School of Mathematical
and Physical Sciences

Department of Mathematics and Statistics

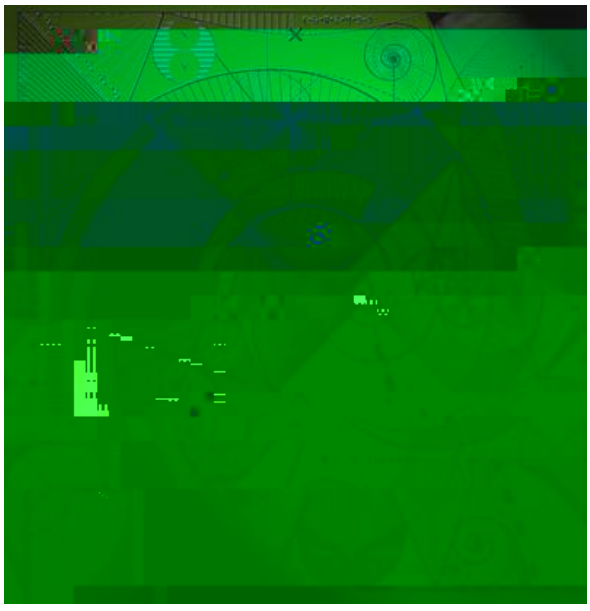
Preprint MPS-2013-05

14 February 2013

Clustering-based improvement of nonparametric functional time series forecasting. Application to intraday household-level load curves

by

Mohamed Chaouch



Clustering-based improvement of nonparametric functional time series forecasting. Application to intraday household-level load curves

Mohamed Chaouch
Centre for the Mathematics of Human behaviour
Department of Mathematics and Statistics
University of Reading, UK
email : *m:chaouch*@reading.ac.uk

February 14, 2013

Abstract

Energy suppliers are facing ever increasing competition, so that factors like quality and continuity of o□ered services must be properly taken into account. Furthermore, in the last few years, many countries are interested in Renewable Energy's (RE) such as solar and wind. RE resources are mainly used for environmental and economic reasons such as reducing the carbon emission. It also be used to reinforce the electric network especially during high peak periods. However, the injection of such energy resources in the Low-Voltage (LV) network can lead to a high voltage constrains. One possible solution is for electricity companies to motivate customers to use thermal or electric storage devices during high-production periods of PV to foster the integration of RE generation into the network. In this paper, we are interested in forecasting household-level electricity demand which represents a key factor to assure the balance supply/demand in the LV network. We propose a novel methodology able to improve short term functional time series forecasts. An application to the Irish smart meter data set showed the performance of the proposed method for forecasting intra-day household level load curves.

Keywords: Household-level forecasting, nonparametric statistics, unsupervised classi-

1 Introduction and Motivations

In recent years we have seen the arrival of new technologies such as Electrical Vehicle (EV) and electric heating as well as the increase of RE sources such as wind and solar. Therefore, the power grid is going through change. In fact, the stochastic nature of the RE sources will lead the power grid to a highly stochastic system. Within this new context, two main problems arise: (1) because of the electrification of appliances and mobility applications, the peak demand will increase and the load curve shape will change. In fact, at the moment a great deal of attention is attracted by EV, both hybrid and not, that will allow users to recharge their vehicles directly at home. It is therefore important to understand and expect what might be the impact on the power grid capacity of this recharging activity. This question has been studied recently by several authors, see for instance [1], [2], for more details. (2) It is well-known that one of the expected solutions to reduce the peak demand is to reinforce the power grid by RE generation. In fact, one can use energy storage sitythesol energy

himself as a producer of PV energy for instance. It is well-known that the development of smart meter and its massive deployment in Europe (80% households will be equipped by 2020) and North America allows us to get individual electricity consumption measures on a very fine time scale.

One-day-ahead forecasting of aggregated electricity demand has been widely studied in statistical literature. Di□erent approaches have been proposed to solve this issue. Time series analysis methods like (S)ARIMA models or exponential smoothing can be found in [5]-[9]. Those based on state-space models in [10]. Machine Learning approaches such as artificial

load segments. In that case, the past load segments are identified by mean of their closeness to some reference load segment which captures some expected qualitative and quantitative characteristics of the segment to be predicted.

