Short-Term Load Forecasting, Profile Identification and Customer Segmentation: A Methodology based on Periodic Time Series.

Marcelo Espinoza, Caroline Joye, Ronnie Belmans, *Senior Member, IEEE*, Bart De Moor, *Fellow, IEEE*

***Abstract*— Results from a project in cooperation with the Belgian National Grid Operator ELIA are presented in this paper. Starting from a set of 245 time series, each one corresponding to 4 years of measurements from a HV-LV substation, individual modelling using Periodic Time Series yields satisfactory results for short-term forecasting or simulation purposes. In addition, we use the stationarity properties of the estimated models to identify typical daily customer profiles. As each one of the 245 substations can be represented by its unique daily profile, it is possible to cluster the 245 profiles in order to obtain a segmentation of the original sample in different classes of customer profiles. This methodology provides a unified framework for the forecasting and clustering problems.**

***Index Terms*— Load Forecasting, Load Modelling, Time Series, Clustering Methods.**

I. Introduction

Within the electricity sector the need for strategic information has become extremely important. Not only accurate forecasts are needed for the short-term operations and mid-term scheduling, but also network managers need to have insight in the type of customers they have to supply as a support for long-term planning. In order to deal with the everyday process of planning, scheduling and unit-commitment [1], the need for accurate short-term forecasts has led to the development of a wide range of models based on different techniques. Some interesting examples are related to traditional time series analysis [2]–[4], and neural networks applications [5]–[9]. At the same time, the unbundling between generation, transmission, distribution and supply induced by the market liberalization has led to network managers being partially blind beyond a certain substation level with respect to the final customers. In these cases, it is required to use indirect techniques to assess the type of demand they have to face [10], [11] in order to support their long-term investment planning. In this context, categories of residential, business and industrial customers have been documented for some locations [12], [13] and usually they are recognized by their "typical" load pattern over a day.

The two problems outlined above, forecasting and profiling, usually have been tackled independently. However, from a manager point of view, the boundary between both problems is irrelevant, and eventually unnecessary. Given a set of measurements, it is possible to produce accurate short-term forecasts and at the same time generate a tool for extracting information on the overall demand structure. In this paper, it is shown how to identify and remove the influence of temperature fluctuations and how to use the forecasting model to identify the type of customer being modelled. We will demonstrate this methodology on a set of 245 time series provided by the Belgian National Grid Operator ELIA, details of which are described below. The methodology is oriented towards handling the problems of short-term forecasting and profile segmentation in a unified framework. In a first stage, the goal is to generate short-term models that can produce accurate forecasts, extract temperature and seasonal effects and identify the type of customer under scrutiny. The goal for a second stage is to partition the set of time series, using clustering algorithms, based on the customer profiles identified using the models from the first stage.

This paper is structured as follows. A general description of the methodology and the data available is presented in Section II. The forecasting problem is addressed in Section III, with the description of the models, the estimation strategy and the forecasting results. Section IV shows the method for computing the daily profiles from the estimated models. This is the key feature of the methodology that makes it possible to have a smooth transition between a forecasting and a clustering problem. The implementation of the customer segmentation using a classic clustering technique is reported in Section V.

II. General Setup

The general setup is described, with information on the data and the general procedure to be followed.

A. General Procedure

The objective is to provide a unified framework for the problems of short-term forecasting and profile segmentation. The initial modelling is based on seasonal time series analysis, using the so-called Periodic Autoregression (PAR) model [14]–[17], which have been used in macroeconomic modelling [18], environmental and hydrological studies [15], and lately in modelling of electricity prices[19]. The stationarity properties, to be defined below, of the estimated autoregression are exploited in order to obtain a very interesting by-product directly interpretable in the context of load profiles. These identified profiles are extracted from each individual model,

M. Espinoza (Research Assistant) and B. De Moor (Full Professor) are with the SCD Research Division, Electrical Engineering Dept. (ESAT), Katholieke Universiteit Leuven, Belgium.

C. Joye is with the Belgian National Grid Operator ELIA.

R. Belmans (Full Professor) is with the ELECTA Division, Electrical Engineering Dept. (ESAT), Katholieke Universiteit Leuven, Belgium.

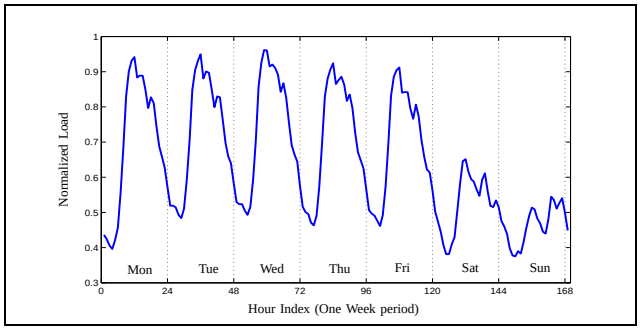

Fig. 1. Example of a Load series within a week. Daily cycles are visible, as well as the weekend effects. Also visible are the intra-day phenomena, such as the peaks (morning, noon, evening) and the night hours.

and an unsupervised clustering technique [20], [21] is implemented in order to obtain a clear assessment of how many different groups of customers exist within the original sample. This segmentation can be used for subsequent planning and management purposes.

B. The Data

The dataset consists of 245 time series, each containing hourly load values from a HV-LV substation within the Belgian grid, for a period of approximately 5 years (from January 1998 until September 2002). The 245 load series differ in their behavior as they represent different types of underlying customers (residential, business, industrial, etc.)[1]. A first linear regression containing only a linear trend is estimated for each substation, to remove any growth trend present in the sample. Finally, the series were normalized using the maximum observed value in order to scale all the series to a range between 0 and 1.

III. Short-Term Forecasting

This section describes the implementation and results of the short-term forecasting problem applied to the sample of 245 different load series using PAR models.

A. Seasonality and Periodic Autoregresive (PAR) Models

Forecasting the load is not straightforward, particularly due to the presence of multiple seasonal patterns in the load series (monthly, weekly, intra-daily). Figure 1 shows an example of a load series in a week, at hourly values starting at 00:00 hrs on Monday, until 24:00 hrs on Sunday. In the literature, it is often found that some subsamples of the load are used to produce short-term forecasts; the subsamples are selected in order to isolate a seasonal pattern (working only with winter, summer, evenings, working-days, etc).

