Available online at www.sciencedirect.com

ScienceDirect

International Journal of Forecasting 26 (2010) 700–711

www.elsevier.com/locate/ijforecast

Functional clustering and linear regression for peak load forecasting

Aldo Goia*, Caterina May[1], Gianluca Fusai[2]

Dipartimento di Scienze Economiche e Metodi Quantitativi (SEMeQ), Università del Piemonte Orientale "A. Avogadro", Via Perrone, 18, 28100 Novara, Italy

Abstract

In this paper we consider the problem of short-term peak load forecasting using past heating demand data in a district-heating system. Our data-set consists of four separate periods, with 198 days in each period and 24 hourly observations in each day. We can detect both an intra-daily seasonality and a seasonality effect within each period. We take advantage of the functional nature of the data-set and propose a forecasting methodology based on functional statistics. In particular, we use a functional clustering procedure to classify the daily load curves. Then, on the basis of the groups obtained, we define a family of functional linear regression models. To make forecasts we assign new load curves to clusters, applying a functional discriminant analysis. Finally, we evaluate the performance of the proposed approach in comparison with some classical models.

Keywords: Short-term forecasting; Out-of-sample; Load curve; Seasonality; Functional regression; Functional clustering; Functional linear discriminant analysis

1. Introduction

Load demand forecasting is becoming more and more important as power generation costs increase and market competition intensifies: accurate forecasts are relevant to energy systems for scheduling generator maintenance and choosing an optimal mix of on-line capacity. The literature on load forecasting considers three main problems: long-term forecasts for system planning, medium-term forecasts for maintenance programs, and short-term prediction for the day-to-day operation, scheduling and load-shedding plans of power utilities. A central role is played by the intra-daily pattern of the load demand, known as the *load curve*, which describes the amount of energy consumed to satisfy the load demand of customers over the course of the day.

* Corresponding author. Tel.: +39 0321 375 319; fax: +39 0321 375 305.

E-mail addresses: aldo.goia@eco.unipmn.it (A. Goia), cmay@mat.unimi.it (C. May), gianluca.fusai@eco.unipmn.it (G. Fusai).

URL: http://semeq.unipmn.it/ (A. Goia, C. May, G. Fusai).

[1] Tel.: +39 0321 375 334; fax: +39 0321 375 305.

[2] Tel.: +39 0321 375 312; fax: +39 0321 375 305.

0169-2070/$ - see front matter © 2009 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2009.05.015

The focus of the present work is on short-term (i.e. around 24 h) forecasting of the daily peak load in a district-heating (or "teleheating") system, which is the maximum of the daily demand for heating. A district-heating system involves distributing the heat for residential and commercial requirements via a network of insulated pipes. We analyze here data on heat consumption in a major Italian centre, Turin, where the district heating is produced through a co-generation system. This technology allows for energy saving and reduces emissions compared to old technologies; it is therefore treated as a renewable energy source and delivers substantial economic and environmental benefits.

In the recent literature concerning prediction in district-heating systems (see, for example, Dotzauer, 2002, and Nielsen & Madsen, 2006, for some applications and references), the algorithms employed are usually similar to those used in the prediction of electrical-power loads. A review of statistical methods for electrical load forecasting has been given by Weron (2006), for instance. These methods are mainly based on ARIMA models, regression models, exponential smoothing, and generalizations of these. Typically, weather variables are used for the prediction of electricity loads, and a great range of modeling approaches are presented in the literature. Developments in forecasting methodologies are also reflected by the contributions in the special issue of the *International Journal of Forecasting* on energy forecasting that appeared in 2008. For instance, Dordonnat, Koopman, Ooms, Dessertaine, and Collet (2008) develop a multi-equation model with time-varying parameters, while Alves da Silva, Ferreira, and Velasquez (2008) propose automated input selection procedures for forecasting models based on neural networks. Whereas electric and wind power, gas consumption and electricity price forecasting are discussed in the cited issue, the present paper deals with heat consumption in a residential area. In this case, the strong correlation between heating and external temperature has led us to a parsimonious model without exogenous variables, which forecasts satisfactorily without requiring the inclusion of weather variables.

In this paper we propose a method based on a functional statistics approach. The functional statistics approach has become the object of an increasing amount of attention on the part of many researchers and practitioners in recent years, since it can be applied when data are collections of discrete observations effected on curves, images or shapes. A survey of these techniques can be found in the monographs of Ramsay and Silverman (2005) and Ferraty and Vieu (2006), for instance. Our data-set consists of hourly observations of the heat consumption over four separated periods, with 198 days in each period, over the years 2001-2005. The load data are registered in megawatts, and have been rescaled to be between 0 and 1. The goal of the present work is to build and estimate models on the basis of the first three periods, and then to evaluate the *out-of-sample* performance of different forecasts of the peak load on the whole fourth period.

First, we define a sample formed by the daily curves of heat consumptions (we will refer to them as "load curves"), and consider a functional linear regression model where the peak demand on a given day is the scalar response and the load curve of the previous day is the functional regressor. Noting that intra-daily effects change with the season within each period, we then propose a methodology for improving the forecasting ability of the functional regression model: we partition the observational curves into homogeneous groups using the functional clustering technique presented by Abraham, Cornillon, Matzner-Lober, and Molinari (2003). The aim of our clustering procedure is to find some characteristic patterns in the data-set that may determine changes in the heat demand peaks. The classification of the load curves into groups is a solution to the problem of modeling the seasonality effect during each period. Then, we estimate a specific functional linear coefficient for each group, obtaining a family of functional regression models. Note that various clustering techniques for classifying similar load patterns have been considered in the literature; for instance, Chicco, Napoli, and Piglione (2001) and Amin-Naseri and Soroush (2006) refer to neural networks clustering; here we use clustering in a functional context instead. In order to assign the new curves to clusters in the forecasting procedure, we use a functional linear discriminant analysis, as was done by James and Hastie (2001). Finally, we compare the out-of-sample performance of our models with classical regression approaches in which the functional nature of the data is not taken into account. All routines are implemented by the R software.

