolution, additional useful information is provided by Figure 57, where these interval sizes are paired with the same period average loads of the curves.

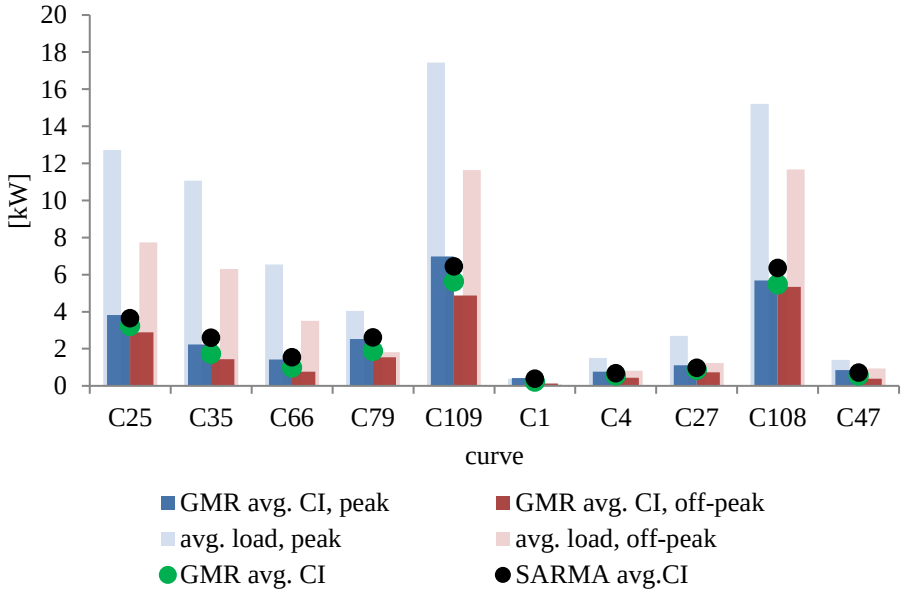
Figure 56: Average confidence interval using SARMA and mixture models for individual curves[135]



Curve	C25	C35	C66	C79	C109	C1	C4	C27	C108	C47
Degree of decrease [%]	11	33	36	28	13	38	18	11	14	24

Source: author's own calculations (R) and figure (Excel).

Figure 57: Average confidence interval using SARMA and mixture models for individual curves (extended figure)



Source: author's own calculations (R) and figure (Excel).

[135] On the figure CI denotes the confidence interval at 95% confidence level.

Following the previous section, it is worth examining the variation of the errors of the SARMA model for each curve, and how much the mixture model could adjust to this by a more consistent modelling of heteroscedasticity. Figures 58-61 show the observations outside the confidence interval in both cases. In some curves the standard deviations of the residuals and the means of conditional standard deviations in Figures 62-71 describe similar tendencies in all instances.[136] In the figures the variation of standard deviations is much smoother – despite having a small sample – compared to the standard deviation of residuals, as the conditional standard deviations are computed (that is, model-based) values. The advantage of the model-based approach is clear, compared to either heuristic methods, or any technique based on residuals generated by a model (such as the previously mentioned kNN-method).

Some relevant results are worth highlighting – without aiming at a comprehensive description – regarding individual curves.

In C25 and C35 it is clear that a spectacular improvement has been achieved regarding weekday-weekend and early morning hours in that the occurrences of residuals outside the confidence interval are more uniform. In the standard errors of C25 it is also easy to see that the winter and summer temperature effects affect the variation of uncertainty during daytime in a different way (more prevalent in morning hours in winter and in the afternoon in summer).

C66 is probably a factory with a ’strict’ working schedule, where occurrences outside the confidence interval using the SARMA model only happen on weekdays in *peak* periods, which was very nicely handled by the model. Besides, summer (otherwise, rather extreme) temperature-dependence results in a somewhat higher standard error in the summer *peak* period.

C79 is special in that the summer behaviour of the curve is highly different from the other periods, primarily regarding the variance. This was the main source of occurrences outside the confidence interval in the SARMA model, and capturing of the ’special’ difference of the covariance structure is nicely presented in this case.

The curve of C109 can also be characterised by summer temperature effect and two daytime *peak* periods in consumption. These two features are reflected in the behaviour of

[136] Only in-sample results are listed here, as the differences between in-sample and out-of-sample results are negligible.

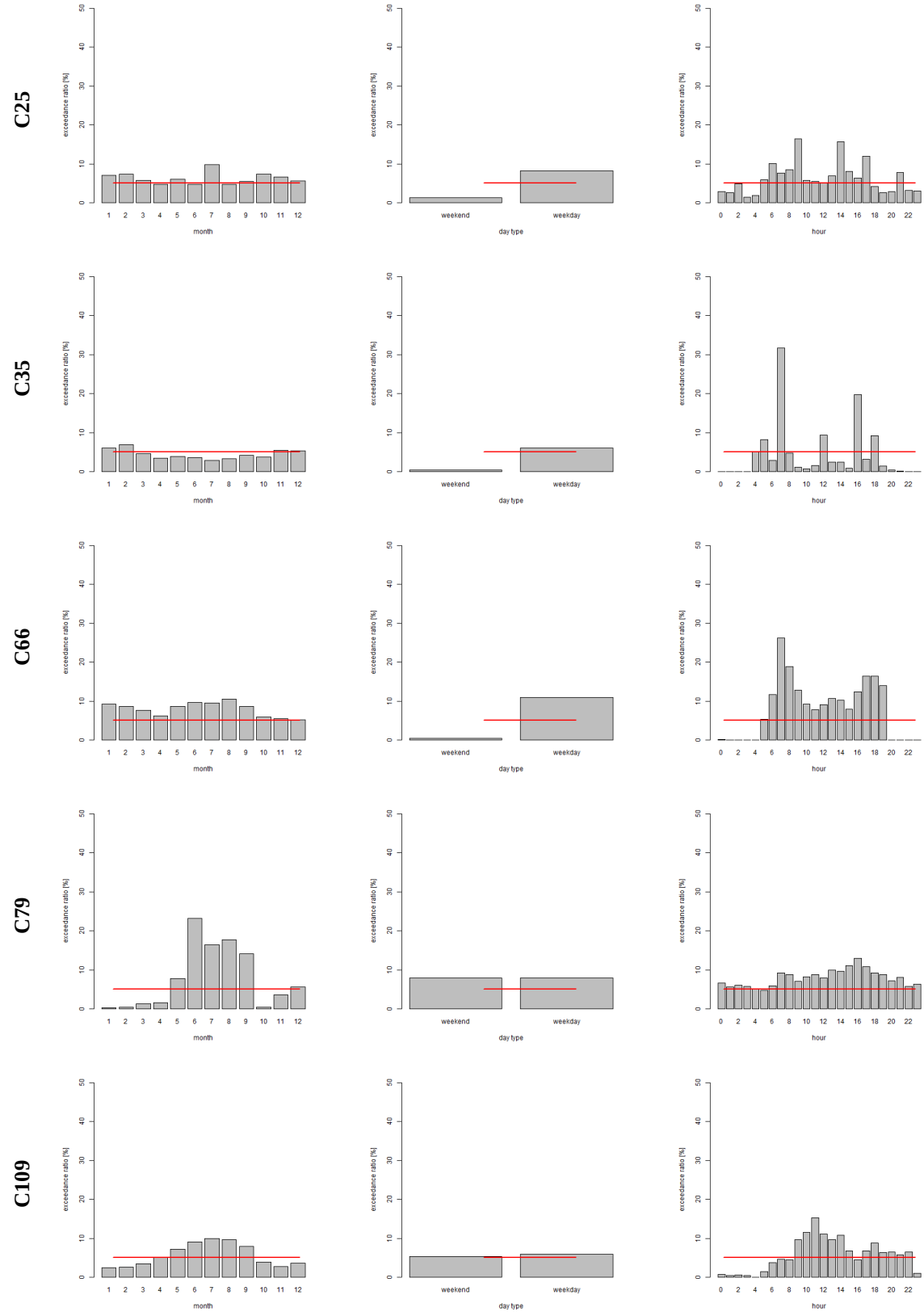
the SARMA model residuals and the means of the conditional standard deviations (anyway, the second *peak* is much smaller).

The higher uncertainty in the evolution of the SARMA model residuals caused by winter temperature-dependence can also be seen in C1. Here, the mixture model would definitely estimate higher standard errors.

As for C27, being a curve with practically no temperature effect, only the different degree of the uncertainties of *peak* and *off-peak* periods can be seen. Concerning C108, the uncertainties of the relatively early morning and late night *peak* periods should be noted.

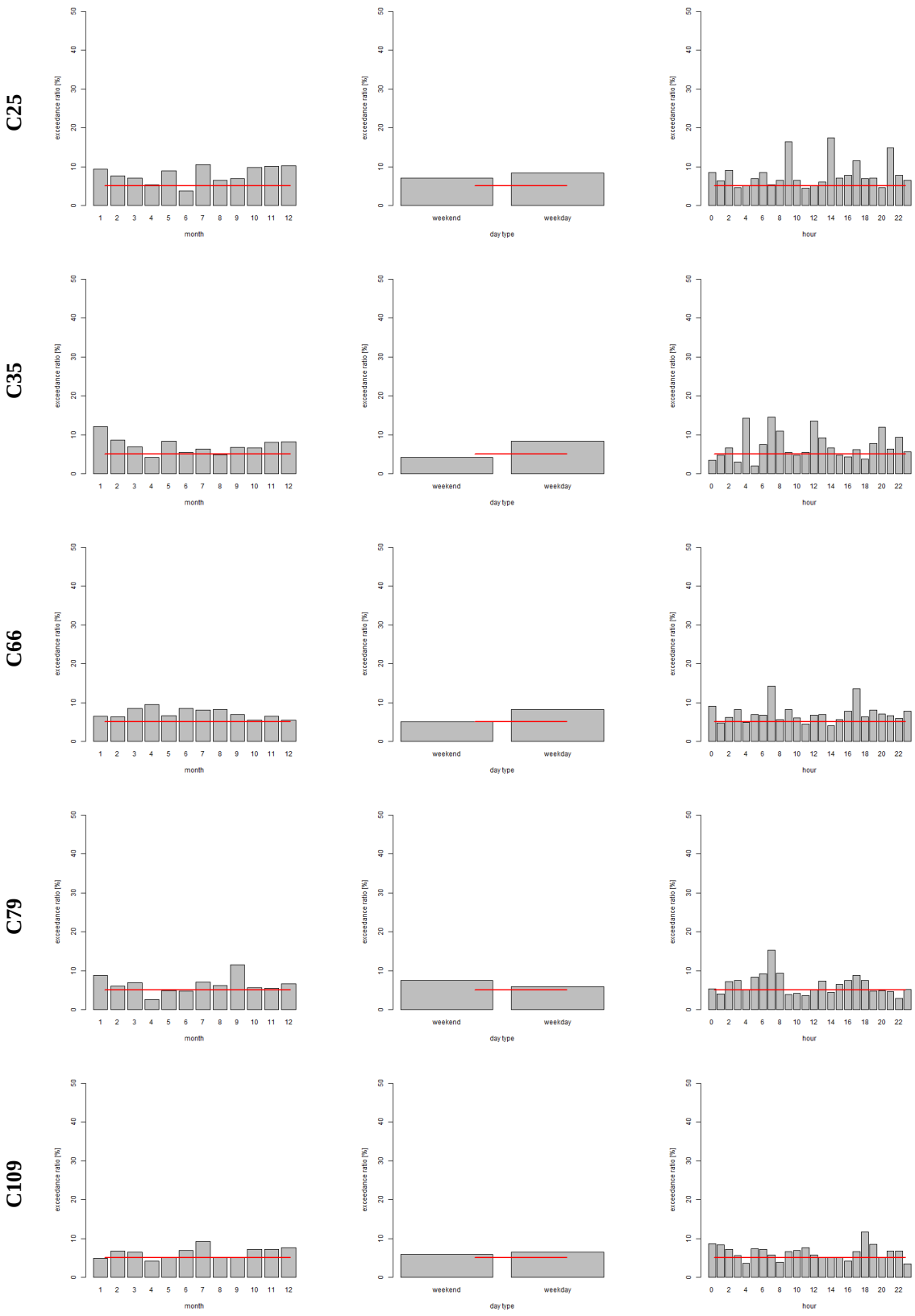
Regarding C47 both temperature-dependence and the uncertainty of the daytime shape are reflected well in the variation of standard errors.

