An optimisation algorithm for matching large scale databases on customers for improved characterisation of electricity consumption

Thibaut Barbier, Robin Girard, Nicolas Kong, François-Pascal Neirac, Georges Kariniotakis, Elena Magliaro

AN OPTIMISATION ALGORITHM FOR MATCHING LARGE SCALE DATABASES ON CUSTOMERS FOR IMPROVED CHARACTERISATION OF ELECTRICITY CONSUMPTION

Thibaut Barbier[1,*], Robin Girard[1], Nicolas Kong[2], François-Pascal Neirac[1], Georges Kariniotakis[1], Elena Magliaro[1]

[1]Centre for Processes, Renewable Energies and Energy Systems (PERSEE), MINES ParisTech, PSL - Research University, Sophia Antipolis, France.
[2]Enedis, Paris La Défense, France
**thibaut.barbier@mines-paristech.fr*

Keywords: database matching, electricity consumption, electricity customers, housings, optimization algorithm

Abstract

This paper presents a method that permits to match customer information from the French DSO Enedis and housing information from the French population census institute INSEE. Our method allows having a list of housings linked to each customer in order to add household and building information to customers. We show with our method improvements in predictions of aggregated load curve indicators compared to the traditional method that averages socio demographic indicators from housing information of the zone covered by measurements. Our results indicate that the proposed algorithm is able to capture efficiently the information of housings in some feeders. This permits to combine the databases of the DSO with external databases that exist from census or other processes. Enriching the information at the level of clients through the proposed automated way is a cost effective approach given the number of customers served by a DSO. This enhanced information can be then the basis to model, analyse and simulate demand in a bottom up approach which can be useful for planning purposes of the distribution networks.

1 Introduction

Most of scientists agree that human being has influenced the climate system, especially due to anthropogenic emissions of greenhouse gases. IPCC [1] showed that in recent decades, climate change has damaged natural and human ecosystems in all continents, oceans and atmosphere. These impacts will continue for centuries, even if anthropogenic emissions are stopped. Climate change will exacerbate current issues in many aspects (with very high confidence), such as health problems, migrations, weather extreme events, and this phenomenon of amplification will increase quickly during the 21st century. Adaptation is a strategy for managing these issues. Also mitigating the magnitude of global warming by lowering those emissions is crucial, in order to delay the risk of severe, widespread and irreversible impacts globally [1].

Electricity represents 18% of the total final energy consumption in the world in 2013 [2] and is expected to rise up to 25% by 2040 [3]. Electricity and heat was 42% of world CO2 emissions in 2013, i.e. 13.8 Giga tons of CO2 [4]. Such awareness has led to consider larger and larger shares of the electricity demand as controllable. This offers many opportunities [5] such as an increased potential for energy saving, the possibility to adapt electricity demand to intermittent renewable energy, or a lowered of peak demand to prevent costly investments on the electricity grid. In order to be able to exploit these opportunities and integrate them in the management and especially the planning processes of the power system, it is necessary to have an in depth knowledge of the characteristics of the demand not only in terms of energy but also in terms of power. However, electricity consumption, especially at the city scale, is the sum of a considerable number of loads, the characteristics of which are often unknown.

Characterizing each contributor of the electricity consumption is not a recent topic. Swan and Ugursal [6] present examples of data required to develop residential energy model. It includes information on the physical characteristics of the dwellings, occupants and their appliances, historical electricity consumption, climatic conditions, and macroeconomic indicators.

DSOs (Distribution System Operators) dispose large amounts of data on energy consumption and aggregated load curves. However, when it goes down to the layer of buildings, information is lacking [6]. Given the number of clients it is very costly to implement such data collection procedure. This type of data is often publicly available by external organizations. It is a challenge to match the information in such external databases to the data in the DSO databases.

In this paper, we present a method that permits to match customer information from the French DSO Enedis and housing information from the French population census institute (INSEE).

We define housing as houses or dwelling places thought of as a group. A housing occupied as a principal residence has a household.

Electricity customer (or customer) materializes a unique contract with the DSO who is paid in order to provide an electricity supply service at a point of common coupling where the customer belongs.

Our method allows having a list of housings linked to each customer in order to add household and building information to customers. We call the information about building and household "socio demographic information". When this information is gathered at district scale or at a zone covered by an electric line we call it "socio demographic indicator".

We show with our method improvements in predictions of aggregated load curve indicators compared to method chosen as reference that averages socio demographic indicators of the zone covered by measurements.

