

Functional Principal Component Analysis as a Versatile Technique to Understand and Predict the Electric Consumption Patterns

Davide Beretta, Samuele Grillo, Davide Pigoli, Enea Bionda, Claudio Bossi, and Carlo Tornelli

Abstract—Understanding and predicting the electric consumption patterns in the short-, mid- and long-term, at the distribution and transmission level, is a fundamental asset for smart grids infrastructure planning, dynamic network reconfiguration, dynamic energy pricing and savings, and thus energy efficiency. This work introduces the Functional Principal Component Analysis (FPCA) as a versatile method to both investigate and predict, at different level of spatial aggregation, the consumption patterns. The method was applied to a unique and sensitive dataset that includes electric consumption and contractual information of Milan metropolitan area. The decomposition of the load patterns into principal functions was found to be a powerful method to identify the physical and behavioral causes underlying the daily consumptions, given knowledge of exogenous variables such as calendar and meteorological data. The effectiveness of long-term predictions based on principal functions was proved on Milan’s metropolitan area data and assessed on a publicly-available dataset.

Index Terms—Electric consumption, functional principal component analysis, FPCA, patterns, analysis, prediction.

I. INTRODUCTION

THE ability to predict the daily electric consumption, both at the distribution and transmission level, in the short-, mid- and long-term, is an essential asset in the evolving scenario of modern cities, where the recent advances in the microelectronics and in the information and communication technologies are finally making possible the transition from the old centralized energy management model to the smart grid paradigm, ever more characterized by the presence of energy generation from distributed sources [1]–[4]. The advantages coming from the prediction of the electric consumption, and the corresponding daily load patterns, are manifold and include, at different levels of spatial aggregation, energy savings, infrastructure planning, and energy pricing [5], [6]. On the one hand, the short term prediction at the distribution level can be exploited to re-configure the electric network on a

timely manner, increasing or decreasing the security level of restricted network regions on the basis of specific needs, and to provide basic information for more customer-based personalized market policies and electric power services, with the combined ultimate scope of preventing load peaks. In the same time window, the prediction at the transmission level can provide useful information to regulate the energy generation and provisioning to macro areas, ultimately allowing to reduce the waste of non-storable electric energy and thus apply favorable energy pricing. On the other hand, the mid- and long-term prediction is beneficial, at both the distribution and transmission levels, to operations planning and infrastructures improvements, with an expected stronger impact when the consumption forecasts are combined with projections on population. In this context, analysis methods that aim at unveiling the physical causes underlying the consumption patterns at different levels of spatial aggregation are of unparalleled importance to gain insights into the mechanisms regulating the electric loads. A variety of different techniques and methodologies has been so far proposed, unraveling the dependency of the daily load patterns on single customers’ behavior as a function of the spatial aggregation [7]–[14]. Among those techniques, the analysis and the classification are generally pursued by means of time-series clustering methods [14]–[19], which suffer from high sensitivity to the choice of the metric [20], while the predictions are based on parametric and non-parametric methods [21], such as multi-regression and auto-regressive historical time series [10], [11], [13], [22]–[24], and exponential smoothing [25]–[27]. In some cases, the prediction follows data dependencies reduction by means of principal component analysis. In all of the aforementioned methods, the physical causes underlying the load patterns remain undisclosed, as the techniques, regardless of being more or less effective, are based on mathematical tools that are transparent to the shape of the analyzed patterns. This work introduces the use of the Functional Principal Component Analysis (FPCA) [28] as an alternative and effective method (i) to investigate the daily electric load patterns at different level of aggregations, with a spatial-aggregation-dependent level of accuracy, providing an unparalleled way to correlate the observed consumption patterns with exogenous causes, and (ii) to predict the daily electric consumption patterns in the short- and long-term. The main contribution of this work is therefore twofold. On the one hand, it shows that the FPCA can be used as a tool to analyze complex dataset

This work has been financed by the Research Fund for the Italian Electrical System in compliance with the Decree of April 16, 2018.

D. Beretta was with RSE SpA, via R. Rubattino, 54, I-20134 Milano (MI), Italy. (e-mail: davide.beretta@empa.ch)

D. Pigoli is with the Department of Mathematics, King’s College, London, WC2R 2LS, UK (e-mail: davide.pigoli@kcl.ac.uk).

S. Grillo, is with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, piazza Leonardo da Vinci, 32, I-20133 Milano, Italy (e-mail: samuele.grillo@polimi.it).

E. Bionda, C. Bossi and C. Tornelli are with RSE SpA, via R. Rubattino, 54, I-20134 Milano (MI), Italy (e-mail: {enea.bionda, claudio.bossi, carlo.tornelli}@rse-web.it).

of electric consumptions¹ and to understand the physical and behavioral causes underlying the load patterns. On the other hand, it shows that the FPCA decomposition can be exploited to predict the electric consumptions in the long-term, with very competitive performances.

II. METHODS

A. Functional Principal Component Analysis (FPCA)

The FPCA is a statistical method to represent functional data in an orthonormal basis of the Hilbert space that consists of the eigenfunctions of the covariance operator. If $f_i(t)$ is the i -th functional data, or curve, of the variable t , according to the FPCA [28]

$$f_i(t) = \mu(t) + \sum_{k=1}^p c_{k,i} \varphi_k(t), \quad (1)$$

where $\mu(t) = n^{-1} \sum_{i=1}^n f_i(t)$ is the average of the input functions and $\varphi_k(t)$ are the eigenfunctions of the sample covariance operator

$$S(t, \tau) = \frac{1}{n-1} \sum_{i=1}^n (f_i(t) - \mu(t))(f_i(\tau) - \mu(\tau)), \quad (2)$$