By following a seasonal-modelling approach, it is possible to incorporate a priori information concerning the seasonalities at several levels (daily, weekly, yearly, etc.) by appropriately choosing the model structure and the estimation method. Examples are stochastic seasonality (Box-Jenkins seasonal ARIMA models [22]), seasonal unit roots [23], nonparametric models with seasonal components [24]; or periodic autoregressions [14], which is the approach applied in this paper.

In simple terms, an autoregression is said to be periodic when the parameters are allowed to vary across seasons. Consider the case of a univariate time series y_t, $t = 1, \cdots, N$, (in this case, the hourly load measurements) available for a sample of $N_d = N/24$ days, corresponding to the N hours. A periodic autoregressive model of order p (PAR(p)) can be written as

$$y_t = C_s + \phi_{s,1} y_{t-1} + \phi_{s,2} y_{t-2} + \cdots + \phi_{s,p} y_{t-p} + \varepsilon_{s,t} \quad (1)$$

where C_s is a seasonally varying intercept term, the $\phi_{i,s}$ are the autoregressive parameters up to the order p, varying across the N_s seasons ($s = 1, 2, \cdots, N_s$). The choice of N_s depends on the frequency of the data and the seasonal pattern under scrutiny. The error term $\varepsilon_{t,s}$ can be a standard white noise with zero mean and variance σ, or it can be allowed to have a variance σ_s corresponding to seasonal heteroskedasticity. Equation (1) gives rise to a system of N_s equations that can be estimated using Ordinary Least Squares (OLS). For further details, the interested reader is referred to [14]–[16].

B. Model Formulation

For this implementation, an approach similar to [19] is followed. Monthly and weekly seasonals are modelled by dummy variables, and the intra-daily seasonal pattern is assumed to be captured by the PAR parameters, i.e. $N_s = 24$ is the number of different "seasons" to be identified using the PAR model. Denote by $y_{h,d}$ the value of the load measured in hour h of day d, with $h = 1, 2, \cdots, 24$ and $d = 1, 2, \cdots, N_d$. A formulation is built where the hourly load $y_{h,d}$ is a function of the last 48 hourly values. The parameter p of the PAR(p) is, therefore, $p = 48$. This value is heuristically defined by trying first $p = 24$, $p = 36$ and finally $p = 48$ in order to obtain a satisfactory model performance and, at the same time, keeping model parsimony. The PAR(48) model applied to the hourly load forecasting problem defines the following set of equations:

$$\begin{aligned}
y_{1,d} &= C_1 + \phi_{1,1} y_{24,d-1} + \phi_{1,2} y_{23,d-1} + \cdots + \varepsilon_{1,d} \\
y_{2,d} &= C_2 + \phi_{2,1} y_{1,d} + \phi_{2,2} y_{24,d-1} + \cdots + \varepsilon_{2,d} \\
y_{3,d} &= C_3 + \phi_{3,1} y_{2,d} + \phi_{3,2} y_{1,d} + \cdots + \varepsilon_{3,d} \\
&\vdots \\
y_{24,d} &= C_{24} + \phi_{24,1} y_{23,d} + \phi_{24,2} y_{22,d} + \cdots + \varepsilon_{24,d}
\end{aligned} \quad (2)$$

This basic PAR template consists of $24 \times 49 = 1176$ parameters. This template is further extended to include exogenous variables to account for temperature effects as well as monthly and weekly seasonal variations, as described below.

C. Exogenous Variables and Estimation

Exogenous variables are added to the system (2) such that each equation also includes dummy variables for the day of the week and the month of the year. The variable $W_d \in \mathbb{R}^7$ is

[1]The selection of the data was managed by ELIA in order to have series without abnormal behavior. In addition, some series contained small episodes of zero values, which were corrected locally.

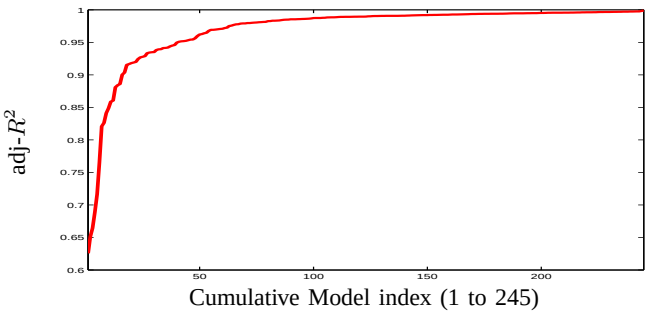

Fig. 2. Adjusted R^2 for the 245 hourly models estimated. Most of the substations have models with high adj-R^2 values.

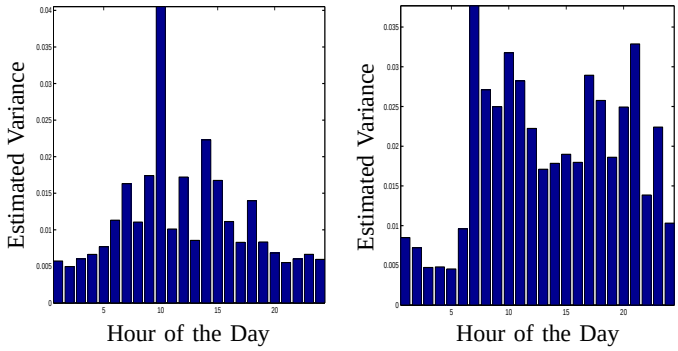

Fig. 3. Comparison of the different daily patterns for the estimated variance of the models, shown here for 2 substations.

a vector of zeroes with a "1" in the position of the day d in the week, e.g. Monday ($W_d = [1, 0, \cdots, 0]$), Tuesday ($W_d = [0, 1, 0, \cdots, 0]$), etc. Analogously, the variable $M_d \in \mathbb{R}^{12}$ is defined as a vector of zeroes with a "1" in the position of the month for which the day d belongs[2]. Additionally, temperature variables are included in order to identify the temperature influence *for each hour of the day*. The temperature variable $T_{h,d}$ is the observed local temperature at a reference location (Ukkel) in Belgium. From $T_{h,d}$, 3 new variables are built to capture the effect of cooling and heating requirements [26] on the load. The variable $\text{CR}_{h,d} = \max(T_{h,d} - 20^o, 0)$ is defined for capturing the cooling requirement, if the ambient temperature is above 20^oC. Analogously, heating and extra-heating variables are defined using $\text{HR}_{h,d} = \max(16.5^o - T_{h,d}, 0)$ and $\text{XHR}_{h,d} = \max(5.0^o - T_{h,d}, 0)$, respectively, with the temperature thresholds taken from standard techniques within the energy industry. Therefore, $T_{h,d}$ has been expanded into a vector $\boldsymbol{\mathcal{T}}_{h,d} = [\text{CR}_{h,d}, \text{HR}_{h,d}, \text{XHR}_{h,d}]$ where each new variable has its own coefficient in the expanded system.