Finally, let us consider some related works which have recently appeared in the literature on forecasting with functional regression. Hyndman and Ullah (2007) propose a robust methodology for forecasting functional time series; their original method, which is applied in the demographic context in particular, also differs from our approach in that they forecast a smoothed function instead of a scalar, as in our problem. In addition, Kargin and Onatski (2008) and, in a electricity load context, Antoch, Prchal, De Rosa, and Sarda (2008), focus on forecasting functional data, which are assumed to be generated by a functional autoregressive process. Regarding the forecasting of scalars, Sood, James, and Tellis (2009) apply functional linear regression to predict the market consumption of new products, showing the advantages of functional techniques in comparison to standard ones. In addition to a different applicative problem, in which our data-set is given by time series, our method differs from theirs because it uses functional clustering not only to describe the data, but also for prediction.

The paper is organized as follows. In Section 2 we describe the data-set, introduce the notations and present the problem considered in this work. In Section 3 we propose and discuss the functional forecasting methodology we adopt. Section 4 is devoted to a numerical comparison of the performances of the models considered with some simple non-functional competitors. Final remarks conclude the paper.

2. The data-set and the forecasting problem

The data-set analyzed in this paper contains hourly observations of a stochastic process $L(t)$, $t \in \mathbb{R}^+$. Here $L(t)$ represents the heating demand at time t for the warming of residential and commercial buildings using the district-heating system described in the introduction. The discretization of this process has been observed over four discrete periods covering the years 2001–02, 2002–03, 2003–04 and 2004–05. Each period consists of 198 successive days, from 15 October to 30 April (29 April for the leap year 2004), according to the Italian regulations for heating distribution. Due to privacy requests from the data supplier, the data have been normalized (that is, rescaled to be between 0 and 1).

The hourly data for the heating demand in three selected weeks have been plotted in Fig. 1. We can clearly distinguish the intra-daily periodical pattern, and we can also note that it evolves over time depending on the season. The variation across each period is due to the climatic conditions of Turin: the winter is rather cold, whereas the climate in autumn and spring is relatively warm.

The repetitiveness of the daily shape is due to a certain inertia in the demand that reflects the aggregate behavior of consumers: analogously to the context of electricity demand, we refer to this intradaily pattern as a daily *load curve*.

The nature of the data suggests, in a natural way, the division of the observed series of hourly heating demand for each period into $n = 198$ *functional observations*, each one coincident with a specific daily load curve. Formally, let $\mathcal{T}_y$ be the yth interval, $y = 1, \ldots, 4$, where the process $L(t)$ has been observed; each $\mathcal{T}_y$ has the same length, and we can divide it into n equal sub-intervals with assigned length 24. Denoting by $L_y(t)$ the restriction of the process $L(t)$ to the yth period and translated on the interval $[0, n \cdot 24]$, we can define the load curve $C_{y,d}$ from period y and day d as

$$C_{y,d} = \left\{ L_y((d-1)\,24 + t),\ t \in [0, 24) \right\}, \quad d = 1, \ldots, 198.$$

Each of these functional data are observed on a finite mesh of discrete hours $h = 1, \ldots, 24$.

Let us now consider the observed daily peak of heat demand defined as

$$P_{y,d} = \max_{h=1,\ldots,24} C_{y,d}(h).$$

The goal of the present study is to forecast the peak $P_{y,d}$ on the basis of the load curve $C_{y,d-1}$ from the previous day. Models have to be estimated on the data belonging to periods $y = 1, 2, 3$, and the forecasting performance has to be evaluated on period $y = 4$.

In Fig. 2, a random selection of load curves is plotted. Typically, a load curve is characterized by an initial peak in the morning, when the consumers need the maximum thermal erogation by the system, and two other local maxima over the rest of the day. A preliminary data analysis shows that the peak load normally occurs around 7 to 8 AM. Moreover, the peak load series of each period y presents a high degree of