defined such that $\langle S(t, \tau) \varphi_k(\tau) | \varphi_k(t) \rangle = \lambda_k$, being $\langle \cdot | \cdot \rangle$ the inner product and λ_k the k -th eigenvalue of S . The $\varphi_k(t)$ are usually called functional principal components and the $c_{k,i}$, which are defined with sign, are given the name of scores. The shape of the $\varphi_k(t)$, together with the sign of the scores, determines the “direction” of variability of the data with respect to the average consumption pattern $\mu(t)$. The absolute value of the score gives information about how much the consumptions deviate from the average. The scores vector $c_{k,i}$, with $k \in [1, p]$, gives the representation of the i -th curve in the new basis. The principal component basis is ordered in the sense that $\sum_i \langle f_i(t) - \mu(t) | \varphi_k(t) \rangle^2 > \sum_i \langle f_i(t) - \mu(t) | \varphi_l(t) \rangle^2$, for every $l > k$. This means that the first p functional principal components provide the best possible approximation (in a square error sense) of the data among all the possible sets of basis functions with p elements. Moreover, by definition of the functional principal components, the variance of the (centered) data projected in the direction of the k -th component is the k -th eigenvalue of the sample covariance operator S . This peculiarity can be exploited to select a number of principal components sufficient to represent the functional data with a certain level of precision, e.g. more than 90% of the variability of the data, or such that the addition of further components does not improve significantly the amount of information reproduced. The cumulative proportion of variability explained by the first p principal components can be computed as $\vartheta(p) = \sum_i^p \lambda_i / \sum_i \lambda_i$. In this work, the $f_i(t)$ is the daily electric consumption pattern of the i -th day of a given electric station or spatial aggregation of electric stations, where $t \in [0, 23]$ is the discrete time

instant, i.e. the hour of the day, at which $f_i(t)$ is evaluated. The FPCA was exploited to identify correlations between the shape of the $\varphi_k(t)$ and exogenous variables such as the day of the week, the month of the year, the temperature, and the relative humidity, to name a few. This part of the analysis was not trivial and required a graphical representation of the components and of the scores as a function of the exogenous variables to visualize trends and behaviors, besides a decent knowledge of the conditions under which the consumption patterns were generated. To the purpose, the present study was extensively supported by graphs showing: i) the first three $\varphi_k(t)$, which were demonstrate to explain more than the 80% of the variability of each consumption pattern analyzed, ii) the daily and monthly distribution of the scores, and iii) the distribution of the scores as a function of the temperature and of the relative humidity. The last figure had particular relevance in the study as it allowed to identify a comfort zone that corresponds to a minimum in the electric consumption.

B. Prediction

The FPCA can be integrated into any time-series predictive model to predict future patterns. The peculiarity of the FPCA-based predictive models is that they predict the future values of the $c_{k,i}$ and thus, since the $\varphi_k(t)$ are orthogonal to each other, they can predict the scores of a selected subset of the $\varphi_k(t)$ without necessarily calculate all the $c_{k,i}$ simultaneously. This property can be exploited to reach a compromise between the complexity of the model and the explained variability predicted. In this framework, the chosen generic predictive model was linear, i.e.,

$$c_{i,k} = \mathbf{x}_{i,k}^T \boldsymbol{\beta}_k + \varepsilon_{i,k} \quad (3)$$

where $\mathbf{x}_{i,k}$ is the vector of the predictors for the k -th FPC score of the i -th day, $\boldsymbol{\beta}_k$ is the vector of the coefficients (to be estimated) associated to the k -th score, and $\varepsilon_{i,k}$ are the zero-mean independent Gaussian errors. The coefficients are estimated via ordinary least squares and the estimated model is used to predict the future scores. The set of predictors used for each score was dynamically selected with a stepwise procedure in order to minimize the Akaike Information Criterion (AIC), i.e.,

$$\text{AIC} = 2p - 2 \ln L, \quad (4)$$

where p is the number of the predictors of the model and L is the corresponding likelihood function. This approach allowed to choose a model showing a good compromise between complexity and accuracy. Since the error model was assumed Gaussian, the likelihood function was proportional to the sum of the square of the residuals. The pool of predictors included the calendar time (number of days passed from the first recorded observation, to account for long term trends), the month of the year (categorical variable represented by eleven indicator variables, to account for seasonal effects), the day of the month (continuous variable to account for inter-month trends), the day of the week (categorical variable represented by six indicator variables, to account for working/festive patterns), and the presence of popular events (three indicator variables to signal the presence of Milan Fashion Week,

¹In our case, we applied the proposed methodology to a dataset which gathers consumption measurements from approximately 3500 MV/LV secondary substations for three years with a 15 min sampling time. This amount of data corresponds to a 350-million-tuples table.

Expo 2015, Milan Design Festival), which might affect the consumption patterns due to the abnormal presence of people in specific geographical areas. The meteorological variables can be included for short-term predictions in case accurate forecasting or other proxies are made available.

In this work, the analysis focused on two different predictions, i.e., i) the daily consumption pattern, and ii) the monthly average energy consumption, which belongs to short- and mid/long-term prediction cases, respectively. The goodness of the prediction of the daily consumption patterns was measured by the Mean Absolute Percentage Error (MAPE), that is

$$\text{MAPE} = \frac{1}{m} \sum_{i=1}^m \left| \frac{x_i - y_i}{x_i} \right| \times 100, \quad (5)$$

where x_i and y_i are the measured and predicted values of the test data set, respectively, and m is the number of predicted values that define the time series in a specific time window. In this context, $m = 24$ is the number of samples that define the daily consumption pattern of a given day of the year. The goodness of the prediction on the monthly average energy consumption was measured by the energy percentage error $\varepsilon\%$, that is

$$\varepsilon\% = \frac{1}{m} \frac{|\sum_{i=1}^m x_i - \sum_{i=1}^m y_i|}{\sum_{i=1}^m x_i} \times 100, \quad (6)$$

where x_i and y_i , and m are again the measured and predicted values, and the number of samples, respectively.