With the inclusion of the exogenous variables, the total number of coefficients to be estimated using a PAR(48) with $N_s = 24$ is $24 \times (49 + 6 + 11 + 3) = 1656$. The augmented system (3) is estimated individually for each one of the 245 time series, using OLS with $t-$tests of significance to keep only those coefficients statistically different from zero[3]. By using the same model template for all substations makes it possible to perform all kinds of comparisons in terms of their parameter estimates values, accuracy obtained, etc.

D. Results: In-sample accuracy

The accuracy of the models, measured by the adjusted-R^2 values obtained for each substation, is shown on Figure 2. Although this is an in-sample indicator, it shows that the model can describe quite well most of the substations, with only 20 substations having an adj-R^2 lower than 0.90. Adding more lags and/or including more external variables may be required to improve the accuracy for these substations.

[2]To avoid exact collinearity between all these variables and the constant terms in the original system (2) only 6 of the W_d components and only 11 of the M_d components are incorporated in the model. This is the standard implementation of dummy variables in any econometric estimation procedure[25].

[3]The computation of the full system (3), for one substation, takes approx. 20 seconds in a Pentium IV computer with 1 Gb RAM.

It is interesting to note that the model captures the behavior of the load series with different accuracy within the day, which has a close relation with allowing individual error terms for each one of the 24 equations, yielding seasonal heteroskedasticity [23]. Figure 3 shows a comparison of the estimated error variance for 2 different substations, with clearly different patterns for the error variance, suggesting these are substations supplying customers with different nature. The left panel shows a rather steady pattern with a high peak around 10AM, the right panel shows more variation in morning and evening hours.

E. Results: Parameter Identification

It is possible to compare different substations by looking into the estimated values of parameters of interest. One example are the parameters related to temperature $\text{CR}_{h,d}$,$\text{HR}_{h,d}$, $\text{XHR}_{h,d}$; the difference between corresponding parameters gives information about temperature sensitivity[4]. As an example, Figure 4 shows the results for heating and cooling requirement variables ($\text{HR}_{h,d}$ and $\text{CR}_{h,d}$, respectively), where its estimated coefficients are depicted as bars, shown here for 6 selected substations. Usually, it is accepted that heating and cooling requirements increase the energy consumption; this is shown by the first substation (Substation A) in Figure 4.

However, other substations show a different sensitivity. Substation B gives an example of inverse sensitivity for cooling requirements: the warmer the temperature above 20^oC, *less* energy is required, i.e., there is no extra consumption, on the contrary, the extra temperature translates into energy saving. Some substations do not show temperature sensitivity either for cooling (Substation C) or heating (Substation F). Other examples show that the maximum effects from temperature-related variables can occur at different hours of the day (Substations D,E).

F. Results: Forecasting

Forecasting using an estimated model is straightforward, although it is known that the performance depends on the forecasting horizon [22]. The simplest scheme is to forecast

[4]Although not shown here because of space constraints, the same analysis can be done for the dummy variables representing months of the year, where specific effects due to "being in January" and "being in July"can be identified.

$$
\begin{aligned}
y_{1,d} &= C_1 + \phi_{1,1} y_{24,d-1} + \phi_{1,2} y_{23,d-1} + \cdots + \phi_{1,48} y_{1,d-2} + \boldsymbol{\alpha}_1^T W_d + \boldsymbol{\beta}_1^T M_d + \boldsymbol{\gamma}_{1,d}^T \boldsymbol{\mathcal{T}}_{1,d} + \varepsilon_{1,d} \\
y_{2,d} &= C_2 + \phi_{2,1} y_{1,d} + \phi_{2,2} y_{24,d-1} + \cdots + \phi_{2,48} y_{2,d-2} + \boldsymbol{\alpha}_2^T W_d + \boldsymbol{\beta}_2^T M_d + \boldsymbol{\gamma}_{2,d}^T \boldsymbol{\mathcal{T}}_{2,d} + \varepsilon_{2,d} \\
y_{3,d} &= C_3 + \phi_{3,1} y_{2,d} + \phi_{3,2} y_{1,d} + \cdots + \phi_{3,48} y_{3,d-2} + \boldsymbol{\alpha}_3^T W_d + \boldsymbol{\beta}_3^T M_d + \boldsymbol{\gamma}_{3,d}^T \boldsymbol{\mathcal{T}}_{3,d} + \varepsilon_{3,d} \\
&\vdots \\
y_{24,d} &= C_{24} + \phi_{24,1} y_{23,d} + \phi_{24,2} y_{22,d} + \cdots + \phi_{24,48} y_{24,d-2} + \boldsymbol{\alpha}_{24}^T W_d + \boldsymbol{\beta}_{24}^T M_d + \boldsymbol{\gamma}_{24,d}^T \boldsymbol{\mathcal{T}}_{24,d} + \varepsilon_{24,d}
\end{aligned}
\tag{3}
$$