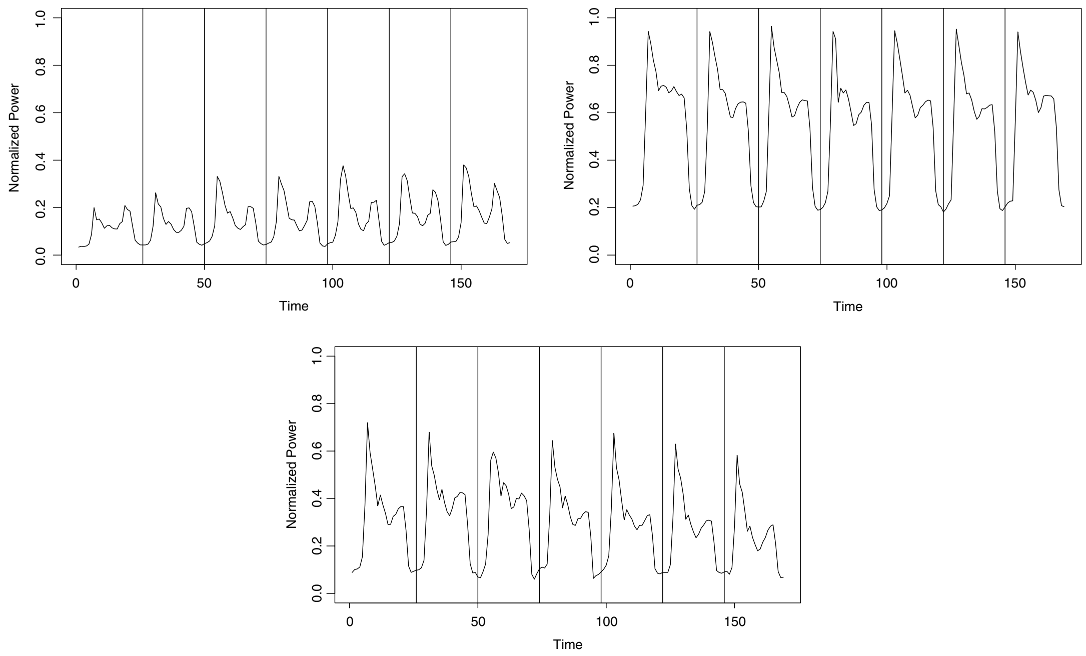

Fig. 1. Heating demand in three selected weeks (the panels contain data from November, January and March 2002–03).

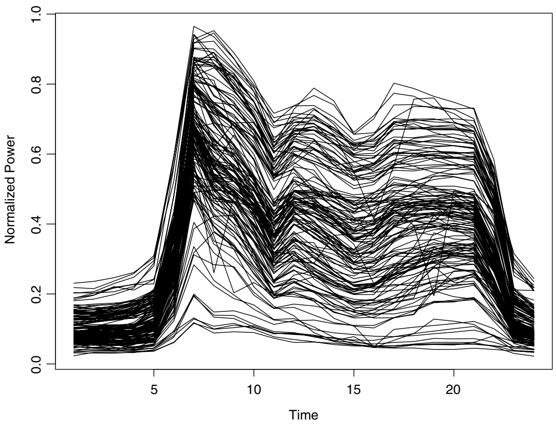

Fig. 2. Daily load curves.

autocorrelation: the mean linear correlation index is 0.983.

In order to better show how the demand for heating evolves during each period, we have summarized our four series by plotting the daily mean data, in Fig. 3. Clearly the process considered is not stationary. Observe also that the *seasonality trend* of each period, which has been represented here by a smoothing spline, has a typical shape. However, it cannot be described by a deterministic model in the out-of-sample procedure. Indeed, we have not been successful in estimating the trend of the fourth period on the basis of the first three, or in predicting it in a satisfactory way. For instance, moving averages or smoothing splines do not provide good results in the forecasting procedure at all. Therefore, we will not consider the deviations from the seasonality trend, and we will solve the problem of seasonal variability using the classification method illustrated in the next section.

In our modeling approach, we will not consider weather variables, such as temperatures, as is often done for short-term load forecasting. The effect of weather variables on peak load prediction is presumably included in the information carried by the heating demand curves in our data-set. Temperature curves may not be such a strong additional driver here, and for this reason we opt for a parsimonious model; we will show in the following sections that our approach is satisfactory for forecasting purposes. Finally, we remark that, in defining models, we will not consider differences between weekdays and weekends or holidays. This assumption, which is confirmed by the data, is reasonable because we are analyzing the thermal heating of civil residences,

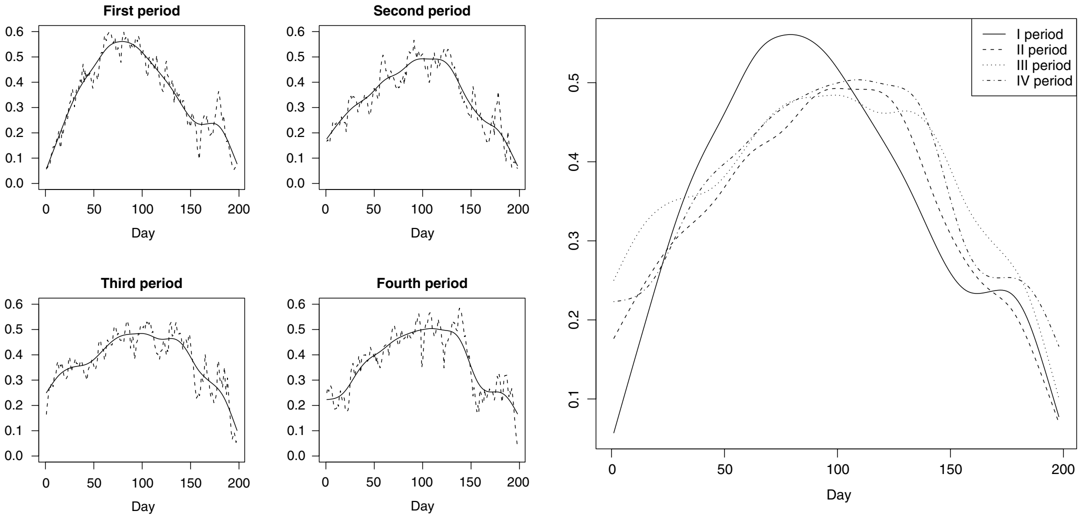

Fig. 3. Daily average of the load series and the seasonality trends.

which does not change considerably depending on the days of the week.

