The "profilograph": a toolbox for the analysis and the segmentation of gas load curves

Patrick Letrémy
SAMOS-MATISSE, University Paris 1 and CNRS
Paris, France
pley@univ-paris1.fr

Eric Esposito, Valérie Laffite, Sally Showk
Research and Development Division
Gaz de France
eric.esposito@gazdefrance.com
valerie.laffite@gazdefrance.com
sally.showk@gazdefrance.com

Marie Cottrell
SAMOS-MATISSE, University Paris 1 and CNRS
Paris, France
cottrell@univ-paris1.fr

Abstract - *The paper presents a method to allocate a class of a Kohonen map to a new customer without knowing anything about the variables used for the classification. In this study, a classification of daily gas load profiles is performed on a panel of Gaz de france customers. Then, we use a multinomial logit model to allocate a class to a new customer according to its additional variables. With this model, it is also possible to infer the probability to belong to each of the Kohonne classes. The main application for Gaz de France is the inference of the daily gas load for any customer.*

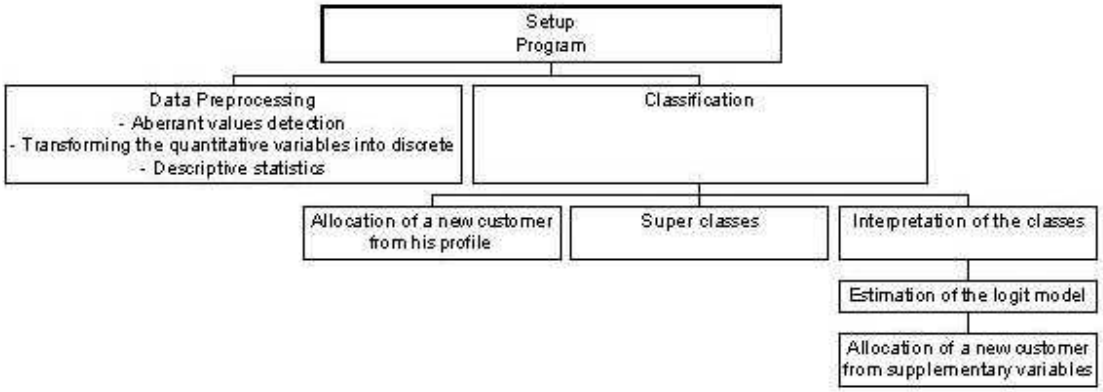
Key words - Kohonen Maps, Load profiles, Logistic regression, Non Ordered Polychotomous Logit Model

1 Context

1.1 The load curves: an important information source

The gas consumption curves (so-called load curves) of the customers make up the main information source about the French Gas Company (Gaz de France). Their study allows us to know the customers' behavior, to anticipate the demand and therefore not only to better manage the supplying but also to achieve tariff optimization.
Taking the large number of customers into account (more than 10 millions by now), it is very difficult to separately analyze the load curves and to synthesize the information in a way which could be easily understood and exploited by all the managers of the Company. In

Figure 1: Structure of the toolbox



collaboration with University Paris 1 and laboratory SAMOS-MATISSE, the Gaz de France Research Division has developed a toolbox which automatically achieves an analysis and a segmentation of the set of load curves. This toolbox is called "Profilograph".

In this operational frame, the toolbox objective is double:

- First, to identify and analyze the different behaviors in term of consumption of the gas customers;
- Secondly, to use the typical behaviors to forecast tomorrow the behavior of any new customer.

1.2 A toolbox adapted to the needs

The "profilograph" is a toolbox which appears to the user as a sequence of command windows which are interactive and modifiable. It automatically ensures the interactions between the modules which achieve an exhaustive analysis (pre-processing, classification, allocation) of the data and provides user-friendly outputs, and especially classification maps of load curves.

In the following, we present the different functionalities of this toolbox. These ones have been implemented and parameterized with the convenient specifications in order to adapt itself to the needs of Gaz de France. First we will specify which data are to be used as inputs for the toolbox, as well as the pre-processing that the user can choose to do.

Section 2 will be devoted to the classification of the customers according to their load curves as well as to the interpretation of the classes.

At last in section 3, we deal with the allocation module, which enables classifying any additional customer into one of the classes which have been defined.

The diagram in figure 1 resumes the different modules which are developed in the toolbox and their chaining up.