In this paper, we are interested in short term forecasting of household-level intra-day electricity load curve. In contrast to aggregated load curves, which are characterised by their seasonality, regularity and sensibility to meteorological conditions, the household load curves are very volatile, their shape depends mainly on the *customer behaviour* and are less dependent to weather conditions. It is easy to see that the presence of customer behaviour, which is difficult to quantify, as a determinant factor of the shape of the individual load curve makes the issue of household-level forecasting difficult to solve. In this paper, we propose an improved version of the approach proposed by [27] adapted to household-level forecasting. The improvement procedure here is based on the use of an unsupervised clustering step of the historical segments which allows us to find segments describing a common consumption behaviour. Then, we use a nonparametric curve discrimination approach to assign a cluster to the last segment. This step allows us to identify segments which will be used to forecast the target segment.

The paper is organised as follows. In Section 2, we introduce the concept of *functional time series* methodology. Then, we summarize the functional wavelet-kernel approach proposed by [27] and describe the methodology proposed in this paper. Section 3 is devoted to an application of our method to intra-day household level load curve forecasting. A comparison study and an extension to 2000 Irish customers load forecasting is given in the same section. Some concluding remarks are given in Section 4.

2 Functional time series forecasting

Let us consider the household electricity demand as a (real-valued) continuous-time stochastic process $X = (X(t);\ t \in \mathbb{R})$. We are interested in the evolution of this process in the future. We suppose that we observe the process X over an interval $[0;T]$ and one would like to predict the behaviour of X on the entire interval $[T;T+\ ]$, where $\ > 0$, rather than at specific time points. To this end we can divide the interval $[0;T]$ into subintervals $[\ell\ ;(\ell+1)\ ]$, $\ell = 0;1;\ldots;k\ \ 1$ with $k = T=\ $, and to consider the (functional-valued) discrete-time stochastic process $\mathbb{S} = (S_n;\ n \in \mathbb{N})$, where $\mathbb{N} = \{1;2;\ldots\}$, defined by

$$S_n(t) = X(t+(n\ \ 1)\ );\quad n \in \mathbb{N};\forall t \in [0;\ ): \tag{1}$$

In this paper we are interested in one-day ahead intra-day load curve forecasting, the segmentation parameter corresponds to the daily electricity demand. In practice,

the electricity demand is recorded at a finite number of equidistance time points within each day, say $t_1; t_2; :::; t_P$, for instance, every half hour (in that case $P = 48$) or every 10 minutes (then $P = 144$). Let us denote by $S_n(t_i)$ the observation at time point t_i, $i = 1; 2; :::; P$, within curve S_n, n 2 N. We denote by

$$S_n = [S_n(t_1); S_n(t_2); :::; S_n(t_P)]; \quad n \text{ 2 N};$$

the segment of the total number of observations of the n-th curve S_n, n 2 N: Therefore, given a "sample" $S_1; S_2; :::; S_L$ of segments, our purpose is then to predict the *whole next segment* S_{L+1}. In other words we want to predict

$$S_{L+1} = [S_{L+1}(t_1); S_{L+1}(t_2); :::; S_{L+1}(t_P)]:$$

This forecasting issue has been a subject of several publication in statistical literature. The Functional Autoregressive (FAR) process has been introduced and studied theoretically by [29] and extensively used in both practical and theoretical studies since then, see [30]-[31] among numerous other contributions. Under the FAR model, the best predictor, $\hat{S}_{L+1}$, of the curve S_{L+1}, given the historical curves $S_1; S_2; :::; S_L$ is the conditional mean of S_{L+1} given the last curve S_L.

