Doctoral Programme

of Management and Business Administration

THESIS BOOKLET

Fruzsina Mák

Volume risk in the power market

Load profiling considering uncertainty

Ph.D. Dissertation

Supervisors:

Beatrix Oravecz, Ph.D.
Senior lecturer
András Sugár, Ph.D.
Associate professor, Head of Department of Statistics

Budapest, 2017

Department of Statistics

THESIS BOOKLET

Fruzsina Mák

Volume risk in the power market

Load profiling considering uncertainty

Ph.D. Dissertation

Supervisors:

Beatrix Oravecz, Ph.D.
Senior lecturer
András Sugár, Ph.D.
Associate professor, Head of Department of Statistics

TABLE OF CONTENTS

1. PREVIOUS RESEARCH EFFORTS AND MOTIVATION

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.

1.1. Relevance of the dissertation

Although the source of the highest potential risks on the energy market is basically price, 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] In the **electricity market** – but also in 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).

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

[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.

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 more 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**.

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[2].

Since consumption itself is **stochastic**, its risk, uncertainty or portfolio effects also need to be taken into consideration in a similar way that 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 the irregular component in consumption.

1.2. Theoretical, methodological background and hypotheses

Based on the literature on profiling, the general framework used for creating consumer profiles and consumer profile groups practically 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. It is essential in feature-

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

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.

The curve features that describe each consumption curve in clustering can be produced in a number of ways. A possible classification of these can be found in Table 1. It is clear that in most cases features that characterise consumption curves are produced on the basis of some daily *representative load curve* (RLC), which, on the one hand, supports the characterisation of individual curves, and on the other hand, supports the creation of consumer profile groups that include those that share similar profiles.

Table 1: 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-**	**frequency-domain-**	**model-**	
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.

Although these representative load curves 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 framework of 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., see e.g. *Pitt* [2000]).

It is important to state that in the so-called mainstream research results, the effect of weather (in most cases, temperature) or the effect of its irregular part is removed from the consumption time series in the course of profiling.

Based on the author's own former research results and its supplementation it can be said that if the behaviour of the irregular component of the time series is relevant from the research perspective, removing the total effect of temperature 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. 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 temperature-independent part of the time series may be a limiting factor regarding possible analyses, relevant methods and conclusions that could be drawn.

Based on the previously summarised research results and on the author's own research experience, 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.

2. RESEARCH METHODOLOGY

As the research methodology applied is supported by the experience stemming from former research efforts on the topic of the dissertation and by the empirical findings of studying the stylized facts of consumption time series, in this section we make a short summary on the conclusions regarding stylized facts and the reasoning on why mixture models can be capable of capturing most of these.

2.1. The examination of stylized facts of consumption time series

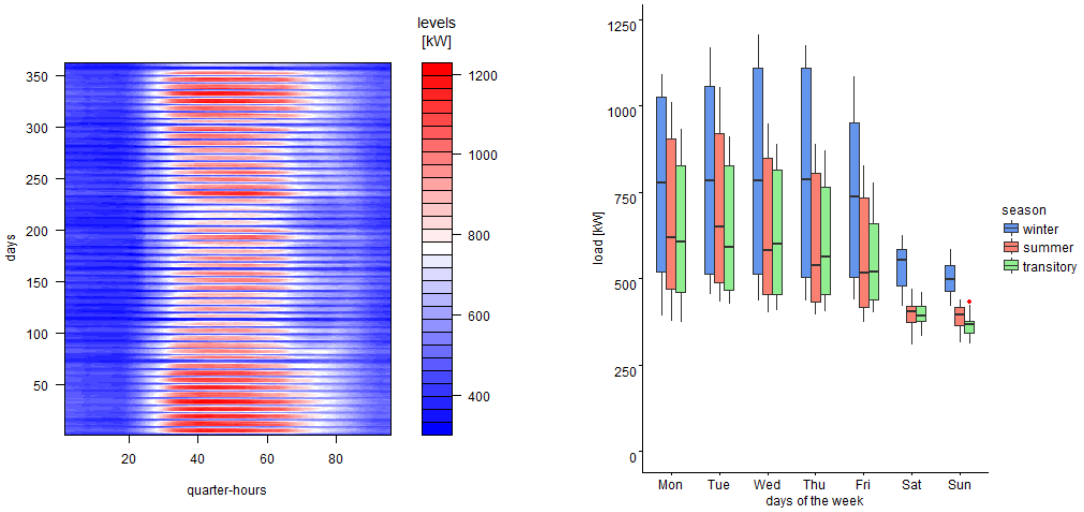
Various individual curves have been examined to discover the main features that can determine the characteristics of a curve and that need to be considered in the identification of the typical consumption pattern. These were typically not classical statistical tests but simpler calculations or figures 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 changes 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.

