REVSTAT – Statistical Journal
Volume 8, Number 2, November 2010, 105–124

# OPTIMIZED CLUSTERS FOR DISAGGREGATED ELECTRICITY LOAD FORECASTING

Authors: Michel Misiti
– Laboratoire de Mathématique, Orsay, France
and Ecole Centrale de Lyon, France
Michel.Misiti@ec-lyon.fr

Yves Misiti
– Laboratoire de Mathématique, Orsay, France
Yves.Misiti@math.u-psud.fr

Georges Oppenheim
– Laboratoire de Mathématique, Orsay, France
and Université de Marne-la-Vallée, France
Georges.Oppenheim@math.u-psud.fr

Jean-Michel Poggi
– Laboratoire de Mathématique, Orsay, France
and Université Paris Descartes, France
Jean-Michel.Poggi@math.u-psud.fr

Received: October 2009 Revised: May 2010 Accepted: June 2010

Abstract:

- To account for the variation of EDF's (the French electrical company) portfolio following the liberalization of the electrical market, it is essential to disaggregate the global load curve. The idea is to disaggregate the global signal in such a way that the sum of disaggregated forecasts significantly improves the prediction of the whole global signal. The strategy is to optimize, a preliminary clustering of individual load curves with respect to a predictability index. The optimized clustering procedure is controlled by a forecasting performance via a cross-prediction dissimilarity index. It can be assimilated to a discrete gradient type algorithm.

Key-Words:

- *clustering; disaggregation; forecasting; optimization; wavelets.*

AMS Subject Classification:

- 62P30, 62H30, 62-07, 62M20.

## 1. INTRODUCTION

This paper is devoted to electricity load forecasting via the disaggregation of the global signal. This disaggregation is based on customer clustering. To clarify, let us split the introduction in two parts. The first paragraph is dedicated to the general context and the second focuses on the specific application.

### 1.1. General context

Regular forecasting of the electrical load demand arises from a multiplicity of sources such as consumer behavior linked to social activities, government regulations and conventions and meteorological factors. Forecasting demand using statistical models with different types of explanatory variables provides accurate results. Those models include different components. There is a trend as well as daily, weekly and annual seasonal components. Calendar affects help take into account public holidays. An additional term may account for the effect of meteorological variables on the electricity load.

A recent special issue of the International Journal of Forecasting, devoted to energy forecasting, presents domain-related papers (see the editorial presentation by Taylor, Espasa [22]). Six papers in this journal provide an overview of the recent strategies for short-term or very short-term electricity load forecasting. The following methodologies: multi-equation model, neural networks or switching models are applied at national level in France, Spain, Australia, Brazil and Great Britain. This paper presents another approach to energy forecasting: a forecasting method based on signal disaggregation via the clustering of individual load curves.

Our goal is twofold. We aim to improve the accuracy of the electricity load forecast and to account for variability in the EDF's customer portfolio. This variability is due to the opening up of the previously nationalized electricity market. One way to deal with this difficulty is to disaggregate the global signal to improve the forecasting performance. Therefore, we need to create customer clusters such that the sum of disaggregated forecasts significantly improves the forecast of the whole global signal. In this paper we propose an optimized clustering scheme controlled through a cross-prediction dissimilarity index and based on a discrete gradient type algorithm.

Clustering has already been used for forecasting in similar electricity cases. Let us briefly present three examples.

The first example uses clustering for short-term peak load forecasting, proposed in Goia *et al.* ([10]). For a given load curve, forecasting is based on a two-stage strategy. A functional clustering is created to classify the daily load curves and then a functional linear regression model is used on each cluster. Next, a new load curve is assigned to the clusters thanks to a functional discriminant analysis.

The second example, proposed by Piao *et al.* ([20]), deals with the prediction of customer load pattern in long duration load profiles. It also starts with clustering based on three daily profiles characteristics and aims at creating classes of load pattern and extracting representative load profiles for each class. Supervised learning methods can possibly be used when a new load curve is treated.

The third example, provided by Espinoza *et al.* ([9]), uses the forecasting step before clustering. Each individual load curve is first modeled using parametric periodic time series. Then, a typical daily profile is extracted from this parametric model for each individual customer. Finally, customer segmentation is obtained from the clustering these typical daily profiles.

The originality of our approach is the inclusion of an optimization step supervised by the forecasting procedure combined with a specific clustering strategy. To complete this introduction, let us present the industrial context of our work.

## 1.2. Industrial context

Load forecasting is a critical task for a company like EDF since it contributes to production planning. Engineers provide at noon, on a daily basis, the next day's consumption forecast. Forecasting is not only useful for short term decisions but also for optimizing production in dams or plants. Models similar to the following simplified version are built. The load $P_t$ is decomposed into three components:

$$P_t = Pi_t + Pd_t + \varepsilon_t \tag{1.1}$$

where $Pi_t$ is a weather independent component containing trend, seasonality and calendar effects, $Pd_t$ a weather dependent component and $\varepsilon_t$ an error term.

