ESSNet Smart Meter

Alexander Kowarik
Johannes Gussenbauer

Matching Australian smart meter data with Austrian household data

Matching the households from the australien smart meter data, SM, with the households from a synthetic Population for Austria, P, was done using a simple algorithm, which proceeds as follows:

1. Select $d_1, ..., d_n$ demographic variables, existing in both data sets, where each variable d_i can take values from a set D_i,$i = 1, ..., n$.
2. This selection naturally splits SM and P into disjoint subgroups SM_i and P_i,$i = 1, ..., \prod_{i=1}^{n} |D_i|$ defined by the values of $d_1, ..., d_n$.
3. For each subgroup P_i, sample with replacement $n = |P_i|$ households from SM_i.
4. If for an i the set SM_i is empty, discard the least significant criteria in $d_1, ..., d_n$ and calculate new subgroups. Do this only for is where $|SM_i| = 0$.
5. Repeat steps 3 and 4 until every household in P was matched with a household in SM.

For the matching the demographic variables *household income group*, *number of occupants*, *number of children 0-10 year*, *number of children 11-17 years*, *number of occupants 70+ years* and *is home during daytime.*

Identifying vaccancies

Identifying vaccancies was done using a variety of different methods. Since there is no data present on which to evaluate the results of each methods, they can only be judged by their methodological soundness. Vaccancy was estimated using the half hourly kwh usage per day. The following method were tested:

1. Using the R-function `auto.arima` to predict kwh usage. A good fit could indicate predictable kwh usage and therefore a vaccant household..
2. Estimating the variance of the daily kwh usage. Low variance could indicate a vaccant houhsehold.
3. Estimating the ratio between the mean kwh usage during day- and night-time.
4. Using a volatitlity measure sd(diff(scale(kwh usage))).
5. Using a cell-wise outlier detection procedure to locate days with 'unusually' low kwh usage.
6. Using estimates from Methods 3., 4. and 5. as well as range, mean and median of daily kwh usage in a random forest model.

For the following descriptions let $x_1, ..., x_48$ be the half hourly kwh usage for a person of a day.

Times series

For the time-series approch the `auto.arima` function from the R-package `forecast` was used. The idea is to fit an ARIMA model to $x_1, ..., x_48$ and estimate the correlation between the real data an the fitted model. A high correlation indicates a predictable pattern which furthermore suggests a vaccant household. This would follow from the assumption that the variance in the kwh usage is generated by the automatic (and deterministic) turn off and turn on behaviour of household appliances, e.q. fridge, heating system, ect. Whereas a low correlation would indicate unpredictable human behaviour.

Variance

Calculating the variance of the $x_1, ..., x_48$ can be used to estimate if a household is vaccant. Since household appliances on standby use only litte kwh, a low variance could indicate a vaccant household whereas a high variance in kwh would be the result of actively using power, e.q. a non-vaccant household.

Day- and night-time

Another indicator for the classification of vaccant households could be the ratio between the mean of half hourly kwh usage during day- and night-time. Day-time was defined from 8 a.m. to 7 p.m. and night time from 8 p.m. to 6 a.m.. A high ratio would implicate a more excessive kwh usage during daytime which would implicate a non-vaccant household.

Volatility measure

Scaling the hourly kwh usage $x_1, ..., x_48$ and estimating the standard deviation of the first differences of the scaled values could also lead to an indicator for vaccant households. A high volatility would result from frequent and large variations in kwh usage which furthermore could point to a non-vaccant household.

Cell-wise outlier detection

Another approach in order to classify vaccant households was the use of a cell-wise outlier detection method. The Method is fully explained in Rousseeuw and Bossche (2016) and usable in R through the package `cellWise`. Given a data matrix $X \in \mathbb{R}^{n \times p}$ the main idea of this method is to use univariate outlier detection methods on each collumn as well es the correlation between every collumn to detect potential outlying cells. One of the advantages of this method are that the method still works if $p > n$, in fact as stated in Rousseeuw and Bossche (2016) the method gains more performance by introducing more dimensions to the data and that it can deal with missing values in the data matrix X.

The method was applied to the total kwh usage per household and day for the year 2013, by using a data matrix $X \in \mathbb{R}^{n \times p}$, with n as the number of households and p as the number of days in 2013. Each cell in X states the total kwh usage of a household and day of the year.

To improve the performance of the method two additional steps were taken before applying the outlier detection method

1. The distribution of total kwh usage per day across the households is skeewed heavily to the right. Since this method internally applies onto each collumn robust estimates for location and scale, which expect the data to be approximately normal, the Box-Cox transformation was initially applied to each collumn. Because zeros can occur in the data we choose to use the two-parametic Box-Cox transformation, which is implemented in the R-package `geoR`.
2. After Box-Cox transformation the data matrix X was split into subsets defined by the number of Occupants in each household. The groupse were labeled *low*,*mid* and *high* and contained of households with 0-2 occupants, 3-4 occupants, and 5+ occupants.

The cell-wise outlier detection method was finally applied on the transformed subsets. Please note, that this method detects upper as well as lower outliers and that in this context only lower outliers are of intereset.

Random forest

References

Rousseeuw, Peter J, and Wannes Van den Bossche. 2016. “Detecting Deviating Data Cells.” *ArXiv*. https://arxiv.org/abs/1601.07251.