Figure 1: *Contour plot* and *boxplot* of a portfolio time series



Source: author's own figures (R).

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.

2.2. Using the mixture model for modelling non-constant covariance structure

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. On the methods applied see, *inter alia*, *Eirola and Lendasse* [2013], *Fraley-Raftery* [2000], *Fraley-Raftery* [2007].

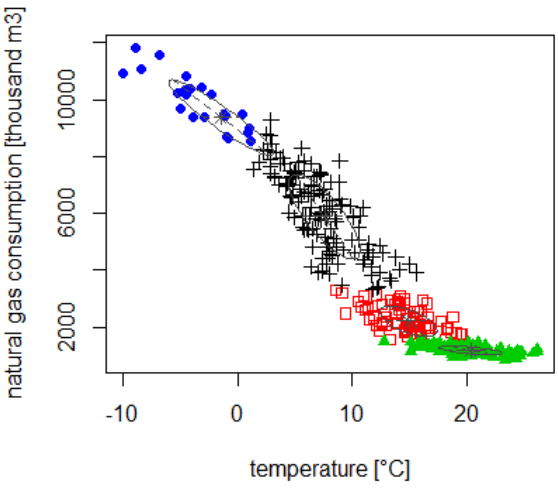
The aim of this dissertation is twofold: on the one hand, it aims at constructing typical, representative consumption patterns; but on the other hand, the uncertainty related to con-

sumption 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 describe 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.

Figure 2: Results of mixture clustering on the example of daily average temperature – natural gas consumption



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

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. Figure 2 is telling of this logic: the relationships 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.

Attention has been drawn to the methodological difficulties in connection with a **previous research result** of this dissertation related to Hungarian **natural gas consumption**[3]. 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.

2.3. Formal description of mixture models and mixture regression

The essence of mixture models can be summarised as follows.

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:

- y_i is a vector of size $(m \times 1)$ containing the attributes of observation i ($i = 1, 2 \ \dots \ n$), n is the number of observations, m is the number of attributes, [4]
- $f(.)$ 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, and the *prior* probability τ_k is the probability of observation i belonging to component k,
- k denotes the components ($k = 1,2, \dots K$), and K is the number of mixture components.

[3] Estimation results are derived using the X13-ARIMA-SEATS seasonal adjustment program.
[4] The term *attribute* used in international, mainly data mining literature is identical with the term *variable* in regression terminology.

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]). 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 showing which observation belongs to which component. 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 $\psi = (\tau_1, \tau_2 \dots \tau_K, \theta_1, \theta_2 \dots, \theta_K)$ denote the parameters to be estimated, that is the *prior* probabilities of the components and the parameters of the distribution of the components. The *likelihood*-function is the following:

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

Given the observations $y = (y_1, y_2 \dots y_n)$ 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:

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

in addition, by 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:

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

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

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 parameter-

ized by mean vector μ_k and covariance matrix Σ_k as 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 the parameters that need to be estimated in the parameter set $\psi = (\tau_1, \tau_2 \dots \tau_K, \theta_1, \theta_2 \dots, \theta_K)$ are the *prior* probabilities and the mean vectors and the covariance matrices per components.

The *Gaussian* mixture regression (GMR) is based on the *Gaussian* mixture model (GMM, see the previous paragraph) by appointing one of the variables, which later becomes the dependent (output) variable, and a regression is written using the other variables as independent variables. From a computational aspect, the mixture regression can be written using the formulas of the weighted least squares method.[5]

The component-based calculations of the conditional mean and conditional standard error are based on conditional mean and conditional standard deviations for each component using *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.

It is shown in the dissertation that the formula for m_{ik} is nothing else but a substitution in the regression equation, and the formula for s_{ik}^2 is identical with the residual variance per component, conditional on the given values of the x_i independent variables.

The expected value and variance (that is, the squared standard error) of the dependent variable can be written as:

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

[5] The notation of mixture regression is slightly different from the notation of classical multivariate regression, but serves practical goals from the perspectives of interpretation and implementation. The introduction of the relationship between the two notations is included in the dissertation. Based on this, the coefficients and the residual variance of the components are as follows (W_k denotes a diagonal matrix of size $(n \times n)$ containing the p_{ik} weights for each component *k*):

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

and

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

where disregarding the weighting above, the classical formulas applied for ordinary least squares are easy to identify.

and

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

The methodological chapter contains a number of formulas and interpretations (especially in connection with mixture regression) 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.