The parameters involved in the two first terms depend on the hour $h$, on the position of the day $d$ within the year and on the day-type. The weather independent part is a linear combination of four sine and cosine terms whose coefficients are mainly functions of $d$, the type of the day (7 types), and $h$:

$$Pi_t = \Pi_{h,y} \sum_{m=1}^{4} a_{h,m} \cos\left(\frac{2\pi\, m d}{c}\right) + b_{h,m} \sin\left(\frac{2\pi\, m d}{c}\right) \tag{1.2}$$

where $c = 326.25$ and $\Pi_{h,y}$ is the load shape at the hour $h$ for the year $y$, which depends on the day-type. The coefficients $a$ and $b$ also depend on the day-type. The weather dependent component is composed of two parts. The first part involves a cooling gradient and a smoothed summer temperature while the second part involves a heating gradient and a smoothed winter temperature. This smoothed temperature can be assimilated to the indoor temperature.

After some fine tuning of the parameters, the quality of the EDF model measured by the Mean Absolute Percentage Error or MAPE, is considered as good. While the quality is satisfactory, it is never good enough during holidays such as Christmas. For instance, an estimation based on a five-year period such as [2000–2005] gives an hour-MAPE around 1.2%. For that same period, the one year forecast hour-MAPE is almost the same value. However, because of the deregulation of the French electricity market the situation has changed since 2007. EDF customers can now switch from one electricity provider to another, bringing instability to the market. As a consequence, the data available for forecasting is evolving. Before deregulation, a 5-year database was trademark for quality forecasting. Because of the new legal and commercial context, only one to two years of good enough quality data is now available to researchers.

### 1.3. Outline

This paper is organized as follows. After this introductory section, Section 2 is devoted to the problem and the data. Section 3 briefly recalls a wavelet based procedure for clustering load curves. Section 4 proposes the optimized clustering for forecasting by disaggregation. Section 5 contains experimental results on real world data and Section 6 presents some perspectives for future work.

## 2. THE DATA AND THE PROBLEM

### 2.1. The data

The data considered in this paper is not the French half-an-hour load consumption but individual commercial customer data. For obvious confidentiality and industrial reasons the French database is partially undisclosed. Moreover, the most recent data is not available. We worked on the [2000–2001] electricity consumption period. Individual power electricity demand curves, anonymous for confidentiality reasons are available for 2309 industrial customers during this period. The sampling period is one hour, leading to 17520 samples.

To highlight the differences among individual curves, let us examine four customer load curves during [2000–2001] time period (see Figure 1).

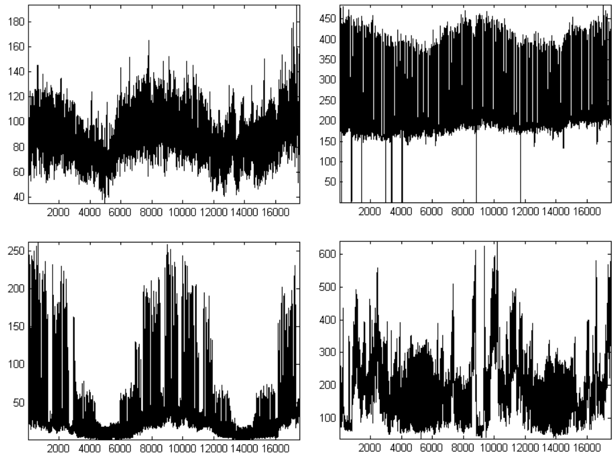

**Figure 1**: Raw data: 4 customer load curves during the 2000–2001 period (load in kW versus time in hours).

The long term shape of the curves differs a lot. It looks climate free for the customer at the bottom right while we can see three different climatic sensitivities on the other graphics. The same 4-customer load curves for one particular week in 2000 are displayed in Figure 2.

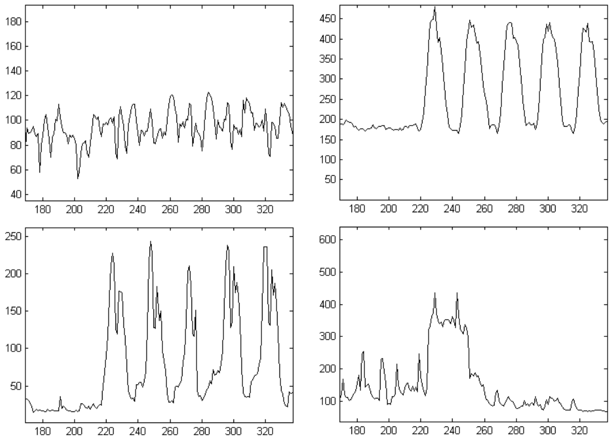

**Figure 2**: Raw data: 4 customer load curves during one week of 2000 (load in kW versus time in hours).

The two customers on the main diagonal of the plot are different from the others. Their graphics do not show any clear ‘social rhythm’, whereas the global shape of the two other customers are similar and display a ‘social rhythm’.