III. DATA

The initial data set comprised: i) the measurements, taken every 15 min, of the average power of 3386 electric secondary MV/LV substations deployed on Milan's metropolitan area; ii) the geo-localization of the substations; iii) the contractual information of the customers connected to 5312 secondary substations deployed on Milan's metropolitan area, including both consumption and generation, and reporting, for each contract, start and expiry date, allocated maximum power and type of customer, i.e., residential, non-residential or public lighting; iv) the meteorological data, including temperature, relative humidity, radiation, rainfall and wind speed, recorded every hour by a number of ARPA's² meteorological stations deployed on Milan's metropolitan area; v) the calendar of national and local festivity, including particular events characterized by massive affluence, i.e., the fashion week, the design week and EXPO 2015; and vi) the geographical borders of Milan's neighborhoods. Data i)–iii) were made available by Milan's Distribution System Operator (Unareti S.p.A.), while data iv) and vi) were freely downloaded from ARPA and Milan's municipality website, respectively. Data v) was generated on the basis of the information freely retrieved from the web. Data covered, almost completely, the period from January 2014 to October 2017, and required some elaboration to take into account for reading failures and for different averaging time windows, local timezone and daylight saving time. Since the ARPA's meteorological stations deployed on Milan's

area were non-uniformly distributed, the weather conditions were averaged over the values measured by all the stations. Data on customers were aggregated to the substation level in order to produce a large database, which could be easily queried, reporting the history of each substation. Considered the volume of the available data (tens of GB), within the framework of an ever increasing and potentially auto-updating dataset, data storage, elaboration and interrogation were approached following big data methodologies, using the Apache Spark framework, and exploiting the Amazon Web Services Simple Cloud Computing and Simple Cloud Storage Service. The study was limited to the population of electric stations showing stable contractual characteristics in the period 2014–2017. A given substation was considered stable with respect to a particular contractual characteristic x , in a specific time window, if the characteristic x , resulting from the aggregation over all the customers connected to that particular substation, satisfied the condition

$$\max(x) - \min(x) < 0.1 \text{ avg}(x), \quad (7)$$

in the time interval considered. In this study, the contractual characteristics considered were: i) contractual consumption, ii) fraction of contractual consumption absorbed by residential customers, iii) fraction of contractual consumption absorbed by public lighting, iv) contractual generation, and v) fraction of contractual generation supplied by photovoltaics. The result of the operation applied to the database for the time interval 2014–2017, returned 3642 stations out of 5312, i.e., approximately 68% of the initial population of substations the contractual characteristics of which were made available by the local electric distribution provider. The population of stable electric stations was used to build the non-normalized probability distribution functions (PDFs) and the normalized cumulative distribution functions (CDFs) of a subset of contractual characteristics, with the aim to visualize the framework of the substations deployed on Milan's metropolitan area from the point of view of the power supplier company. This allowed to define a window for each of the contractual characteristics that takes into account for the most representative substations of Milan's metropolitan area, and to identify the presence of substations showing particular contractual characteristics, the analysis of which could explain specific load patterns. In particular, the PDFs revealed the existence of a number of substations having contractual consumption (almost) entirely absorbed by residential, non-residential, or public lighting customers. Most of the electric stations display aggregated contractual consumption in the interval 500–1500 kW and fraction of consumption due to residential customers in the range 30–70%. Data also reveal that the fraction of consumption absorbed by public lighting is less than 1% of the contractual consumption, that the contractual generation is always less than 10 kW, and that the contractual generation is entirely provided by photovoltaics. Since the power measurements provided by the DSO covered only a subset of the stable electric stations registered in the customers contracts database, the FPCA of the load patterns was limited to the power curves of a subset of 2523 electric stations out of the 3386 initially available. Different level of spatial aggregations were consid-

²ARPA is the "Agenzia Regionale per la Protezione Ambientale", i.e., Regional Agency for Environmental Protection.

ered, i.e., the single electric station, the neighborhood and the whole Milan’s city area. To this purpose, the data on the electric stations were aggregated to the NIL (Nucleo di Identità Locale, i.e. neighborhood) and Milan levels, respectively. In each case, a database with the unique identification code of the entity (substation, NIL or Milan), the aggregated power measurements, the date and day time of the recorded events, the spatially-averaged weather conditions and the aggregated contractual characteristics of the entity was built, allowing again to easily access the aggregated data by performing simple database queries. As the FPCA aimed at understanding the complexity of the electric load patterns by unveiling the exogenous causes that underlie the consumptions, a series of filters were applied to the set of electric stations to extract a number of populations of stations showing specific contractual characteristics, stable in the time interval 2014–2017. These populations were given the name of residential (RES), non-residential (NRS), public lightning (PLT), photovoltaic (PVG), mixed (MIX), NIL and city (CTY), respectively, and the criteria used to select them are reported in Table I. A filtering process was necessary to remove all the stations affected by incomplete and/or corrupted data. This process followed three distinct steps: i) deletion of days characterized by incomplete data, i.e., with less than 24 entries per day per electric station. The substations having a number of incomplete days higher than 20% of the available calendar days were removed completely from the sub-populations; ii) deletion of days characterized by corrupted data. The RES, NRS and PVG were considered corrupted if the power reading was equal to 0 kW at least once within the day. The sub-population of PLT was considered corrupted if the power reading was equal to 0 kW for the whole day. The corresponding substations were deleted from the sub-populations if the number of corrupted days was higher than 10% of the available calendar days; iii) deletion of the substations showing non-corrupted data in less than 1095 days (approximately three years of data). The number of stations belonging to each population after the filtering is reported in Table I under the column “FPCA”. They represent the final dataset over which the FPCA described in the following was applied to.

Table I
LIST OF POPULATIONS CONSIDERED IN THIS WORK, ALONG WITH THEIR CONTRACTUAL CONSUMPTION AND CHARACTERISTICS, AND THEIR NUMEROSITY.

Pop.	kW	Contractual characteristics	no.	FPCA
RES	500–1500	$P_{\text{res}} \geq 90\%$ & $\frac{P_{\text{gen}}}{P_{\text{con}}} \leq 10\%$	12	7
NRS	500–1500	$P_{\text{res}} \leq 10\%$ & $\frac{P_{\text{gen}}}{P_{\text{con}}} \leq 10\%$	70	27
PLT	> 0	$P_{\text{plt}} \geq 90\%$	3	1
PVG	> 0	$\frac{P_{\text{gen}}}{P_{\text{con}}} \geq 30\%$	3	2
MIX	> 0	$40\% \leq \frac{P_{\text{gen}}}{P_{\text{con}}} \leq 60\%$	12	8
NIL	—	—	82	82
CTY	—	—	1	1

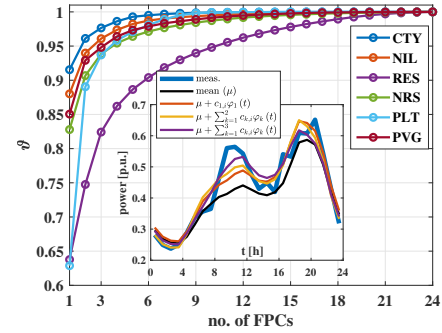


Figure 1. Cumulative proportion of explained variability ϑ , averaged over the population set, per population and plotted as a function of the ordered number of functional principal component. The inset plot shows the effect of progressively adding each FPC with the corresponding score to the mean.

