ESSNet Smart Meter

Alexander Kowarik
Johannes Gussenbauer

Matching Australian smart meter data with Austrian household data

Matching the households from the Australian 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 vacancies 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. Vacancy 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 vacant household..
2. Estimating the variance of the daily kWh usage. Low variance could indicate a vacant household.
3. Estimating the ratio between the mean kWh usage during day- and night-time.
4. Using a volatility measure.
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.

Time series

For the time-series approach 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 vacant 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 behavior of household appliances, e.q. fridge, heating system, etc. Whereas a low correlation would indicate unpredictable human behavior.

Figure 1. shows the KWH usage of specific days and households with right bars showing the absolute correlation. A high correlation should indicate a vacant household and vice versa. As it can be seen this methods produces questionable results since the kWh usage takes on similar forms for cases with low and high correlation.

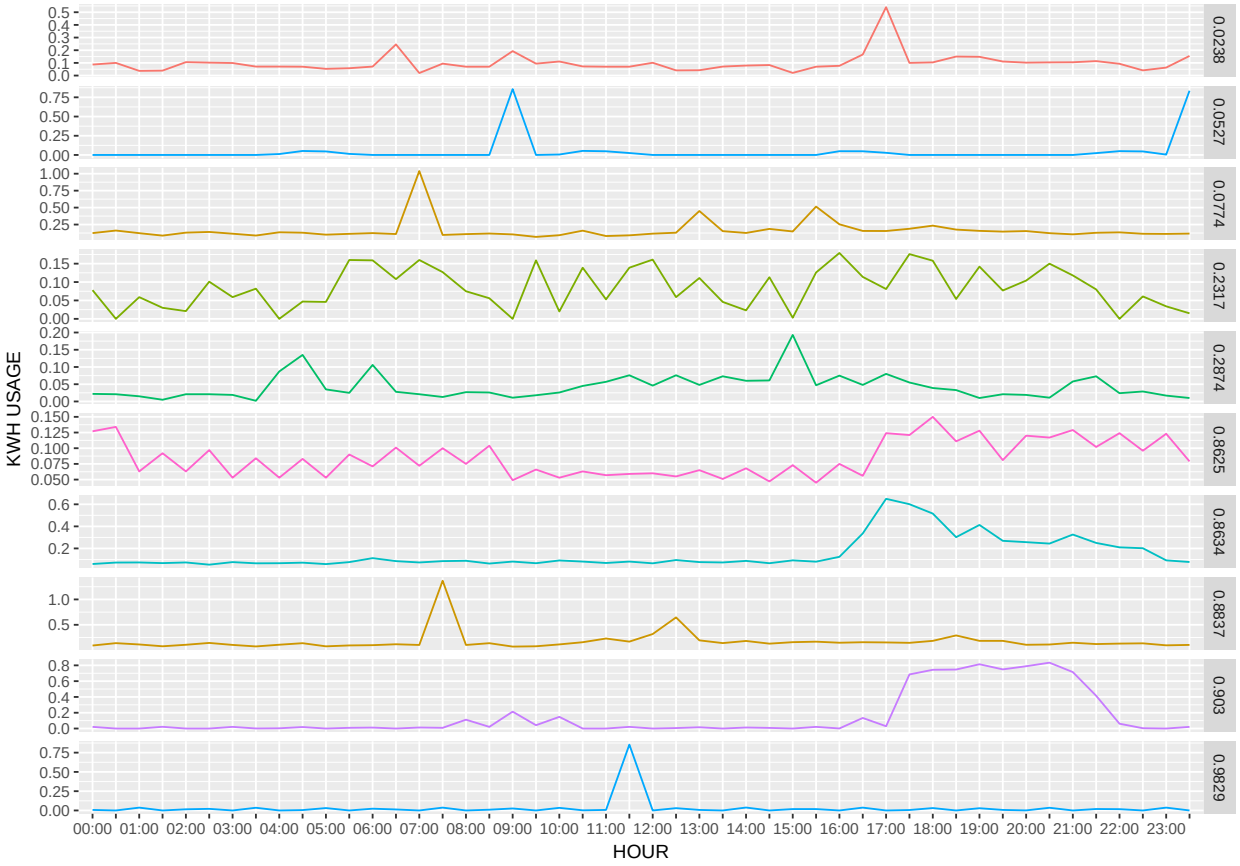

Figure 1: Examples using time series approach.