Weekends are easy to detect and the main differences appear in the middle of the work days. The bottom left graphic, displays a bimodal shape instead of a single peaked curve.

## 2.2. Aggregated versus disaggregated

The disaggregation based forecasting problem goes as follows. Let us denote by $X_i(t)$ the value of the load curve of the $i$th customer at time $t$ and we consider the aggregated electricity consumption signal:

$$S(t) = \sum X_i(t)\ . \tag{2.1}$$

The aggregated forecast is obtained by modeling and forecasting the signal:

$$\widehat{S}_{\text{aggr}}(t) = \widehat{S}(t)\ . \tag{2.2}$$

Associated with any partition of clustered individuals, we can define the consumption of each cluster $g$:

$$S_g(t) = \sum_{i \in g} X_i(t)\ . \tag{2.3}$$

Then, the disaggregated forecast is obtained by modeling and forecasting the signal within each cluster $\widehat{S}_g(t)$ and then summing over all clusters:

$$\widehat{S}_{\text{dis}}(t) = \sum_{g} \widehat{S}_g(t)\ . \tag{2.4}$$

We restrict our attention to this particular form of aggregation (2.4) which may not be the optimal combination. It could be interesting to consider a weighted sum and optimize the weights. However, in this study we specifically focus on the clustering issue and preserve the percentage of the load curve associated with each cluster.

Our challenge is to find the best partition of individuals. This partition has to be as accurate as possible from a forecasting perspective so that the latter will perform significantly better than the aggregated forecast.

Why can we expect such a result? Let us present two useful results already validated in narrower context.

A first indication is provided by the simplest statistical inference problem: the estimation of the mean $\mu$ of a variable $Y$ on a given population using the random sample mean $\overline{Y}$. It is an unbiased estimator of $\mu$ of variance $\sigma^2(Y)/n$ where $n$ is the sample size. Using stratified representative sampling with respect to a given partition, the associated stratified estimator's variance (which is the disaggregated one) is reduced to the within variance over $n$: $\sigma_w^2(Y)/n$, which is always smaller than the variance of $\overline{Y}$.

A second indication comes from a simple result stated for two clusters and true for more clusters. Let us denote $X_t$ and $Y_t$, two sequences corresponding to generic signal pairs $X_{i,t}$ and $X_{j,t}$ associated with two different clusters. Assume that $X_t$ and $Y_t$ are two sequences of stationary square integrable random variables and define

$$S_t = X_t + Y_t \; . \tag{2.5}$$

Then denoting by

$$\widehat{Z}_t = E(Z_t \mid Z_{t-1}, Z_{t-2}, ..., Z_1) \tag{2.6}$$

the conditional mean of $Z_t$ given its own past. Let us define the two error indices

$$Err_{\text{aggr}} = E(S_t - \widehat{S}_t)^2 \tag{2.7}$$

and

$$Err_{\text{dis}} = E(S_t - \widehat{X}_t - \widehat{Y}_t)^2 \, . \tag{2.8}$$

So, if $X_t$ and $Y_t$ are independent then

$$Err_{\text{dis}} \leq Err_{\text{aggr}} \; . \tag{2.9}$$

If signals corresponding to two different clusters are independent and the conditional mean (in fact actually an accurate estimation) is used to predict, then the disaggregated forecast is of better quality than the forecast on the whole global signal.

Let us give a proof of that result. Starting from the definition of $Err_{\text{aggr}}$ and since $S_t = X_t + Y_t$, we get

$$Err_{\text{aggr}} \geq E\big(S_t - E(S_t \mid X_{t-1}, Y_{t-1}, ..., X_1, Y_1)\big)^2 \, . \tag{2.10}$$

Independence between the $X$'s and $Y$'s leads to

$$E\big(X_t - E(X_t \mid X_{t-1}, Y_{t-1}, ..., X_1, Y_1)\big) = E\big(X_t - E(X_t \mid X_{t-1}, ..., X_1)\big) \tag{2.11}$$

as well as the equation obtained by permuting $X$ and $Y$ in (2.11). Adding up these two equations and taking squares of both sides, we obtain

$$E\big(S_t - E(S_t \mid X_{t-1}, Y_{t-1}, ..., X_1, Y_1)\big)^2 = E\big(S_t - \widehat{X}_t - \widehat{Y}_t\big)^2 = Err_{\text{dis}} \; .$$

Therefore, with (2.10) we obtain inequality (2.9).