IV. RESULT

A. Load pattern analysis

Figure 1 shows ϑ as a function of the number of $\varphi(t)$ for each population, together with an example of pattern reconstruction as a function of the number of $\varphi(t)$. ϑ increases with the spatial aggregation level, and the lowest ϑ is observed for the RES population, which is subjected to more variability and is thus harder to reproduce with a small number of $\varphi(t)$. Since, on average, more than 80% of the variability of each population can be explained by the first three $\varphi(t)$, the study was limited to the first three $\varphi(t)$ only. The amount of data generated is cumbersome. Therefore, the analysis that follows is limited to a single element per population (see Table I), herein denoted with the same name of the population it belongs to, exception made for the population MIX, which is not here reported. The patterns were individually normalized with respect to the maximum power measurement displayed by each of them in the entire data set, i.e., the years 2014–2018. The normalization was crucial as it allowed for a direct comparison between the values of the scores of each $\varphi_k(t)$, even if belonging to different electric stations or level of spatial aggregation. The $f_i(t)$ and $\mu(t)$, together with the first three $\varphi(t)$, and the effect of the scores of the first three $\varphi(t)$ on $\mu(t)$, are grouped by population and shown in a series of sub-panels in Figure 2.

While the $f_i(t)$ of the non-aggregated electric stations belonging to different populations are very different among each other, the daily patterns of the spatially aggregated consumptions, i.e., NIL and CTY, are very similar. This reflects a rapid loss in the capacity of resolving specific patterns, ascribed to particular contractual characteristics, as soon as the spatial aggregation proceeds towards larger areas. In all the analyzed cases, $\varphi_1(t)$ is very similar to the corresponding $\mu(t)$ and the variability that it adds to the average profile does not significantly affect the dynamics of the daily consumption profile, which is mostly determined by the combined effect of $\varphi_2(t)$ and $\varphi_3(t)$. The distribution of the scores as a function of the day of the week, the month of the year, and the meteorological conditions are grouped by population and shown in a series of sub-panels in Figure 3. In the following, the FPCA of each population is briefly discussed in detail.

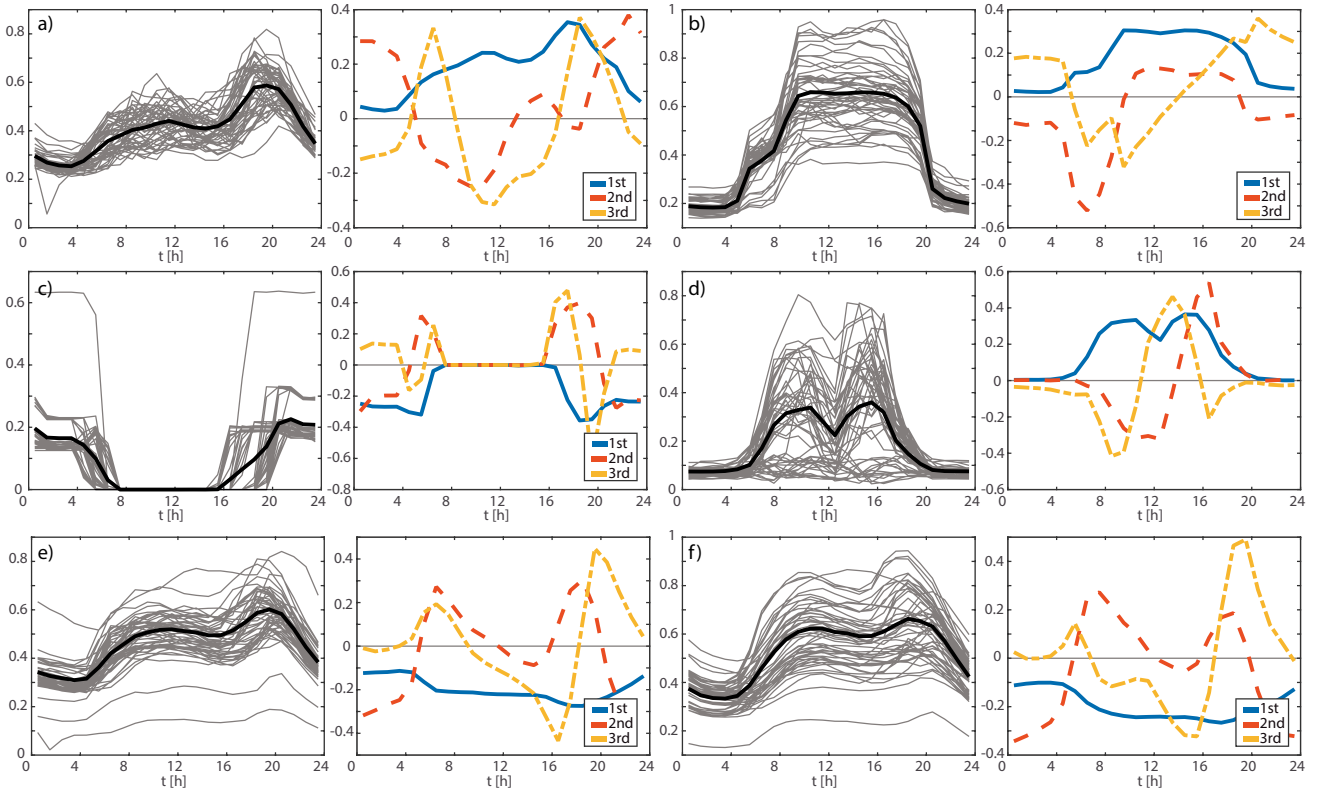


Figure 2. The panel figure shows, on the left hand-side subplots, the daily average power load $\mu(t)$ superimposed on sample daily power load patterns and, on the right hand-side subplots, the first three FPCs (i.e., $\varphi_1(t)$ solid blue line, $\varphi_2(t)$ dashed red line, and $\varphi_3(t)$ dashed yellow line) of the analyzed elements taken from each population: a) RES, b) NRS, c) PLT, d) PVG, e) NIL, and f) CTY.