2 Data and preprocessing

2.1 Different types of analyzed data

The load curves which are studied with the "Profilograph" are real-valued quantitative data of consumption which have been read at time intervals which can be variable (daily, monthly). Their volume can be very important, with respect to the number of lines (number of customers) as well as to the number of columns (number of readings). Moreover the load curves can have missing values.
The toolbox also uses so-called supplementary (or additional) data, different from the consumption ones, which can be qualitative or quantitative. That ones give other information about each customer: his characteristics (geographical and climatic zones, nature of activity), his contract (yearly level, tariff, options, etc.). They are not used during the classification stage, but provide enrichment to the interpretation of the classes which are built.

2.2 Data preprocessing

The toolbox allows managing all the mentioned kinds of variables. However, some of the modules (interpretation and allocation using additional variables) use qualitative variables as inputs. So a module is designed to transform the quantitative variables into discrete ones according to some percentiles chosen by the user.
Moreover, for global analysis, the toolbox contains a module for descriptive statistics (frequency tables for qualitative variables and elementary statistics for the quantitative ones).
At last, the aberrant values detection module allows finding, list and replacing the erroneous values of the continuous variables of consumption by missing values. To do that, it uses a threshold which can be parameterized by the user.

3 Classification and interpretation of the classes

3.1 Classification of the load curves

The load curves, or consumption profiles, are classified in a non supervised way according to the Kohonen algorithm [3], [5], [2].
The existence of missing values in consumption curves is managed by the toolbox. For that, the Kohonen algorithm is used over all the load curves, including those which have missing values. Those curves are classified according to the components which are not missing.
The segmentation aims to group customers who have the same consumption profile, and therefore the same load curve, whatever the consumption level can be. For that, this module proposes to normalize the data by dividing each consumption data by the customer consumption mean value over the studied period.
The classification outputs are displayed on one- or two-dimensional maps. These maps give a glimpse of the set of all the classes and of all the code vectors which are considered as the representative of the classes.
The outputs are diverse:

- SAS tables: class of each customer, code vectors of the classes, number of customers in each class, etc.

Figure 2: Classification of the consomption customers

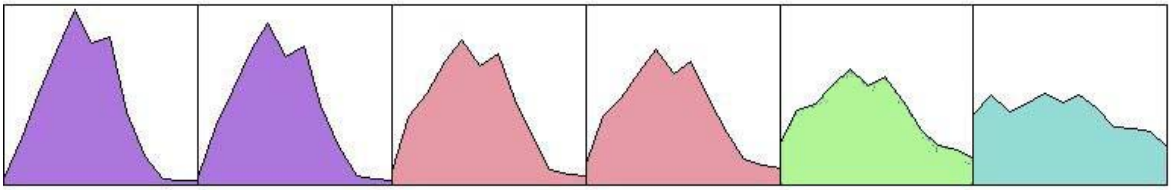
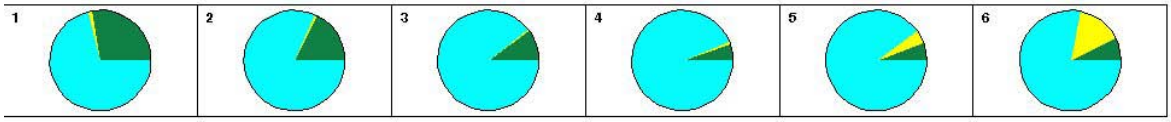


Figure 3: Distribution of the tariff over the 6 classes



- Visualization: Kohonen map, code vectors and confidence regions (around the code vector), class mean values, etc.

The segmentation can be refined by grouping the Kohonen classes which have close code vectors into "super classes". The super classes are defined by a Hierarchical Ascending Classification (HAC) achieved on the code vectors. Their number is fixed by the user, who has at his disposal the dendrogram of the HAC.