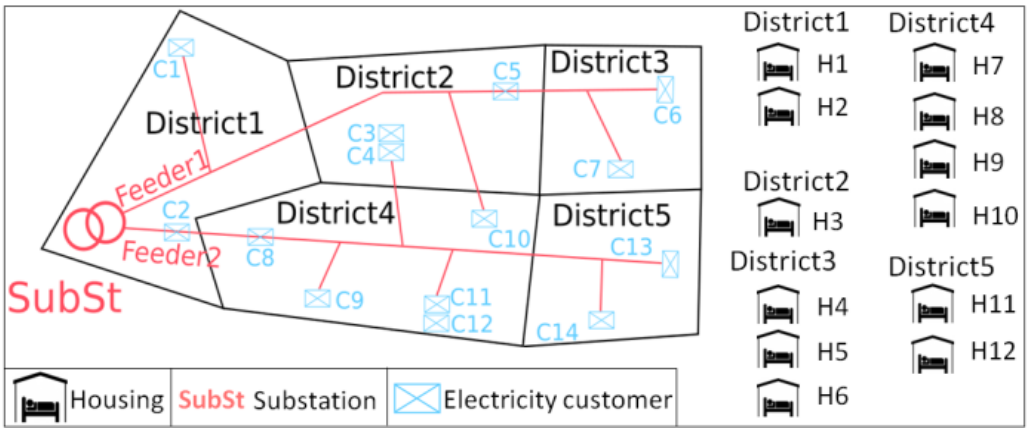

Fig.1: available information on the left (DSO) and on the right (population census database)

2 Illustration of the problem

Figure 1 illustrates an area of 5 districts where the electricity customers are supplied by the different feeders of a same substation. This is typical information we can have from DSO's databases. Some customers have the same address (e.g. C3-C4 and C11-C12): it represents different customers in the same building.

For each district we can have from external databases, like the ones resulting from population census, the list of the housings. Electricity customers can be either residential or not. If a customer is residential, it implies that it is linked to one or multiple housings. But also a single housing can have different electricity contracts and so can be linked to different customers.

Our method proposes to match customers and housings, in order to have a list of housing linked to residential customers. At least, as we know from the DSO which customer is linked to which feeder we can have a list of housing per electric feeder

3 Problem solving method

3.1 Estimation of common parameters

Electricity customers and housings have different characteristics listed in Table 1. This table explains the difference between customer and housing in terms of information.

Table 1 List of the different characteristics of customers and housings

Database entity	Useful characteristics to characterize electricity consumption
Electricity customer	Subscribed power, annual consumption, text address, district, type of customer (e.g. residential, tertiary artisan or shopkeeper).
Housing	Number of inhabitants, type (flat, house), surface, number of rooms, age of the building, category (vacant, secondary residence, principal residence), type of heating (electric, fuel, gas, wood, other), district.

We can notice that housings and electricity customers have no common parameters except that we both know the district where they are.

The idea is to create common parameters between the two databases in order to measure a distance and be able to match them.

We choose to estimate the yearly energy consumption of housings as a linear model of surface, presence of an electric heating, age of building, number of occupants and number of rooms. The model is fitted with aggregated information on each district, which was presented in a previous work [7].

3.2 *Matching data with the common parameters*

The problem set is as following: we want to match the housings and the residential electricity customers. Housings' annual electricity consumption was estimated with the method described in the previous section. We define below the list of parameters to be matched from the two considered databases:

$x_{i\in\{1,2,\dots,n\}}$ represents the estimated electricity annual consumption of the housings, n is the number of housings in the considered district. x_i is the ith housing.

$y_{j\in\{1,2,\dots,p\}}$ represents the residential annual consumption of the electricity customers. p is the number of customers in the considered district. y_j is the jth electricity customer.

The problem of matching can be formulated as a distance (here chosen to be quadratic) minimization problem between the matched entities:

$$\min_{(\alpha_{ij})_{(i,j)\in\{1,2,\dots,n\}x\{1,2,\dots,p\}}} [\sum_j^p(\sum_i^n \alpha_{ij}x_i - y_j)^2] \quad \textbf{(1)}$$

S.t.:

$\alpha_{ij} \in \{0,1\}, \forall\ (i,j) \in [1\colon n]x[1\colon p]$ **(2)**

$\sum_j^n \alpha_{ij} \geq 1\ \forall i \in [1\colon n]$ **(3)**

$\sum_i^n \alpha_{ij} = 1\ \forall j \in [1\colon p]$ **(4)**

where:

- $(\alpha_{ij})_{(i,j)\in[1:n]x[1:p]}$ is the matching unknown: $\alpha_{ij} = 1$ if x_i is matched with y_j, 0 else (2);
- (3) is the constraint that each housing is linked to at least an electricity customer;
- (4) is the constraint that each electricity customer is linked an unique housing.