It is worth mentioning that the discussion in subsections IV-A1–IV-A4 are referred to representative secondary substations, and that they apply to any substation of the same set, or to same level of spatial aggregation.

1) *RES*: The $f_i(t)$ shows two typical peaks, around 12:00PM and 08:00PM, already observed in literature [17], [29]. The distribution of the scores of $\varphi_1(t)$ follows the trend of the consumptions, which are higher than $\mu(t)$ during the weekends and in the colder months, while lower than $\mu(t)$ in the milder months, exception made for the month of July which is characterized by consumptions slightly higher than the one of the adjacent months of June and August. The trends suggest a correlation between $\varphi_1(t)$ and the presence at home, which usually determines higher power loads. Since $\varphi_1(t)$ explains only $\approx 65\%$ of the variability of the $f_i(t)$ (see Figure 1), and since $\varphi_2(t)$ and $\varphi_3(t)$ are mostly related to the dynamics of the consumptions, the analyzed RES station is subjected to large daily variability. The scores of $\varphi_2(t)$ and $\varphi_3(t)$ follow similar distributions. Their combined effect is to increase the consumption levels during the day and to decrease the consumptions levels at night, slightly flattening the consumption patterns and shrinking them towards the central hours of the day. This phenomenon, which strongly occurs during the weekends, can be ascribed to more intense home activities, and to the use of air conditioning systems. The hypothesis is further supported by the trend observed in the monthly distributions of the scores of $\varphi_2(t)$, which is positively peaked in the summer, reflecting an intensive

use of conditioning system in the late evening and a lower presence at home during the day. The monthly distribution of the scores of $\varphi_3(t)$ is of no trivial interpretation. The contour plots on the meteorological dependencies reveal the existence of a “comfort zone” in the temperature interval 15–25 °C and relative humidity range 20–60%, where the scores of $\varphi_1(t)$ and $\varphi_2(t)$ are negative and the electrical consumptions reach a minimum.

2) *NRS*: The $f_i(t)$ shows the typical profile of tertiary commercial activities, with high consumption levels extending from the early morning to the late evening [17], [29]. The distribution of the scores of $\varphi_1(t)$ follows the consumptions, which are above $\mu(t)$ during the weekdays, and in general in the colder months of the year, exception made for the month of July, when they reach levels similar to the ones in winter. The observed trend suggests for a correlation between $\varphi_1(t)$ and the regular commercial activities, which include the use of air conditioning systems. Since $\varphi_1(t)$ explains more than 80% of the data variability, $\varphi_2(t)$ and $\varphi_3(t)$ play a marginal role. Nevertheless, $\varphi_2(t)$ has a peculiar shape negatively peaked between 5:00AM and 10:00AM in the morning. The trend of its daily scores, always negative but Monday and Sunday, and the corresponding monthly trend, negatively peaked in the summer, let suppose a direct correlation between $\varphi_2(t)$ and the clock-in time, which correspond to an intensive use of air conditioning systems. The effect of $\varphi_3(t)$ on $\mu(t)$ is very small when compared to the first two $\varphi_k(t)$ and of no trivial interpretation. The same “comfort zone” observed for