Figure 58: The ratio of observations outside the confidence interval (CI95) for individual curves (SARMA model)



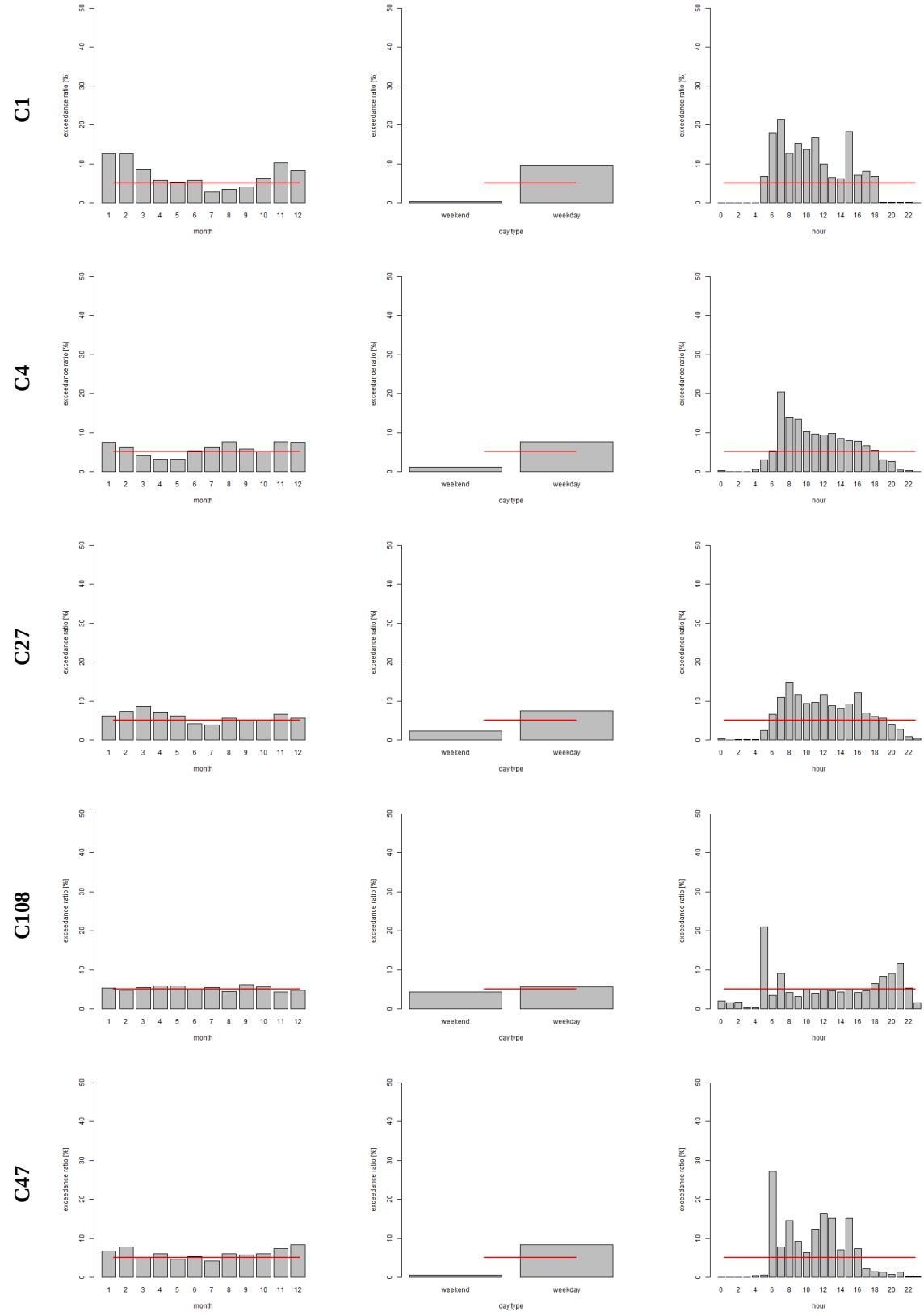
Source: author's own calculations (R) and figures(R).

Figure 59: The ratio of observations outside the confidence interval (CI95) for individual curves (mixture regression)



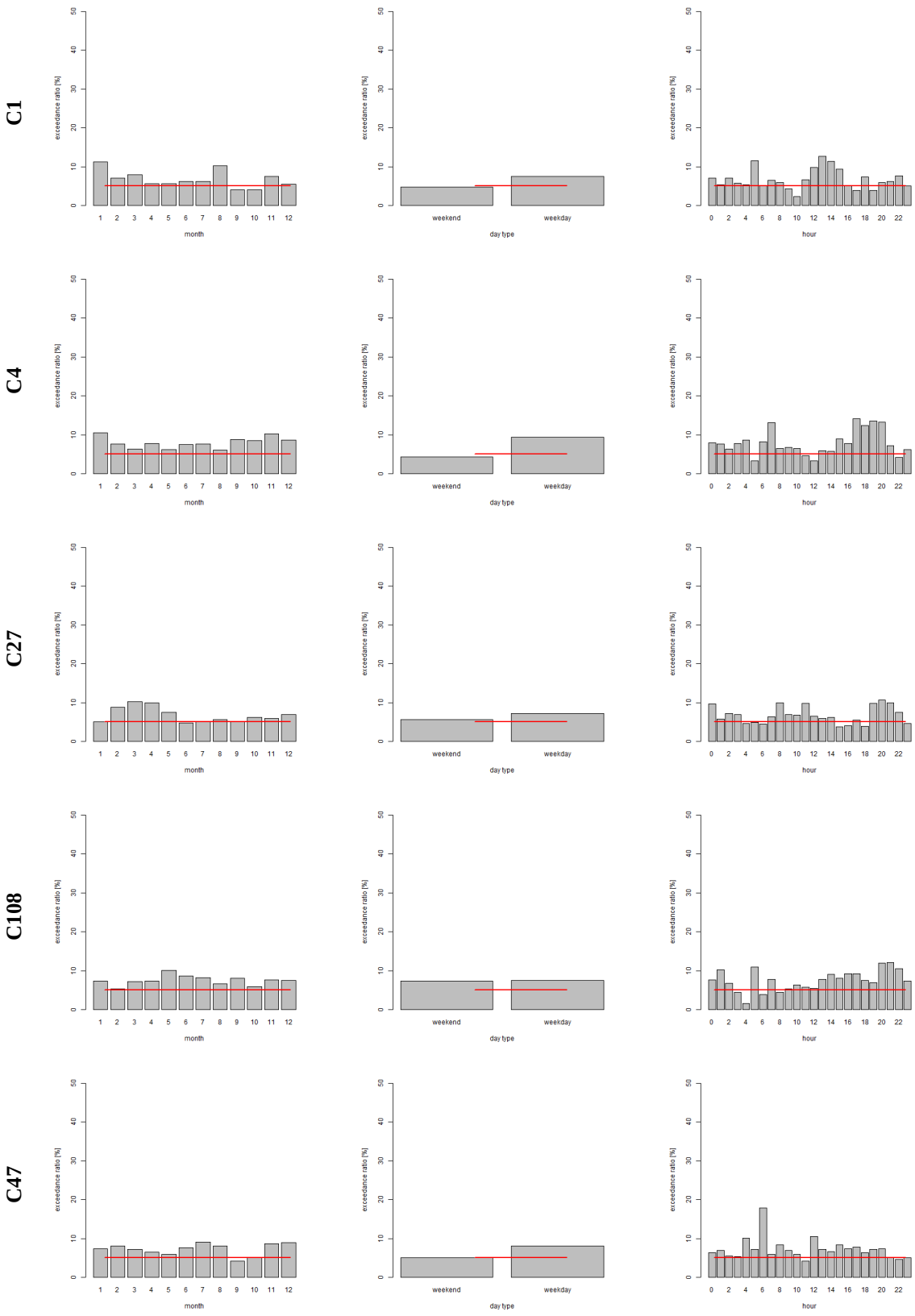
Source: author's own calculations (R) and figures (R).

Figure 60: The ratio of observations outside the confidence interval (CI95) for individual curves (SARMA model)



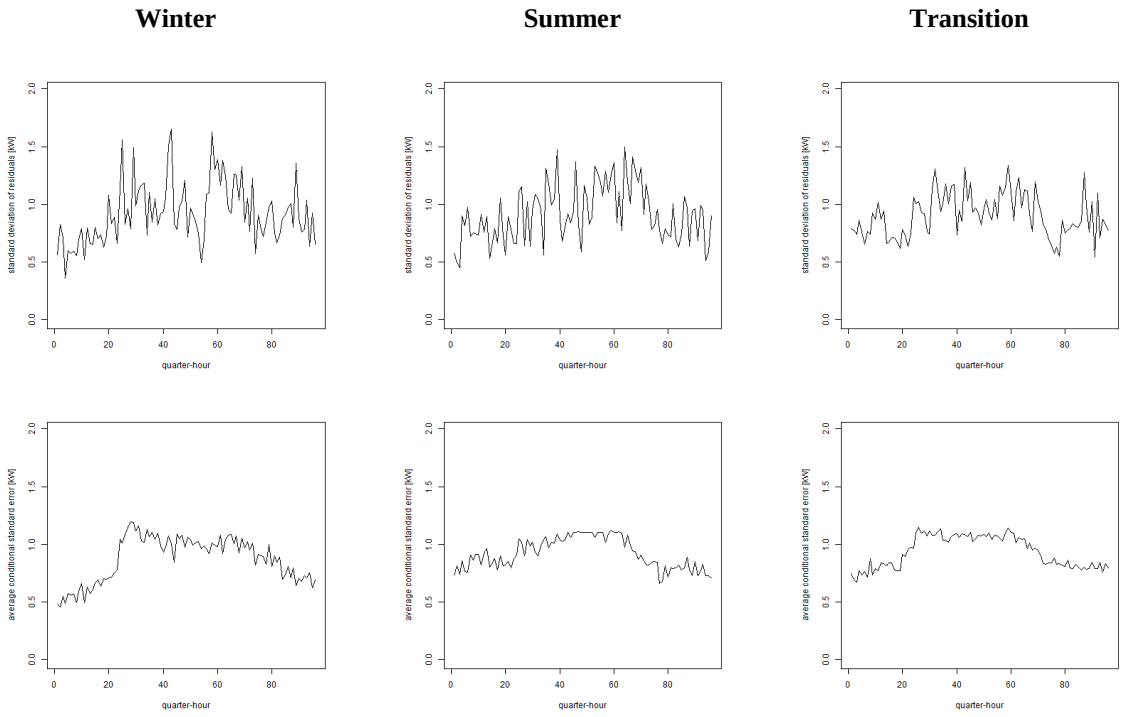
Source: author's own calculations (R) and figures (R).

Figure 61: The ratio of observations outside the confidence interval (CI95) for individual curves (mixture regression)



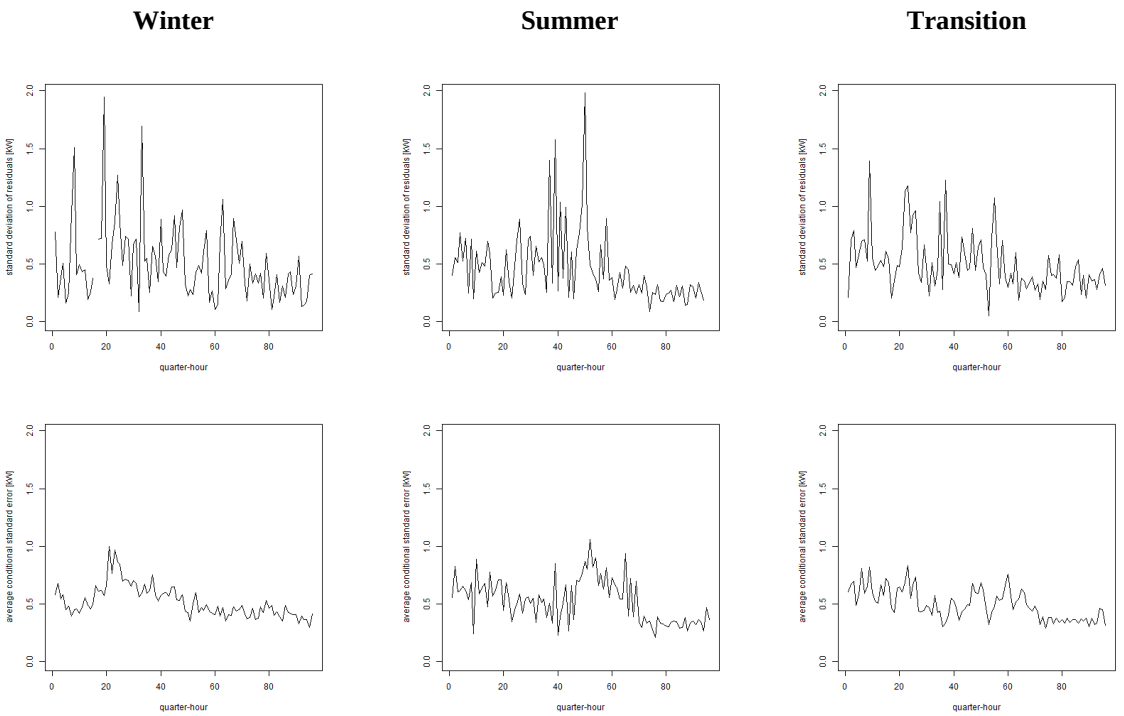
Source: author's own calculations (R) and figures (R).

Figure 62: The standard deviation of mixture regression residuals and the average conditional standard deviations for C25 (in-sample results, weekdays)



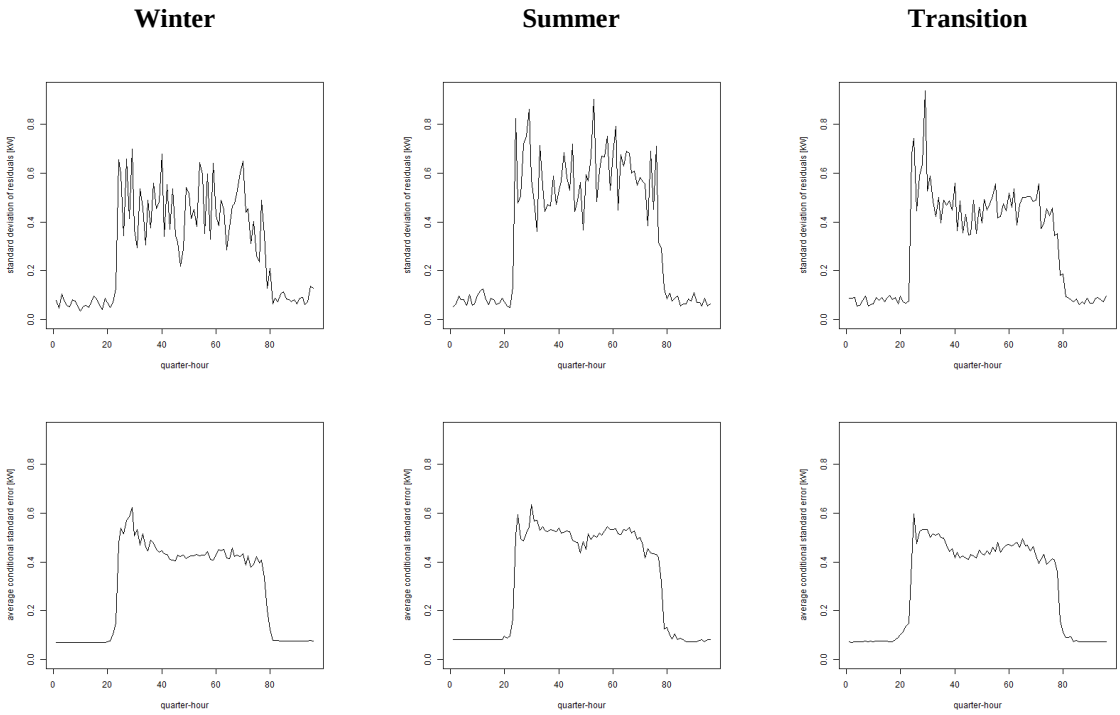
Source: author's own calculations (R) and figures (R).