2.1 Functional wavelet-kernel approach (FWK)

A nonparametric approach based on kernel method has been developed by [27] to solve the same forecasting issue. In contrast to the FAR model, authors in [27] supposed that the regressiTJ0 6824 Tavarast to the FARve

household load curves do not contain such kind of similarity between days. For that reason, we use in this step, an unsupervised classification method to identify days describing a common consumption behaviour pattern. (b) Assign to the last observed segment S_L to the most "appropriate" cluster. The main purpose of this step is to find days that contain the same information as the last observed day. In other words, we look, in the historical segments, for those that describe a similar behaviour as what I observe today. (c) Apply the FWK method to forecast the segment S_{L+1} by using segments that belong to the cluster obtained in step (b). The following algorithm describes in more detail how we improve the FWK forecasts using clustering and curve discrimination approaches.

- Step1: *Unsupervised curve classification*

  Suppose that we have L 1 historical segments $S_1;S_2;:::;S_{L\ 1}$. In this step we are interested in splitting *automatically* these L 1 curves into M clusters, say $G_1;G_2;:::;G_m;:::;G_M$. Because we do not have at hand any categorial response variable and the data set are clearly of functional nature then this problem can be seen as an *unsupervised* curves classification. Since the number M of clusters is unknown in our case, then the unsupervised curves classification problem becomes harder to solve. In the statistical literature few authors gave a solution to that problem. These contributions are mainly restricted to the works by [33]-[34] and [35] in which k-means techniques for classification analysis are extended to curves data. In this paper we used the hierarchical algorithm proposed by [36]. The reader is referred to [36] to get more about this algorithm and the methodology behind.

- Step2: *Curve discrimination*

  The curve-discrimination step can be stated as follows. Given the historical segments $S_1;S_2;:::;S_{L\ 1}$, then from step1 we know in which cluster each segment belongs to. Let us denote by $G_`$ the cluster of the segment $S_`$. Assume that the

(4) can be seen as a regression function. Therefore, a nonparametric estimator of these probabilities has been proposed in [37]. For all $m \in 1;2;:::;M$,

$$\hat{p}_m(S_L) = \frac{\sum_{\ell:G_\ell = G_m} K_h(D(S_L;S_\ell))}{\sum_{\ell=1}^{L-1} K_h(D(S_L;S_\ell))}:$$

Therefore, say G_m, the cluster corresponding to the highest probability. We suppose that $G_m = \{S_1^{(m)};S_2^{(m)};:::;S_{K(m)}^{(m)}\}$, where $S_d^{(m)}$, $8d = 1;2;:::;K(m)$ are the segments that belong to the cluster G_m and $K(m)$ is the total number of segments in G_m.

- Step3: *Forecasting*

  Using results obtained in step2, we can now build the following sample of segments $\{S_d^{(m)};S_{d+1}^{(m)}\}_{d=1;:::;K(m)}$, where $S_d^{(m)}$ is the segment (corresponding to day d) that belongs to the cluster G_m and $S_{d+1}^{(m)}$ is the segment observed at the day $d+1$. Observe that $S_{d+1}^{(m)}$ doesn't necessarily belongs to the cluster G_m. Recall that our target is to forecast the segment $S_{L+1\ +1}$. Therefore wem

Figure 1: First sample of residential customer's load curve.