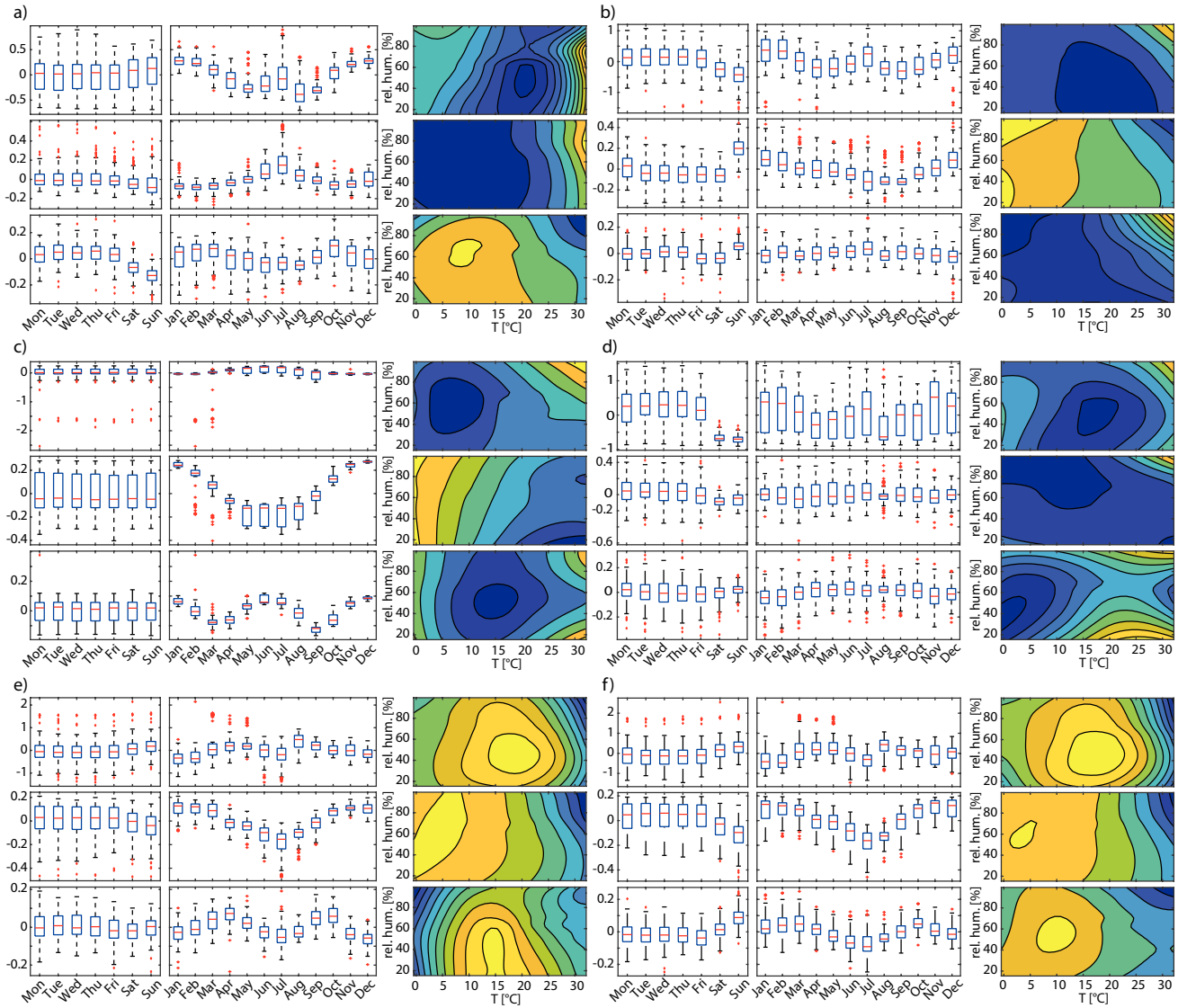


Figure 3. Weekly, monthly and weather-dependent trends of the scores of representative elements belonging to the a) RES, b) NRS, c) PLT, d) PVG, e) NIL, and f) CTY populations. The rows of each subplot refer to an FPC (from top to bottom, φ_1 , φ_2 , and φ_3) while the columns of each subplot refer to weekly, monthly, and weather-dependent distributions of the scores of the corresponding FPC. In contour plots, cold (warm) colors are assigned to negative (positive) values.

the RES station is found for the analyzed NRS station.

3) *PLT*: The $f_i(t)$ shows a unique profile, with consumptions limited to the dark hours of the day. As expected, the scores of $\varphi_1(t)$, $\varphi_2(t)$ and $\varphi_3(t)$ do not follow any daily trend. However, a monthly trend that resembles a sinusoid is evident for all of them. This trend is typically observed in phenomena that depend on the revolution of the Earth around the Sun, as are the hours of daylight. While $\varphi_1(t)$ captures the shift of the consumption with respect to $\mu(t)$, the combined effect of $\varphi_2(t)$ and $\varphi_3(t)$ is to shrink (or enlarge) the width of the daily time with zero load. It is worth observing that i) the switch from zero to max load is not a step function, but a ramp with a certain steepness that varies through the year; ii) that the median of the distribution of the scores of $\varphi_2(t)$ reaches a minimum during the months characterized by longer days; and iii) that the median of the distribution of the scores of $\varphi_3(t)$ follows a sinusoidal trend with frequency two times the

one of $\varphi_2(t)$. Incidentally, it turns out that the scores of $\varphi_3(t)$ reach their minimum in the months characterized by the switch to/from the daylight saving time. Therefore, while $\varphi_2(t)$ can be correlated to the number of daylight hours, which directly affects the overall time over which the public lighting turns on, $\varphi_3(t)$ can be ascribed to the switching to daylight saving time. As expected, no direct correlation between the principal components and the temperature and the relative humidity is observed.

4) *PVG*: The $f_i(t)$ shows a profile that is similar to the one observed in the NRS substation, with an additional peculiar cuspid around midday that demonstrates the effect of the self-generated and -consumed electricity. The daily and monthly distributions of the scores of $\varphi_1(t)$ pretty well resemble the trend observed for the NRS substation and thus, following the same reasoning adopted in that case, $\varphi_1(t)$ can be logically correlated to regular commercial activity. The distribution of

the scores of $\varphi_2(t)$ follows the dynamics of the two peaks of $f_i(t)$, while the one of $\varphi_3(t)$ follows the height of the cuspid. While the interpretation of $\varphi_2(t)$ is not trivial, the monthly distribution of the scores of $\varphi_3(t)$ are such that the cuspid is less deep during the summer. This could be ascribed to a stronger electric absorption that occurs during the warmer months because of the air conditioning systems, regardless of a larger photovoltaic generation. In fact, it has to be noted that the actual generation installed on the analyzed electric station might be much lower than the registered contractual generation.

5) *NIL and CTY*: The $f_i(t)$ of the spatially aggregated substations have much in common, with a first peak in the late morning and a second peak in the early evening. The first three $\varphi_k(t)$ are similar to each other, and their scores follow the same weekly and monthly trends. On the one hand, the median of the weekly distribution of the scores of $\varphi_1(t)$ is such that the consumptions are above the average level during the weekdays, and below the average level during the weekends, while the scores of $\varphi_2(t)$ and $\varphi_3(t)$ are such that the dynamics of the consumption patterns in the early morning and in the evening are mostly affected during the weekends, with a reduction of the morning consumption and an increase of the consumption in the late hours. On the other hand, the monthly distributions of the scores indicate consumptions above the average in cold and hot months, exception made for August, which is characterized by reduced industrial activity and presence, and thus lower consumptions. It is interesting observing the existence of a “comfort zone” even in the case of spatially aggregated consumptions, and that this zone is again in the temperature interval 15–25°C and relative humidity range 20–60%.

B. Prediction

1) *A comparison on a benchmark dataset*: The comparison between the proposed method and some existing work on long-term forecast of electricity load is here presented, using a public domain dataset made available for the EUNITE competition [30], consisting in electricity load demand values recorded every two hours in Eastern Slovakia in 1997 and 1998. This dataset has already been used for the long-term forecasting in [31], where data recorded in 1997 are used as training set to predict electricity load demand in 1998. In [31], authors report the performance of several predictive methods using five indices, i.e., the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE), the Normalized Mean Square Error (NMSE), the Relative Error Percentage (REP) and the Pearson Product-Moment Correlation Coefficient (PPMCC). We applied the proposed method to this dataset, using data from 1997 for the estimation of the daily function principal components and the prediction models for the correspondent scores. We then predict the electric load demands for 1998 using the first $K = 4$ functional principal components, since they explain more than 99% of the variability in the training set. The accuracy of the prediction is measured with the same indices used in [31] and the result can be seen in Table II. The long-term predictor based on

daily function principal component scores outperforms all the methods reported in Table I in [31] for all the considered indices.