As a conclusion, the two previously stated results suggest that it may be useful to disaggregate the global signal to significantly improve forecasting. Our idea is to find a good tradeoff between homogeneity within clusters and quality of the model's estimation. Homogeneity increases while the quality decreases with a higher number of clusters. Hence a three-step strategy:

1. Preprocessing individual customer data using wavelets;
2. Primary customer clustering with numerous homogeneous clusters;
3. Aggregation using stepwise optimization algorithm based on a dissimilarity index linked to a cross-prediction error and a discrete gradient type algorithm.

First, let us provide some additional information on the basic forecasting model. Then, we will develop the three-step strategy.

## 2.3. Eventail-like forecasting model

The aim of this paragraph is to clarify the internal forecasting procedure to the non-initiated reader while avoiding detailed information. Let us emphasize the fact that the error reduction via the new scheme is solely due to clustering optimization. Indeed, we do not perform ad-hoc adaptation of the model design strategy to the obtained clusters.

We circumscribe this paper to a single 'black-box' method used to design the forecasting model, starting from a given time series. Let us explain that we will take full advantage of a fully automatic version of EDF operational model called Eventail. Eventail is designed to predict the aggregated electricity consumption. Bruhns *et al.* ([6]) give a detailed description of a non-linear forecasting model of French electricity load in use at EDF. This model allows for different levels of seasonality and weather dependence. As previously stated, daily, weekly and annual components of the endogenous variable are considered, along with exogenous variables such as temperature, cloud cover, calendar events as well as a long-term trend. The mid-term model is a highly parameterized climate-free SARIMA model additively corrected with a weather dependent term. This model delivers an accurate forecast.

Let us note some results. The forecasting performance on the sample of 2309 customers, measured by the long-term MAPE (for Mean Absolute Percentage Error) is about 4.06% for the global aggregated signal. Meanwhile, on the same sample, the completely disaggregated forecasting performance reaches 2.94%.

**Remark 2.1.** We will not provide comparisons of Eventail with other forecasting methods. This would be interesting (see for example Hippert *et al.* ([14]) for recent statistical time series tools for load forecasting), however since Eventail is the current operational tool, it is regularly improved in order to take into account the new characteristics of the load curve. Let us mention that Bruhns *et al.* ([6]) describe the forecasting model already used at EDF for mid-term load forecasting and provide a comparative study of various alternatives. Also, a more recent discussion on how to handle changes in customer behavior in a similar context can be found in Dordonnat *et al.* ([8]) who describe a forecasting model based on time-varying processes, specifically a periodic state space model.

## 3. CLUSTERING USING WAVELETS

The aim of the preliminary step is to build basic clusters (often called super customers hereafter) based on our sensibly assembled customers. The key idea is to take advantage of the hierarchical multiresolution structure of wavelet decomposition (see Misiti *et al.* [18]) for clustering signals. Simply put, wavelets allow us to write each individual signal as the sum of orthogonal signals: a coarser approximation at large scale (low frequency) and additional details at different resolutions, of decreasing scales. The approximation at level $j$ roughly represents the local mean signal on intervals of length $2^j$ while the detail at level $j$ contains fluctuations around this local mean on the same corresponding intervals. Let $n$ be the common length of the $p$ series individually denoted by $X^{(i)}$. Then, for a given orthogonal wavelet $\psi$, each time series can be decomposed at level $J$ (which is at most the integer part of $\log_2(n)$). This leads to:

$$X^{(i)} = A_J^{(i)} + \sum_{j=1}^{J} D_j^{(i)}\,, \tag{3.1}$$

where $A_k^{(i)}$ and $D_k^{(i)}$ denote respectively the approximation and the detail, at level $k$, of the signal $X^{(i)}$.

The procedure, described in Misiti *et al.* ([19]), is a hybrid scheme mixing regularization and filtering approaches, according to James and Sugar's ([15]) terminology. Let us describe this scheme. First, there is individual denoising using a signal-adapted wavelet basis, then a projection on a one common wavelet basis to get a huge dimensionality reduction effect (see Biau *et al.* [5]). Then each customer is characterized by coefficients. The last step of the process is the clustering of the customers using Ward's method with squared Euclidean distances, in order to preserve distances between signals through wavelet coefficients encoding. We generate hierarchies of partitions corresponding to different numbers of clusters and various wavelet representations, that are typically approximations of decreasing resolution level.