Figure 63: The standard deviation of mixture regression residuals and the average conditional standard deviations for C25 (in-sample results, weekends)



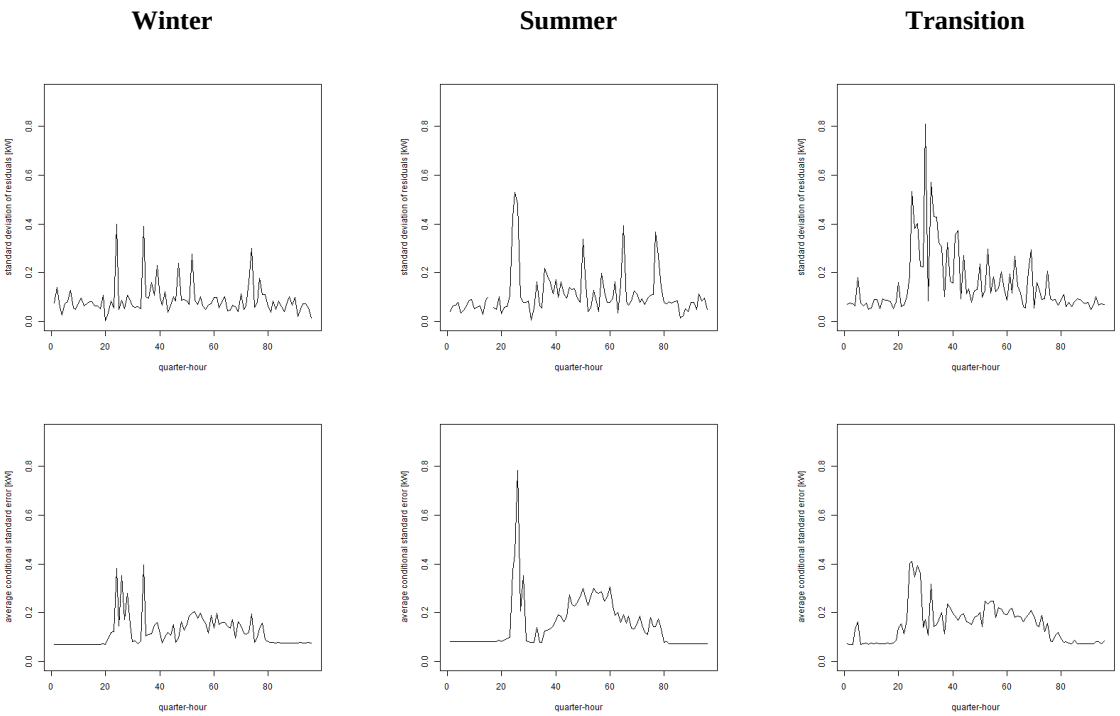
Source: author's own calculations (R) and figures (R).

Figure 64: The standard deviation of mixture regression residuals and the average conditional standard deviations for C66 (in-sample results, weekdays)



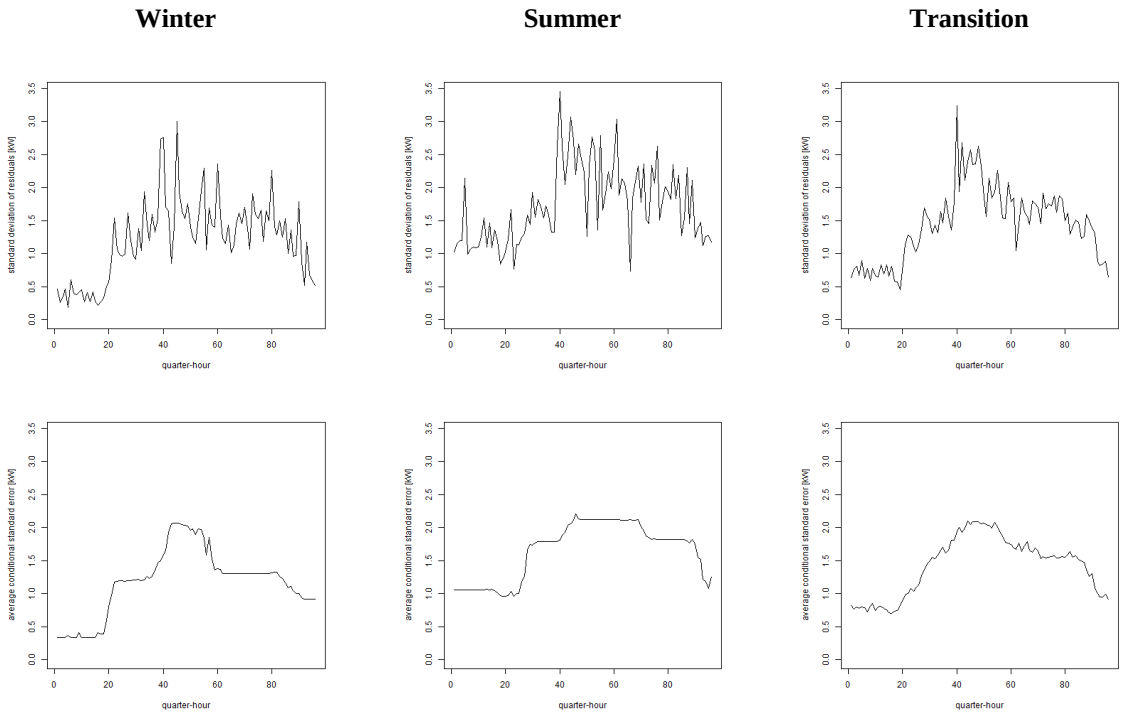
Source: author's own calculations (R) and figures (R).

Figure 65: The standard deviation of mixture regression residuals and the average conditional standard deviations for C66 (in-sample results, weekends)



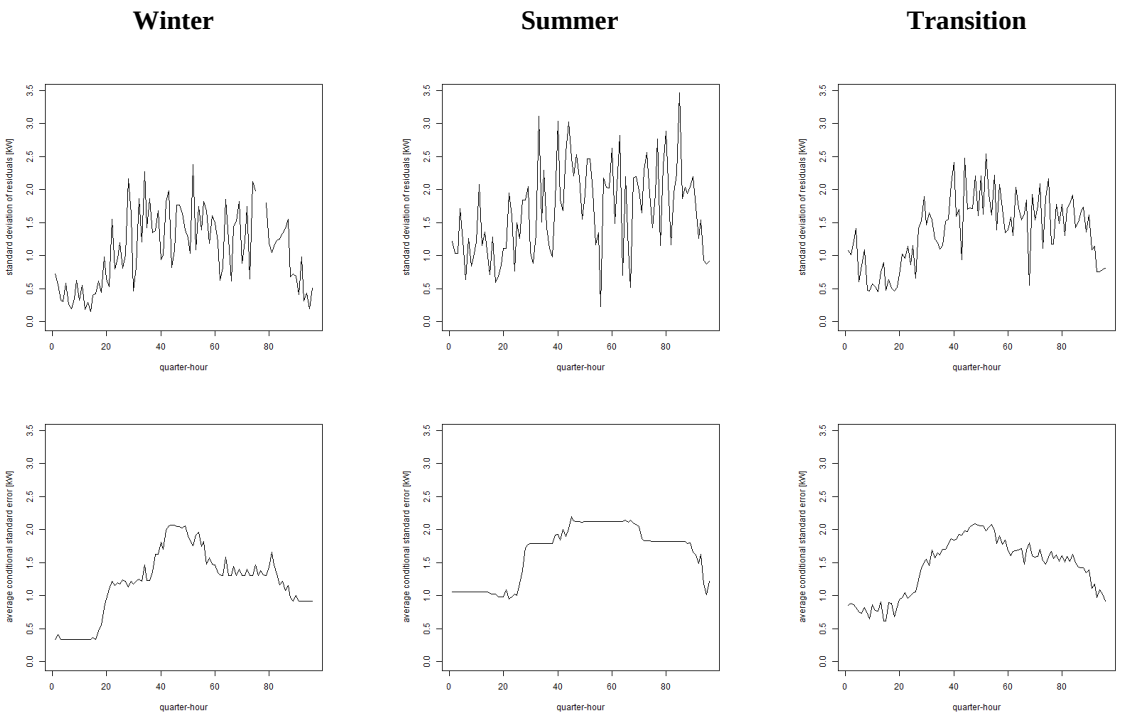
Source: author's own calculations (R) and figures (R).

Figure 66: The standard deviation of mixture regression residuals and the average conditional standard deviations for C109 (in-sample results, weekdays)



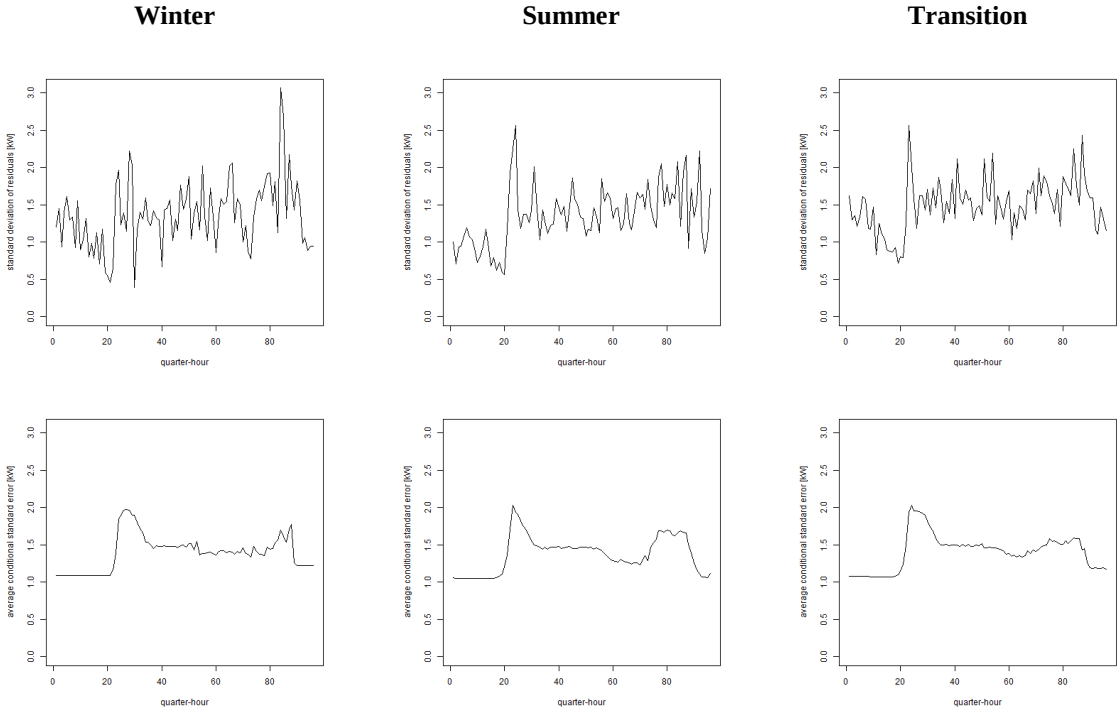
Source: author's own calculations (R) and figures (R).

Figure 67: The standard deviation of mixture regression residuals and the average conditional standard deviations for C109 (in-sample results, weekends)



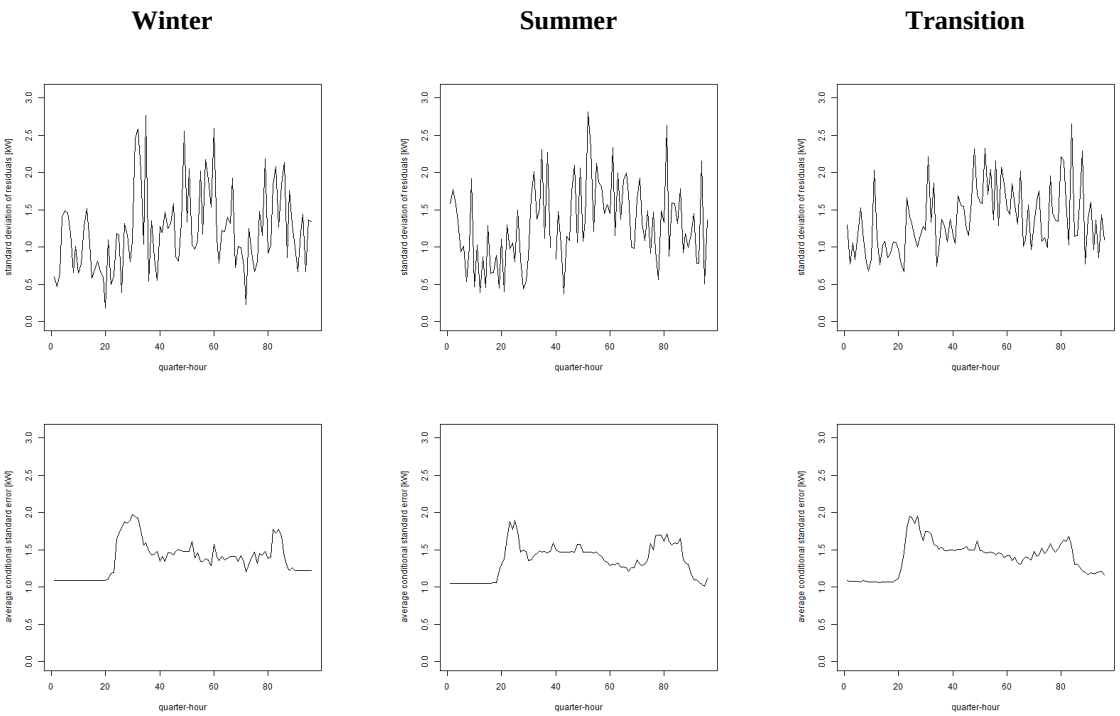
Source: author's own calculations (R) and figures (R).