Table II
PERFORMANCE OF THE PROPOSED METHOD IN THE CASE OF THE DATASET FROM EUNITE COMPETITION. THE FPCA-BASED PREDICTION OUTPERFORMS ALL THE METHODS PRESENTED IN TABLE I OF [31]. AMONG THE MANY METHODS PRESENTED IN [31], ONLY THE RESULTS OF THE BEST PERFORMING ONE (I.E., SVP+SVB) ARE HERE REPORTED.

	MAE	MAPE	NMSE	REP	PPMCC
FPCA	27.7	4.7	0.1	5.7	0.94
SVP + SVB	43.0	7.0	0.5	8.7	0.88

The final model for the score of the first principal component, which explains more than 93% of the variability in the training set, includes as predictors the month, the day of the week, the day of the month and the interaction between the day of the month and the month (meaning that the coefficient associated to the day of the month changes from month to month).

2) *Results for electricity consumption in the Milan area*: The FPCA method has been used to predict the power consumption profiles of each substation. As the aim of the present paper is to show the potential benefits in using FPCA, efforts have not been put in optimizing the selection of the training set, of the prediction model, and of the predictors. The forecast was carried out on a long-term basis. The period 2014–2015 was used to predict the hourly power profiles for the whole 2016 for each selected substation.

The predicted and the actual daily load curve of a given day is shown in Figure 4a–f per each representative element of each population studied in this work. Each plot shows the day with the lowest daily MAPE. The corresponding yearly MAPEs are shown in the second column of Table III. The prediction model predicts and reproduces the actual load curves of the population RES with good accuracy. The errors for PLT and PVG recommend that further analysis and customization should be dedicated to better predict these kinds of substations. It also worth noting that the prediction occurred once for the whole year. The error on the prediction does not decrease with the spatial aggregation, i.e., with the loss of resolution on contractual characteristics. This goes against what ϑ would suggest (see Figure 1), which increases with the spatial aggregation, and might be consequence of the choice of the linear model as a good compromise between complexity and accuracy.

The percentage error on the average monthly energy is reported in Table III per month per population. The monthly energy estimates RES and NRS have $\varepsilon\% < 10\%$, while the error on PLT, PVG, NIL and CTY is larger and depends on the month of the year.

V. CONCLUSION AND PERSPECTIVES

This work demonstrates that the FPCA is a versatile technique that can be exploited to study the electric consumption patterns at various level of spatial aggregation. The FPCA

Table III

YEARLY MAPES (SECOND COLUMN) AND MONTHLY ENERGY PERCENTAGE ERROR FOR EACH REPRESENTATIVE SUBSTATION IN EACH POPULATION.

Population	MAPE	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
RES	7.63	-2.97	-1.41	-4.69	-1.04	3.68	7.99	-1.32	1.75	1.68	-0.75	2.69	1.53
NRS	10.38	2.14	2.74	5.72	-3.02	-4.66	5.02	0.61	-3.34	-5.10	-3.50	-2.76	2.67
PLT	34.30	-27.73	31.45	20.61	0.27	0.74	0.76	0.93	7.71	-0.05	0.02	-0.10	10.86
PVG	39.19	-60.11	-30.55	-44.97	-62.63	-23.74	-31.91	-35.93	-36.64	-18.99	-18.13	-52.27	-48.15
NIL	9.41	-8.47	-6.75	37.62	-2.82	-1.47	-17.26	1.73	1.29	5.17	4.01	-16.88	5.82
CTY	12.52	-27.67	-22.97	24.60	-15.26	-10.75	-24.37	-2.95	-7.03	12.80	8.24	-10.93	5.05

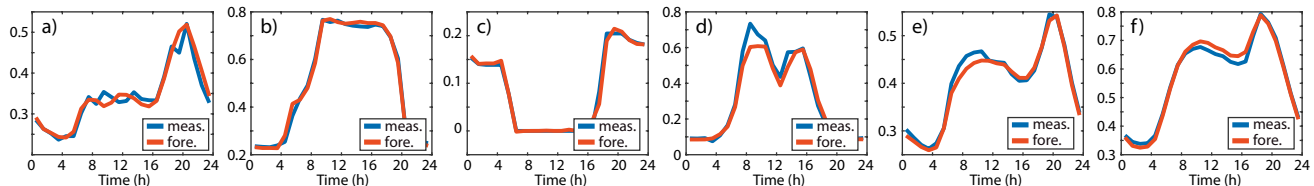


Figure 4. Predicted (red) and actual (blue) daily load curves of the day with lowest daily MAPE of each representative substation in each population: (a) RES on 04/08/2016, (b) NRS on 03/11/2016, (c) PLT on 07/23/2016, (d) PVG on 06/17/2016, (e) NIL on 04/10/2016, and (f) CTY on 05/21/2016.

exploits unique properties that allows to establish a compromise between complexity and amount of variability explained by decomposing, or predicting, the daily patterns on the basis of a selected limited number of $\varphi_k(t)$. To demonstrate the capabilities of the method, the FPCA was applied to three years of unique and sensitive historical data of electric consumption at the distribution level. The first three functional principal components were found sufficient to explain more than 80% of the variability of the data per electric station type and/or spatial aggregations, while the correlation between the principal components and the exogenous causes was rapidly lost after the first two components. The first two principal components were found strongly correlated with the calendar periodicity and the weather conditions, the latter allowing to identify a “comfort region” where the consumptions reach a minimum. A linear prediction algorithm, based on the FPCA decomposition method and chosen to minimize the Akaike Information Criterion (AIC), demonstrated that FPCA-based linear models have interesting capabilities in predicting time series both in the short- and in the mid/long-term. This opens the doors to further investigations that will aim at understanding advantages and disadvantages of FPCA-based linear models with respect to other more common models, such as ARIMA and Neural Networks.

