Published in proceedings of Entreprise Research Innovation (ENTRENOVA) 2016, Rovinj, Croatia, 8-9 September 2016, which should be cited to refer to this work.

Multi-method approach to compare the socio-demographic typology of residents and clusters of electricity load curves in a Swiss sustainable neighbourhood

Francesco Cimmino, Joelle Mastelic, Stephane Genoud, University of Applied Science Western Switzerland, Entrepreneurship & Management Institute. Switzerland

Abstract

A sustainable neighbourhood was built Switzerland by one of the leaders in this field. Half of the 400 apartments have been equipped with smart meters delivering big data on energy consumption (electricity, water, heating…). The company would like to know if it is possible to link socio-demographic typology of residents with energy consumption patterns. To answer this question we present in this article a multi-method approach combining qualitative analysis, frequently used in marketing (multiple correspondence analyses), and quantitative analysis from applied statistics to answer this question. First, we have conducted a survey among the residents of the sustainable neighbourhood to gather socio-demographic data, and then we have proposed a marketing typology of residents. In parallel, we have analysed load curves with statistical models (clustering factors, hermano beta models, coincidence factors, som, expert practice) to see if there are patterns of energy consumption and to determine groups of similar load curves. Then we have compared the discrepancies in the composition of the groups between both methods. This study is based on a single case study generating a new research hypothesis: the typology of residents based on socio-demographic data can be linked to energy consumption pattern of a household.

Keywords: applied statistics, typology, energy, cluster, sustainable consumption

JEL classification: Mathematical economics

JEL classification of the paper should be submitted according to the classification scheme available at the link http://www.aeaweb.org/journal/jel_class_system.php

Acknowledgments (if any):

Introduction

Research on energy conservation in the residential area is often technology driven: for instance what are the new isolation materials to be able used to save energy in buildings? In this paper we focused our research mainly on the demand side and would like to better understand the role of the residents on the energy consumption. In particular, we would like to know who are the residents under study (socio-demographic typology). We also would like to analyse the energy consumption of these residents (electricity load curves) to see if we can detect patterns of consumption. Then, the goal is to compare the different classifications: on one side the typology of residents and on the other side the groups of electricity load curves. Is there a link between both classifications?

So our research question is: "What would be the best methodology to be able to compare both types of classifications (residents and energy consumption)?"

To be able to compare both classifications, we need different types of data: (1) socio-demographic data of residents and (2) energy consumption data. Due to privacy rules, it is not always easy to work on both data sets and to link them. Utilities for instance, collected through smart meters, hold big data on electricity load curves. However we don't know the consumers except from their names and postal addresses. On the other side, cities know who live in a particular neighbourhood but do not have data on energy consumption of the residents.

To be able to get both types of data sets, we have worked in partnership with the building company in a particular sustainable neighbourhood composed of 400 apartments, half of them are equipped with smart meters. We then have conducted a survey in the neighbourhood to get the socio-demographic data of the residents.

Socio-economic Typology of Residents

To realise the socio-demographic typology, we have chosen the multiple correspondence analyses (MCA) (Abdi & Valentin 2007),which is a statistical technique to analyse categorical data. It is a generalisation of the most common technique: the principal component analysis. To compute this analysis, we have selected four socio-demographic variables: the training level, the annual income, the type of work and the household's composition. In this analysis, the apartments where we have more than 20% of non-answers are deleted from the data set. For the rest of non-answers, we use the methodology of regularised interactive MCA (Josse, Chavent, Liquet, *& all.* 2012) to complete the data set.

Clusters of Load Curves

A load curve represents the hourly consumption of dwellings. These data are derived from electrical measurements for each apartment. These data, thus increasing over time, are in kWh. In order to analyse them in more details, we have applied to this data a first derivative to obtain data in kilowatt for a period of one hour.
These are strategic data, in particular when they are displayed with a granulometry at the minute. With this type of data, it is then possible to analyse the behaviour of each inhabitant (f.i: switching of the light in the toilet, cooking, absence of the

residents...). In this article, we did not need to work with a granulometry at the minute because hourly aggregation allowed sufficient methodological results.