Figure 68: The standard deviation of mixture regression residuals and the average conditional standard deviations for C108 (in-sample results, weekdays)



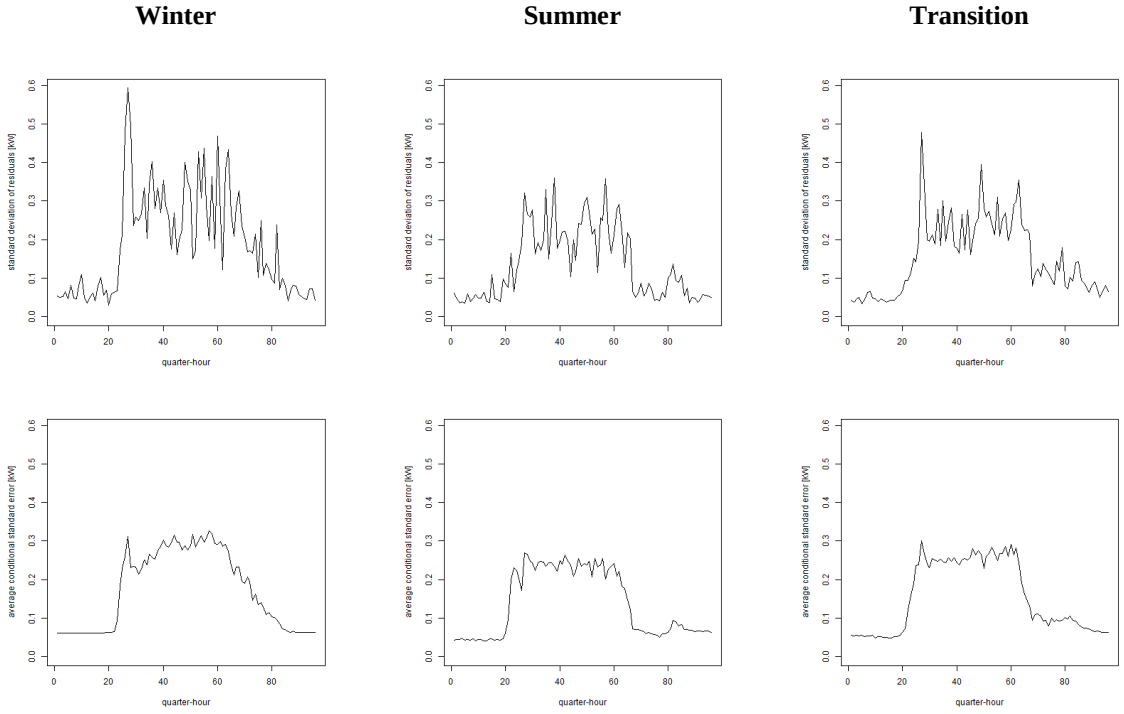
Source: author's own calculations (R) and figures (R).

Figure 69: The standard deviation of mixture regression residuals and the average conditional standard deviations for C108 (in-sample results, weekends)



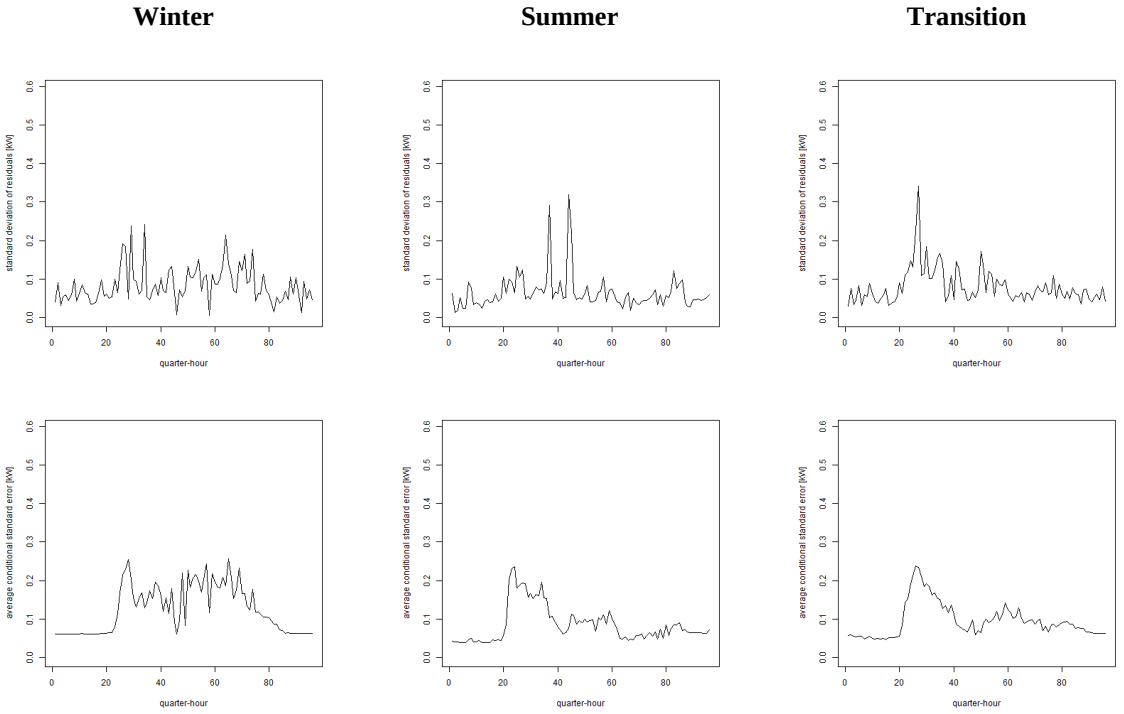
Source: author's own calculations (R) and figures (R).

Figure 70: The standard deviation of mixture regression residuals and the average conditional standard deviations for C47 (in-sample results, weekdays)



Source: author's own calculations (R) and figures (R).

Figure 71: The standard deviation of mixture regression residuals and the average conditional standard deviations for C47 (in-sample results, weekends)



Source: author's own calculations (R) and figures (R).

4.2.2.3. Summary of results and conclusions

This section has examined to what extent the constant standard error produced by classical time series regression models is inconsistent with empirical experience. Using the logic shown in the previous section it was checked how the standard deviation varies in different time periods, and what the ratio of observations outside the confidence interval is. If the hypothesis of heteroscedasticity is correct, at a 95% confidence level 5% of the observations should be outside the confidence interval at all times (with random differences allowed). The results unanimously show that in certain periods the number of observations outside the confidence interval exceed the 5% tendentiously, while in other periods it is far behind it. In this behaviour, however, an obvious regular, seasonal tendency can be noticed. Though it is different by curve, on the whole it is true, that it is multiple seasonality that characterises each individual curve, describing not only the characteristics of consumption, but also its uncertainty and dispersion.

It was investigated whether the derivation of standard errors estimated by mixture regression and the confidence intervals are consistent with each other. It was found that the method described is capable of capturing the heteroscedasticity in time series, and the ratio of observations outside the confidence interval is much more uniformly distributed around the ratio indicated by the confidence level. The errors estimated by mixture regression are slightly different from the errors produced using classical techniques (though there are no major differences in goodness of fit), but this is probably due to the fact that mixture regression can capture nonlinear relationships.

Based on the above, neither hypothesis H3, nor H4 can be rejected. That is, based on hypothesis H4, volume risk is not constant in time, the risk changes in time depending in different exogenous variables, seasonal and calendar effects, that is, it is characteristically higher:

- **on weekdays than at weekends and on holidays,**
- **in weekday *peak* periods than in *off-peak* periods, and**
- **in periods when consumption is weather-dependent.**

Although it seems obvious in hindsight, it is an almost unexpected result that regarding a great number of curves the highest risk was found in the morning ramp and the evening setback.

According to hypothesis H3 it is true that assuming the constant dispersion of errors, the volume risk is underestimated in some periods and overestimated in

others. In connection with this, it was an advantageous outcome that with the estimation of the appropriate (heteroscedastic) model, it was possible to give a much narrower confidence interval for the expected value of the dependent variable on the whole. It is also an empirical experience that the standard errors of classical time series regression are so great, and result in such wide confidence interval that are not meaningful and are inapplicable in practice.

In the variation of standard errors, the position of intraday *peak* periods was continually observable; also, how winter and summer temperature affects the uncertainty of both morning ramp and daytime *peak* periods. Regarding the magnitude of standard errors (whether investigated as aggregated, see Figures 56-57 or seasonally separated, see Figures 62-71) it can be said that depending on the curve, the uncertainty of *off-peak* periods is often half of or even smaller than the uncertainty of *peak* periods. Comparing the uncertainty of these *off-peak* periods with the results estimated by the classical time series regression, the differences in magnitude are even more observable.

SUMMARY OF THE KEY FINDINGS OF THE DISSERTATION

The dissertation focussed on consumer profiles and the modelling of volume risk stemming from the irregular behaviour in the variation of consumption. This chapter summarises what conclusions can be formulated on the basis of empirical results.

A) The examination of stylized facts of consumption time series

Various individual curves have been examined to discover the main features that can determine the characteristics of a curve that need to be considered in the identification of the typical consumption pattern. These were typically not classical statistical tests but simpler calculations or figures that are relatively rarely used in the concise description and characterisation of a curve.

The research results regarding this can be summarised briefly as follows:

- *Contour plots* have been used to examine the **distribution** of load values **throughout a whole year** to explore information such as:
    - how the level of *peak* period, *off-peak* period, weekday and weekend load and the daily position of *peak* periods change throughout the year,
    - to what extent the load is influenced by public holidays,
    - what conclusions can be drawn about the effects of the temperature,
    - and in the case of which curves it is apparent that the so-called '*illumination effect*' (caused by the sunset) can clearly be observed, a phenomenon whose such transparent detection in an empirical study is unprecedented – it is usually only referred to by relying on heuristics.
- *Scatter plots* were used to reveal **weather-**dependency, especially how much it differs by curve in different seasons or on days of the week; besides, how load values are **grouped and clustered** as a function of the temperature.
- *Boxplots* and descriptive statistics (mean, measures of dispersion, skewness and kurtosis) were used to analyse the **distribution** of loads **within a day**, on weekdays and weekends, moreover, in winter, summer and transition periods. They were used to check:
    - how stable or unstable the intraday distribution is by curve,

- how intraday distribution is modified by various seasonal factors or the temperature, and
- when stochastic shocks have a greater role in the case of different curves.

Considering the results of the research, it can be said that the **highest ratio of the variance of the electricity consumption curves can be explained by intraday seasonality, therefore hypothesis H1 cannot be rejected**. On these grounds it has been concluded that creating typical daily profiles – which is common practice – is basically a fine technique, though **typical consumption patterns are not necessarily formed according to the daily shapes and their efficient modelling** is not necessarily carried out along that.

B) Using the mixture model for creating typical consumption patterns

Contrary to traditional techniques, in this dissertation it is not daily load curves that are clustered, but quarter-hourly times (as observations) according to the resolution of the time series, which in turn provide the basis for results that are called typical. In the *Gaussian Mixture Model* (GMM) estimated by the so-called *Expectation-Maximization* (EM) procedure, those times belong to a cluster whose values appear together with the greatest likelihood in the same cluster. The aim of this dissertation is twofold: on the one hand, it aims at constructing typical, representative consumption patterns; but on the other hand, the uncertainty related to consumption is also interesting. For this reason, the co-movement between consumption and its lags or potentially available exogenous variables also receive emphasis.

The above mentioned model-based clustering methodology can handle many of the problems that have surfaced in connection with profiling in a more efficient way than the other well-known techniques, hence:

- Classifying individual consumption patterns has been performed without the need to preadjust time series, among others, for example:
  - removing *outlier*s, or
  - removing the effect of (irregular, extreme) temperature or other weather effect.
- The technique is essentially **multivariate** which – as opposed to the techniques often used in practice and in academic studies – groups the time series values not in

themselves, but together with some variables that describes temporal attributes or seasonality (e.g. weather). This way, of course, the extension with other exogenous variables remains an option. This opportunity does not really appear in most profiling methods, due to the difficulties of preadjustment, among others.

Given the construction of the mixture model, the advantage of the method is that both interaction effects between variables and the nonlinearity are captured without an explicit definition of these effects. This was revealed in the covariance-matrix estimations that were different by cluster, because involving these effects is usually supported by the underlying assumption that the **covariance structure of the variables is not constant in the whole sample**. The latter property is especially important in the regression application of the model, because the simultaneous handling of the (**nonlinear** or **interaction**) effects between the variables and **heteroscedasticity** require more attention. The results are convincing regarding both profiling and the modelling of uncertainty in that the stylised facts that characterise consumption can be fundamentally captured and quantified.