In this work, for the sake of simplicity a linear model is used to predict future scores. Also, the model is selected in an automatic way with no fine-tuning for the different types of substation. More complex models can of course be identified for the scores prediction and this can be scope for future work. Moreover, we expect that different predictive models will be needed for the different types of substations and this model building can be informed by the analysis we presented. Additional features can also be introduced, in particular meteorological variables.

To conclude, this work assesses the FPCA as an innovative and powerful method to investigate and predict the electric consumption patterns at any spatial aggregation level and

opens the doors to further studies that aim at optimizing the algorithm for predictive purposes.

REFERENCES

- [1] D. Meadows, J. Randers, The limits to growth: the 30-year update, Routledge, 2012.
- [2] S. Greengard, The internet of things, MIT press, 2015.
- [3] K. Zhou, C. Fu, S. Yang, Big data driven smart energy management: From big data to big insights, *Renew. Sustain. Energy Rev.* 56 (2016) 215–225.
- [4] V. Mayer-Schönberger, K. Cukier, Big data: A revolution that will transform how we live, work, and think, Houghton Mifflin Harcourt, 2013.
- [5] S.-C. Chan, K. M. Tsui, H. Wu, Y. Hou, Y.-C. Wu, F. F. Wu, Load/price forecasting and managing demand response for smart grids: Methodologies and challenges, *IEEE Signal Process. Mag.* 29 (5) (2012) 68–85.
- [6] A. R. Khan, A. Mahmood, A. Safdar, Z. A. Khan, N. A. Khan, Load forecasting, dynamic pricing and dsm in smart grid: A review, *Renew. Sustain. Energy Rev.* 54 (2016) 1311–1322.
- [7] A. Capasso, W. Grattieri, R. Lamedica, A. Prudenzi, A bottom-up approach to residential load modeling, *IEEE Trans. Power Syst.* 9 (2) (1994) 957–964.
- [8] C. F. Walker, J. L. Pokoski, Residential load shape modelling based on customer behavior, *IEEE Trans. Power App. Syst. PAS-104* (7) (1985) 1703–1711.
- [9] A. Henley, J. Peirson, Non-linearities in electricity demand and temperature: parametric versus non-parametric methods, *Oxford bulletin of economics and statistics* 59 (1) (1997) 149–162.
- [10] E. Valor, V. Meneu, V. Caselles, Daily air temperature and electricity load in Spain, *Journal of applied Meteorology* 40 (8) (2001) 1413–1421.
- [11] A. Pardo, V. Meneu, E. Valor, Temperature and seasonality influences on Spanish electricity load, *Energ. Econ.* 24 (1) (2002) 55–70.
- [12] M. Christenson, H. Manz, D. Gyalistras, Climate warming impact on degree-days and building energy demand in Switzerland, *Energ. Convers. Manage.* 47 (6) (2006) 671–686.
- [13] R. F. Engle, C. Mustafa, J. Rice, Modelling peak electricity demand, *Journal of forecasting* 11 (3) (1992) 241–251.
- [14] L. Hernández, C. Baladrón, J. M. Aguiar, L. Calavia, B. Carro, A. Sánchez-Esguevillas, D. J. Cook, D. Chinarro, J. Gómez, A study of the relationship between weather variables and electric power demand inside a smart grid/smart world framework, *Sensors* 12 (9) (2012) 11571–11591.
- [15] M. Lindèn, J. Helbrink, M. Nilsson, D. Pogojan, J. Ridenour, A. Badano, Categorisation of electricity customers based upon their demand patterns, *CIREN-Open Access Proceedings Journal* 2017 (1) (2017) 2628–2631.

- [16] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, M. E. Webber, Clustering analysis of residential electricity demand profiles, *Appl Energy* 135 (2014) 461–471.
- [17] I. Panapakidis, M. Alexiadis, G. Papagiannis, Electricity customer characterization based on different representative load curves, in: 2012 9th International Conference on the European Energy Market, IEEE, 2012, pp. 1–8.
- [18] T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas, M. Kolehmainen, Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data, *Applied Energy* 87 (11) (2010) 3538–3545.
- [19] G. Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping, *Energy* 42 (1) (2012) 68–80.
- [20] T. W. Liao, Clustering of time series data—a survey, *Pattern recognition* 38 (11) (2005) 1857–1874.
- [21] L. Suganthi, A. A. Samuel, Energy models for demand forecasting—a review, *Renewable and sustainable energy reviews* 16 (2) (2012) 1223–1240.
- [22] J. Hosking, R. Natarajan, S. Ghosh, S. Subramanian, X. Zhang, Short-term forecasting of the daily load curve for residential electricity usage in the smart grid, *Applied Stochastic Models in Business and Industry* 29 (6) (2013) 604–620.
- [23] J. Massana, C. Pous, L. Burgas, J. Melendez, J. Colomer, Short-term load forecasting in a non-residential building contrasting models and attributes, *Energy and Buildings* 92 (2015) 322–330.
- [24] J. Moral-Carcedo, J. Vicens-Otero, Modelling the non-linear response of spanish electricity demand to temperature variations, *Energy economics* 27 (3) (2005) 477–494.
- [25] E. S. Gardner Jr, Exponential smoothing: The state of the art, *Journal of forecasting* 4 (1) (1985) 1–28.
- [26] W. Christiaanse, Short-term load forecasting using general exponential smoothing, *IEEE Trans. Power App. Syst.* PAS-90 (2) (1971) 900–911.
- [27] J. W. Taylor, Short-term electricity demand forecasting using double seasonal exponential smoothing, *Journal of the Operational Research Society* 54 (8) (2003) 799–805.
- [28] J. O. Ramsay, B. W. Silverman, *Functional Data Analysis*, 2nd Edition, Springer Series in Statistics, Springer, 2005.
- [29] S.-I. Yang, C. Shen, et al., A review of electric load classification in smart grid environment, *Renewable and Sustainable Energy Reviews* 24 (2013) 103–110.
- [30] European network on intelligent technologies for smart adaptive systems. URL <http://neuron.tuke.sk/competition/>
- [31] L. Ghelardoni, A. Ghio, D. Anguita, Energy Load Forecasting Using Empirical Mode Decomposition and Support Vector Regression, *IEEE Trans. Smart Grid* 4 (1) (2013) 549–556.