3.2 Interpretation of the classes

This module is used after the load curves have been classified in order to better explain the obtained classes. The segmentation is crossed with the additional variables. The interpretation program projects the supplementary qualitative variables on the Kohonen classes. It computes and represents their distribution for each class in the form of frequency pies. Therefore, for each class represented by its code vector, it is possible to make obvious its characteristic features.
Figure 2 shows a one-dimensional Kohonen map (with 6 units, 1700 iterations and 4 [1] "super classes" which classifies a database containing one year of monthly consumptions (12 values from September to August) with 3122 customers and figure 3 shows the distribution of an additional explanatory variable (the kind of tariff). One can see that the profiles are ordered along the map starting from the more depending on the climate (season-sensitive classes 1 and 2) to the less depending one (class 6). The construction of 4 super-classes supplies a regrouping of classes 1 and 2, and of classes 3 and 4. These groupings gather very similar code vectors. The distribution of the tariff (with 3 levels) shows for example that the proportion of "green" tariff decreases for lightly season-sensitive classes and increases for strongly season-sensitive ones, the contrary being true for the "yellow" tariff.

4 Allocation of additional customers to the Kohonen classes

Once the classification stage is achieved, it is necessary to classify new customers, who do not belong to the learning set in one of the Kohonen classes.

[1] We choose 4 superclasses according to the classical choice by Gaz de France

In concrete term, two cases can occur:

- The consomption profile of a new customer is known
- The consomption profile is unknown. The allocation is achieved thanks to supplementary variables that are known for the customer.

4.1 Allocation when the consumption profile of the new customer is known

Let us consider the first case. The profile is known but can possibly contain missing values for the observed period. Let us denote $y_i = (y_{i1}, y_{i2}, \ldots, y_{iN})$ the consumption profile of customer i, and E_i the set of non missing components of y_i.
We have to compare the consumption profile of this new customer to each code vector provided by the Kohonen algorithm. This comparison uses the restricted Euclidean distance defined by:

$$d_{ik} = \sqrt{\sum_{j \in E_i} (y_{ij} - w_{kj})^2}.$$

where

- d_{ik} is the distance between the consumption profile of customer i and the code vector of class k,
- w_{kj} is the j-th component of the code vector of class k.

The membership probability p_{ik} for customer i to belong to class k can be computed by

$$p_{ik} = \frac{\exp(-d_{ik})}{\sum_{j=1}^{K} \exp(-d_{ij})}$$

where d_{ij} is the distance between the profile of customer i and the code vector of class j and K is the total number of classes.
The customer i will be classified into the Kohonen class which has the nearest code vector (for the previously defined distance), that is in class k where

$$k = \arg\max p_{ij} = \arg\min d_{ij}.$$

4.2 Allocation when the new customer profile is not known

Here it is necessary to classify a new customer into one of the Kohonen classes, without using his profile, since it is not known. This case is the more frequent in practice, since most of the customers' gas-meter is read only once or twice a year.
Let us assume that for all the customers (as well those who belong to the learning database as the new ones), we have at disposal p supplementary variables as mentioned in section 2.1. Individual i is described a vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.
The customers used in the learning database have been split out according to their consumption profiles into K classes, by using a Kohonen algorithm.
Then, one has to estimate the probability for a new customer to belong to a class only from the supplementary variables. The chosen model is a non ordered polytchotomous logit model, since the variable to explain (membership probability to a Kohonen class) has more than two non ordered modalities (there are more than two classes !).

4.2.1 Definition of the non ordered polychotomous logit model

Using the non ordered polychotomous logit model as discriminating tool has been proposed by Schmidt and Strauss, [6]. It is an extension of the binary logit model, which is often used in the studies of appetence or attrition. The difference comes from the fact that the variable to explain can take more than two modalities, and overall that these modalities are not naturally ordered, contrarily to the case of a score for example, where the order of the values is meaningful. However, the model uses the same theoretical frame, since it is estimated by using the maximum likelihood principle, [4], [7].

One has to choose a class as a reference class, suppose that it is the class K. Then the non ordered polychotomous logit model is written as

$$\frac{P(k/x)}{P(K/x)} = \exp(x \cdot \beta_k)$$

for $k = 1, 2, \ldots, K-1$, where the $\beta_k \in R^p$ are the model parameters and $P(k/x)$ is the probability that an individual belongs to class k, given it is described by x.

4.2.2 Estimation of the non ordered polychotomous logit model from the learning database

Procedure CATMOD of SAS software is designed to estimate this kind of model by using the maximum likelihood principle, [1]. As class K is arbitrarily defined as a reference class, procedure CATMOD is used to estimate the $K-1$ values for each observed x:

$$\frac{P(1/x)}{P(K/x)}, \frac{P(2/x)}{P(K/x)}, \ldots, \frac{P(K-1/x)}{P(K/x)}$$

In the following, the notations are simplified by writing p_k instead of $\frac{P(k/x)}{P(K/x)}$.