Attention has been drawn to this in connection with a **previous research result** of this dissertation related to Hungarian **natural gas consumption**. It was also stated that either the removal of the **total effect of temperature** or the **irregular effect of temperature** may have many unfavourable consequences, especially if we also intend to model the uncertainty of consumption time series (in the previous case the temperature effect can often not be separated appropriately using the regression decomposition logic; in the latter case, what is removed is the heteroscedastic feature). In addition, theoretical considerations also support the idea that the weather-dependent part of consumption should not be separated (as temperature has a great influence on the value of consumption, and even its uncertainty), but instead, some multivariate technique needs to be used.

It has been shown that the parameters of the mixture model components (the mean and of the covariance matrix of the multivariate normal distribution) can be understood as extracted information which helped cluster and group various individual consumption curves. The results have been **compared** with a technique that may be regarded classical. Measuring distance was performed using the so-called **Kullback-Leibler divergence** which at the same time can be used to measure the distances of the components of each curve.

The formation of profile groups has been performed in this paper especially to prove and illustrate better information extraction. The results have shown that groups are formed rather according to fundamental features that describe consumption, such as, for example the weekday *peak-off-peak* consumption ratio, the level of weekend consumption compared to weekdays, the nature of temperature-dependency (the latter may also influence the seasonal *peak-off-peak* ratio), the position of the *peak* period within a day, etc.

The method has many **favourable features from methodological aspects**. The **typical** consumption that represents clusters can be obtained naturally as a mode (the mean) of the estimated multidimensional normal distribution components. This releases the often occurring problem of what the typical value to represent the cluster should be (it is usually the mean that is used), because the typical, characteristic value is basically the mode. In the same way, it is rather a methodological advantage that the mixture model is not sensitive to having a small sample, as – being a model-based technique – it recognises structure. This feature has already been taken advantage of in the calculations. Another advantage is that the choice of the optimal number of clusters may be selected objectively, through model selection criteria.

A difference, not so much in methodology but rather in approach, is that in the classical case the various category-type variables are basically *dummy* variables encoded in 1-0 values. Conversely, in the mixture model these roles are taken over by components (specifically, component memberships marked z_i replaced by the *posterior* probability of belonging to a component marked p_{ik} during the estimation, see the chapter on methods) – as a consequence, the category-type information can be exploited not only regarding the expected value but also for the description of dispersion.

Based on all of the above results it has been concluded that the profile group formation based on the mixture model gives much more realistic results compared to classical techniques. Besides the numerous advantages of mixture models, they perform grouping with the observations considering not only the expected value, but also the dispersion, that is, basically the uncertainty or risk. Therefore, the related hypothesis H2 has not been rejected either.

C) Using heuristic and classical stochastic time series methods to measure uncertainty of consumption

The investigations that have been performed in this dissertation to measure the irregular behaviour by curve will be described here. Based on the standard errors calculated using classical time series techniques, SARMA and PAR regression, confidence intervals were produced. It has been found that assuming constant standard deviation, on the whole, the uncertainty of each curve is estimated fairly well, but in certain periods the risk is over- or underestimated, and thus the assumption of a constant confidence interval does not fit the empirical findings. This was examined by investigating the **ratio of observations that are outside of the 95% confidence interval**. If the interval is 'correct/appropriate', for every month, weekend and weekday and every (quarter-)hour 5% of the observations should be outside the interval (of course, random deviations are allowed). Experience, however, shows that while depending on the curve, it is generally true that in *peak* periods, in morning ramps and evening setbacks, on weekdays, and in the summer and winter, this is well beyond 5%, and at other times, it is much lower.[137]

Studying the standard deviations of **errors (residuals) in classical time series regressions** the following conclusion has been reached regarding the time-dependent risk of consumption:

- the uncertainty of *peak* period consumption is higher,
- the risk of *off-peak* period consumption is lower,
- in many curves the morning ramps and evening setbacks have the highest uncertainty,
- in periods when consumption is weather-(temperature-)dependent, the risk of consumption is typically higher *ceteris paribus*.[138]

Experience may differ by curve, nevertheless, they are perfectly consistent with the results reached by the calculation of heuristic measures (risk index) that are often used in practice; the major advantage of the model-based approach is its well-grounded nature (see for example the issue of omitted variables, the handling of time-dependency, etc.).

[137] The performance of PAR regression can be suprising in the sense that this method estimates periodically varying autocovariance through periodically (quarter-hourly) alternating autoregressive coefficients, which could partly deal with heteroscedasticity but based on the results of this paper, this time-dependent autocovariance did not prove to be a satisfactory solution.

[138] Weather (temperature) – as we know – is a stochastic variable in itself, and the gain of not removing the effect of temperature in profiling is reflected in this important statement.

More accurate or grounded statements than the above cannot be formulated, due partly to the fact that ‘grouping’ residuals (based on seasons, days of the week) is not forward-looking in any way; moreover, as a consequence of the noisy, hectic nature of the calculated results mainly questionable statements can be made even using time series models.

Hypothesis H3 then has not been rejected, that is, depending on the time, risk is either under- or overestimated for each curve in classical regression approaches that assume constant standard deviation for the error term.

D) Using mixture models to measure the uncertainty of consumption

Based on the summary of experiences, the regression application of the mixture model (which was also used for profiling) has produced so-called **conditional, time-dependent standard errors** and confidence intervals that are in line with the risks of consumption.

It has been investigated how much the confidence intervals produced based on mixture regression meet the requirements (the 95% confidence intervals have also been calculated for mixture regression), or in different terms, how much the standard deviation of errors (e.g. calculated for hours, weekdays/weekends, months) is consistent with the standard errors calculated on the basis of mixture regression.

Based on the results, it has been shown that mixture regression can represent the time-dependent uncertainty of consumption (the ratio of observations that are outside the confidence interval is much more consistent than with classical models), and generally they are roughly identical with the expectations formulated on the basis of heuristic measures and SARMA model residuals. The source of differences in this case may basically be that the SARMA model is linear, while mixture regression is not, therefore, a better capturing of nonlinear relationships may produce slightly different results.

An advantage of using mixture regression is that standard errors **can be written as functions of the independent variable, that is, with the condition of the independent variables**, in this way, writing the seasonal behaviour of the uncertainty of the consumption with the same variables as the seasonal behaviour of the consumption curve itself. The standard errors reflect not only which periods show higher uncertainty within the day, week or year, but also though to a different degree for each curve, that the winter temperature increases rather uncertainty of the morning periods, while the summer increases the uncertainty of the afternoon periods.

Based on classical and mixture regression calculations, **H4 hypothesis has been accepted, that is, it is true that the consumption risk is typically not constant in time, it is higher on weekdays, in *peak* periods and also in weather-dependent periods; that is, it is characterised by multiple seasonality, as is consumption itself.**

The **importance of the** results lies in that

- the profile and the uncertainty of consumption (modelling of volume risk) is performed in a unified framework,
- the application of mixture regression offers promising results, and its energy market use can be regarded relatively new,
- in such regression applications of the mixture model, the *backtest* of the results has not occurred in any earlier study. The regression application itself (and certain steps of the clustering) is not directly performed by using a publicly available *R Project* package, therefore, its implementation also formed part of the study.

Beyond what has been discussed in the formulated hypotheses, it is an important result that the **width of the confidence interval produced by mixture regression is (though to a different degree in each curve) much smaller than what is arrived at using classical techniques**. The latter result is also important because the confidence interval that is produced in time series models is in the majority of the cases very wide, and is not really suitable for practice. This chapter has also examined how the average width of the confidence interval changes in *peak* and *off-peak* hours, compared to the average loads of this period.

Because of the averaging obviously only approximations are possible, but most certainly the uncertainty of prices must have similar features. This inevitably draws the attention to specific outstanding goals of demand side management and their necessity; thinking about the smoothing of the consumption curve, for example, or the decreasing of the balancing energy costs resulting from actual-planned deviations or even the decreasing of *peak–off-peak* ratio resulting from shifts in energy use.

It definitely needs to be added to the evaluation of the results that the methods examined and used have made it possible to not only investigate and measure when the uncertainty of consumption is highest, but also to what extent. It is also essential for the potential applications (whether in connection with classical field or the field of demand side management).

AVENUES FOR FURTHER RESEARCH AND APPLICATION IN PRACTICE

There have been various references in the empirical part of the paper to the use of the results in practice, and in connection with this to new directions for further research. These will be summarised in this chapter.

A possible further research opportunity may be the examination of profile groups based on mixture models on hundreds or thousands of curves, and the completion of comparative analysis with classical techniques. In the dissertation, the emphasis is much more on the uncertainty-related evaluation than on an analysis of such great amount. It is definitely worth examining how much the profile grouping changes as a result of the fact that the mixture model essentially extracts the whole information from the curve. **The emphasis is from the methodological – and practical – perspective on better information extraction and the exploitation of capturing uncertainty**. In the course of such an extended study, of course, a number of questions may arise in connection with grouping; such as choosing the optimal number of clusters, examining the indicators that evaluate the appropriacy of the result of the clustering, etc. It is necessary to examine these in dealing with such a huge dataset.

In practice, it is often a problem that the time series is not available for the whole year in the case of individual consumption curves. As the mixture model is less sensitive to **small sample size**, it is worth examining – within sensible limits – whether it is more efficient or whether it yields more applicable results compared to more sample size sensitive solutions in such cases where the information is available only for a fraction of a year.

Although it is true that mixture regression produced consistent standard errors with residuals, globally, the ratio of observations outside the confidence interval is still higher than what is expected based on the confidence level (however, it has been shown that the performance of SARMA models is roughly similar). It is worth investigating if working with **the mixture of other distributions** instead of the normal distribution produces better results. The fact that the ratio of observations outside the confidence interval is higher than what is explained by the confidence level points to the necessity of using fat-tailed distributions. Testing this hypothesis and seeking a general, easy-to-apply technique for such an amount of a heterogeneous set of curves is definitely an exciting research task.

Another possible direction for further research involves the inclusion of **further weather (or other types of) variables** besides temperature – even for the handling of the phenomenon mentioned above. Regarding weather variables it is of course inevitable that their quality is appropriate, because even the literature is not uniform in this respect – such as in the case of temperature (often not even concerning the existence of the relationship). Nevertheless, as the effect of the temperature is by far the strongest, in longer term planning or in the planning of the yearly consumption of a consumer, temperature may be enough. The inclusion of other variables can only have real benefits if it is, for example, separately measured energy use that needs to be modelled. All of this, of course, requires appropriate technical infrastructure – even in terms of the frequency of metering (recording data) in the case of both consumption and exogenous variables.

In this dissertation it has often been stressed that 'classical' profiling techniques applied on curves after having removed weather effects show fewer options for progress; especially considering that **the weather-dependent part of consumption is more difficult to influence**, and is more price-inflexible. These results in connection with measuring weather-dependent uncertainty are definitely a useful starting point for related studies.

What is definitely a promising opportunity for further development in the future is the examination of the **portfolio effect** mainly with regard to modelling the correlation between error terms. An approximate estimation of this may be the calculation of linear correlation coefficients for various periods. Based on the example in Appendix F) it is deemed likely that the degree of the diversification of volume risk is time-dependent, as the correlation of residuals[139] is also time-dependent. Nevertheless, quantification may be possible in the framework of the mixture model. As mixture models estimate the components of variables with different covariance structures, and this is transformed to errors as well, it may ease modelling of the co-movement of errors, covariance – that is, essentially the portfolio effect – in one single step.

For every statistical model it is important to evaluate the out-of-sample performance. This dissertation provided only limited opportunities to do so, as only yearly curves were available. The evaluation of static forecasts (that is forecasts for one period ahead) has essentially taken place; therefore, an especially interesting field is the creation and evaluation of dynamic **forecasts** (that is forecasts for multiple periods ahead).

[139] Even its significance, or the lack of it.

Besides the above, there are further potential fields of research that are slightly different from the focus of this dissertation, but need to be mentioned here. The chapter on the previous research results has – for example – mentioned a technique (also empirically reproduced here) where each daily curve has been modelled as a mixture of the normal distribution density functions. The method can potentially **estimate the time of *peak*** within a day. Nowadays it is gaining an important role as there are many such tendencies (such as the spread of electric cars) that – if they gain greater volume – can fundamentally reshape system level daily profile with the shifting of daily *peak*s – both in time and magnitude.

It is worth noting that the formulation of the first hypothesis of this paper was induced by the fact that profiling uses basically daily profile curves. As the highest ratio of the variance explained of the curves is by intraday seasonality, these techniques do not provide such misleading results in the case of electricity curves. As a consequence it is worth examining **other energy sectors** (such as natural gas, where for many consumers the heating effect is dominant) not only from the perspective of profiling, but also regarding volume risk, where an important proportion of the variance is not dominated by the intraday seasonality, but by the weather. Here, mixture model based profile may have an even greater benefit compared to classical techniques than what has been shown in this paper.

Likewise interesting is the field of examining the uncertainty of the **supply side** (in electricity markets, basically the power plants) in addition to the demand side. The difficulty here often lies in the fact that in the case of weather-dependent suppliers it is necessary to have local, *onsite* weather data (wind speed, solar radiation, cloud coverage, humidity, etc.) measured on the place of production; as the information from classical meteorological data services is often not quite appropriate. At the same time, the fundamental exploration of nonlinearity or the interaction effects between variables and the simultaneous quantification of uncertainty (see for example the evaluation of the reliability of production schedules) is a requirement here as well, and the examples for the simultaneous modelling of both on the supply side is scarce.

In connection with this, it is also important to **match both the demand and the supply** side both in profile and in the uncertainty of the profile. It is especially important to highlight here the increasing spread of domestic smart metering in the future, where the quantity of data to become available – with the more exact knowledge of the behaviour of

small consumers – will provide useful additional information on the evaluation of domestic (household) energy production projects.

