QÜESTIIÓ, vol. 25, 3, p. 509-520, 2001

PRACTICAL DATA MINING IN A LARGE UTILITY COMPANY*

GEORGES HÉBRAIL*

We present in this paper the main applications of data mining techniques at Electricité de France, the French national electric power company. This includes electric load curve analysis and prediction of customer characteristics. Closely related with data mining techniques are data warehouse management problems: we show that statistical methods can be used to help to manage data consistency and to provide accurate reports even when missing data are present.

Keywords: Data mining, statistical methods, data warehouses, applications, load curve analysis, utility company

AMS Classification (MSC 2000): 62-07, 62P30, 62M45, 62M10, 62H30

* This is a reprinted version of the article published in Compstat 2000 Proceedings (ISBN 3-7908-1326-5).

* Electricite de France, R&D Division. 1, Av. du Général de Gaulle. 92141 Clamart, France.

– Received May 2001.
– Accepted November 2001.

1. INTRODUCTION AND SHORT PRESENTATION OF EDF

Electricité de France (EDF) is the French national electric power company which is in charge of the generation, transmission and distribution of electric power in France. In 1999, the electric power market was deregulated in France so that now EDF does not have anymore the monopoly on the market of large customers. In parallel with the deregulation, EDF develops its activity in other countries in Europe and in the world. EDF is a large company, gathering nearly 115 000 employees with around 30 million customers in France.

Acting in a deregulated environment is quite different from the previous situation where EDF had a monopolistic position on the French electric power market. The company currently develop new tools based on data mining techniques to be able to face this competition. Actually, EDF developed in the past many tools in order to optimize its activity and reduce the cost of generation, transmission and distribution of electric power: many of these tools designed to solve technical problems are adapted to solve commercial problems.

In this paper, we present the main developments of data mining techniques at EDF in the area of customer relationship management. Data mining techniques are also used in other domains, like power plant maintenance, electrical network operation, human resource management, but are not described here. In the paper, we will not make a special distinction between the terms «*data mining*» and «*statistical*» methods since in our opinion most of efficient data mining techniques are based on well-known statistical methods, the main difference being related to their scalability to large volumes of data.

In Section 2, we present the development of data mining techniques for dealing with electric power load curves of individual customers, i.e. curves representing their electric power consumption. In particular, we present a software developed in our Division which performs clustering of such curves interactively: this software is based on the Self Organizing Map (SOM) approach (see Kohonen (1990)). In this section, we also present a new approach we are studying to analyze long time series associated with the consumption of one customer: pattern extraction techniques are applied to build a symbolic description of the time series.

Section 3 is devoted to the presentation of data mining techniques applied to predict missing data in customer databases. Some fields in customer databases are partially filled, like for instance the possible use of electric power for water heating. Logistic regression methods are applied to predict missing information, from customers with non missing data.

In Section 4, applications related to interactions between statistical methods and data warehouses are described. We show that standard statistical methods can be used for two very useful goals: detection and characterization of records which fail consistency

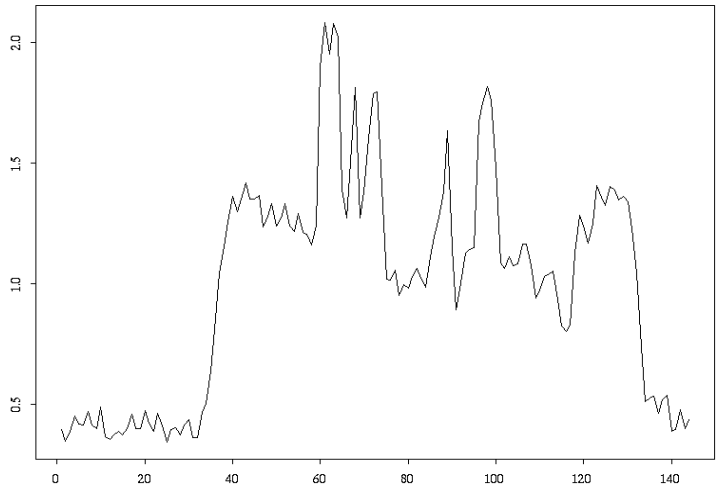


Figure 1. Example of a daily load curve.

checking of the data warehouse, and production of accurate reports even when data is missing, by statistical adjustment.