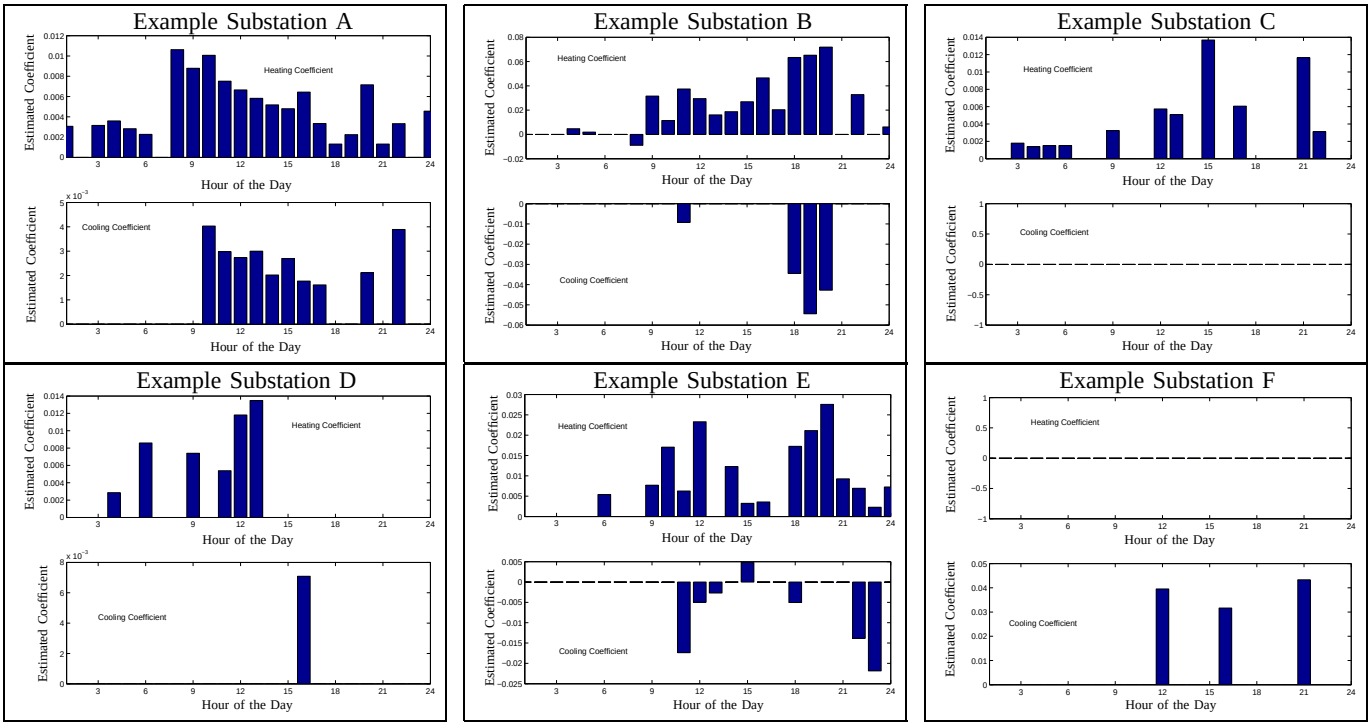

Fig. 4. Parameters Identification, example for 6 substations. The estimated coefficients for Heating (top) and Cooling (Bottom) requirements show different types of sensitivities across substations. Those with zero value are not statistically significant. Maximum effects can occur at different hour of the day, and also with different sign: sometimes cooling increases the load, sometimes the need for cooling has a negative effect on the load.

the first out-of-sample load value using all information available, then wait one hour until the true value of this forecast has been observed, and then forecast the next value again using all available information ("one-step-ahead prediction"). However, planning engineers require forecasts with a longer time horizon, at least a full day in advance. In this case, it is required to predict the first out-of-sample value using all the working sample, then predict the second value out-of-sample accounting from this first prediction, and so on. In practice, it is reasonable to stop this iterative process after 24 hours and update the information with actual observations.

In these forecasting exercises, the external variables are assumed to be known. This is not a problem for the calendar variables, although external weather forecasts [27] should be used instead of actual temperature values. Nevertheless, using actual values for temperature as inputs for the load forecasting helps to assess the model performance without additional error sources. On the other hand, using different temperature values leads to simulation exercises, where the aim is to look at what would be the load if the temperature pattern changes.

Given the estimated variance of the error terms for each equation, confidence intervals for the predictions at the 95% level are computed using standard inference techniques [28]. These confidence intervals have a different width over the day, as the error variance changes. In addition, clearly the confidence intervals becomes wider if more points are being forecasted iteratively without any update in between. This is presented in Figure 5 for 3 selected substations with very different behavior. Each row represents a substation, where the left panel shows the observed load series for a period of 96 hours starting on a Sunday. The center panel shows the forecasts and confidence intervals using the One-step-ahead prediction mode. The observed load series (dashed line) is compared with the forecasted values (thick line). The confidence intervals are also indicated (thin lines). The right panel shows the situation for an Iterative-prediction mode with update every 24 hours.

The best performance is obtained when using the one-step-ahead mode, which implies an update every hour with the actual observations. The iterative-forecasting with updates every 24 hours is less optimal, but depending on the substation, it can still provide good predictions. In order to quantify the