APPENDICES

A) Statistical software packages and the most important functions used for calculations

The majority of calculations in this dissertation were performed using the **R Project** statistical software package. This section will give an overview of what *packages*, functions were used. This is important to have a clearer view of the methodological arc of the dissertation; in addition, there are elements of the calculations for which it was necessary to implement commands and calculations in so-called user-defined functions (for example: the regression approach that builds on the mixture model). There are empirical results in the dissertation that were not arrived at by using R, which will also be highlighted here.

The most important methods used for a great proportion of the empirical studies in this paper were realised by the following *packages* (Chapter 4):

- for SARMA regressions (*seasonal autoregressive moving average regression*) the *package* '*stats*'[140] was used,[141]
- PAR regressions (*periodic autoregressive regression*) were made using *package* '*partsm*'[142],
- *Gaussian* mixture models were estimated by the EM (*Expectation-Maximization*) procedure using the *package 'mclust'*[143],
- the commands of the EDDA discriminant analysis based on the *Gaussian* mixture model are the author's own functions building on the results of '*mclust*' as this is not an in-built function in the R *package*,
- the commands of the regression based on the *Gaussian* mixture model are the author's own functions building the results of *'mclust'*, as there are no functions for regression applications of the *Gaussian* mixture model in the R *package*.

[140] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html
[141] In a few cases – in the earlier phases of the study – in the making of SARMA regressions, **Gretl** was used (http://gretl.sourceforge.net/), which is excellent software for econometric, regression solutions, and is also open source.
[142] https://cran.r-project.org/web/packages/partsm/index.html
[143] https://cran.r-project.org/web/packages/mclust/index.html

Further functions of the empirical results with smaller emphasis or functions used in the representation, description of previous research results found in the literature are the following:

- fitting of the mixture density functions (Section 3.3) was made using the *package* '*mclust*',
- *package* '*stats*' was used for K-Means clustering (Section 3.2.3) and hierarchical clustering (Section 4.1),
- *package* '*graphics*'[144] was used to make *contour plot*s (Section 1.4),
- *boxplot*s were made using *package* '*ggplot2*'[145] (Section 1.4),
- the *package* '*monomvn*'[146] was used for the calculation of the Kullback-Leibler divergence in hierarchical clustering; the estimation applied to the mixture of normal distributions was based on the author's own function using *Hershey-Olsen* [2007].

The results of the irregular effects of temperature (and their effects on seasonal adjustment) were produced by X13-ARIMA seasonal adjustment method – also from one of the previous studies by the author, using the **X13-ARIMA-SEATS**[147] seasonal adjustment software (there is also an interface *package* called *'seasonal'*[148] in R, through which the functions of X13-ARIMA-SEATS can be called); Figure 27 (Section 3.1.3) used to visualise heteroscedasticity is the result of the SARIMA regression made in the **EViews**[149] econometric software.

[144] https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/00Index.html
[145] https://cran.r-project.org/web/packages/ggplot2/index.html
[146] https://cran.r-project.org/web/packages/monomvn/index.html
[147] https://www.census.gov/srd/www/x13as/
[148] https://cran.r-project.org/web/packages/seasonal/index.html
[149] http://www.eviews.com/home.html

B) Empirical example of the natural gas consumption data of Budapest – some calculation results

This appendix contains some results of the K-Means and mixture clustering described in Chapter 3 on the natural gas consumption data of Budapest.

Table 19: BSS/TSS ratios in K-Means clustering on the example of daily average temperature – natural gas consumption

Number of clusters	2	3	4	5	6	7	8	9	10
BSS/TSS (%)	79.7	90.2	94.4	96.4	97.8	98.2	98.3	98.7	99.0

Source: author's own calculations (R) and table.

Table 20: BIC criteria in mixture clustering on the example of daily average temperature – natural gas consumption

Number of clusters	1	2	3	4	5	6
BIC	17 127.6	16 328.3	16 221.0	**16 213.2**	16 254.4	16 304.1

Source: author's own calculations (R) and table.

Table 21: Cluster centroids in mixture clustering on the example of daily average temperature – natural gas consumption

Variable / Cluster	1. (blue)	2. (black)	3. (red)	4. (green)
temperature [°C]	-1.45	7.57	14.89	20.31
natural gas consumption [thousand m^3]	9 376.74	5 889.05	2 211.55	1 184.04

Source: author's own calculations (R) and table.

Table 22: The distribution of days among clusters and months on the example of daily average temperature – natural gas consumption (mixture clustering)

Cluster/Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	Total
1. (blue)	36.4	31.8	0	0	0	0	0	0	0	0	4.6	27.3	100.0
2. (black)	16.4	15.0	17.9	5.0	0	0	0	0	0	7.1	20.7	17.9	100.0
3. (red)	0	0	7.6	29.1	26.6	0	0	0	10.1	26.6	0	0	100.0
4. (green)	0	0	0	0	8.1	24.2	25.0	25.0	17.7	0	0	0	100.0

Source: author's own calculations (R) and table.

Table 23: The distribution of days among clusters and weekdays/weekends on the example of daily average temperature – natural gas consumption (mixture clustering)

Cluster/Day of the week	Weekday	Weekend	Total
1. (blue)	72.7	27.3	100.0
2. (black)	74.3	25.7	100.0
3. (red)	67.1	32.9	100.0
4. (green)	71.0	29.0	100.0

Source: author's own calculations (R) and table.

C) Examination of stylized facts of load time series

The appendix contains some supplementary results related to Section 1.4 on the examination of stylized facts of load curves. Table 24 shows the descriptive statistics of the weekly time series of the Hungarian system load, which is worth examining in parallel with the *boxplot*s found in Section 1.4.

Table 24: Descriptive statistics of the weekly time series of the Hungarian system load in daily resolution[150]

Measure	Season	Mon	Tue	Wed	Thu	Fri	Sat	Sun
$\bar{Y}$ [MW]	winter	5 396.45	5 501.18	5 539.88	5 408.05	5 340.32	5 124.29	4 887.36
	summer	5 151.63	5 314.92	5 142.26	4 837.76	4 929.93	4 476.53	4 230.46
	transition	4 899.53	4 883.30	4 881.34	4 869.16	4 795.88	4 384.30	3 827.43
Q_1 [MW]	winter	4 821.07	5 043.25	5 015.85	4 891.91	4 914.03	4 796.97	4 488.83
	summer	4 443.42	4 731.33	4 759.73	4 519.87	4 586.45	4 237.23	3 861.42
	transition	4 480.57	4 707.46	4 663.91	4 601.99	4 561.74	4 209.40	3 694.68
Me [MW]	winter	5 798.47	5 835.29	5 883.33	5 772.15	5 638.47	5 283.88	4 997.30
	summer	5 492.63	5 568.60	5 352.26	4 992.08	5 038.09	4 573.68	4 455.15
	transition	5 202.25	5 098.33	5 117.09	5 072.33	4 978.13	4 481.00	3 862.72
Q_3 [MW]	winter	5 900.50	5 911.02	6 064.95	5 916.05	5 698.83	5 521.76	5 236.65
	summer	5 766.76	5 901.37	5 512.43	5 232.89	5 362.40	4 795.27	4 555.81
	transition	5 311.14	5 238.27	5 241.25	5 277.33	5 182.99	4 634.17	4 097.67
σ [MW]	winter	753.89	621.33	672.01	674.40	594.00	488.13	524.51
	summer	754.41	694.19	519.59	490.81	541.02	416.50	452.93
	transition	596.95	477.43	506.92	520.90	474.31	377.57	368.16
V [%]	winter	13.97	11.29	12.13	12.47	11.12	9.53	10.73
	summer	14.64	13.06	10.10	10.15	10.97	9.30	10.71
	transition	12.18	9.78	10.38	10.70	9.89	8.61	9.62
α_3	winter	-0.93	-0.87	-0.80	-0.86	-0.93	-0.51	-0.29
	summer	-0.78	-0.62	-0.85	-0.84	-0.76	-0.69	-0.67
	transition	-1.03	-1.05	-1.07	-0.97	-0.99	-0.78	-0.58

Source: author's own calculations (R) and table.

Table 25 shows the between variance / total variance ratio (in the Hungarian literature the so-called H^2-measure) values of months, days and hours/quarter-hours as grouping variables for some curves, not only for portfolios, but for individual curves as well.[151]

[150] Descriptive statistics in Table 24 are the following: mean, lower quartile, median, upper quartile, standard deviation, relative standard deviation, and α_3 measure of assymetry. As these are all familiar measures, only the latter two will be explained here. Relative standard deviation is the ratio of the standard deviation and the mean. As it measures dispersion in percent, the various curves are comparable in terms of where the degree of dispersion is highest. α_3 is the so-called moment-coefficient of skewness, its positive values mean (*positively)* right skewed, and negative values mean (*negatively)* left skewed distribution, its values near zero indicate approximately symmetrical distribution.

Table 25: Variance ratios explained by seasonal variables for curves

Curve	Grouping variable			
	Months	**Days**	**Hours**	**Quarter-hours**
C25	3.45	49.76	11.48	11.87
C66	0.46	18.63	52.52	52.84
C96	4.94	0.51	69.90	70.19
C109	12.52	1.38	62.14	62.84
Portfolio	3.66	23.72	48.07	48.37
System load	9.68	11.59	60.84	61.50

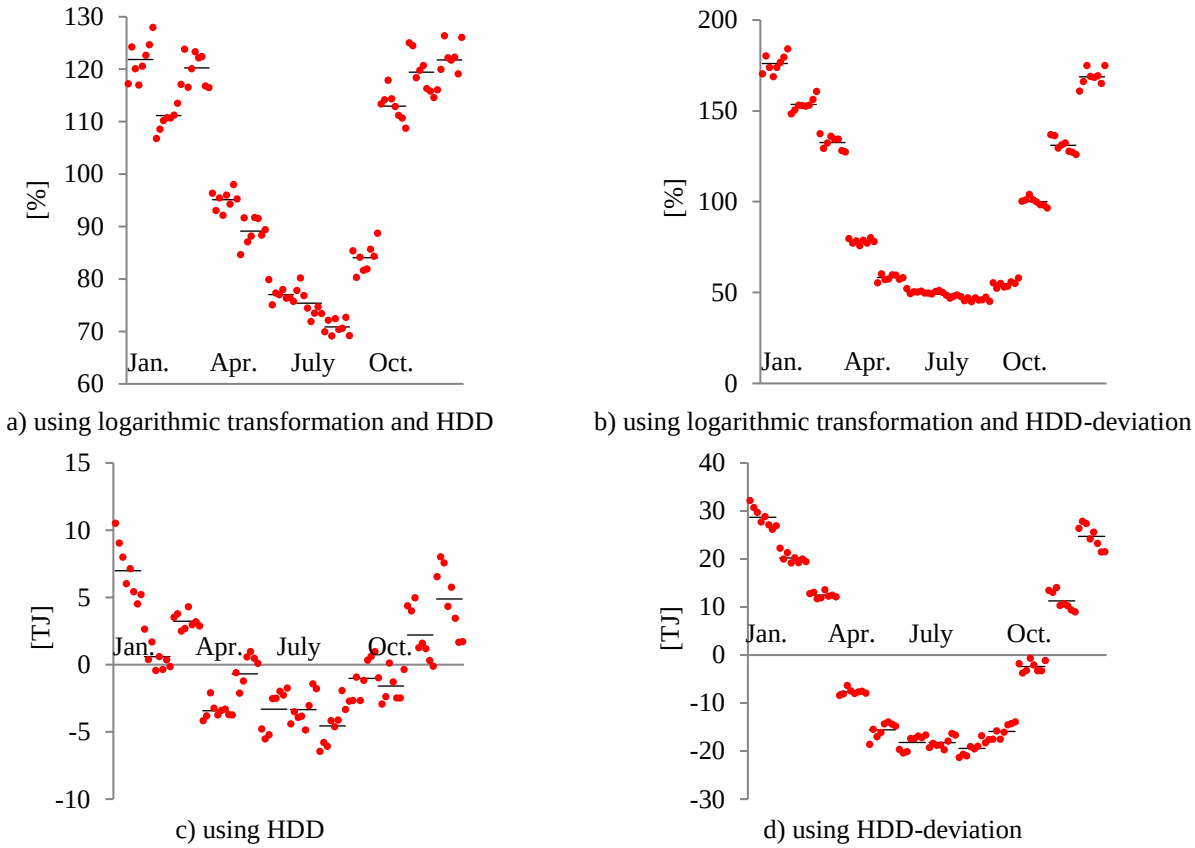
Source: author's own calculations (R) and table.

[151] This task needs to be thought of as a relationship where the continuous variable is the curve time series and the grouping variables are months, days of the week, and hours, quarterhours. Based on this, it is possible to calculate the total, within and between error sum of squares, based on which it can be identified what proportion of the variance of the curve's time series is explained by the grouping variable.

D) SI ratios in the seasonal adjustment of national gas consumption

The appendix describes the SI ratios produced in the seasonal adjustment framework shown in Chapter 2; depending on how the dependent variable was included (using logarithmic transformation or not), and with the different treatments of degree days (using HDD or HDD-deviation). For a better comparability of results, for each degree day variable only one coefficient (beta parameter) was estimated in each model setup, thus the increase of more estimated coefficients does not distort the results here regarding the best fit.[152] The following figure shows the SI ratios calculated for four different model setups[153].

Figure 72: SI ratios in different seasonal adjustment model setups



a) using logarithmic transformation and HDD

b) using logarithmic transformation and HDD-deviation

c) using HDD

d) using HDD-deviation

Source: author's calculations (X13-ARIMA-SEATS, Excel) and author's own diagrams (Excel).