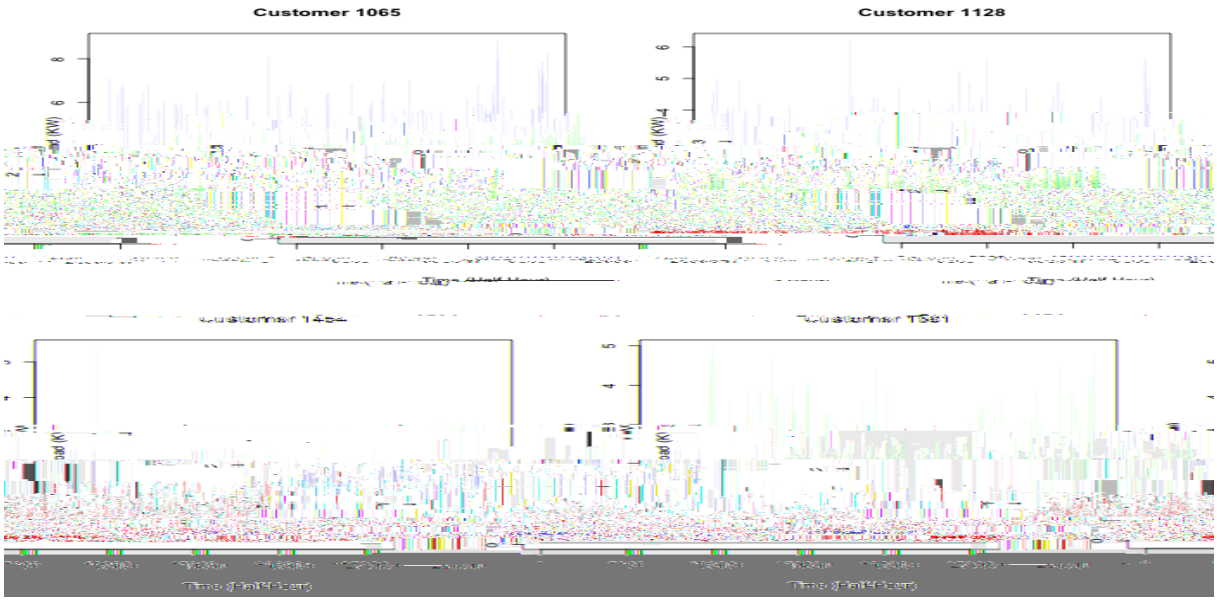


Figure 2: Second sample of residential customer's load curve.

3.2 An illustration of CFWK approach to customer 1016

In this section, we focus on the application of the CFWK method to one randomly chosen customer. We take as example the customer number 1016 in the Irish data. Later, we suggest to extend the results to the entire sample of 2000 customers. Figure 3 (a) shows the original time series which represents the half-hourly electricity demand of this cus-

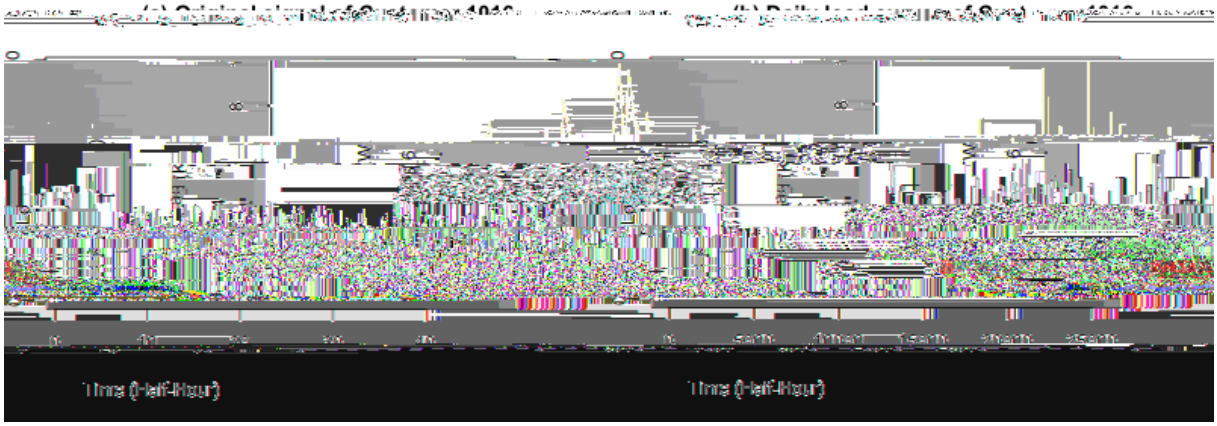


Figure 3: (a) Half-hour electricity demand of customer 1016 between 14/07/2009 and 31/12/2010. (b) A sample of 535 daily load curve (segments) of the same customer.

tomer between 14/07/2009 and 31/12/2010. In Figure 3 (b), we split the original signal into daily load curves ($P = 48$) in order to be able to apply the proposed functional approach. Thus we obtain a sample, say $S_1; S_2; :::; S_{535}$, of 535 daily load curves (segments). One can easily observe, from Figure 3 (b), that the electricity demand for that customer is very low between 00:00 and 07:00. Then, the demand increase around 07:30 which corresponds to the morning activity in the household. During the day, consumption decreases in the most of days. Finally, we can observe the classical evening peak demand between 19:00 and 20:00.