Overview of the approaches

We chose two principal families of approaches to cluster the load curves: (1) the expert approach and (2) the statistical approach.

(1) The expert approach is based on an interview of a Swiss expert in the energy field. We asked, "what are, from your point of view, the best methods to classify load curves?" The expert identified three principal methods: the average, the max and the dup (DUP).

(2) For the statistical methods, we based our choice on the TSclust R package (Montero &Vilar 2014) because it is one of the most exhaustive tools for temporal series clustering. We made a selection based on the most appropriate models, and we chose: Euclidian distance, dynamic warping distance (DWR), correlation distance (Spearman and Kendall), periodogram distance and spectral density distance. We completed the battery of models with the periodic analysis distance (Dudek, Hurd, Wojtowicz, *et al.* 2013) and the self-organizing map (Wehrens, Buydens& others 2007).

Methodology

First, we conducted a survey among the residents of a Swiss sustainable neighbourhood to gather socio-demographic data. We then we have proposed a marketing typology of residents. With this information, we have made a MCA.

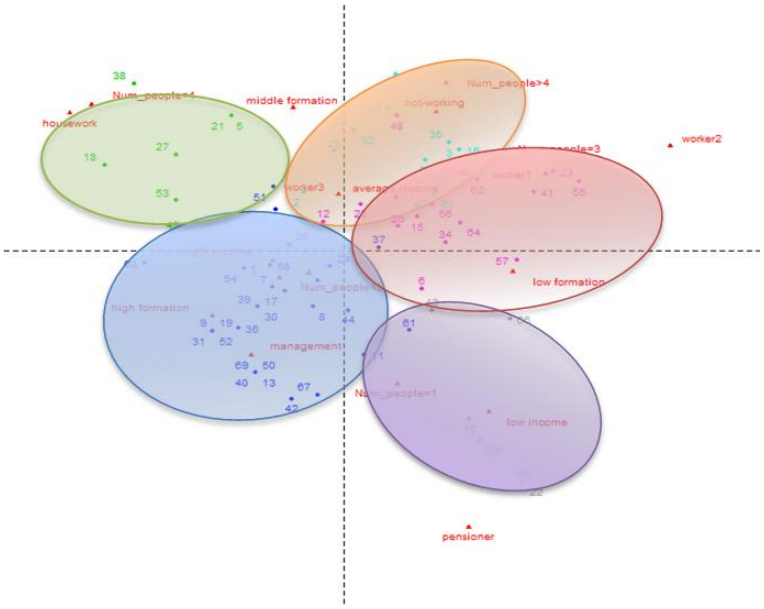
Then we analysed the load curves. For each method described in the cluster of load curve, we followed the same methodology. We measured the distance/similarity between the data. With this distance/similarity values we have made a hierarchical clustering. The number of groups is given by the Dunn index (Dunn 1974), which teste the combinations of the compactness and separation of each cluster. We can have thefore have the optimal number of groups. The only exception is the Kohonen (SOM): this method creates directly groups of clusters.

After the classification, we have created a contingency table between the results of MCA groups and each clustering of load curves. To compute this relation, we have used two statistical tests: the Chi-square test and the Cramer's V test. The Chi-square test (Yates 1934) is used to determine whether there is a significant association between the two variables. The Cramer's V is a measure of the association for nominal variables. Effectively it is the Pearson Chi-square statistic rescaled with values between 0 and 1, as follows: 0 indicates independency; 1 indicates perfect associations.

Results

We obtained 5 groups in the MCA procedure. This method can synthetize almost 36% of the total information.

Figure 1
MCA of socio-demo variable



Source: Author's illustration

The typology of residents is composed of five groups:

Blue: household with a higher education, a higher income, a management work and composed of fewer than 3 people.
Orange: households composed of more than 4 people, with middle education and average income, manly families.
Green: households composed of 4 people, and average education.
Red: households with average income and a low education, most of them are workers, and composed of three people in the household.
Violet: one-person households, with low income, part of them are pensioners.