2. ELECTRIC LOAD CURVE ANALYSIS

2.1. Typical available data

Electric power sales for a customer are represented by a time series (also called a *load curve*) figuring the electric power consumption for every period of time. Availability of such data deeply depends on the type of customer. General public small customers (like residential ones) are poorly described since a communicating meter is too expensive regarding to their consumption: for these customers there are only a few points of the curve every year.

For larger customers, a communicating meter is often available for many reasons: the billing is done every month, the consumption is high and justifies the communicating meter investment, a detailed record of consumption is necessary because prices depend on the period. For these customers, a load curve is available figuring points every 10 minutes. Figure 1 shows an example of such a curve for a period of one day. The curve is generally available all over the year. In the rest of this section, we describe applications when such data is available.

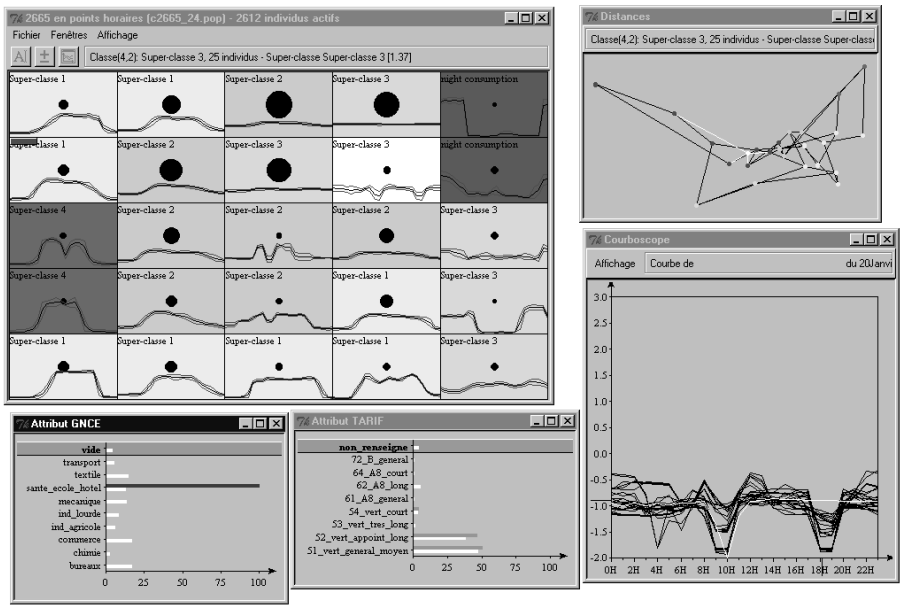


Figure 2. Main screen of the Courboscope software.

2.2. Goals of load curve analysis

Load curves have been analyzed at EDF for more than 20 years, with different goals of study. Analyzed curves may correspond to individual customer curves or to aggregates over a geographical sector. Applications may be either understanding the behavior of customers in relation to some external characteristics, or predicting consumption in a short or long term perspective.

In this paper, we focus on the exploratory data analysis of a set of load curves, each curve being associated with a single customer. The goal is to understand customer consumption behavior all along the year, in relation to his usage of electric power, to the pricing policy, and to the characteristics of the customer (mainly its activity).

2.3. The «Courboscope» software

The «Courboscope» (*courbe* is the French word for curve) is a software which has been developed a few years ago in our R&D Division. It is described in detail in Debrégeas and Hébrail (1998). This software is a tool for clustering sets of curves.

Curves are here assumed to be described by the same number of points (for instance 24 points for a daily hourly measured curve). Each curve is thus considered as a point in a p-dimensional euclidian space (3^{24} for daily curves). The euclidian distance is used as a measure of dissimilarity between curves. If this distance is not accurate, a preliminary process normalizes the curves so that the euclidian distance becomes appropriate.

Curves are also described by additional attributes, such the name of the customer, his activity, the day in the week, the month, the season, the pricing option, ...