This problem is a classical quadratic function to minimize (1). However, (2) is a non-continuous constraint. Solving method by relaxing the constraint (2) in an interval and putting additional constraints to force solution to be closed to boundaries 0 and 1 where tested, but without any success: solutions where not closed to the boundaries.

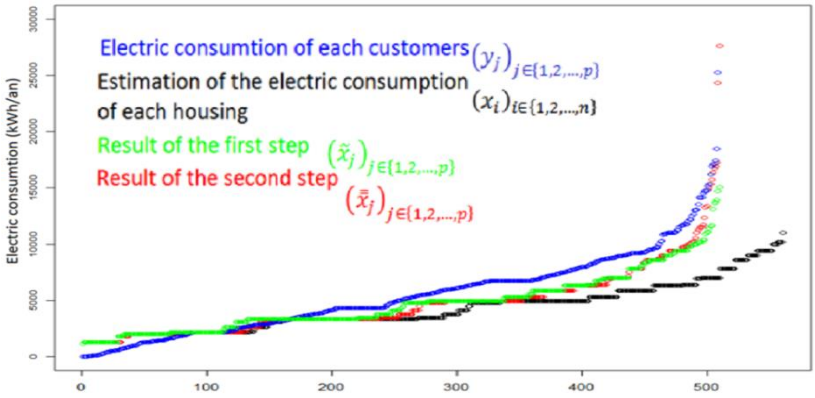

Fig. 2 Illustration of the different steps of the matching between customers and housings

Testing each possibilities of $(\alpha_{ij})_{(i,j)\in[1:n]x[1:p]}$ leads to too long calculations as the number of possibilities is $2^{n.p}$.

The approximate method of resolution chosen is as following: we first set the initial state by ordering data to match and make a random gathering in order to have the same number of data to match $y_{j\in\{1,2,\dots,p\}}$ and $\tilde{x}_{j\in\{1,2,\dots,p\}}$, as illustrated figure 2. Then we optimize the global distance by moving the x_i that minimizes the global distance $\sum_{j=1}^p(\tilde{x_j} - y_j)^2$. We finally stop when the global distance remains the same. As shown figure 3, housings are now linked to customers.

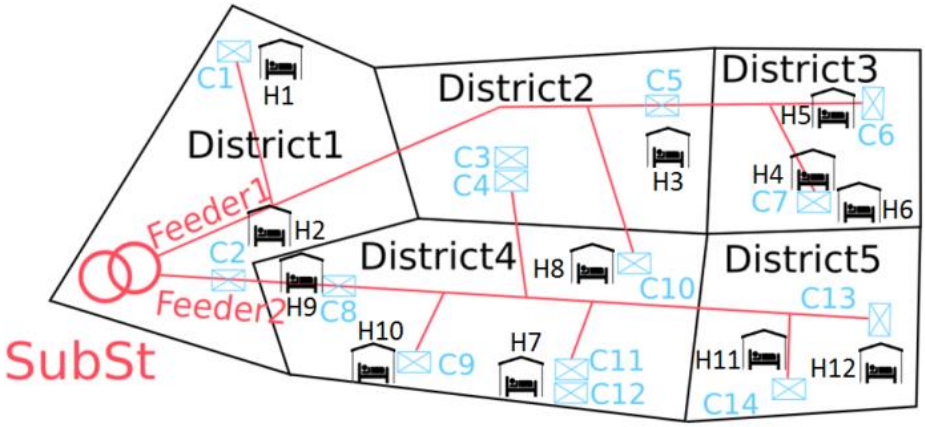

Fig. 3 example of result of matching: housings are now linked to customers.

4 Case study and interest of the method

4.1 *Data and reference of comparison*

Data were collected every 10 minutes by Enedis (main French DSO) for the whole year 2012. We used 59 feeder measurements from the region of Lyon, in France. For each feeder we have the list of customers supplied, whose characteristics are listed in Table I. The characteristics of the feeders are as following:

Table 1 Characteristics of the feeders data set used

Indicator	Min value	Mean value	Max value
Number of customers supplied	330	3400	7000
Fraction of residential consumption	42%	67%	88%
Number of districts supplied	2	5	13

For each district supplied by the feeders, we have also a list of housings with some characteristics listed in Table I. This data are provided by the French institute of statistics (INSEE).