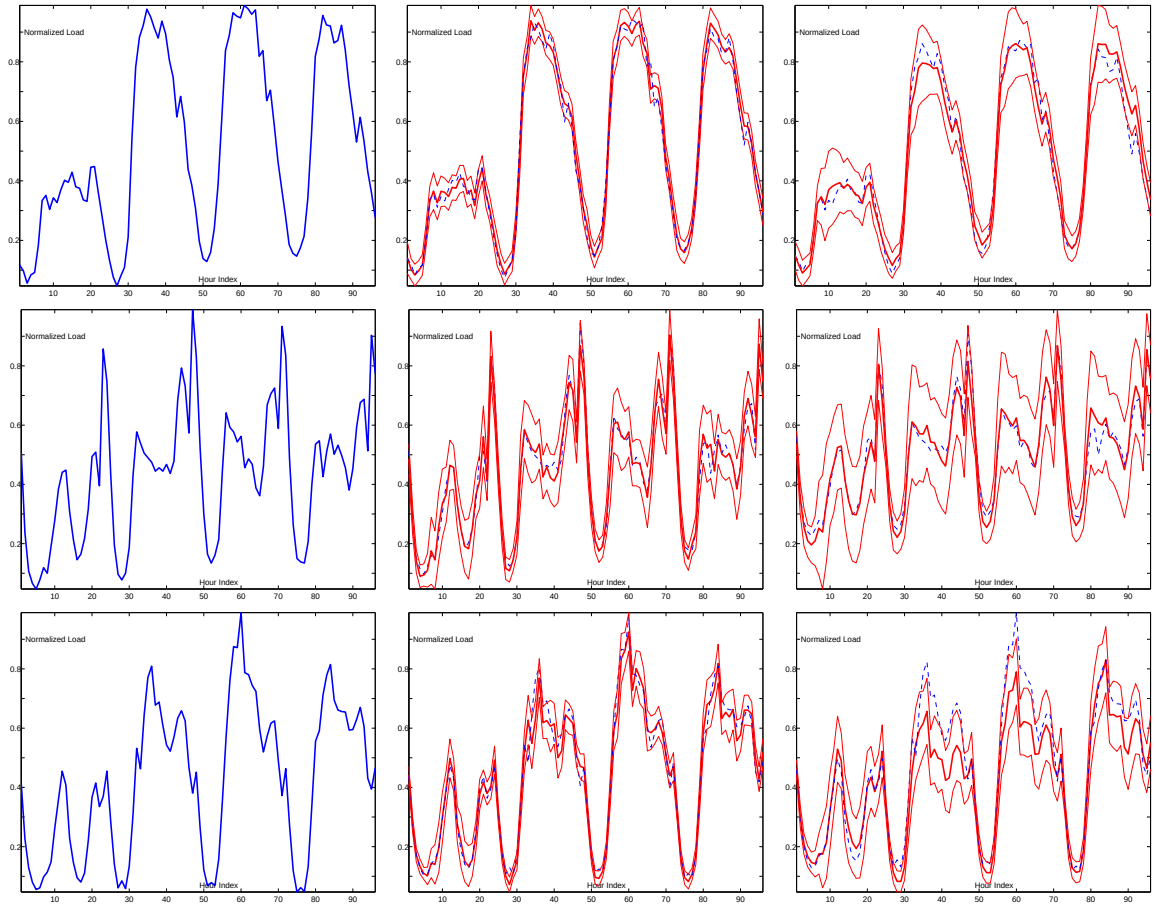

Fig. 5. Out-of-sample predictions for 3 selected substations. Within each row, representing a substation, the observed load series (left) and its forecasts under different updating modes: One-step-ahead (center) and Iterative-forecasts with update after 24 hours (left). On each panel, the forecasts (thick line), the observed value (dashed) and the 95% confidence interval (thin lines) are depicted.

forecasting performance of the models over the different 245 time series, the Mean Absolute Percentage Error (MAPE) and the Root-Mean Squared Error (RMSE) are computed for the following out-of-sample forecasting exercises:

- Case I: One-step-ahead prediction, for a 7 days-period.
- Case II: Iterative prediction with update every 24 hours, for a 7 days period.

Table I shows how many substations are included within each category of different error levels. Clearly, the PAR(48) model template can produce excellent forecasting results (less than 3% error in this setting) for 238 load series when working with iterative predictions with updates every hour (one-step-ahead prediction, Case I). If the update is made every 24 hours (Case II), then 166 time series have their errors below 3% . The model performance can be improved by adding more terms into the PAR(p) formulation.

TABLE I

MODELS PERFORMANCE FOR DIFFERENT FORECASTING CASES.

	Number of Substations for which prediction error is less than					
	$<1\%$	$<3\%$	$<5\%$	$<8\%$	$<10\%$	$<20\%$
Case I						
MAPE	206	238	241	241	242	245
RMSE	201	238	241	242	242	245
Case II						
MAPE	13	166	189	205	214	245
RMSE	8	166	200	229	234	245

IV. TYPICAL DAILY PROFILES

The definition of a Typical Daily Profile for each substation from the parameters of the system (3) is described in this section.

A. Equivalent Vectorial Notation and Convergence

By defining a vector $\boldsymbol{Y}_d = [y_{1,d} \quad y_{2,d} \quad y_{3,d} \quad \cdots \quad y_{23,d} \; y_{24,d}]^T \in \mathbb{R}^{24}$, containing the load information for the 24 hours of day d, it is possible to write (3) as

$$\Phi_0 \boldsymbol{Y}_d = \boldsymbol{C} + \Phi_1 \boldsymbol{Y}_{d-1} + \Phi_2 \boldsymbol{Y}_{d-2} + \Phi_3 \boldsymbol{X}_d + \boldsymbol{\varepsilon}_d \quad (4)$$

with Φ_0, Φ_1, Φ_2 and Φ_3 containing the coefficients ϕ of the system (3) [14]. The matrix $\boldsymbol{X}_d$ contains all exogenous variables for temperature and calendar information. The system is now written in a Vector Auto-Regression (VAR) form with 2 lagged values for $\boldsymbol{Y}_d$ [28], and it is easily seen that the original PAR(48) is equivalent to a VAR(2) formulation. Once the system (3) has been estimated, all coefficients of the matrices Φ_0, Φ_1, Φ_2 and Φ_3 are known. The next-day forecasts can be simply written as

$$E_d[\boldsymbol{Y}_{d+1}] = \hat{\boldsymbol{Y}}_{d+1} = \Phi_0^{-1}\{\boldsymbol{C} + \Phi_1 \boldsymbol{Y}_d + \Phi_2 \boldsymbol{Y}_{d-1} + \Phi_3 \boldsymbol{X}_{d+1}\}$$

where E_d is the expectation taken at time d. The matrix Φ_0 is always invertible as it is a lower triangular matrix with ones in the main diagonal. Applying this formulation iteratively for n days, a multi-step ahead prediction can be obtained. The above equation requires the knowledge of the values of $\boldsymbol{X}_{d+1}$ on day d. At least the calendar information is always available for the future, and for the temperature information one should rely on the best available weather-temperature forecasts. In any case, the information contained in the variables $\boldsymbol{X}_d$ are exogenous (to the load) as they are capturing seasonal effects to the load itself. Thus, a very interesting question is to check what happens to the load when these variables are defined to be zero, or in other words, when all seasonal effects are captured and removed from the load model[5].

Setting $\boldsymbol{X}_d = 0$, the system becomes

$$\Phi_0 \boldsymbol{Y}_d = \boldsymbol{C} + \Phi_1 \boldsymbol{Y}_{d-1} + \Phi_2 \boldsymbol{Y}_{d-2} + \boldsymbol{\varepsilon}_d. \tag{5}$$

If this equation is used in iterative-forecasting mode for n periods ahead, it converges, under stability conditions, to a unique value $\boldsymbol{Y}^*$, which can be computed as

$$\boldsymbol{Y}^* = \{\Phi_0 - \Phi_1 - \Phi_2\}^{-1} \boldsymbol{C}. \tag{6}$$

Convergence is achieved if and only if the VAR system (5) is stationary, i.e., if the equation

$$|\Phi_0 - \Phi_1 z - \Phi_2 z^2| = 0 \tag{7}$$

has all its roots z_i outside the unit circle [28].