These results greatly supported the implementation of the *Mixture Regression* in the *R Project* statistical software package. The commands of the regression based on the *Gaussian* mixture model are the author's own functions building on the results of the package *'mclust'*, as there are no functions for regression applications of the *Gaussian* mixture model in the *R package*.

3. KEY FINDINGS OF THE DISSERTATION AND AVENUES FOR FURTHER RESEARCH

There are so-called classical solutions in the chapters on profiling and measuring uncertainty of this dissertation, which on the one hand help the exploration of empirical findings, and on the other hand serve as *benchmark* and help comparison with the new results. As the key findings of this dissertation can be evaluated in relation to classical results, the main consequences are summarised in the light of this. In the end of this section we provide a short summary regarding the potential applications and future research topics building on or evoked by the results and conclusions of this dissertation.

3.1. Using the mixture model for creating typical consumption patterns

In the empirical parts of the dissertation 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.

3.2. Using heuristic and classical stochastic time series methods to measure uncertainty of consumption

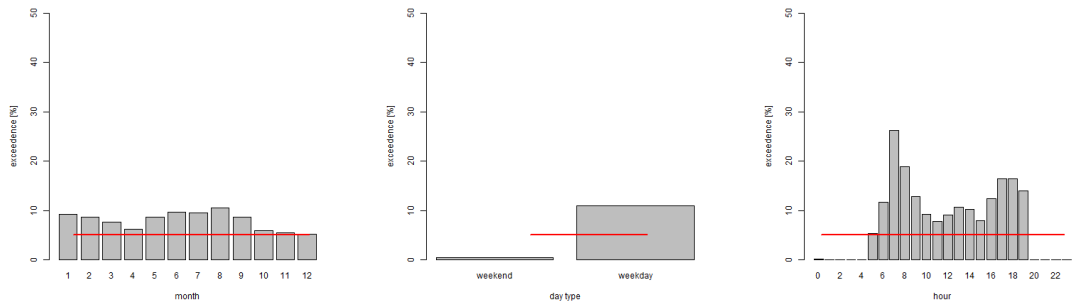
The investigations that have been performed in this dissertation to measure the irregular behaviour by curve will be described here. Based on the standard errors calculated using classical time series techniques, SARMA and PAR regression, confidence intervals were produced. It has been found that assuming 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.[6]

[6] 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.

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*.[7]

Figure 3: Ratio of observations outside the confidence interval in an individual curve



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

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.).

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 when 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.

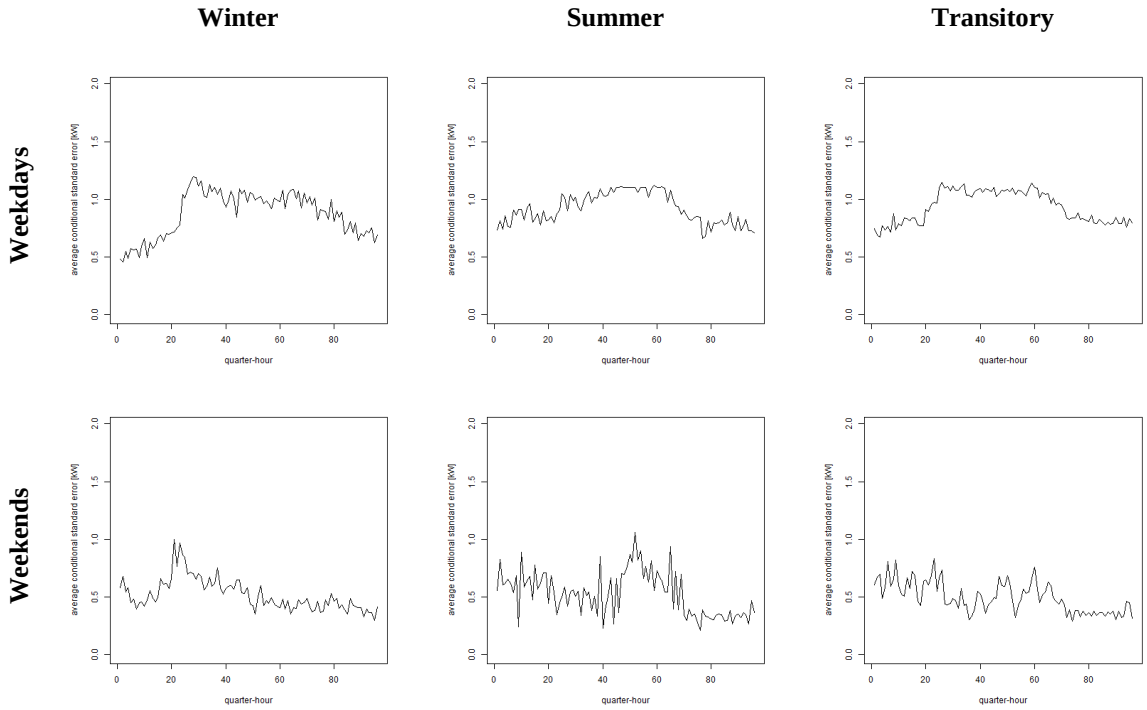
[7] Weather (temperature) – as we know – is a stochastic variable in itself, and the gain of not removing the effect of temperature in profiling is reflected in this important statement.