The user's task associated with this software is to build clusters of curves, where curves belonging to the same clusters have similar shapes. The clustering method is the SOM (Self-Organized Map) method proposed by Kohonen, see Kohonen (1990). This method, based on a neural network approach, produces clusters organized in a spatial way so that clusters which are close in the map have similar global shapes (see the upper left window in Figure 2). This special arrangement of clusters is quite interesting because it allows to perform clustering with a large number of clusters (for instance $10 \times 10 = 100$ clusters): the result is still readable.

Beyond construction of clusters, the software helps the user to give an interpretation to each cluster. A sophisticated and easy to use interface enables the user both to examine details of each cluster (lower right window in Figure 2), and to characterize each cluster with external attributes (lower left windows).

This software is used by load curve analysts who bring out typical shapes of curves (usually on a basis of daily or weekly curves), associated with some external information, typically a tariff, an activity, or an electric power usage. The result of this process is what we call an *interpreted map*, where a label is associated with each cluster cell.

Interpreted maps can be used by other (operational) people through another module of the software which is designed to classify new curves into the interpreted map. Main applications here are default detection and pricing advice. Two kinds of expertise are thus mixed together: the analyst expertise by the means of the interpreted map, and the expertise of people on the field through this classification module, which enables them to check if the use of the euclidian distance does not get completely crazy.

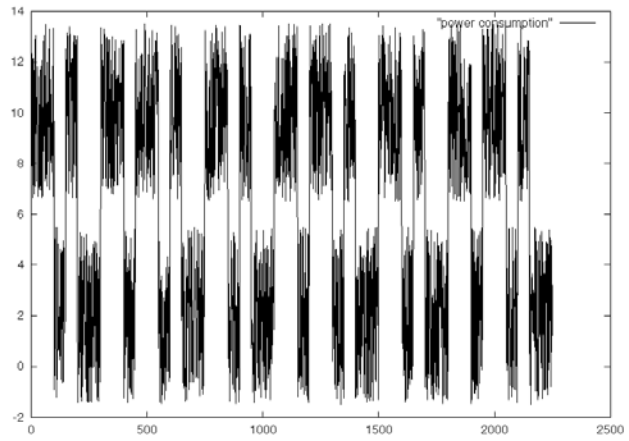
The Courboscope software is not dedicated to electric power load curves but is generic. It can be used with any type of curves described by any additional attributes.

2.4. Towards a symbolic analysis of curves

The approach described in the previous section shows several restrictions:

– all curves have to be of the same length and described by points at the same time positions,

Initial one year long curve
(only 2300 of the 8760 are represented below)



Symbolic representation of the one year long curve

Typical curves associated with symbols

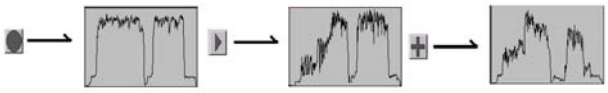


Figure 3. Example of a symbolic representation of curves.

– long curves (for instance time series describing one month or one year) cannot be treated because the euclidian distance in a 8000 dimension space is not meaningful,
– multi-dimensional curves (i.e. several distinct values available for every time tick) cannot be processed easily.

We are currently working on another approach where long curves (i.e. time series) are transformed into a symbolic form (see Hugueney and Hébrail (1999)). The basic process is the following:

– selection of a window width (manually or guided by a Fourier transform),
– generation of a set of fixed length curves obtained by sliding a fixed length window all along the long curve,
– clustering of the fixed length curves with the Courboscope, in order to obtain clusters of fixed length curves,

– transformation of the long curve into a succession of symbols, each symbol being associated with a cluster of fixed length curves.

Figure 3 shows an example where a curve of one year long (8760 hourly points) is replaced by a succession of 26 from a set of 3 different symbols. Here, each symbol represents a two week typical curve.

Once a long curve has been transformed into such a symbolic form, several methods can be used, in an exploratory perspective:

– visualization of the symbolic representation: this provides a very synthetic view of the long curve,
– application of data mining methods such as sequence analysis, in order to discover frequent subsequences (see Agrawal and Srikant (1995)),
– application of multivariate data analysis methods in the case of multi-dimensional curves.

Manual editing of the symbolic representation is also very interesting. Editing can be done at two different levels:

– edition of the sequence of symbols, the application here is to simulate a change in the customer's activity schedule,
– edition of the curves associated with symbols, the application here is to simulate a change in the customer's equipment.