B. Typical Daily Profile Definition

The convergence condition (7) is verified for each of the 245 substations. This is not surprising, since an autoregressive model defined with a "correct" order leads to a stationary formulation, otherwise it can not be used as a stable forecasting model. A model that does not satisfy the convergence condition should be allowed to include extra lag terms, in order to write a VAR with a higher order until it achieves stationarity. In this dataset, every load series has its own convergence vector $\boldsymbol{Y}^*$, computed from the estimated model coefficients contained in Φ_0, Φ_1, and Φ_2. As each vector $\boldsymbol{Y}_d$ contains daily information of the load, the $\boldsymbol{Y}^*$ convergence vector, computed after all seasonal effects have been removed, can be interpreted in terms of daily load information: it contains information about the typical daily profile of the load in absence of seasonal and temperature information. Therefore, we define the Typical Daily Profile (TDP) as follows.

Definition: The Typical Daily Profile (TDP) $\boldsymbol{Y}^*$ of a sample load series $\boldsymbol{Y}_d \in \mathbb{R}^{24}, d = 1, \cdots, N_d$ is the convergence vector of a VAR(q) system obtained from an equivalent PAR(p) after extracting all exogenous information.

The definition requires the obtained VAR(q) system to be stationary, a condition attainable in the process of defining the order of the PAR(p) process. It is also an empirical definition, as it is based on a statistically sound procedure which is the estimation of a set of autoregressions. The TDP can be used as a representation of the corresponding substation for which a PAR(p) model was initially computed. The main advantage of the TDP is that it provides a representation, in terms of a daily load profile, which is independent of the seasonal and calendar variations present in the load. In other words, the difference between the TDP and the actual observed daily load profile for a given day is attributable only to the seasonal and calendar effects for that particular day.

[5]Removing temperature and seasonal effects is a usual task in long-term grid management, to compute year-to-year growth trends, to identify how much of the yearly peak was due to weather, scenario analysis, etc.

C. Typical Daily Profiles in the current sample

The dataset contains information on 245 substations. Each substation can be estimated using the PAR(48) model template, and its TDP can be computed, leading to a set of 245 TDPs. To have an assessment of the difference between each TDP and its underlying load history, Figure 6 shows 8 substations for which the TDP (thick line) is compared with actual daily load profiles observed over 500 days randomly selected from the dataset. For each substation, the TDP captures the behavior of the load that is not attributable to seasonal and calendar variations. It is also clear that the daily behavior of these substations are not the same, with peaks located at different hours of the day. Using TDPs is thus a simple and powerful procedure for comparing the profiles of substations. Once these profiles have been identified, a natural question to ask is how many different types of profiles there are in the sample. If it is assumed that a different profile means a different type of underlying customer (or group of customers), the question translates into customer segmentation, or customer clustering.

V. Clustering using Typical Daily Profiles

This sections describes the implementation of a clustering exercise. Having identified the 245 Typical Daily Profiles (TDP), they are used as representations for the original load series. Unsupervised clustering aims at identifying different groups or patterns in a data sample, doing so without external information from the user. In this setting, the aim is to know how many different types of load profiles have to be considered, having no a priori information about the particular composition of the demand beyond each substation level.

A. Implementation

Clustering is a wide concept within statistics and machine learning [20], [21]. In plain terms, the goal of a clustering algorithm is to identify groups of "similar" data within the dataset, without using external information, and assign each datapoint into (at least) one of the groups, or clusters. In this application K-means is used [29], a classic clustering technique available as an standard function in many mathematical software packages. As a preprocessing step, Principal Components Analysis [30] is applied to the data in order to reduce the dimensionality of the problem. It is found that by keeping 9 principal components it is possible to recover 99% of the original set of TDPs. The application of K-means requires the user to give the number of desired clusters N_C as input parameter to the algorithm. For this case, N_C is tested from 2 to 15, which is a reasonable range, as empirical research has identified a similar number of different

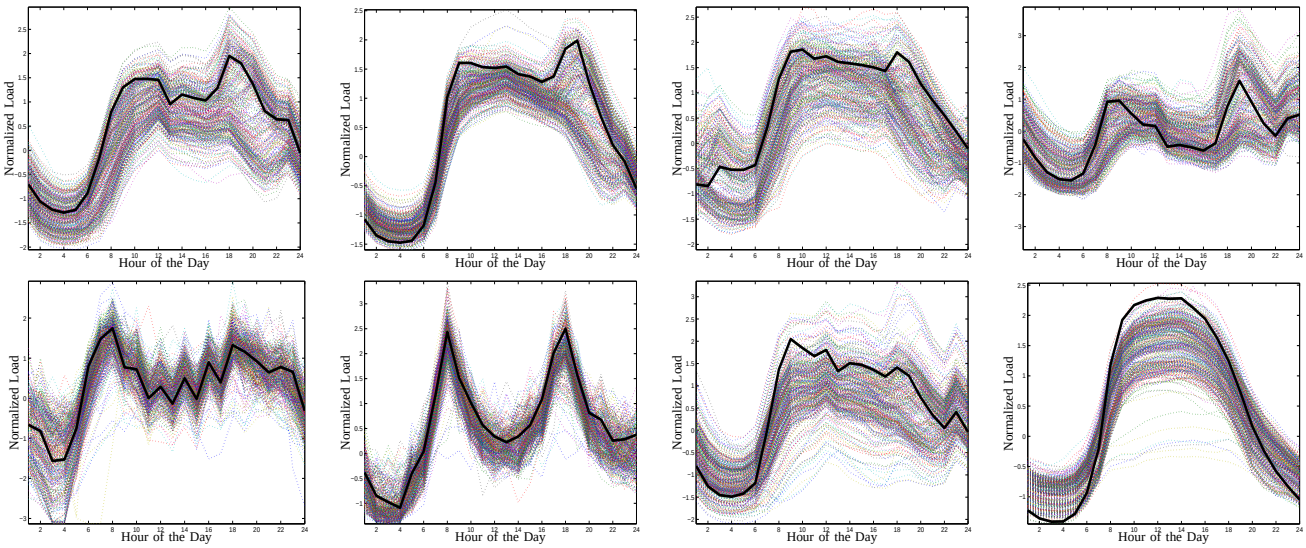

Fig. 6. TDPs for selected 8 substations. Each panel shows the TDP of a substation (thick line) in relation to the corresponding daily loads (dotted lines) observed over 500 random days.