For the final step, considering any partition $P$ obtained by clustering data $Z$ and for a given number of clusters, we can compute the following usual variance ratio quality index:

$$I_Z(P) = \frac{\mathrm{Var}_b(Z,P)}{\mathrm{Var}_w(Z,P)} \,, \tag{3.2}$$

where $\mathrm{Var}_b(Z,P)$ and $\mathrm{Var}_w(Z,P)$ denote respectively the variance between clusters and within clusters. This quality index allows us to compare two partitions based on two different signal representations but it depends heavily on the number of clusters. For instance, let us say $P'$ is a finer partition obtained from $P$. Then $I_Z(P') \geq I_Z(P)$. Since we have to compare partitions with different number of clusters, we will choose the one leading to the best normalized variance ratio index:

$$I_Z^N(P) = \frac{\mathrm{Var}_b(Z,P)}{C(P)\cdot \mathrm{Var}_w(Z,P)} \,, \tag{3.3}$$

where $C(P)$ is the number of clusters within partition $P$.

This index is similar to the statistic of Calinski and Harabasz ([7]), considered as a 'good competitor' (see for example Tibshirani *et al.* [23]). It allows us to select a convenient number of clusters as well as a critical level of wavelet decomposition (simulated examples, electricity data processing and further details can be found in Misiti *et al.* [19]).

In our electrical context, in an earlier study we obtained various partitions using this clustering scheme but without taking into account the forecasting objective. The most interesting partitions are made of 15 to 19 clusters and highlight wavelet approximation coefficients at level 6 (around 2 coefficients a week) and detail coefficients at level 2 (around 5 coefficients a day). These partitions reached a forecasting performance of 2.75% long-term MAPE which is better than the fully aggregated or the fully disaggregated forecasts. However, partitions describes in this paragraph cannot be improved with the optimization process described in the next section.

Therefore, hereafter we will work from this initial pre-processing. We will select wavelet approximation coefficients at level 6 in order to get the load curve's global shape. We will relax unsupervised clusters constraints. This means that we will start with a large number of clusters and step by step aggregate them with an optimization criterion supervised by predictability. According to the variance ratio, 90 clusters are sufficient to assume strong homogeneity.

## 4. OPTIMIZED CLUSTERING DIRECTED BY FORECASTING

### 4.1. A multistage procedure

The proposed optimized clustering scheme is as follows:

1. *Wavelet preprocessing.*
   Customer characterization through wavelet representation of each signal after standardization using approximation coefficients at level 6.

2. *First clustering* around numerous centroids.
   A minimum of 90 clusters regrouping homogeneous customers. Each cluster is represented by its aggregated signal.

3. *Iterative optimization.*
   The starting point being the described initial partition, an optimization process supervised through cross-prediction dissimilarity index is run. A discrete gradient type procedure based on $D$ matrix (defined in the next section) explores the set of partitions.

### 4.2. Cross-prediction dissimilarity

To qualify a specific aggregation we use cross-prediction dissimilarity between elements. Those elements can be either individual or aggregated signals. This dissimilarity index between $X_k$ and $X_j$ is based on the following idea. The model fitted on past observations of $X_j(t)$ is used to predict the future of $X_k(t)$ and vice-versa. In our specific electrical context, let us denote by

$$\mathit{forec}_{k|j}^{2001} = \mathit{forecast}\big(X_j^{2000}, X_k^{2001}\big)\,, \tag{4.1}$$

the forecasts of $X_k$ on the year 2001 (the test period) obtained from the model fitted on $X_j$ on the year 2000 (the learning period). The fitted model is based on the Eventail-like design tool. Then, the associated error is defined by:

$$E_{k|j} = \mathit{error}\big(X_k^{2001}, \mathit{forec}_{k|j}^{2001}\big)\,. \tag{4.2}$$

Then a natural symmetric measure of dissimilarity is given:

$$D = (D_{j,k}) = \big((E_{k|j} + E_{j|k})/2\big)\,. \tag{4.3}$$

To fairly rescale the $X_k$ and $X_j$ load curves for testing and for estimating the index $D$ is based on errors obtained from the *l1*-normalized versions (*i.e.* signals summing to 1) .

### 4.3. Zooming in on the optimization step

The iterative optimization of the initial partition is supervised through the cross-prediction dissimilarity. It can be adapted to the forecast horizon as well as to the error criterion. The iterative optimization is based on discrete gradient via a neighborhood definition through dissimilarity between an element and a cluster induced by the matrix $D$. The basic step is an iterative exploration of elements. These elements are always candidates for cluster change, using nearest $D$-neighbors. It should be noted that the partition evolves and that the basic step consists of moving an element from one cluster to another. Therefore, this process generates a non monotonic sequence of partitions, which is not a hierarchical approach. This sequence of partitions evolves through element assignment modifications. The number of clusters decreases slowly along the iterations. A cluster disappears only when it is empty. The optimization scheme goes as follows:

1. *Compute matrix* $D$ of dissimilarities between elements;
2. *Compute dissimilarities* between each element and the current clusters using $D$ and a linkage function (the minimum for example);
3. *Select* a neighbor: a couple ($E$,$C$), an element $E$ candidate to move to a cluster $C$;
4. *Test* the new affectation gain for the disaggregated forecast associated to the resulting partition
   - *if* the error does not decrease *then*
     - *if* there are candidates *then* select the next one and *go to* step 4
     - *else end* (no improvement by moving an element from a cluster to another)
   - *if* the error decreases *then* modify partition and *go to* step 2.