The application of the matching algorithm between customers and housings for each district permits to derive a list of housings per feeder. For example in figure 3, the matching procedure for Feeder1 permits to identify that this feeder supplies housings H1, H3, H4, H5, and H6. With this list of housings it is possible to characterize socio-demographically each feeder. This is done through averaging the values of the parameters (as per Table I) of the associated feeder housings. At the scale of the feeder we also calculate reference socio-demographic indicators, which are the sum of average values of socio demographic indicators for each district weighted by the ratio of residential customers in each district.

4.2 *Model to explain thermo sensitivity using socio-demographic indicators*

Thanks to the previous section we have associated to each feeder additional useful information resulting from the reference calculation and the matching. Now we use this information to model the thermo-sensitivity per feeder. The thermo-sensitivity chosen here is the linear fit of consumption points in ordinate and temperature points on the abscissa, by only selecting points whose temperature are below a certain threshold of non-heating. Figure 4 shows the thermo-sensitivity plot of a feeder. In this figure we see the link between electricity and temperature, which represents well the French behaviour: because of electric heaters the consumption rises with temperature decrease. This behaviour is similar to other intermediate temperature European countries (averages temperature between 9°C and 14°C) [8]. This information can be useful for a DSO in order to size its grid as in this case peak of consumption is due to cold temperatures.

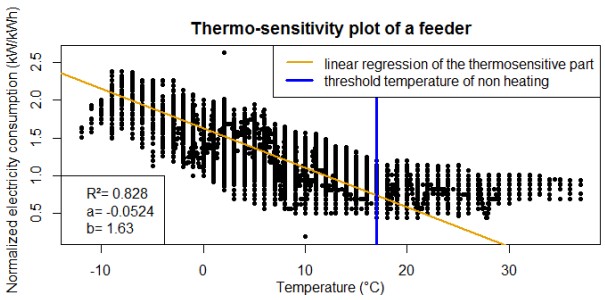

Fig. 4 example of thermo-sensitivity plot of a feeder

In order to quantify the impact of socio-demographic indicators on the consumption thermo-sensitivity, we use a linear model as shown figure 5.

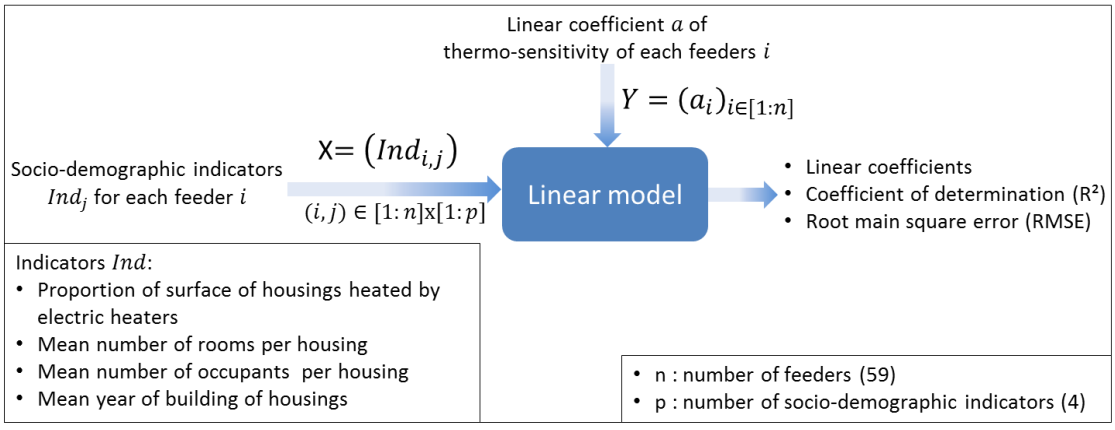

Fig. 5 Illustration of the different steps of the matching between customers and housings

4.3 *Quantitative interest of matching to explain thermo-sensitivity*

We display prediction errors made by the linear model fitted on the data available.

We compare the errors of thermo-sensitivity predicted by using socio-demographic indicators taken as reference (average values in each district covered by a given feeder) and by using socio-demographic indicators from matching algorithm.

As shown figure 6, the error of prediction is 0.8% less important with matched data compared to reference average data.