To validate our method, we split this sample in two parts. Firstly, denoted by $L = \{S_1; S_2; :::; S_{170}\}$, a learning sample containing daily load curves from 14/07/2009 to 31/12/2009. This sample will be used to build clusters and find the "optimal" bandwidth h. The second part, denoted by $T = \{S_{171}; S_{172}; :::; S_{535}\}$, is the test sample which will be used to compare our forecasts to the observed daily load curves for the period between 01/01/2010 to 31/12/2010 (365 days). Each segment in the test sample T is forecasted independently. In fact, to forecast the segment S_{171} we use as historical segments $S_1; S_2; :::; S_{169}$ and the last observed segment is S_{170}. Then to forecast the segment S_{172}, we consider the historical data $S_1; S_2; :::; S_{170}$ and the last observed segment now is S_{171} (the true one and not its forecast). This procedure will be repeated until we forecast all segments that belong to the test sample T. Based on the sample $S_1; S_2; :::; S_{170}$ of segments, the goal now is to forecast the segment S_{171} (which corresponds to the 1st Jannuary 2010) using CFWK approach. To this end, the following steps are taken:

1. *How many clusters do we have?*

Based on segments $S_1; S_2; :::; S_{169}$ and using the hierarchical algorithm proposed by [36],

Figure 4: Clusters obtained for cluster 1016.

we find three clusters which are represented in Figure 4 (a)-(c). The median daily pro-

Table 1: Type of days within each cluster

	Cluster 1	**Cluster 2**	**Cluster 3**
Mon.	3	**17**	4
Tue.	8	**9**	8
Wed.	1	**15**	8
Thu.	4	**15**	5
Fri.	4	8	**12**
Sat.	**22**	1	1
Sun.	**9**	5	**10**
Total	51	70	48

2. *Which cluster to be assigned to the last observed segment $S_L = S_{170}$?*

A nonparametric curve discrimination method introduced by [37] has been used to assign a cluster for each last observed segment S_L in the training sample. In this example the last segment S_{170} corresponds to the load curve observed on 31/12/2009. Our main task is to predict the corresponding class (which will be in our case cluster 1, 2 or 3) for this segment. To apply the discrimination method explained in sub-section 2.2 several tuning parameters should be fixed. The kernel is chosen to be quadratic and the optimal bandwidth is chosen by the cross-validation method on the k-nearest neighbors (see [23], p. 115 for more details). Another important parameter needs to be fixed is the semi-metric D(*;*). In this example, because of the roughness of the load curves, we used a semi-metric computed with the functional principal components analysis (see [38]) with an optimal dimension equal to 2. The optimality here was measured with respect to the rate of misclassified curves obtained within the learning sample (17% in this case). Finally, the discrimination method assigned the cluster 1 to the segment S_{170}. This result looks to be compatible with the shape of the load curve of the segment S_{170} presented in Figure 5. In fact, since 31/12/ 2009 is a Christmas Holiday, the customer behaviour in that period is expected to be the same as on the week-end. We can easily see, from Figure 5, the absence of the small peak demand usually observed at 07:30 on working days. We also observe the presence of two important peaks during the day: the first one around 12:30 which corresponds to lunch time and another more important one around 15:30.