This scheme can be adapted to parallel computations simply through a better organization of the candidates' examination in the more internal loop. Parallel capacities could also be used to explore multistart versions of the algorithm. However, these aspects are out of the scope of this paper which focuses on the question of the possible usefulness of disaggregation.

## 5. EXPERIMENTAL RESULTS

### 5.1. Performance results

Starting from 90 clusters, the optimized partition reaches the performances measured by long-term and short-term MAPE, given by Table 1.

**Table 1**: Performances of optimized partition starting from 90 clusters.

| | Aggregated | Disaggregated | Gain |
|---|---|---|---|
| MAPE long-term (LT) | 4.06% | 2.39% with 19 clusters | 41.13% |
| MAPE short-term (ST) | 2.47% | 1.51% with 28 clusters | 38.86% |

The procedure can be stopped at any step of the optimization process, therefore, improving the previous acceptable solution. The 195 step process with an error rate gain of 41%, is illustrated on Figure 3. This error reduction largely and obviously improves the optimization process, which starts with 90 clusters and ends with 19 clusters.

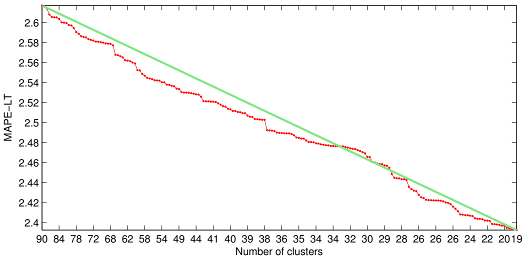

**Figure 3**: Optimization process: from 90 to 19 clusters leading to a gain of 41%.

### 5.1.1. About wavelet preprocessing

The first step of the global procedure (wavelet preprocessing and initial clustering using wavelets) is important. Indeed if one performs directly a hierarchical clustering of the original 2309 customers using the dissimilarity matrix $D$ and then optimizes the associated 90 clusters partition, the MAPE-LT error criterion stabilizes around 2.7% instead of 2.5%.

### 5.1.2. About the optimization step

The optimization step is also important. Indeed, starting from the 90 cluster partition, if one constructs the hierarchy of partitions (by hierarchical clustering using $D$), it is difficult to select a critical number of clusters (see Figure 4) and the MAPE-LT error criterion remains about 2.6% instead of 2.5%.

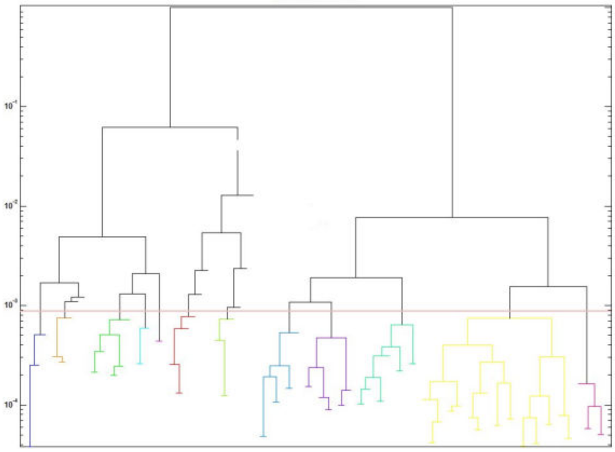

**Figure 4**: Dendrogram: hierarchical clustering using $D$.

## 5.2. About the number of basic customers

Finally, let us mention that the number of initial clusters (taken here to be equal to 90) is an important parameter, especially when the method is used for a significantly large number of customers. Indeed, the actual performance is slightly improved by increasing the number of clusters. The initial 2.39% performance on 90 clusters reaches 2.31% with 200 clusters and even 2.26% for 500 clusters, therefore increasing the reduction rate from 41.1 to 44.3%.

## 5.3. Clusters interpretation

In this paragraph, we will focus on the 19 clusters resulting from the final optimized partition. For example, Figure 5 presents cluster 1 made up of 10 super customers. It superimposes the 10 super customers consumptions with the average consumption of the cluster. The top graphic represents the year 2000 while the bottom graphic zooms in on the first quarter of that year (January, February and March 2000).

Let us note that the extreme regularity and homogeneity of the final 19 average cluster curves is remarkable. This can be explained by the fact that those curves are perfectly suited for forecast using the Eventail model. In other words, the optimization algorithm produces curves well adapted to Eventail black box forecast method.