Once curves have been edited at the symbolic level, a simple process can reconstitute a predicted curve which reflects changes done at the symbolic level. The real-world application here is to simulate changes in customer behavior in relation to pricing options.

3. PREDICTION OF CUSTOMER CHARACTERISTICS

In this section, we describe data mining applications which are not specific to electric power distribution. These applications are related to the marketing activity of electric power sales within general public customers.

There are two main ideas underlying these applications:

– information which is necessary for a good marketing activity cannot be obtained for all general public customers: the cost is prohibitive,
– customer databases contain enough data to infer missing information using statistical methods.

We only do here a short presentation of these applications since they are quite standard in the data mining field.

3.1. Brief overview of customer databases

EDF sells electric power to nearly 30 million general public customers in France. There are around 100 local customer databases, each database describing an average of 300 000 customers of the same geographical area. These databases have been designed primarily for billing purposes: they contain information about the contract, the consumption and the premises where power is delivered.

Some years ago, these databases have been enriched by additional fields to improve customer relationship and to allow marketing actions to be more efficient. This fields are filled out with the stream of contacts between the company and its customers. Consequently, they are currently partially filled.

In parallel, national surveys are carried out in order to acquire more information on a small number of customers. These surveys also contain a copy of information available in our databases for the corresponding customers (identity of each customer is lost for ethical reasons).

3.2. Prediction of customer characteristics

Data mining techniques are used to predict some customer characteristics when they are missing. The basic method we use is logistic regression.

Two different problems are addressed:

– application of logistic regression within a local database: partially filled fields define the prediction variables and the learning sample gathers customers with non missing values.
– application of logistic regression to predict values of fields available in surveys but not in customer databases. Values are predicted for all customers in each local database.

3.2.1. Prediction within local databases

As written before, local databases describe an average of 300 000 general public customers. Some fields describing them are partially filled because these fields have been introduced in the database only some years ago. Depending on the database, these fields are filled out for 10% to 50% of the customers. This is usually enough to build a prediction model from a sample gathering the customers with non missing values.

Several fields —like for instance the presence/absence of electric heating of water— have been predicted successfully with error rates from 10 to 30%, depending on the field and on the local database. This precision is enough to feed marketing campaigns.

It is important to note that one model is built for each local database, in order to take into account specific local characteristics of the populations. A special attention has to be given to the fact that in this case the sample is biased. A reweighting is necessary and is based on a stratification of the local database.

3.2.2. Projection of national survey data to local customer databases

The other application is prediction (within each local customer database) of fields which are available only in national surveys. Survey data contain both answers to questionnaires and customer data extracted from our databases: this enables to build prediction models applicable to local customer databases.

In this case, the sample is not made of customers from the local database on which the prediction is performed. Actually, customers of national surveys are picked up randomly from all local databases. So, there is only a very small number of customers from each local database in the survey. This number is not sufficient to build a model for each database. A model is built from the whole survey and applied to the different local databases.

We have done preliminary experiments which show that prediction is possible but with an error rate which is still too high for our needs. We are currently running more experiments on this problem to improve the method.

4. STATISTICS AND DATA WAREHOUSES

In this section, we examine the joint use of statistical techniques and data warehouse facilities. We are currently working in two directions:

– quality of data in *relational* data warehouses,
– estimation of missing data in *multi-dimensional* data warehouses.

All applications described in this section correspond to very basic statistical methods but are really useful. The only difficulty is that the volume of data may become very large: algorithms must be scalable.

4.1. Quality of data warehouses

In this section, we focus on data warehouses where the database model is the *relational* model, i.e. the database is a collection of standard data tables.

First, simple statistics are systematically computed on every table of the data warehouse in order to give a diagnosis of the quality of data. The following basic statistics appear to be very useful to detect errors: number of values, number of different values, and number of missing values for every column. Barcharts, boxplots and histograms on some specified variables (like the contract type or the power consumption) are also of good help, especially to detect outliers which often correspond to errors in the database.