3. Forecasting via functional linear models

3.1. Basic model

We first consider a functional linear regression model where the scalar response is the daily peak of heating $P_{y,d}$ at day d defined in Section 2, with $d = 2, \ldots, 198$, and the functional regressor is the load curve $C_{y,d-1}$ at day $(d-1)$:

$$P_{y,d} = \int_0^{24} \beta(t) C_{y,d-1}(t) \mathrm{d}t + \varepsilon_{y,d} \qquad y = 1, \ldots, 4,\ d = 2, \ldots, 198, \tag{1}$$

where $\beta(\cdot)$ is the *weight regression function* (or *functional coefficient*) which has to be estimated. We assume that $\varepsilon_{y,d}$ in Eq. (1) is a sequence of i.i.d. centered random variables with finite variance. This model represents a generalization of the classical linear multivariate regression: here the regressor is a curve, namely an element of an infinite dimensional space, instead of a real random vector. Conditions for the identifiability of model (1), that is, the existence and the uniqueness of the functional coefficient, are discussed by Cardot, Ferraty, and Sarda (2003).

Many methods have been proposed for estimating $\beta(\cdot)$; here we use the functional regression estimation procedure illustrated by Cardot, Ferraty, and Sarda (1999), which adapts the well known principal components regression method (see e.g. Jolliffe, 2004) to the functional framework. The method involves projecting the functional observations on a space of finite dimension, spanned by the first q eigenfunctions of the empirical covariance operator of the functional variable. Hence, the estimation of the functional coefficient is obtained by inverting the covariance operator of the functional regressor in such a finite dimensional space.

In our study, the estimation of $\beta(\cdot)$ is performed using the data belonging to periods $y = 1, 2, 3$. A delicate point is the choice of q; since the data are autocorrelated, in order to select the optimal value of q we evaluate the forecasting performances, varying q, through a rolling hold-out sample procedure. In practice, we define a sequence of rolling training sets with length 350. For each day in the forecasting set, formed by the last 241 days, the preceding 350 are used to obtain the prevision, and thus evaluate the daily square forecasting error. Fig. 4 displays the mean square forecasting error from the rolling out-of-sample forecast procedure when q varies between 1 and 20. We can see that the forecasting ability of the model improves as we increase q from 1 to 6, but it does not change substantially with larger values of q; hence, we fix $q = 6$. An analogous result has been found empirically by Hyndman and Booth (2008); in

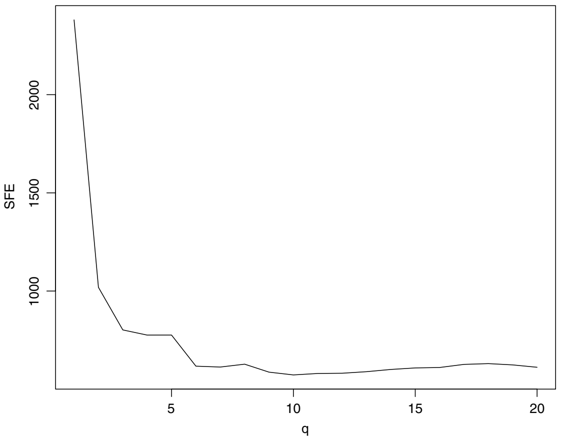

Fig. 4. Mean square forecasting errors obtained by a rolling hold-out sample for values of q between 1 and 20.

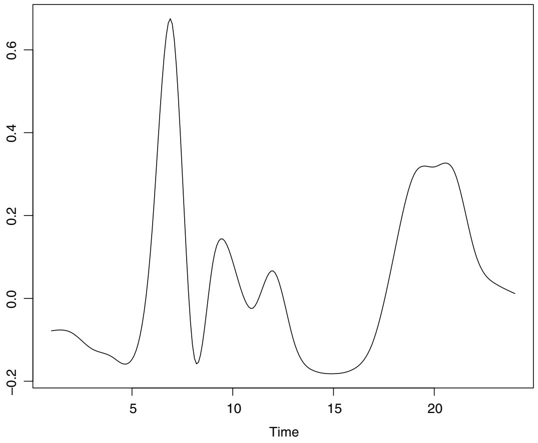

Fig. 5. Estimated weight regression function $\widehat{\beta}$ in model (1).

the context of overparameterized regression problems, Greenshtein and Ritov (2004) and Greenshtein (2006) refer to this phenomenon as "persistence in linear predictor selection".

The R^2 determination coefficient of the model is equal to 0.943. In Fig. 5 we plot the estimated functional coefficient $\widehat{\beta}(\cdot)$, interpolated with a cubic smoothing spline. Observing the graph, we can see that there are two maxima, located at approximately 7–8 AM and 7–9 PM, which correspond to the morning and evening peaks respectively.

By using the estimated model (1), we forecast the peak load for each day $d = 2, \ldots, 198$ in the fourth period; the predicted values, compared with the true ones, are plotted in Fig. 6. In order to evaluate the performance of the model, we analyze the distribution of the absolute percentage errors, defined as $\left|P_{4,d} - \widehat{P}_{4,d}\right| / P_{4,d}$, where $\widehat{P}_{4,d}$ is the predicted peak load on day d in period $y = 4$. These errors vary between 0.02% and 93.69%, and have a mean equal to 6.96% (the classical MAPE index) and a standard deviation equal to 10.80%. In addition, the first, second and third quartiles are respectively 1.94%, 3.87% and 7.56%; and on 90% of days considered the absolute error does not exceed 13.09%. Clearly, the presence of extreme values in the forecasting error distribution produces a MAPE larger than the median. In fact, we note that the forecasting performance of the model is quite heterogeneous during the entire period. A synthesis of the distributions of the absolute percentage errors for each month is plotted in Fig. 7: from this we can see that the model performs well in the winter months December, January and February (even though errors are greater than 30% on two days in these months); while forecasting in April provides the largest absolute errors. This is reasonable due to the high climatic irregularity which is typical of the spring season.