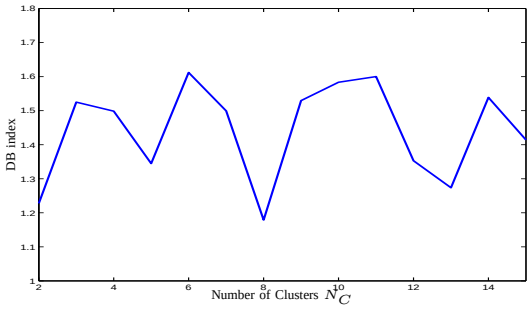

Fig. 7. Davies-Bouldin Validity Index. The local minimum at $N_C = 8$ shows that a partition containing 8 clusters can be selected.

profiles [10], [12]. In order to choose the "best" clustering, performance or validity indices are typically used [31]. In this paper the so-called Davies-Bouldin (DB) validity index [32], which is a function of the ratio of the sum of within-cluster scatter to between-cluster separation, is applied. For clusters denoted $Q_i, i = 1, \ldots N_C$, the DB index is

$$DB = \frac{1}{N_C} \sum_{j=1}^{N_C} \max_{l \neq j} \frac{S(Q_j) + S(Q_l)}{d(Q_j, Q_l)} \tag{8}$$

where $S(Q_k)$ is the (average) distance within cluster Q_k and $d(Q_j, Q_l)$ is the distance between clusters Q_j and Q_l. The "optimal" number of clusters N_C is the one for which the DB validity index shows a minimum value.

B. Clustering Results

A local minimum for the Davies-Bouldin validity index is found at $N_C = 8$ clusters (Figure 7). The 8 different clusters are represented on Figure 8. According to interpretations by industry experts, the sample contains an important quantity of profiles with "residential" behavior, particularly clusters 1 and 5. Clusters 4 and 7 can be related to "commercial" or "business" activities. Cluster 1 captures a profile with equal peaks in the morning and evening, and a low demand in between. Clusters 3, 6 and 8 capture different variants of substation with very low demand during daytime, as e.g. street lightning or other industrial activities for which electrical energy is used at night. Possibly a more detailed characterization of the profiles based on the clustering exercise can be achieved by applying more complex techniques, or by defining an ad-hoc clustering technique for load profiles, to take into account e.g. the unbalanced presence of different profiles in the sample. Although the present exercise can be a start for industry managers to draw conclusions on the current sample, it is certainly an interesting research topic for further development.

VI. Conclusion

The general problems of short-term load forecasting and profile identification can be addressed within a unified framework by using the proposed methodology based on the use of Periodic Autoregressive (PAR) models. Starting from a single PAR model template containing 24 seasonal equations,and using the last 48 load values within each equation, it is possible to estimate a model suitable for short-term forecasting. A comparison between substations can be made based on the identified parameters, as shown in the temperature sensitivity example. In addition, by exploiting the stationarity properties of the PAR model, it is possible to compute a convergence vector that can be interpreted as a Typical Daily Profile. This convergence vector is computed from the estimated coefficients of the PAR model.

This methodology is successfully applied within the framework of an applied project in cooperation with the Belgian National Grid Operator ELIA. In a sample of 245 substations, each one a time series containing about 40,000 datapoints, estimated PAR models can generate short-term forecasts with a satisfactory accuracy. After individual PAR models are estimated, their convergence vectors are computed and the original sample can now be represented by 245 Typical Daily Profiles. This set of 245 Typical Daily Profiles can be used for

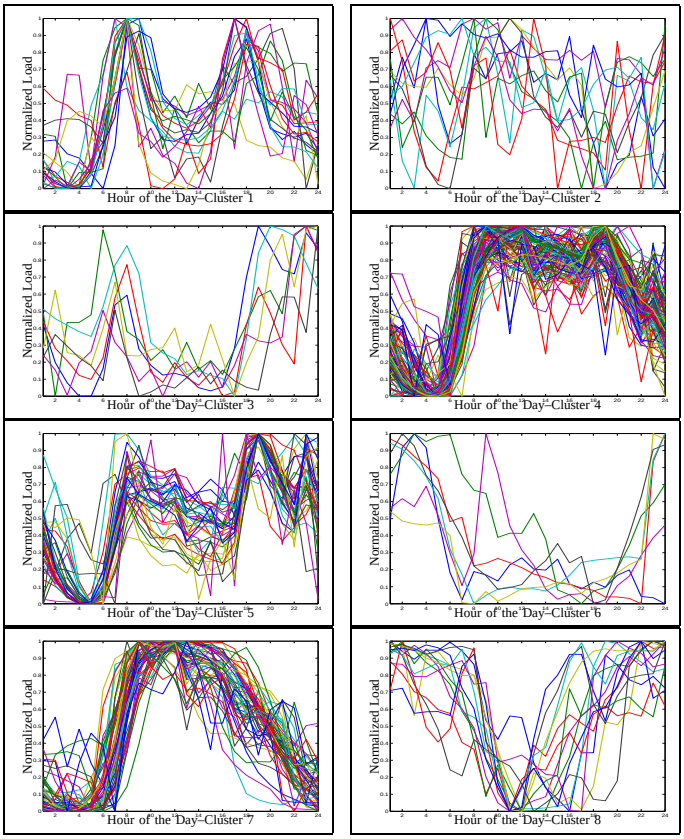

Fig. 8. Clustering Exercise. Using K-means and the DB validity index it is possible to identify 8 clusters in the set of 245 TDPs.

clustering, in order to quantify how many different groups or classes of profiles can be identified within the sample. Using a classic clustering technique, it is possible to identify 8 different clusters, capturing the different types of profiles within the sample. This information can be used for further specific refinement of the forecasting models, such as building ad-hoc models for each specific cluster. It can be a starting point for the application of more complex clustering techniques.