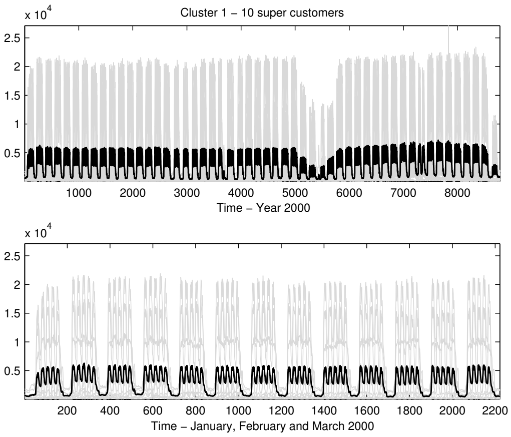

**Figure 5**: Cluster 1: average consumption (dark curve) and individual consumption of the super customers (light curves). Top: the year 2000. Bottom: zoom in on the first three months of 2000.

To get extra information on cluster 1, let us look at Figure 6 and its 11 graphics. It displays the 198 individual customer consumptions leading to the 10 super customers of the optimized partition. Each one of the ten first plots displays a super customer consumption together with the average consumption. The last plot displays the aggregated signal.

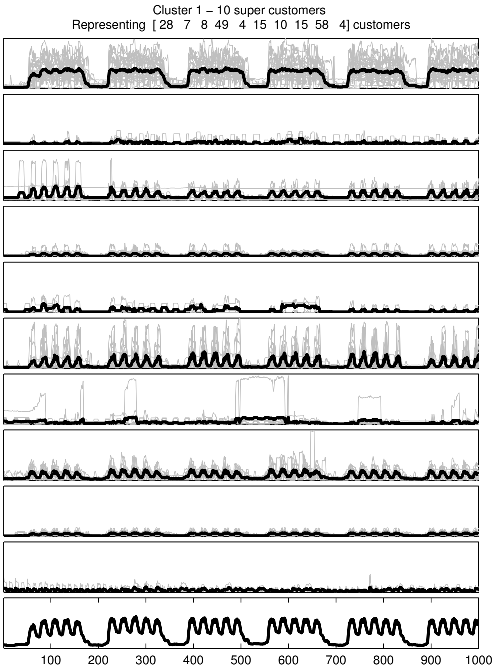

**Figure 6**: For cluster 1, ten top plots representing individual consumptions (light curves), leading to 10 super customers (dark curves). Bottom plot: total cluster consumption.

So, despite a great heterogeneity of the customers within each cluster, the signals associated with the final clusters are very stable and easy to predict using Eventail.

## 5.4. Validation results

The optimization procedure can be modeled using the three following steps:

1. Starting from the $N$ individual customers, performing a discrete wavelet transform DWT at a given level $j$ of the $N = 2309$ signals $S$, normalizing in *l1*-norm and clustering the resulting signals. As a result:

$$(K, P_K) = \mathit{MSC}(N, j; S) \ ,$$

which leads to $K$ super customers associated with a partition $P_K$. This multiscale clustering (MSC) step involves selecting level of decomposition $j$ (typically $j = 6$ or $j = 4$) as well as choosing $K$, the number of clusters (usually $K = 60, 90, 200, 500$).

2. Computation of $D(e, e')$ (for elements $e$ and $e'$) then computation of $D_c(e, c)$ (for an element $e$ and a cluster $c$) and the optimization leads to:

$$(k, \widetilde{P_k}) = \mathit{Opt}(K, P_K) \ .$$

3. Expansion of the $k$ clusters of the $K$ super customers over the $N$ initial individuals to produce:

$$(k, P_k) = \mathit{Exp}(k, \widetilde{P_k}, S) \ .$$

Suppose now that, three years of observations are available for a subset of customers. Then, we can select the parameters (choice of $K$ in particular) and estimate the quality according to the following validation principle. The first two years are used to design the clusters and the last year is used as the test sample.

Unfortunately, a subset of only 1482 customers is available during the three considered years. Quality estimations are as follows. On the learning set (year 2001), quality gain is about 23% since we go from 3.14% to 2.31%. On the test set (year 2002) there is an 8% reduction, from 6.82% to 6.32%. Nonetheless, the disaggregation procedure still provides significant gain. Of course this must be considered with caution since the year 2002 seems to be much more difficult to predict using Eventail: the MAPE forecast error of the aggregated predictor increases from 3.14% to 6.82% and is perhaps not representative.

## 6. FUTURE WORK

Let us briefly present some future possible developments.

First of all, alternatives to the current algorithm could be studied, the main difficulty being to cope with the computational burden. A scheme better suited for parallelism could be developed. A divisive strategy instead of data aggregation could be used to optimize the forecasting objective. It would start with the whole population and iteratively segment the current subgroup. Segmentation could be completed according to a 2 or 3-means clustering using approximation coefficients.