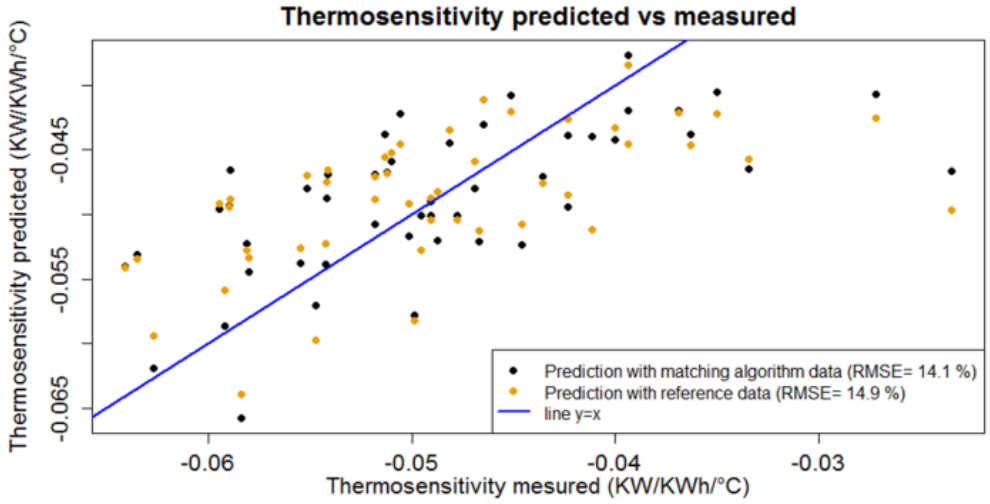

Fig. 6 Thermo-sensitivity prediction errors for each feeder with reference data and data from matching algorithm

5 Conclusion

In this work we propose a method to match housing information with residential customers. We test the interest of this method: we compare errors of prediction of thermo-sensitivity of electric feeders with socio-demographic data from average values of the zone and from data resulting from the matching algorithm. We show that the matching algorithm improves prediction, meaning that the matched information is closer to real customers' housing information than the averaged values. Figure 7 shows the distribution of the relative improvement for each feeder.

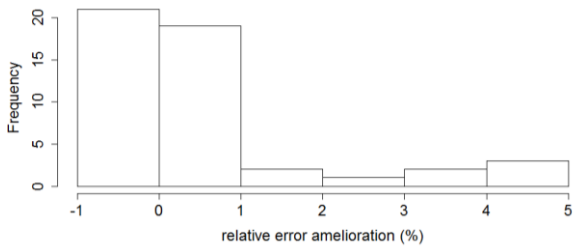

Fig. 7 Histogram of the relative amelioration of errors with data from matching method compared to reference average data.

Our results indicate that the proposed algorithm is able to capture efficiently the information of housings in some feeders. This permits to combine the databases of the DSO with external databases that exist from census or other processes. Enriching the information at the level of clients through the proposed automated way is a cost effective approach given the number of customers served by a DSO. This enhanced information can be then the basis to model, analyse and simulate demand in a bottom up approach which can be useful for planning purposes of the distribution networks. Further work will focus on investigating the local characteristics of the feeders for which thermo-sensitivity is well predicted, on testing the matching algorithm on higher amount of data and multiple years and regions to validate the results. This method allows also estimating thermo-sensitivity for scales from districts to regions.

6 Acknowledgements

We thank Enedis for providing the data for this work.

7 References

[1] Intergovernmental Panel on Climate Change., 'Climate Change 2014. Synthesis Report. Summary for Policymakers' (2014), pp. 2–17

[2] International energy Agency., 'Key word Energy Statistics' (2015), p. 28

[3] International energy Agency. 'World Energy Outlook 2015. Executive Summary' (2015), p.5

[4] International energy Agency. 'Key trends in CO2 emissions' (2015), p.6

[5] Siano, P.: 'Demand response and smart grids. A survey', Renewable and Sustainable Energy Reviews, February 2014, **30**, pp. 461–478

[6] Swan, L. G., Ugursal, V. I.: 'Modeling of end-use energy consumption in the residential sector: a review of modeling techniques', Renewable and Sustainable Energy Reviews, 2009, **13**, (8), pp. 1819–35

[7] Barbier, T., Girard, R., Neirac, F. P., *et al.*: 'A novel approach for electric load curve holistic modelling and simulation'. Proc. Int. Conf. MedPower 2014, Athens, Greece, Nov 2014, (IET), pp. 1-8.

[8] Bessec, M., Fouquau, J.: 'The non-linear link between electricity consumption and temperature in Europe: A threshold panel approach', Energy Economics, 2008, **30**, (8), pp. 2705–21