3. *Day-head forecasting and validation criteria*

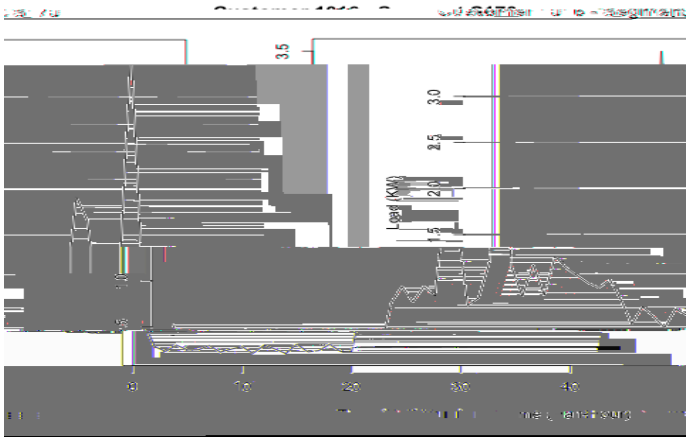


Figure 5: Last observed segment in the training sample for customer 1016: $S_L = S_{170}$ which corresponds to 31/12/2009.

Recall that our main purpose in this example is to forecast the half-hour load curve of the 1st of January 2010 which corresponds to the segment S_{171}. Using results obtained in step 1 and 2 we can then consider the following sample $S_d^{(1)}; S_{d+1}^{(1)}{}_{d=1;2;\dots;51}$, where 51 is the number of segments in cluster 1. Therefore, the forecast of S_{171} is obtained as follow

$\hat{S}$

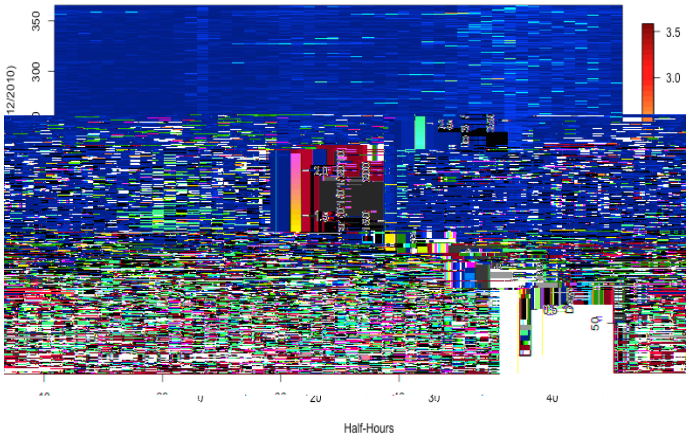


Figure 6: Half-hour Absolute Errors (HHAE) obtained by CFWK method (case of customer 1016).

Figure 7: Distribution (by month) of the daily median absolute errors (DMAE) obtained by CFWK and FWK (case of customer 1016).

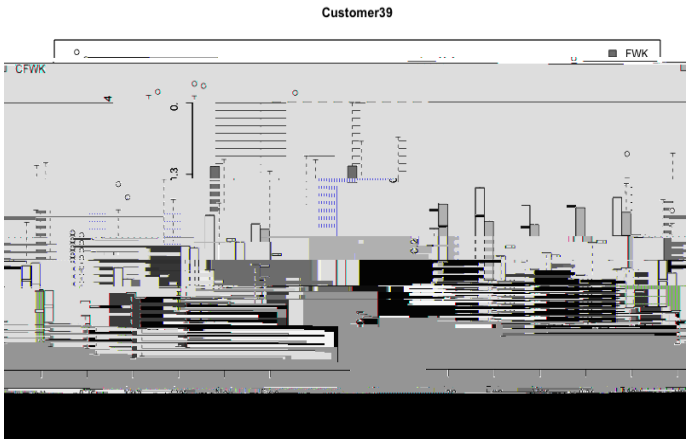


Figure 8: Distribution (by month) of the daily median absolute errors (DMAE) obtained by CFWK and FWK (case of customer 39).