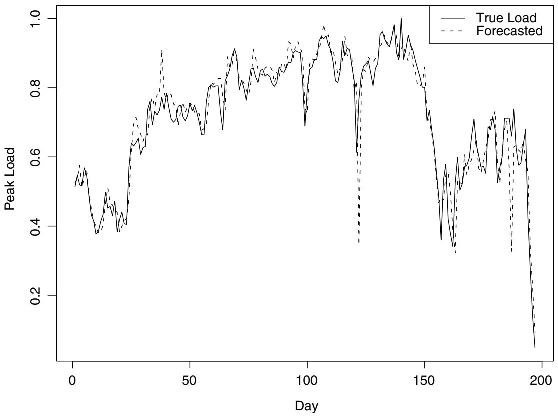

Fig. 6. Forecasted peak load values from the functional linear model.

3.2. Models based on curve classification

The difficulty of defining a satisfactory forecasting model is mainly due to the high variability of the heating demand during each period. In order to reduce this variability, we stratify the set of load curves into a few homogeneous groups exhibiting similar demand

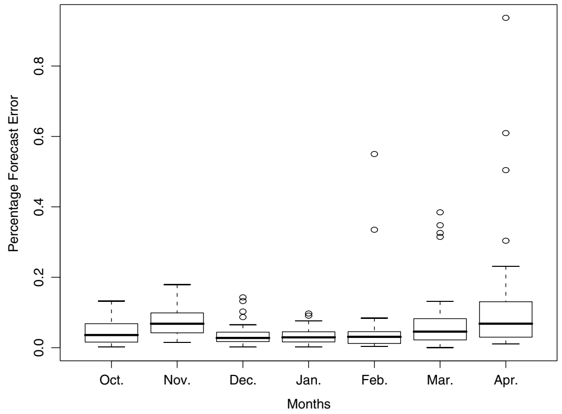

Fig. 7. Absolute percentage error distributions of the estimated functional regression model (1).

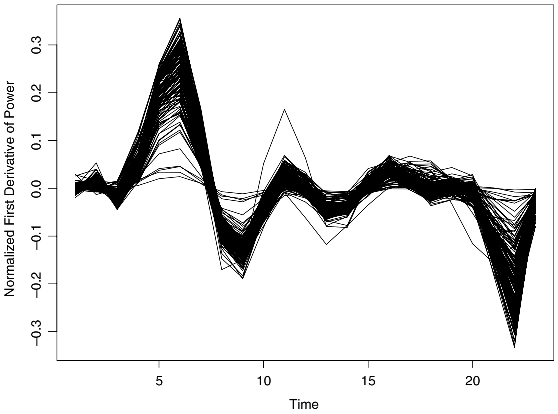

Fig. 8. First derivative of the load curves.

patterns, thus obtaining clusters which maximize the variance between the groups and minimize the variance within. Then, we use this idea to improve the forecasting ability of the functional linear regression model discussed in Section 3.1.

The first step is to group the load curves on the basis of some similarities: we apply the unsupervised functional clustering procedure proposed by Abraham et al. (2003), which takes into account the functional nature of the data. Note that this classification method can be generalized to the case in which the thermal loads are observed on an unequally spaced temporal mesh, a situation in which a classical multivariate classification method could not be used in any case.

The second step is to associate a specific functional linear regression model with each group: in this way we obtain a family of functional models. The optimal number of groups (and therefore models) is established by inspecting the determination coefficient R^2.

Finally, in the last step, we forecast the peak using an out-of-sample procedure; to assign new curves to groups we use the functional linear discriminant analysis proposed by James and Hastie (2001).

Let us now illustrate the above steps in detail, after which we will discuss the forecasting performance of the models obtained.

Step 1 – *Clustering*

The aim of the unsupervised functional classification is to detect characteristic patterns of the load curves, which may determine changes in the peak load demand. The classification of the sample of curves may involve the original data, and may also involve their derivatives. An initial analysis has revealed that the clustering of the original curves is not significant for our forecasting purposes, while the first derivatives of the functional data provide more suitable results. The procedure has two steps: first, the derivatives of the functional data are represented by a B-spline basis, then the coefficient vectors associated with each curve are clustered by a standard K-means algorithm.

In practice, we consider the first differences of the original data of the first three periods; in fact, since we have hourly average load data, we can assume that the measurement error of the original data is negligible. We fit the first derivative of the load curves using a regression spline (see e.g. De Boor, 2001). According to the literature dealing with spline functions, we use cubic splines. In addition, we fix 8 equispaced knots, since the intervals are *a priori* of equal importance. Therefore the knots are placed every 160 min (i.e. 2 h 40 min, 5 h 20 min, 8 h 00 min and so on). Hence, we associate with each curve the vector of coefficients in the B-splines basis representation; it follows that the first derivative of every load curve is summarized by a 12-dimensional vector. In Fig. 8 the first derivatives of a random sample of load curves are plotted. These quantities are relevant because they provide additional information on the intra-daily heating demand excursion, which can differ depending on the season.