Another line of work on electrical data could be to further develop forecasting with wavelet methods (see Antoniadis *et al.* ([2]), Amin Ghafari, Poggi ([3])). The aim would be to adapt the models to the clusters using a similar method to the one described in Hathaway, Bezdek ([13]) or more recently clusterwise linear models proposed in Gruen, Leisch ([11]).

Also, we could take advantage of external meteorological and economical information as diagnostic and performance measurement tools. Eventually, the whole procedure should integrate parameters' data-driven choices: the wavelet and the representation basis, the obtained partition and the adaptation of the model to cluster specificities.

Last but not least, theoretically we could explore how to maximize the profits of the disaggregation method in general conditions.

## ACKNOWLEDGMENTS

This work is part of a scientific collaboration between EDF Clamart R&D, OSIRIS department and Orsay University. The authors thank EDF for the problem as well as the data. The authors also thank Sophie Lamarche and Alain Dessertaine for helpful discussions, and Chris Roth and Dominique Sorbier for careful reading. Lastly, we thank the anonymous referees for their useful suggestions and meaningful comments.

## REFERENCES

[1] Abraham, C.; Cornillon, P.; Matzner-Lober, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines, *Scand. J. Stat.*, **30**(3), 581–595.

[2] Antoniadis, A.; Paparoditis, E. and Sapatinas, T. (2006). A functional wavelet-kernel approach for time series prediction, *J. of the Royal Stat. Soc., Series B*, **68**, 837–857.

[3] Aminghafari, M. and Poggi, J.M. (2007). Forecasting time series using wavelets, *Int. Journ. of Wavelets, Multiresolution and Inf. Proc.*, **5**(5), 709–724.

[4] Biau, G.; Bunea, F. and Wegkamp, M. (2005). Functional classification in Hilbert Spaces, *IEEE Trans. Inf. Theory*, **51**(6), 2163–2172.

[5] Biau, G.; Devroye, L. and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces, *IEEE Trans. on Inf. Theory*, **54**, 781–790.

[6] Bruhns, A.; Deurveilher, G. and Roy, J.S. (2005). A non linear regression model for mid-term load forecasting and improvements in seasonality, *Proceedings of the 15th Power Systems Computation Conference 2005, Liege Belgium.*

[7] Calinski, R.B. and Harabasz, J. (1974). A dendrite method for cluster analysis, *Comm. Stat.*, **3**, 1–27.

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

[9] Espinoza, M.; Joye, C.; Belmans, R. and De Moor B. (2005). Short-term load forecasting, profile identification, and customer segmentation: a methodology based on periodic time series, *IEEE Transactions on Power Systems*, **20**(3), 1622–1630.

[10] Goia, A.; May, C. and Fusai, G. (2009). Functional clustering and linear regression for peak load forecasting, *Int. J. of Forecasting*, to appear.

[11] Gruen, B. and Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R, *Computational Statistics and Data Analysis*, **51**(11), 5247–5252.

[12] Hastie, T.; Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer.

[13] Hathaway, R.J. and Bezdek, J.C. (1993). Switching regression models and fuzzy clustering, *IEEE Trans. Fuzzy Systems*, **1**(3), 195–203.

[14] Hippert, H.S.; Pedreira, C.E. and Souza, R.C. (2001). Neural networks for short-term load forecasting: A review and evaluation, *IEEE Transactions on Power Systems*, **16**(1), 44–55.

[15] James, G. and Sugar, C. (2003). Clustering for sparsely sampled functional data, *JASA*, **98**, 397–408.

[16] Kaufman, L. and Rousseeuw, P. (2005). *Finding groups in Data, an introduction to Cluster Analysis*, Wiley.

[17] Mallat, S. (1998). *A Wavelet Tour of Signal Processing*, Academic Press.

[18] Misiti, M.; Misiti, Y.; Oppenheim, G. and Poggi, J.-M. (2007). *Wavelets and Their Applications*, Hermes Lavoisier, ISTE Publishing Knowledge.

[19] Misiti, M.; Misiti, Y.; Oppenheim, G. and Poggi, J.-M. (2007). Clustering signals using wavelets, F. Sandoval *et al.* (Eds.): *IWANN 2007*, Lecture Notes in Computer Science, 4507, 514–521, Springer.

[20] Piao, M.; Lee, H.G.; Park, J.H. and Ryu, K.H. (2008). Application of Classification Methods for Forecasting Mid-Term Power Load Patterns in Advanced Intelligent Computing Theories and Applications, D.-S. Huang *et al.* (Eds.): *ICIC 2008, CCIS 15*, Springer, 47–54.

[21] Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*, Springer.

[22] Taylor, J.W. and Espasa, A. (2008). Editorial Energy forecasting, *Int. J. of Forecasting*, **16**, 561–565.

[23] Tibshirani, R.; Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic, *J. of the Royal Stat. Soc., Series B*, **63**(2), 411–423.