3.3. 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.

Figure 4: Average conditional standard errors in mixture regression in an individual curve



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

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 non-linear 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, 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

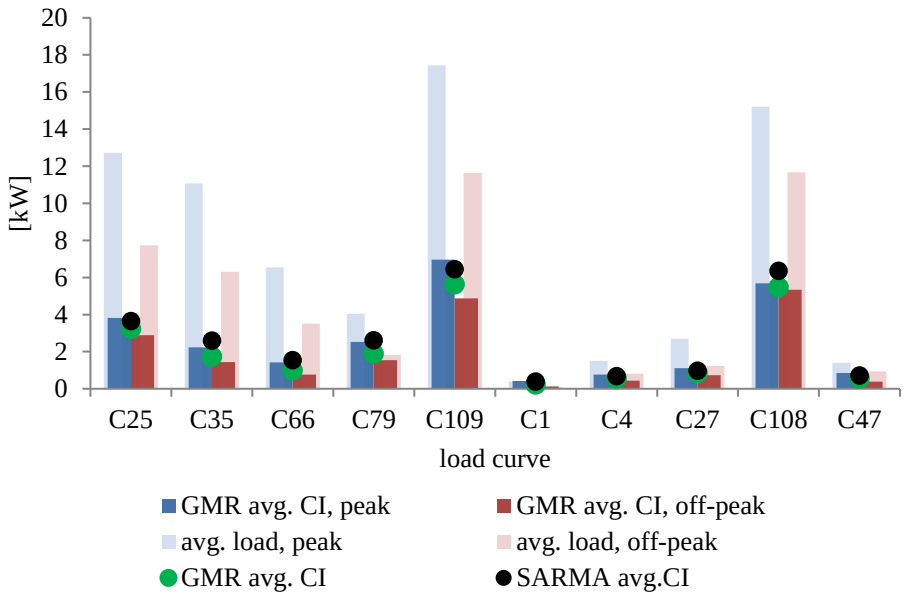
- the profile and the uncertainty of consumption (modelling of volume risk) is performed in a unified framework,
- the application of mixture regression issues in promising results, and its energy market use can be regarded relatively new,
- in such regression applications of the mixture model, the *backtest* 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 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 classical 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, their necessity, thinking about, for example, the smoothing of the consumption curve, or the decreasing of the balancing energy

costs resulting from actual-planned deviations or the decreasing of *peak–off-peak* ratio resulting from shifts in energy use.

Figure 5: Average confidence interval in SARMA and mixture regression in individual curves[8]



Load 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).

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).

3.4. 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 rather 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**

[8] On the figure CI denotes the confidence interval at 95% confidence level.

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 the 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 influ-**

ence, 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[9] 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).

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

[9] Even its significance, or the lack of it.

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 non-linearity 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.

4. REFERENCES

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.

Box, G. E. P. – Jenkins, G. M. [1970]: Time Series Analysis: Forecasting and Control. Holden Day. San Francisco.

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.

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

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.

Hamilton, J. D. [1994]: Time Series Analysis. Princeton University Press. Princeton, New Jersey.

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.

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.

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.

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 energiatőzsdeadatok 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.

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.*

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.

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.

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.

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.

The full list of references can be found in the dissertation.

5. PUBLICATIONS

5.1. 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 energiatőzsdeadatok 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.

5.2. 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.