Secondly, we carry out a project addressing the problem of detecting errors or inconsistencies in relational data warehouses. This project is based on a database point of view: a set of constraints which should be satisfied by the database is defined. A program generates for each table of the database the set of records which violate the constraints. Here, statistical methods (see Morineau (1984) for instance) are used to characterize error records against others, using all columns of the table. This helps the user to understand the reason of errors: for instance, it may reveal that records in error correspond to bills of October or to customers of some geographical area.

4.2. Estimation of missing data in data warehouses

In Section 3.2.1, we have seen that logistic regression is used to predict values for customers showing missing values on some partially filled fields. This is prediction of missing data at the detailed level. In this section, we address the problem of estimation of aggregated data from a database containing missing values at the detailed level.

Data warehouses now offer a *multi-dimensional* model of data in addition to the relational model. This new model allows to define data structures which are matrices with several dimensions. In real-world applications, usual dimensions are *Product*, *Store*, and *Time*. For each product, store, and date, data usually consist in sales defined by several numerical values such as the number of sold items, the amount of the transaction, ... For each dimension, hierarchies are defined in order to build syntheses of data, for instance: products are grouped into types of products, brands, or store departments; stores are grouped into geographical areas and regions; time is aggregated at the day, month, quarter, and year levels.

Once sales data have been collected from stores, the user can use the dimension hierarchies to build interactively any aggregated array of data: for instance the total sales of food by region and month for year 1999. The database system ensures that the response time is good even if the volume of data is huge (millions of transactions for a large corporate department store). This facility assumes that all data are available, i.e. there are no missing values in the detailed data.

Within our local databases, we are currently building such a multi-dimensional database centered on electric power sales. This database is intended to be used by marketing people in order to have an idea of the sales by geographical area, time period, and

type of customer. As mentioned in Section 3, there are many missing values in the customer characteristics which are not necessary for billing activity. This is the case for information about usages of electric power (heating, water-heating, type of appliances, ...). At the detailed level of data, prediction can be done with a logistic regression. At aggregated levels, prediction of aggregate values can also be done with statistical methods but the approach is different: it is related to sample adjustment to take into account the fact that non missing values constitute a biased sample.

We are currently working on the problem of adding statistical inference capabilities to multi-dimensional databases, in order to be able to query the database as if there were no missing values. An error has to be defined and evaluated for each query, depending on the level of aggregation of the query.

5. CONCLUSION

In this paper, we have presented several applications of data mining and statistical methods in the context of a large utility company.

Some of them are rather specific to electric power distribution, or by extension to utilities. The methods in this case are related to the analysis of load curves of large customers. The Courboscope software allows to analyze short period curves (day or week) to understand customer behavior. Symbolic curve analysis is intended to analyze longer curves (one year for instance): it is an active research area for us.

There are also other applications of statistical methods related to load curves, in order to predict the consumption demand in the future. These applications were not described in this paper and are based on ARIMA and neural network models (see for instance Mangeas and Muller (1997)).

Other applications presented in the paper are related to general public customers. Better knowledge of these customers is achieved by analysis of customer databases and national surveys. Here statistical methods allow prediction of customer characteristics either at a detailed or aggregated level. Work on projecting survey data on customer databases is an active research area for us.

6. REFERENCES

Agrawal, R. & Srikant, R. (1995). «Mining Sequential Patterns, *Proceedings of 11 th Int'l Conf. on Data Engineering (DE'95)*, Taipei, Taiwan.

Debrégeas, A. & Hébrail, G. (1998). «Interactive Interpretation of Kohonen Maps Applied To Curves», in *KDD'98, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, New-York, 179-183, AAAI Press.

Hugueney, B. & Hébrail, G. (1999). «Construction de descriptions symboliques de courbes numériques». *EDF R&D Division internal report*, n° HI-23/99-025.

Kohonen, T. (1995). *Self-organizing maps*, Berlin: Springer.

Mangeas, M. & Muller, C. (1997). «An automatic search of feed forward neural network architecture based on genetic algorithms: application to the short term load forecasting». In *ISAP'97, Proceedings of the International Conference on Intelligent System Applications for Power Systems*, Corea.

Morineau, A. (1984). «Note sur la Caractérisation Statistique d'une Classe et les Valeurs-test». *Bulletin du CESIA*, vol. 2, n° 1-2, Paris.