Daily Median Absolute Error (SDMAE) defined as follow:

$$\mathrm{SDMAE}_d = \mathrm{Median}\ \left(\mathrm{DMAE}_d^{(1)}; \mathrm{DMAE}_d^{(2)}; \ldots; \mathrm{DMAE}_d^{(2000)}\right) :$$

Figure 10 displays the distribution, for each month, of the SDMAE errors provided by CFKW and FKW approaches. Table 2 gives numerical summary of results obtained in Figure 10. For instance, if we take the January 2010 as an example, one can observe that, with the CFWK (resp. FWK) approach, 50% (of the 2000 customers in the panel) have a daily median absolute error (DMAE) less than 0.206 KW (resp. 0.222 KW) and 75% of them have a DMAE errors between 0.195 KW and 0.215 KW (resp. 0.211 and 0.235). The same analysis might be made for the other months. Table 2 shows that CFWK approach is more e□cient than the FWK one.

Figure 9: Distribution (by month) of the daily median absolute errors (DMAE) obtained by CFWK and FWK (case of customer 708).

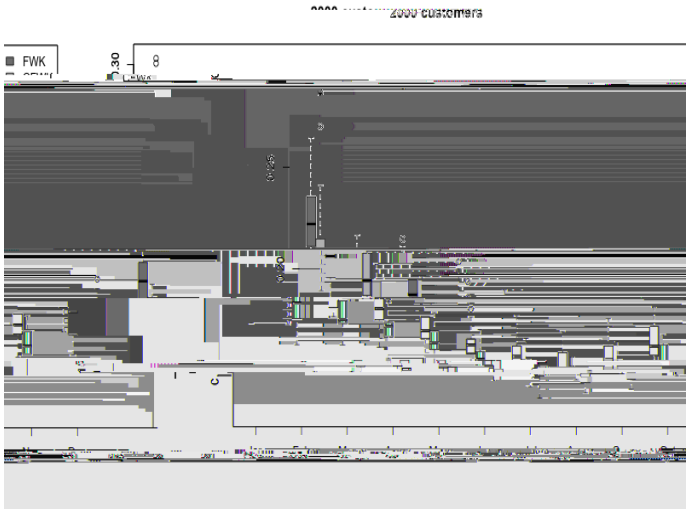


Figure 10: Distribution (by month) of the Sample Daily Median Absolute Errors (SDMAE) obtained by CFWK and FWK.

4 Conclusion

In this paper, a new approach for forecasting functional time series has been proposed. An application to short-term intra-day household-level load curve forecasting has shown the performance of the proposed methodology. The idea behind the use of a classification step is mainly to get a reasonable assumption of stationarity for our time series. Moreover, because the intra-day individual load curve shape is mainly a□ected by the consumption behaviour of the customer and there is no evidence to identify a common pattern between days we used an unsupervised classification method to find

Table 2: Distribution (by month) of the Sample Daily Median Absolute Errors (SDMAE) obtained by CFWK and FWK.

	CFWK				FWK			
	Mean	$Q_{0:25}$	$Q_{0:5}$	$Q_{0:75}$	Mean	$Q_{0:25}$	$Q_{0:5}$	$Q_{0:75}$
Jan.	0.209	0.195	0.206	0.215	0.226	0.211	0.222	0.235
Feb.	0.183	0.175	0.183	0.193	0.196	0.187	0.193	0.205
Mar.	0.177	0.173	0.174	0.184	0.189	0.181	0.187	0.194
Apr.	0.171	0.166	0.172	0.174	0.181	0.175	0.180	0.188
May	0.164	0.160	0.163	0.167	0.171	0.165	0.169	0.175
Jun.	0.156	0.151	0.155	0.159	0.163	0.157	0.162	0.166
Jul.	0.151	0.147	0.151	0.156	0.158	0.152	0.158	0.164
Aug.	0.148	0.144	0.148	0.151	0.155	0.151	0.154	0.158
Sep.	0.151	0.148	0.151	0.156	0.158	0.153	0.156	0.162
Oct.	0.156	0.152	0.153	0.163	0.164	0.157	0.162	0.172
Nov.	0.164	0.158	0.163	0.170	0.174	0.169	0.173	0.178
Dec.	0.178	0.169	0.180	0.186	0.193	0.182	0.198	0.203

similar segments. The numerical results obtained showed that the clustering based approach works very satisfactorily and outperforms the functional wavelet-kernel time series predictor. We note the proposed methodology might be improved by using some daily exogenous functional random variables, like internal/external daily temperature and sunshine curves. Other discrete variables, such as surface of the property, number of electric appliances and number of occupants can also be taken into account which might a□ect daily individual load demand.