To classify the 594 curves of the first three periods into G clusters ($G \geq 2$), with G fixed, we apply the K-means algorithm to the vectors of coefficients associated with each curve. This algorithm chooses

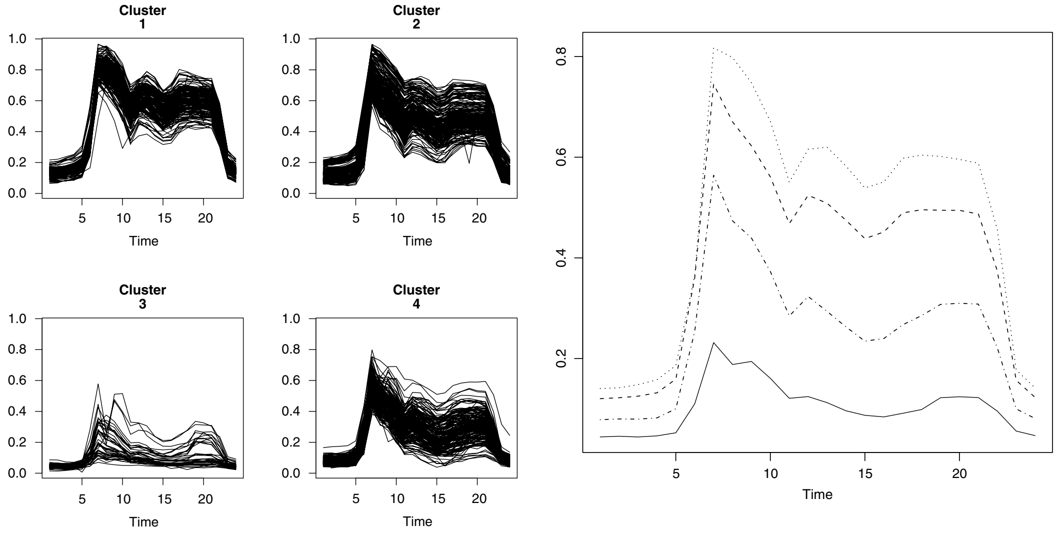

Fig. 9. Clusters based on first derivatives and the corresponding centers.

G centers and groups vectors on the basis of the nearest center; centers and the related groups are obtained iteratively in such a way as to minimize the sum of the Euclidean distances of all vectors from the corresponding centers (for details see Hartigan & Wong, 1979). Repeating this procedure for every G, we partition the sample into blocks for building the forecasting models.

Step 2 – Defining the family of models

For every clustering dimension G, we build a family of models as follows. We associate a functional linear model (1) with each group $j = 1, \ldots, G$, and we estimate the functional coefficients $\beta_j(\cdot)$, following the same method as in Section 3.1 and using $q = 6$ for each model.

Since the goal of our clustering is to improve the performance of the functional model, we choose the optimal number of clusters G by evaluating the determination coefficient R^2 corresponding to the G-dimensional family of regression models. If we analyze R^2 when G varies, we note that its value increases from 0.946, when $G = 2$, to 0.952, when $G = 4$; it then decreases for $G \geq 5$. Thus, we partition the sample into $G = 4$ groups.

The clusters of curves and the corresponding centers, which provide representative values for each group, are represented in Fig. 9. Note that performing the classification based on first derivatives permits us to discriminate the curves based on both the level of the consumption of heating and some typical shapes. Clusters 1 and 2 contain respectively 177 and 211 curves, and come principally from the winter season; even if these curves are apparently similar, the clusters have different centers and very different derivatives. Cluster 3 consists of 41 curves which are principally observed in October and April, and is associated with days with little demand. The remaining 162 curves belong to the fourth group and are observed in Autumn and Spring, characterizing days of medium consumption.

Step 3 – Forecasting

In order to forecast the peaks of the fourth period, we need to assign the new curves to clusters. To this end, we apply a functional linear discriminant analysis; this technique involves representing the functional data in a suitable finite dimensional space, and then applying the classical linear discriminant analysis to the associated coefficient vectors. Here the training-set of the linear discriminant model is formed using the same coefficients which were used for defining the clusters. The effectiveness of our discriminant procedure is evaluated via cross-validation on the first three periods, and the misclassification error turns out to be equal to 6.26%. To assign the curves of the fourth period to clusters, we summarize the derivatives of each one using a 12-dimensional vector following the procedure described in Step 2; hence, we compute the probability of

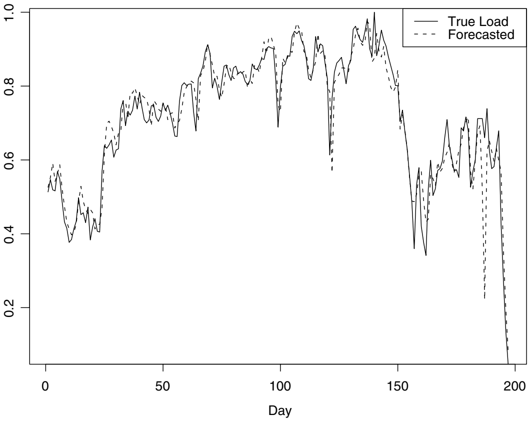

Fig. 10. Forecasted peak load values with the family of functional linear models defined by clustering.

assignment to the clusters using the estimated linear discriminant model. In this way, we assign 98 curves to cluster 1, 36 curves to cluster 2, 9 curves to cluster 3, and 54 curves to cluster 4.

Once we have made the classification, in the out-of-sample exercise we assign the load curve $C_{4,d}$ to the corresponding cluster j for each day d, and then we predict the peak demand $P_{4,d+1}$ for the successive day by applying the functional linear regression model based on the estimated functional coefficient $\widehat{\beta}_j(\cdot)$. The forecasted and observed peak series are plotted in Fig. 10.