The explanatory variables can be quantitative or qualitative. Procedure CATMOD can use continuous values, by using the *direct*. However, in practice, using continuous quantitative variables can involve numerical errors, when they take too many values. So the continuous variables are discretized into a finite number of values before introducing in the model. So the number of possible instances of vector x is finite.
The procedure is applied to the learning database, which contains all the variables, the measures of consumption and the supplementary variables. A new column is added to the table, with the number of the Kohonen class of each customer.
For each possible x, procedure CATMOD provides the estimates of parameters $\beta_k \in R^p, k = 1, \ldots, K-1$ and after that, we have to solve the system (1):

$$\begin{cases} \frac{p_1}{p_K} = \alpha_1 \\ \frac{p_2}{p_K} = \alpha_2 \\ \vdots \\ \frac{p_{K-1}}{p_K} = \alpha_{K-1} \end{cases}$$

where $\alpha_k = \exp(x\beta_k)$ for $k = 1, \ldots, K-1$.

This system has K unknown parameters but can be solved since the probabilities satisfy equation (2):

$$p_1 + p_2 + \ldots + p_K = 1.$$

System (1) can be written

$$\left\{ \begin{array}{lcl} (1+\alpha_1)p_1 + \alpha_1 p_2 + \ldots + \alpha_1 p_{K-1} & = & \alpha_1 \\ \alpha_2 p_1 + (1+\alpha_2)p_2 + \ldots + \alpha_2 p_{K-1} & = & \alpha_2 \\ \vdots & & \\ \alpha_{K-1} p_1 + \alpha_{K-1} p_2 + \ldots + (1+\alpha_{K-1} p_{K-1} & = & \alpha_{K-1} \end{array} \right.$$

With matrix notations, this system is written $MP = A$, with

$$M = \begin{pmatrix} 1+\alpha_1 & \alpha_1 & \ldots & \alpha_1 \\ \alpha_2 & 1+\alpha_2 & \ldots & \alpha_2 \\ \vdots & & & \\ \alpha_{K-1} & \alpha_{K-1} & \ldots & 1+\alpha_{K-1} \end{pmatrix}$$

and

$$P = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{K-1} \end{pmatrix}, A = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{K-1} \end{pmatrix}$$

By computing the inverse of matrix M, (using the function GINV which deals with singular and non singular matrices), we compute the probabilities $p_1, p_2, \ldots, p_{K-1}$ as a function of x and deduce p_K from equation (2).

4.2.3 Application to new customer data

For each new customer i described by $x_i = (x_{il}), l = 1, \ldots, p$ and for each class k, one computes the probability that this customer belongs to class k. The new customer is allocated to the class for which the probability is maximum.

5 Conclusion

The possible applications of this toolbox are numerous in the Gaz de France Company. Nowadays, it is mainly used to identify the consumption behaviors of the portfolio of the customers of the Company. It allows building a statistical segmentation of customers according to their profiles, which fill up and enrich the segmentations which were done until now in an internal way by "experts". It is an interactive toolbox for analysis from preprocessing stage to final segmentation. The main originality of this toolbox lies in its allocation module which allows identifying the annual consumption profile of every new customer. Then it is possible to estimate the daily consumption, by taking the corresponding values of the code vector of its class as a prediction.

References

[1] P.Allison (1999), *Logistic Regression Using The SAS System. Theory and Application*, Cary, NC, SAS Institute Inc.

[2] M.Cottrell, S.Ibbou, P.Letrémy, P.Rousset (2003) Cartes auto-organisées pour l'analyse exploratoire de données et la visualisation, Journal de la Société Française de Statistique, **tome 144, no 4**, p. 67-106.

[3] T.Kohonen (1995) *Self-Organizing Maps*, Springer Series in Information Sciences, **Vol 30**, Springer.

[4] G.Maddala (1983), *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press.

[5] E.Oja, S.Kaski (1999), *Kohonen Maps*, Elsevier.

[6] P. Schmidt, R.P.Strauss (1975), The Prediction of Occupation Using Multiple Logit Models *International Economic Review*, **vol. 16, no 2.**

[7] K.Train (1986), *Qualitative Choice Analysis*, The MIT Press.