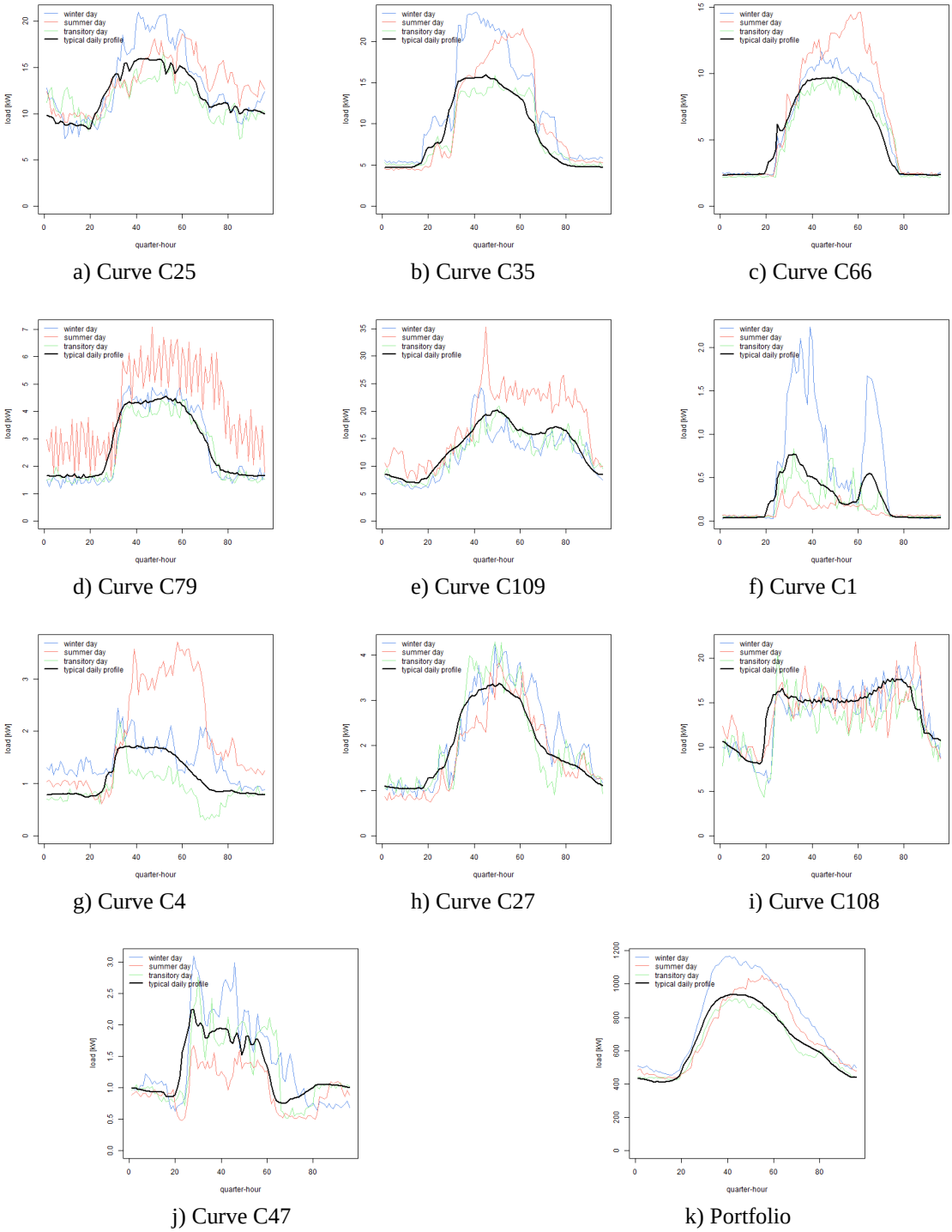
[152] Whether the dependent variable needs to undergo logarithmic transformation, and how degree-day should appear can be tested for within the framework of the software package through model selection criteria. Based on this, the recommended model is the one that appears in the main text (logarithmic transformation and the estimation of the monthly different HDD-deviation coefficients; see Figure 24 in the main text and Figure 72. b) here).

[153] In an additive model (in the above example when the dependent variable value is not logarithmised) these appear removing the trend component not as executing a ratio but as a difference – this problem in using precise terms will be disregarded here (and is disregarded in the literature as well).

E) Typical daily profiles and weekly load time series figures

This appendix contains some time series figures of curves studied in Chapter 4.

Figure 73: Typical daily profiles of individual curves and the portfolio



a) Curve C25 b) Curve C35 c) Curve C66

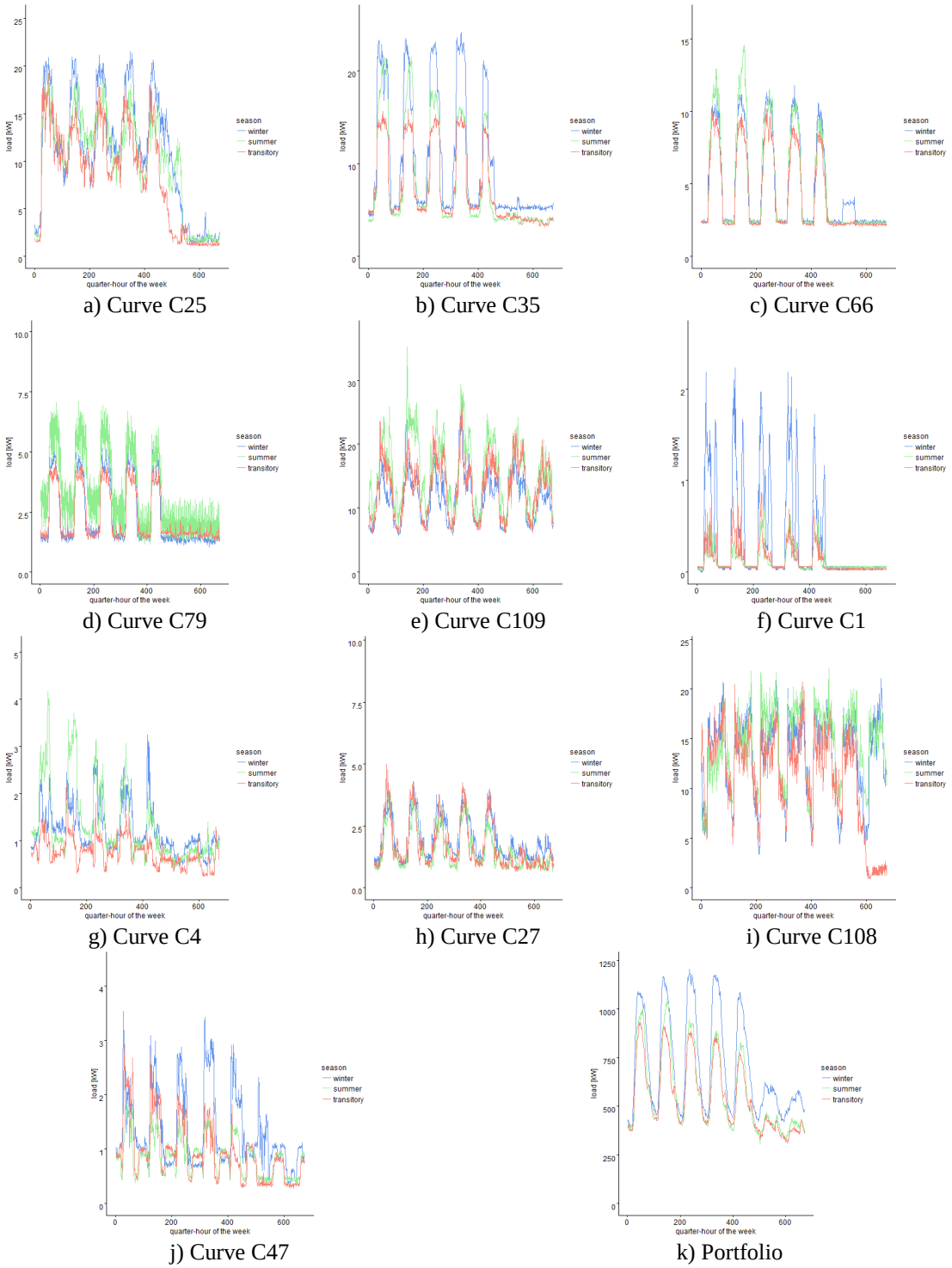
d) Curve C79 e) Curve C109 f) Curve C1

g) Curve C4 h) Curve C27 i) Curve C108

j) Curve C47 k) Portfolio

Source: author's own calculations (R) and author's own figures (R).

Figure 74: Weekly time series of individual curves and the portfolio by season



a) Curve C25 b) Curve C35 c) Curve C66

d) Curve C79 e) Curve C109 f) Curve C1

g) Curve C4 h) Curve C27 i) Curve C108

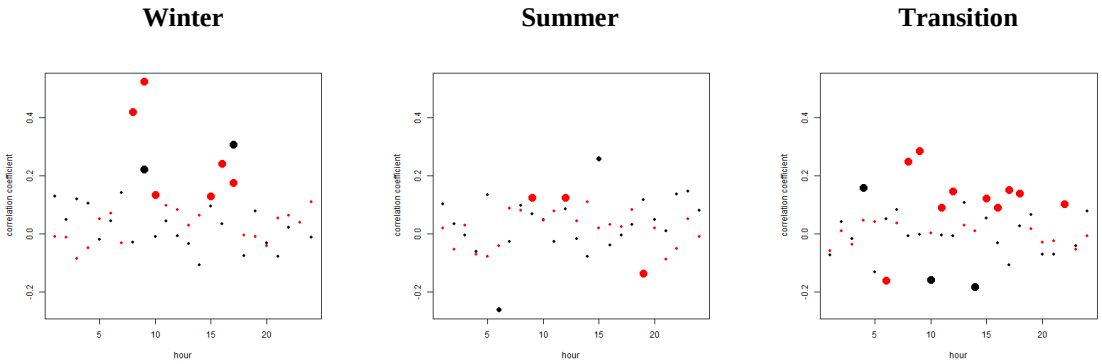
j) Curve C47 k) Portfolio

Source: author's own calculations (R) and author's own figures (R).

F) The diversification of volume risk

Figure 75 shows the linear correlation coefficient values between the normalised (by conditional standard deviation[154]) residuals of C35 and the portfolio that appears many times in the dissertation. Larger dots indicate the significant coefficients at a 5% significance level (weekdays are marked red, weekends black). It can be stated that the values of these coefficients differ significantly from zero especially on weekdays, in *peak* periods, and they typically mark a positive relationship.

Figure 75: Linear correlation coefficient values between the standardised residuals of the individual curve C35 and the portfolio[155]



Source: author's own calculations (R) and author's own figures (R).

The linear correlation coefficient value for the whole sample is 0.0671 (p-value = 0.0000), which shows weak, positive (but significant) relationship between the residuals of individual curve and the residuals of the portfolio, which thus blurs seasonal trends.

Obviously, linear correlation coefficients measure the strength of the relationship appropriately if the relationship between the variables is linear. What can be stated nevertheless is that the non-diversifiable part of the volume risk may be smaller or larger depending on the period, and assuming constant strength of relationship its value will be over- or underestimated.

The comprehensive study and modelling of the above is beyond the scope of this dissertation, however, this short (rather illustrative) example draws the attention efficiently to the fact that it is especially the periods with greater standard deviation whose risk can be decreased in a lesser extent by diversification, which is definitely to be considered in modelling.

[154] This way heteroscedasticity does not influence the results.
[155] Weekdays are marked red, weekends are black.

Examples to this phenomenon – as a kind of stylised fact – cannot be found either in academic or practical studies; as a consequence, there is no publicly available model to capture this feature.

The strength of the relationship in this example was made weather-(independent variable-) dependent in a rather *ad-hoc* way. In the same way, the time-dependent strength of a relationship may be calculated with *wavelet* transform, which also reveals in which frequency the relationship (co-movement) exists. A likewise appropriate technique may be the application of the so-called local correlation coefficients (see: *Tjostheim–Hufthammer* [2013]), for which there are a number of examples in finance. The problem with this is that the 'localised' prefix is to be understood not temporally, but regarding the domain of the variable. Notwithstanding the above, a solution may be imagined even in the framework of mixture models.

REFERENCES

2007. évi LXXXVI. törvény a villamos energiáról

Banfield, J. D. – A. E. Raftery, A. E. [1993]: Model-based Gaussian and non-Gaussian clustering. *Biometrics*. 49 pp. 803–821.

Baudry, J.-P. – Raftery, A. E. – Celeux, G. , Lo, K. – Gottardo, R. [2010]: Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*. 19 (2) pp. 332-353.

Biernacki, C. – Celeux, G. – Gérard, G. [2003]: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*. 41 pp. 561-575.

Bollerslev, T. [1986]: Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*. 31 (3) pp. 307-327.

Box, G. E. P. – Jenkins, G. M. [1970]: Time Series Analysis: Forecasting and Control. Holden Day. San Francisco.

Brealey, A. R. – Myres, C. S. [2005]: Modern vállalati pénzügyek. Panem Kft. Budapest.

Breusch, T. S. [1978]: Testing for Autocorrelation in Dynamic Linear Models. *Australian Economic Papers*. 17 pp. 334–355.

Carpaneto, E. – Chicco, G. – Napoli, R. – Scutariu, M. [2003]: Customer Classification by Means of Harmonic Representation of Distinguishing Features. Paper for *IEEE Bologna Power Tech Conference*, June 23th-26th, Bologna, Italy.

Carpaneto, E. – Chicco, G. – Napoli, R. – Scutariu, M. [2006]: Electricity customer classification using frequency-domain load pattern data. *Electrical Power and Energy Systems*. 28 pp. 13-20.

Chicco, G. [2012]: Overview and performance assesment of the clustering methods for electrical load pattern grouping. *Energy*. 42 pp.68-80.