The forecasting results are promising: the absolute percentage errors have a mean value of 6.55% with a standard deviation of 9.65%. With respect to the model described in Section 3.1, we thus have a MAPE reduction and a contraction in global variability. The improvement involves the whole error distribution: the quartiles are 1.69%, 3.59% and 7.31%, and the 90th percentile is 12.90%. These values are considerably smaller than the corresponding values when forecasting is performed using a unique regression model. We also observe that the maximum error is reduced to 72.97%, compared to the previously obtained 93.69%.

To better understand the behavior of the family of models, we can analyze the forecasting ability within each cluster; the means and the standard deviations of the absolute forecast errors are collected in Table 1. We observe that the main source of unpredictability is curves in group 3, and is related to the month of April. Note the excellent performances of the models in clusters 1 and 2.

Table 1
Forecasting performances of the functional linear model in each cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
MAPE	3.61%	5.77%	32.68%	8.06%
Stand. dev.	3.85%	3.73%	24.07%	9.28%

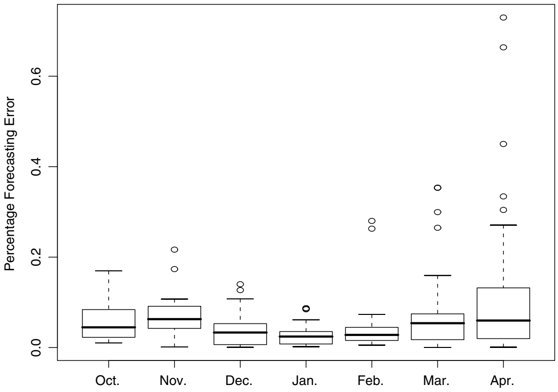

Fig. 11. Absolute percentage error distributions of the functional regression models obtained by clustering.

In Fig. 11 we plot the distribution of errors in the different months. By a comparison with the box-plot of Fig. 7 we can observe the improvement in forecasting ability from using the multiple models.

4. Out-of-sample comparative analysis

In this section we evaluate the out-of-sample performances of the functional models proposed in Section 3 in comparison with some simple alternative models; the aim is to discover whether the functional approach is convenient from a practical point of view.

A elementary approach, followed by a practitioner who observes the high autocorrelation of the daily peak load series, may consist of the following forecasting model:

$$P_{y,d} = P_{y,d-1} + \varepsilon_{y,d}. \tag{2}$$

We will refer to Eq. (2) as the “Naif model”. Even though it is very simple, this is a standard procedure among practitioners. Moreover, the model performs moderately well: the mean of absolute percentage errors is 8.20% and the standard deviation

Table 2
Least square estimates of the selected multiple regression model.

Variables	OLS estimate	Std. error
Constant	34.424	4.727
$C_{d-1}(1)$	−0.435	0.134
$C_{d-1}(5)$	−0.232	0.114
$C_{d-1}(6)$	0.190	0.049
$C_{d-1}(7)$	0.521	0.028
$C_{d-1}(15)$	−0.304	0.057
$C_{d-1}(18)$	−0.174	0.077
$C_{d-1}(21)$	1.102	0.076

equals 16.88%. In particular, predicted values are good during the Winter, when the seasonality level has a quite regular behavior; however, the forecasts deteriorate considerably in Autumn and Spring.

As a natural refinement of the "Naif model", we can consider a multiple regression model in which the peak load $P_{y,d}$ is the response variable, and the 24 hourly heating demands $C_{y,d-1}(1), \dots, C_{y,d-1}(24)$ are the predictors. We estimate a reduced model on the first three periods using a forward step-wise selection procedure; hence we obtain a multiple regression model with 7 predictors, which correspond to the load demand at hours $h = 1, 5, 6, 7, 15, 18, 21$. The least square coefficients estimates (all significant at the 5% level) of the selected model and their standard errors are summarized in Table 2. We note that, under the standard hypothesis on the error distribution, the regression is significant at the 5% level, with R^2 (and the adjusted R^2) equal to 0.951.

In terms of forecasting ability, the multiple regression model performs better than the "Naif" one, with a mean and standard deviation of absolute percentage errors equal to 7.21% and 15.05% respectively.

One criticism of the previous approach is that the model contains several correlated variables: even if it is known that multi-collinearity does not seriously jeopardize the forecasting ability of a model, and in fact often enhances it, the high collinearity between the regressors suggests an appeal to classical regression on principal components (Jolliffe, 2004). The technique consists of reducing the dimensionality of the regression model by using the principal components of the predictors $C_{y,d-1}(1), \dots, C_{y,d-1}(24)$, which are the most important in terms of explained variance. If we extract the principal component of the 24 hourly load demands, we find that the first one explains 94.9% of the variability; to fix the number of principal components and introduce them as regressors, we adopt a forecasting criterion based on a hold-out sample procedure. The mean square forecasting error decreases rapidly when we pass from 1 to 6 principal components; thereafter, adding regressors does not significantly improve the performance. We obtain a model where the predictors are the first 6 principal components, which explain around 99% of the hourly load demands of the day $(d - 1)$; the regression model, significant at the 5% level, has a determination coefficient R^2 (and adjusted R^2) equal to 0.947. This technique does not provide a better forecasting performance than the multiple regression previously illustrated: the percentage absolute errors have a mean equal to 7.50% and a standard deviation equal to 16.52%.

In Table 3, some statistics regarding the absolute percentage forecasting error distribution of the models discussed in the paper are summarized. We note the superiority of the functional models with respect to the other classical regression models: the mean and the median of absolute errors are smaller, and there is a reduction in the standard deviation. Moreover, we observe that the functional models permit a reduction in the maximum error. Finally, using a family of functional linear models allows the best performances in the class of regression models analyzed.