Acknowledgment

The author would like to thank Scottish and Southern Power Distribution SSEPD for support and funding via the New Thames Valley Vision Project (SSET203 - New Thames Valley Vision) - funded through the Low Carbon Network Fund

[11] R. Lamedica, A. Prudenzi, M. Sforna, M. Caciotta and V. Cencelli, “A neural network based technique for short-term forecasting of anomalous load periods”, *IEEE Trans. Power Syst.*, vol. 11, no. 4, pp. 1749-1756, 1996.

[12] S. R. Abbas and M. Arif, “Electric load forecasting using support vector machines optimized by genetic algorithm”, *INMIC’06 IEEE Multitopic Conference 2006*, Dec. 2006, pp. 395-399.

[13] J. Fidalgo and M. A. Matos, “Forecasting Portugal global load with artificial neural networks”, *ICANN2007-International Congress on Artificial Neural Networks 2007*, pp. 728-737.

[14] J. M. Poggi, “Prévision non paramétrique de la consommation électrique”, *Rev. de Stat. Appliquée*, vol.12, no.4, pp. 83-98, 1994.

[15] V. Lefieu, “Modèles semi-paramétriques appliqués à la prévision des séries temporelles. Cas de la consommation d’électricité”, *Ph.D. dissertation*, 2007.

[16] Y. Xia, H. Tong and W. K., Li, “An adaptive estimation of dimension reduction space”, *Jour. of Roy. Stat. Soc. B*, vol. 64, no. 3, pp. 363-410, 2002.

[17] S. Fan and R. J. Hyndman, “Short-term load forecasting based on a semi-parametric additive model”, *IEEE Trans. Power Syst.*

[23] F. Ferraty and Ph. Vieu, *Nonparametric Functional Data Analysis. Theory and Practice*, Springer-Verlag, New-York, 2006.

[24] J. M. Vilar, R. Cao and G. Aneiros, "Forecasting next-day electricity demand and price using nonparametric functional methods", *Inter. J. Electri. Power and Energ. Syst.*, vol. 39, no. 1, pp. 48-55, 2012.

[25] A. Goia, C. May and G. Fused, "Functional clustering and linear regression for peak load forecasting", *Inter. J. Forecast.*, vol. 26, no. 4, pp. 700-711, 2010.

[26] J. Antoch, L. Prchal, M. R. De Rosa and P. Sarda, "Electricity consumption prediction with functional linear regression using spline estimators", *J. Appl. Stat.*, vol. 37, no. 12, pp. 2027-2041, 2010.

[27] A. Antoniadis, E. Paparoditis and T. Sapatinas, "A functional wavelet-kernel approach for time series prediction", *J. Roy. Stat. Soc. B*, vol. 68, pp. 837-857, 2006.

[28] E. Paparoditis and T. Sapatinas, "Short-Term Load Forecasting: The Similar Shape Functional Time Series Predictor", *Preprint*, 2012.

[29] D. Bosq, "Linear procD. ⊞⊞E.aeprsting:

[36] S. Dabo-Niang, F. Ferraty and P. Vieu, “Mode estimation for functional random variable and its application for curves classification”, *Far East J. Theoret. Statist*, vol. 18, no. 1, pp. 93-119, 2006.

[37] F. Ferraty and P. Vieu, “Curves discrimination: a nonparametric functional approach”, *Comput. Stat. Data Anal.*, vol. 44, pp. 161-173, 2003.

[38] P. Hall and M. Hosseini-Nasab, “On properties of functional principal components analysis”, *J. Roy. Stat. Soc.*, vol. 68, no. 1, pp. 109-126, 2006.

[39] A. Antoniadis, E. Paparoditis and T. Sapatinas, “Bandwidth selection for functional time series prediction”, *Stat. Probab. Letters*, vol. 79, pp. 733-740, 2009.