Chicco, G. – Napoli, R. – Piglione, F. – Postolache, P. – Scutariu, M. – Toader, C. [2005]: Emergent electricity customer classification. *IEE Proceedings – Generation, Transmission and Distribution.* 152 (2) pp. 164-172.

Cont, R. [2001]: Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance.* Vol. 1, pp. 223-236.

Cover, T. M. – Thomas, J. A. [1991]: Elements of Information Theory. New York: Wiley.

Dempster, A. P. – Laird, N. M. – Rubin, D. B. [1977]: Maximum Likelihood from Incomplete Data Via the EM-Algorithm. *Journal of Royal Statistical Society B.* Vol. 39. pp. 1-38.

Dickey, D. A. – Fuller, W. A. [1979]: Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*. 74 (366) pp. 427–431.

Eirola, E. – Lendasse, A. [2013]: Gaussian Mixture Models for Time Series Modelling, Forecasting, and Interpolation. *Advances in Intelligent Data Analysis XII. Lecture Notes in Computer Science.* (8207) pp. 162-173

Elexon [2013]: Load Profiles and their use in Electricity Settlement. Guidance. Version 2.0. Elexon Ltd. UK.

Engle, R. F. [1982]: Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica*. 50 (4) pp. 987-1007.

Espinoza, M. – Joye, C. – Belmans, R. – De Moor, B. [2005]: Short-Term Load Forecasting, Profile Identification and Customer Segmentation: A Methodology based on Periodic Time Series. *IEEE Transactions on Power Systems*. 20 (30) pp. 1622-1630.

Fraley, C. – Raftery, A. E. [2000]: Model-Based Clustering, Discriminant Analysis, and Density Estimation. Technical Report no. 380. Department of Statistics, University of Washington.

Fraley, C. – Raftery, A. E. [2007]: Model-based Methods of Classification: Using the mclust Software in Chemometrics. *Journal of Statistical Software.* 18 (6) pp. 1-13.

Friedman, J. H. [1991]: Multivariate Adaptive Regression Splines. *The Annals of Statistics.* 19 (1) pp. 1-67.

Godfrey, L. G. [1978]: Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables. *Econometrica.* 46 pp. 1293–1301.

Hamilton, J. D. [1994]: Time Series Analysis. Princeton University Press. Princeton, New Jersey.

Heo, Y. – Zavala, V. M. [2012]: Gaussian process modeling for measurement and verification of building energy savings. *Energy and Buildings*. 53 pp. 7-18.

Hershey, J. R. – Olsen, P. A. [2007]: Approximating the Kullback-Leibler divergence between Gaussian mixture models. *IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* (4) pp. IV-317-IV-320.

Hino, H. – Shen, H. – Murata, N. – Wakao, S. [2013]: A Versatile Clustering Method for Electricity Consumption Pattern Analysis in Households. *IEEE Transactions on Smart Grid.* 4 (2) pp. 1048-1057.

Howden, S. M. – Crimp, S. [2001]: Effect of Climate and Climate Change on Electricity Demand in Australia. *CSIRO Sustainable Ecosystems*. Canberra.

Hunyadi, L. – Vita, L. [2003]: Statisztika közgazdászoknak. Aula Kiadó Kft., Budapest.

Hylleberg, S. – Engle, R. F. – Granger, C. W. J. – YOO, B. S. [1990]: Seasonal Integration and Cointegration. *Journal of Econometrics*. 44. pp. 215–238.

Jaques, J. – Preda, C. [2013]: Functional data clustering: a survey. Research Report. *Research Centre Lille – Nord Europe.*

Junghans, G. [2015]: Portfolio risk management in a highly complex multi-regional market: Case stusy of Baltic market. *2nd Annual Intelligent Risk and Portfolio Optimisation for the Energy Markets*. 22nd-23rd September 2015, Berlin, Germany.

Kerékgyártó, Gy. – L. Balogh, I. – Sugár, A. – Szarvas, B. [2008]: Statisztikai módszerek és alkalmazásuk a gazdasági és társadalmi elemzésekben. Aula Kiadó Kft., Budapest.

Kwiatkowski, D. – Phillips, P. C. B. – Schmidt, P. – Shin, Y. [1992]: Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics.* 54 (1–3) pp. 159–178.

Leng, T. K. – Cheong, C. W. – Hooi, T. S. [2014]: Impact of global financial crisis on stylized facts between eenrgy markets and stock markets. *Proceedings of the 3rd International Conference on Mathematical Sciences.* AIP Conf. Proc. Vol. 1602, pp. 994-1001.

Levy, G. [2013]: Electricity contract risk with portfolio effects. *EnergyRisk risk-net/energy-risk.* Technical Paper. pp. 40-46.

Li, X. – Bowers, C. P. – Schnier, T. [2010]: Classification of Energy Consumption in Buildings With Outlier Detection. *IEEE Transactions on Industrial Electronics*. 57 (11) pp. 3639-3644.

Liao, T. W. [2005]: Clustering of time series data – a survey. *Pattern Recognition*. 28 pp. 1857-1874.

Ljung, G. M. – Box, G. E. P. [1978]: On a Measure of a Lack of Fit in Time Series Models. *Biometrika*. 65 (2) pp. 297–303.

Lo, K. L. – Wu, Y. K. [2003]: Risk assesment due to local demand forecast uncertainty int he compoetitive supply industry. *IEE Proceedings – Generation, Transmission and Distribution.* 150 (5) pp. 573-581.

Macedo, M. N. Q. – Galo, J. J. M. – de Almeida, L. A. L. – de C. Lima, A. C. [2015]: Demand side management using artificial neural networks in a smart grid environment. *Renewable and Sustainable Energy Reviews.* 41 pp. 128-133.

Maddala, G. S. [2004]: Bevetés az ökonometriába. Nemzeti Tankönyvkiadó, Budapest.

Mák, F. [2014a]: Egységgyöktesztek alkalmazása szezonalitást is tartalmazó idősorok esetében energiatőzsde-adatok példáján. *Statisztikai Szemle.* 92 (7) pp. 647–679.

Mák, F. [2014b]: Analyzing Interrelated Stochastic Trend and Seasonality on the Example of Energy Trading Data. *Society and Economy.* 36 (2) pp. 233-261.

Mák, F. [2015]: Az időjárás véletlen hatásának szerepe a szezonális kiigazítás során, a hazai földgázfogyasztás példáján. *Statisztikai Szemle.* 93 (5) pp. 417–441.

Manfren, M. – Aste, N. – Moshksar, R. [2013]: Calibration and uncertainty analysis for computer models – A meta-model based approach for integrated building energy simulation. *Applied Energy.* 103 pp. 627-641.

Marossy, Z. [2010]: A spot villamosenergia-árak elemzése statisztikai és ökonofizikai eszközökkel. PhD értekezés. *Budapesti Corvinus Egyetem*. Budapest.

Mathieu, J. L. – Price, P. N. – Kilicote, S. – Piette, M. A. [2011]: Quantifying Changes in Building Electricity Use, With Application to Demand Response. *IEEE Transactions on Smart Grid*. 2 (3) pp. 507-518.

McKenna, S. A. – Fusco, F. – Eck, B. J. [2014]: Water demand pattern classification from smart meter data. *Procedia Engineering*. 70 pp. 1121-1130.

McLachlan, G. J. – Basford, K. E. [1988]: Mixture Models: Inference and Applications to Clustering. Marcel Dekker.

McLachlan, G. J. – Krishnan, T. [1997]: The EM Algorithm and Extensions. Wiley.

McLachlan, G. J. – Peel, D. [2000]: Finite Mixture Models. Wiley.

Mutanen, A. – Ruska, M. – Repo, S. – Järventausta, P. [2011]: Customer Classification and Load Profiling Method for Distribution Systems. *IEEE Transactions on Power Delivery*. 26 (3) pp. 1755-1763.

Panapakidis, I. P. – Alexiadis, M. C. – Papagiannis, G. K. [2012]: Load Profiling in the Deregulated Electricity Markets: A Review of the Applications. *2012 9th International Conference ont he European Energy Market.* pp. 1-8.

Panapakidis, I. P. – Papadopoulos, T. A. – Christoforidis, G. C. – Papagiannis, G. K. [2014]: Pattern recognition algorithms for electricity load curve analysis of buildings. *Energy and Buildings*. 73 pp. 137-145.

Pitt, B. [2000]: Applications of Data Mining Techniques to Electric Load Profiling. PhD Thesis. *University of Manchester Institute of Science and Technology.*

Povinelli, R. J. – Johnson, M. T. – Lindgren, A. C. – Ye, J. [2014]: Time Series Classification Using Gaussian Mixture Models of Reconstructed Phase Spaces. *IEEE Transactions on Knowledge and Data Engineering.* 16 (6) pp. 779-783.

Ramanathan, R. [2003]: Bevezetés az ökonometriába alkalmazásokkal. Panem Kft., Budapest.

Räsänen, T. – Voukantsis, D. – Niska, H. – Karatzas, K. – Kolehmainen, M. [2010]: Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy*. 87 pp. 3538-3545.

Shekofteh, Y. – Almasganj, F. [2013]: Feature Extraction Based on Speech Attractors int he Reconstructed Phase Space for Automatic Speech Recognition Systems. *ETRI Journal.* 35 (1) pp. 100-108.

Singh, R. – Pal, B. C. – Jabr, R. A. [2010]: Statistical Representation of Distribution System Loads Using Gaussian Mixture Model. *IEEE Transactions on Power Systems*. 25 (1) pp. 29-37.

Srivastav, A. – Tewari, A. – Dong, B. [2013]: Baseline building energy modeling and localized uncertainty quantification using Gaussian mixture models. *Energy and Buildings*. 65 pp. 438-447.

Subbarao, K. – Lei, Y. – Reddy, T. A. [2011]: The Nearest Neighborhood Method to Improve Uncertainty Estimates in Statistical Building Energy Models. *ASHRAE Transactions*. 117 (2) pp. 459-471.

Sugár, A. [1999a]: Szezonális kiigazítási eljárások (I.). *Statisztikai Szemle*. 77 (9) pp. 705–721.

Sugár, A. [1999b]: Szezonális kiigazítási eljárások (II.). *Statisztikai Szemle*. 77 (10-11) pp. 816–832.

Sugár A. [2011]: A hőmérséklet hatásáról a villamosenergia- és gázfogyasztás magyarországi példáján. *Statisztikai Szemle*. 89 (4) pp. 379–398.

Tjostheim, D. – Hufthammer, K. O. [2013]: Local Gaussian correlation: A new measure of dependence. *Journal of Econometrics.* 172 pp. 33-48.

Tsekouras, G. J. – Hatziargyriou, N. D. – Dialynas, E. N. [2007]: Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers. *IEEE Transactions on Power Systems*. 22 (3) pp. 1120-1128.

Tsekouras, G. J. – Kotoulas, P. B. – Tsikeris, C. D. – Dialynas, E. N. – Hatziargyriou, N. D. [2008]: A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. *Electric Power Systems Research*. 78 pp. 1494-1510.

Váradi, K. [2012]: Likviditási kockázat a részvénypiacokon. PhD értekezés. *Budapesti Corvinus Egyetem.* Budapest.

Varian, H. R. [2004]: Mikroökonómia középfokon. *Akadémiai Kiadó Zrt.* Budapest.

Verdú, S. V. – García, M. O. – Senabre, C. – Marín, A. G. Franco, F. J. G. [2006]: Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps. *IEEE Transaction on Power Systems*. 21 (4) pp. 1672-1682.

References from the Internet:

Treatment of special weather effects:

https://en.wikipedia.org/wiki/Wind_chill
https://en.wikipedia.org/wiki/Humidex

Methodological terms:

mathworld.wolfram.com/RelativeEntropy.html
https://en.wikipedia.org/wiki/Hermite_interpolation

Data used in analyses (see also in the corresponding chapters):

ec.europa.eu/eurostat/
https://www.fgsz.hu
https://www.hupx.hu
https://www.mavir.hu
www.varaljamet.eoldal.hu/cikkek/climate_budapest.html

Heuristic measurement of volume risk:

www.enoro.com/

Others:

https://wikiszotar.hu/

PUBLICATIONS

A) Publications in Hungarian in the field of the dissertation

Reviewed journal:

Mák, F. [2011]: Egységgyöktesztek alkalmazása strukturális törések mellett a hazai benzinár példáján. *Statisztikai Szemle*. 89 (5) pp. 545–573.

Mák, F. [2014]: Egységgyöktesztek alkalmazása szezonalitást is tartalmazó idősorok esetében energiatőzsde-adatok példáján. *Statisztikai Szemle*. 92 (7) pp. 647–679.

Mák, F. [2015]: Az időjárás véletlen hatásának szerepe a szezonális kiigazítás során, a hazai földgázfogyasztás példáján. *Statisztikai Szemle*. 93 (5) pp. 417–441.

B) Publications in English in the field of the dissertation

Reviewed journal:

Mák, F. [2014]: Analyzing Interrelated Stochastic Trend and Seasonality on the Example of Energy Trading Data. *Society and Economy*. 36 (2) pp. 233-261.

C) Most important unpublished works

Mák, F. – Sugár, A. – Tóth, B. [2009]: A magyarországi gázpiaci és gázszolgáltatásokhoz köthető egyéb termékek keresleti elemzése. *Tanulmány a Fővárosi Gázművek Zrt. részére.*

Mák, F. – Sugár, A. [2011]: Lokális hőigényekre vonatkozó előrejelzési modellek. *Tanulmány az E.On Energiaszolgáltató Kft. részére.*

Mák, F. [2015]: Hőmérsékletkorrekciós modell. *Tanulmány és modell dokumentáció a Magyar Energetikai és Közmű-szabályozási Hivatal részére.*