Then we compare the clusters coming from MCA with the clusters of the load curves in the following table:

Table 1
Test between Socio-Demo and Load curve clusters

	P.values of Chi square test	Phi of the Cramer test
Cluster Average	0.20	0.39
Cluster Average/Max	0.27	0.40
Cluster DUP	0.27	0.40
Cluster Periodic Mean 24	0.36	0.35
Cluster SOM	0.06	0.42
Cluster Peridogramm	0.15	0.40

Cluster Spectral density	0.01	0.48
Cluster DWR	0.63	0.30
Cluster Euclidian	0.09	0.60
Cluster Kendal	0.14	0.44
Cluster Spearman	0.50	0.32

Source: Authors' (2016)

Discussion

We obtain three "winning" methods: Euclidian, SOM and spectral estimation to answer the research question. But this result has some limitations: the Chi Square is a parametric test and below 50 observations, the p-values can have some distortion. The number of missing values in the MCA can have an impact on the classification of the apartments. Also the short data collection period for load curves is another limit in the research projects.

Conclusion

We propose a methodological approach to compare two rankings with very different approaches. The first socio-demographic tools and the second: clustering applied to time series.

The comparison of several classifications with several different methods shown similarities between these rankings. The goal was not to check if the demographic ratings and rankings clustering made sense. It is therefore possible that we got a positive significance between for some comparison, but that either a false significance because the clustering of time series obtained by the Euclidian method comes the closest family's socio-demographic, while the Euclidian method does not take into account the optimal sequence alignment (same consumption, but offset 2 h).
In a next article we will quantitatively analyse the behaviour of the inhabitants, on data with a granulometry at the minute. We will verify that we have comparable rankings, which would, to identify socio-demographic classes with loads curve either profiler socio-demographic residents on the basis of information consumption.

References

Abdi, H. and Valentin, D. (2007) Multiple correspondence analysis.*Encyclopedia of measurement and statistics*.651–657.

Dudek, A., Hurd, H., Wojtowicz, W. and Wojtowicz, M.W. (2013) *Package 'perARMA'*. [Online] available at: http://watson.nci.nih.gov/cran_mirror/web/packages/perARMA/perARMA.pdf (Accessed: 13 May 2016).

Josse, J., Chavent, M., Liquet, B. and Husson, F. (2012) Handling missing values with regularised iterative multiple correspondence analysis. *Journal of classification*. 29 (1), 91–116.

Montero, P. and Vilar, J.A. (2014) TSclust: An R Package for Time Series Clustering. *Journal of*. [Online] available at: http://www.jstatsoft.org/v62/i01/paper (Accessed: 22 September 2015).

Wehrens, R., Buydens, L.M. and others (2007) Self-and super-organizing maps in R: the Kohonen package. *J Stat Softw*. 21 (5), 1–19.

Gower, J., Lubbe, S. and Roux, N. l. (2011) Multiple Correspondence Analysis, in Understanding Biplots, John Wiley & Sons, Ltd, Chichester, UK.doi: 10.1002/9780470973196.ch8

Yates, F, (1934). Contingency Tables Involving Small Numbers and the χ^2 Test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217–235. http://doi.org/10.2307/2983604

Regarding the authors

Francesco Maria Cimmino: research assistant. Masters in Statistics and Economics, University "Sapienza" of Rome (2012) and Masters in Econometrics for the banking and finance, Aix-Marseille School of Economics (2014). Email: francesco.cimmino@hevs.ch

Joelle Mastelic: Marketing professor in the Entrepreneurship and Management Institute (IEM). She works mainly on the social adoption of technologies in the field of energy. She manages the Energy Living Lab that performs applied research projects and mandates for economic operators and public authorities. Email: joelle.mastelic@hevs.ch

Dr Stéphane Genoud: Professor, Head of Energy Management Unit in the Entrepreneurship & Management Institute, HES-SO Valais / Wallis. PhD in Economics (University of Neuchâtel). Heads of Bachelor course "Energy Management". He is the founder of several businesses in the field of energy markets, energy audit and has a 20 years' experience in this field. Email: stephane.genoud@hevs.ch