Table 3
Out-of-sample performances by absolute percentage error of the estimated models.

APE	Naif model	Multiple regression	Princ. comp. regression	Functional regression	Clust. funct. regression
Mean	8.20	7.21	7.50	6.96	6.55
St. Dev.	16.88	15.05	16.52	10.80	9.65
Median	4.28	3.79	3.92	3.87	3.59
Max.	189.10	165.22	198.74	93.69	72.97

5. Conclusions

In this paper we have proposed a method based on functional data analysis to approach a specific problem in short-term forecasting in a district-heating system. A family of functional linear models, selected by means of curve classification procedures, has been used to make peak load forecasts: the daily peak of heating demand is predicted on the basis of the "load curve" of the previous day. The technique presented, which generalizes the classical multiple regression model, is relatively simple and takes into consideration the functional nature of the problem considered. In fact, the method used could also be implemented when load curves are observed at a larger number of points in time, and also when the points are not equally spaced, a situation in which a multivariate approach could not be employed. The proposed model has shown a good performance in comparison with competing non-functional models. Moreover, functional clustering and functional linear discriminant analysis allow us to take into account the way in which the intra-daily pattern evolves over time without introducing a cumbersome specification with dummy variables or other complex procedures. The forecasting results using the functional techniques are promising.

From a technical point of view, weather variables such as temperature could be added in the functional model as predictors; however, because of the information contained in the heating data of the problem considered, a parsimonious model without requiring the inclusion of exogenous variables turns out to be satisfactory in the forecasting exercise. The functional method may be also extended to a non-parametric approach by employing, for instance, non-parametric regression models, or by using non-parametric unsupervised classification, as illustrated by Ferraty and Vieu (2006). There could be future interesting extensions of the functional techniques discussed in this paper in various directions. It could be possible, first, to assume that the errors are dependent on time, which may also better capture middle season months like April and October; second, to build forecasts of the entire daily load curve; and, finally, to provide distributional forecasts (as, for example, in Hyndman & Fan, 2008).

Acknowledgements

We would like to thank R. Angelini and G. Arduino (AEM Torino SpA), who provided the data-set. We also express our gratitude to the Editor and to the three anonymous Referees, whose comments have improved this paper. Professor G. Fusai would also like to acknowledge grant support to the project from the Fondazione CRT Torino.

References

Abraham, C., Cornillon, P. A., Matzner-Lober, E., & Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, *30*, 581–595.

Alves da Silva, A. P., Ferreira, V. H., & Velasquez, R. M. G. (2008). Input space to neural network based load forecasters. *International Journal of Forecasting*, *24*(4), 616–629.

Amin-Naseri, M. R., & Soroush, A. R. (2006). A hybrid neural network model for daily peak load forecasting using a novel clustering approach. In *Proceeding of the 10th IASTED international conference on artificial intelligence and soft computing. 2006* (pp. 104–109).

Antoch, J., Prchal, L., De Rosa, M. R., & Sarda, P. (2008). Functional linear regression with functional response: Application to prediction of electricity consumption. In *Functional and operatorial statistics*. Springer-Verlag.

Cardot, H., Ferraty, F., & Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters*, *45*, 11–22.

Cardot, H., Ferraty, F., & Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, *13*, 571–591.

Chicco, G., Napoli, R., & Piglione, F. (2001). Load pattern clustering for short-term load forecasting of anomalous days. In *IEEE porto power tech proceedings*: *vol. 2*. Piscataway, NJ, USA: IEEE.

De Boor, C. (2001). *A practical guide to splines*. Berlin: Springer-Verlag.

Dordonnat, V., Koopman, S. J., Ooms, M., Dessertaine, A., & Collet, J. (2008). An hourly periodic state space model for modelling French national electricity load. *International Journal of Forecasting*, *24*(4), 566–587.

Dotzauer, E. (2002). Simple model for prediction of loads in district-heating systems. *Applied Energy*, *73*(3), 277–284.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis*. New York: Springer.

Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under L1 constraint. *The Annals of Statistics*, *35*(5), 2367–2386.

Greenshtein, E., & Ritov, Y. (2004). Persistence in high-dimensional linear prediction selection and the virtue of overparameterization. *Bernoulli*, *10*(6), 971–988.

Hartigan, J. A., & Wong, M. A. (1979). A K-means clustering algorithm. *Journal of Applied Statistics*, *28*, 100–108.

Hyndman, R. J., & Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, *24*(3), 323–342.

Hyndman, R. J., & Fan, S. (2008). *Density forecasting for long-term peak electricity demand*. Working paper 06/08, Department of Econometrics and Business Statistics, Monash University.

Hyndman, R. J., & Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*, *51*, 4942–4956.

James, G. M., & Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society*, *63*, 533–550.

Jolliffe, I. T. (2004). *Principal component analysis*. New York: Springer.

Kargin, V., & Onatski, A. (2008). Curve forecasting by functional autoregression. *Journal of Multivariate Analysis*, *99*, 2508–2526.

Nielsen, H. A., & Madsen, H. (2006). Modelling the heat consumption in district heating systems using a grey-box approach. *Energy and Buildings*, *38*, 63–71.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York: Springer.

Sood, A., James, G., & Tellis, G. (2009). Functional regression: A new model for predicting market penetration of new products. *Marketing Science*, *28*(1), 36–51.

Weron, R. (2006). *Modeling and forecasting electricity loads and prices: A statistical approach*. Chichester: Wiley.