Acknowledgments This work was supported by grants and projects for the Research Council K.U.Leuven (GOA-Mefisto 666, GOA-Ambiorics, several PhD/Postdocs & fellow grants), the Flemish Government (FWO: PhD/Postdocs grants, projects G.0240.99, G.0407.02, G.0197.02 (power islands), G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, ICCoS, ANMMM; AWI;IWT:PhD grants, GBOU(McKnow)Soft4s), the Belgian Federal Government (Belgian Federal Science Policy Office: IUAP V-22; PODO-II (CP/01/40), the EU(FP5-Quprodis;ERNSI, Eureka 2063-Impact;Eureka 2419-FLiTE) and Contracts Research/Agreements (ISMC/IPCOS,Data4s,TML,Elia,LMS,IPCOS,Mastercard). B. De Moor and R. Belmans are full professors at the K.U.Leuven, Belgium. The scientific responsibility is assumed by its authors.

REFERENCES

[1] E. Mariani and S. Murthy, *Advanced Load Dispatch for Power Systems*, ser. Advances in Industrial Control. Springer-Verlag, 1997.

[2] N. Amjady, "Short-term hourly load forecasting using time-series modeling with peak load estimation capability," *IEEE Trans. Power Syst.*, vol. 16, no. 4, pp. 798–805, 2001.

[3] R. Ramanathan, R. Engle, C. Granger, and C. Vahid-Aragui, F.and Brace, "Short-run forecasts of electricity load and peaks," *International Journal of Forecasting*, no. 13, pp. 161–174, 1997.

[4] S.-J. Huang and K.-R. Shih, "Short term load forecasting via ARMA model identification including non-gaussian process considerations," *IEEE Trans. Power Syst.*, vol. 18, no. 2, pp. 673–679, 2003.

[5] H. Steinherz, C. Pedreira, and R. Castro, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans. Power Syst.*, vol. 16, no. 1, 2001.

[6] A. Khotanzad, R. Afkhami-Rohani, and D. Maratukulam, "ANNSTLF-artificial neural network short-term load forecaster-generation three," *IEEE Trans. Power Syst.*, vol. 13, no. 4, pp. 1413–1422, 1998.

[7] K.-H. Kim, H.-S. Youn, and Y.-C. Kang, "Short-term load forecasting for special days in anomalous load conditions using neural networks and fuzzy inference method," *IEEE Trans. Power Syst.*, vol. 15, no. 2, pp. 559–565, 2000.

[8] L. Mohan Saini and M. Kumar Soni, "Artificial neural network-based peak load forecasting using conjugate gradient methods," *IEEE Trans. Power Syst.*, vol. 17, no. 3, pp. 907–912, 2002.

[9] D. Fay, J. Ringwood, M. Condon, and M. Kelly, "24-h electrical load data-a sequential or partitioned time series?" *Neurocomputing*, no. 55, pp. 469–498, 2003.

[10] H. Liao and D. Niebur, "Load profile estimation in electric transmission networks using independent component analysis," *IEEE Trans. Power Syst.*, vol. 18, no. 2, pp. 707–715, 2003.

[11] S. Heunis and R. Herman, "A probabilistic model for residential consumer loads," *IEEE Trans. Power Syst.*, vol. 17, no. 3, pp. 621–625, 2002.

[12] J. Jardini, C. Tahan, M. Gouvea, and S. Ahn, "Daily load profiles for residential, commercial and industrial low voltage consumers," *IEEE Trans. Power Delivery*, vol. 15, no. 1, pp. 375–380, 2000.

[13] E. Carpaneto, G. Chicco, R. Napoli, and M. Scutariu, "Customer classification by means of harmonic representation of distinguishing features." IEEE Bologna Power Tech Conference, 2003.

[14] P. Franses and R. Paap, *Periodic Time Series Models*. Oxford University Press, 2003.

[15] B. Troutman, "Some results in periodic autoregressions," *Biometrika*, vol. 66, pp. 219–228, 1979.

[16] A. McLeod, "Diagnostic checking of periodic autoregression models with applications," *The Journal of Time Series Analysis*, vol. 15, no. 2, pp. 221–223, 1994.

[17] D. Osborn and J. Smith, "The performance of periodic autoregressive models in forecasting seasonal U.K. consumption," *Journal of Business & Economic Statistics*, vol. 7, pp. 117–127, 1989.

[18] D. Osborn, "The implications of periodically varying coefficients for seasonal time-series processes," *Journal of Econometrics*, vol. 48, pp. 373–384, 1991.

[19] G. Guthrie and S. Videbeck, "High frequency electricity spot price dynamics: An intra-day markets approach," New Zealand Institute for the Study of Competition and Regulation, Tech. Rep., 2002.

[20] P. Berkhin, "Survey of clustering data mining techniques," Accrue Software Inc., Tech. Rep., 2002.

[21] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.

[22] G. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day, 1970.

[23] S. Hylleberg, *Modelling Seasonality*. Oxford University Press, 1992.

[24] L. Yang and R. Tschernig, "Non- and semiparametric identification of seasonal nonlinear autoregression models," *Econometric Theory*, vol. 18, pp. 1408–1448, 2002.

[25] J. Johnston, *Econometric Methods*, ser. Economics Series. McGraw-Hill, 1991.

[26] R. Engle, C. Granger, J. Rice, and A. Weiss, "Semiparametric estimates of the relation between weather and electricity sales," *Journal of the American Statistical Association*, vol. 81, no. 394, pp. 310–320, 1986.

[27] J. Taylor and R. Buizza, "Neural network load forecasting with weather ensemple predictions," *IEEE Trans. Power Syst.*, vol. 17, no. 2, pp. 626–632, 2002.

[28] J. Hamilton, *Time Series Analysis*. Princeton University Press, 1994.

[29] J. Tou and R. Gonzalez, *Pattern Recognition Principles*. Addison-Wesley, 1974.

[30] I. Jolliffe, *Principal Components Analysis*, ser. Springer Series in Statistics. Springer-Verlag, 1986.

[31] A. Hardy, "On the number of clusters," *Computational Statistics & Data Analysis*, vol. 23, pp. 83–96, 1996.

[32] D. Davies and D. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 1, pp. 224–227, 1979.