Variance

Calculating the variance of the $x_1, ..., x_{48}$ can be used to estimate if a household is vacant. Since household appliances on standby use only little kWh, a low variance could indicate a vacant household whereas a high variance in kWh would be the result of actively using power, e.q. a non-vacant household.

Figure 2. shows the KWH usage of specific days and households with the right bars showing the estimated variance. A high variance should indicate a non-vacant household and vice versa. This method performs a bit more convincingly than the time series approach, but cases with high variance and (seemingly) deterministic changes in kWh usage can still occur, e.q. the third graph from the bottom.

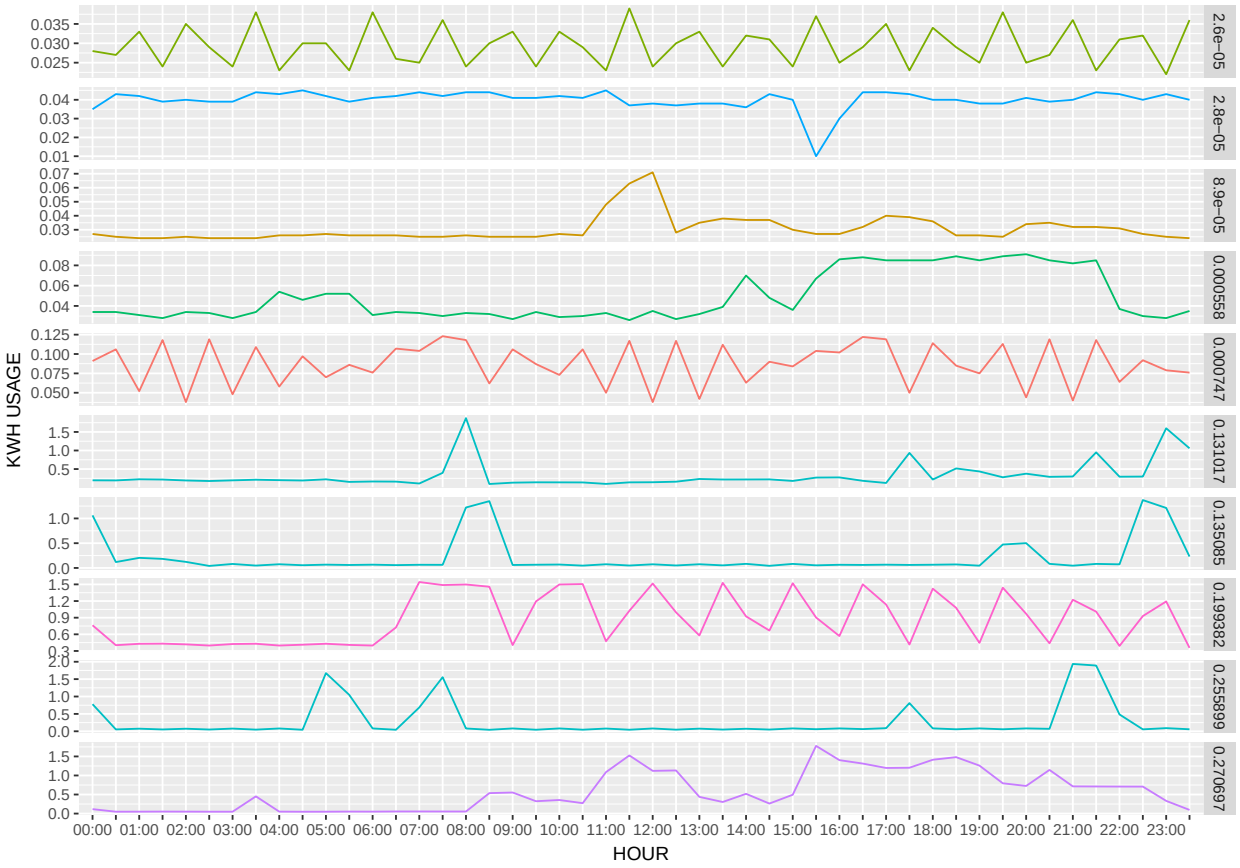

Figure 2: Examples using variance approach.

Day- and night-time

Another indicator for the classification of vacant 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-vacant household.

Figure 3. shows the KWH usage of specific days and households with the right bars showing the ratio between the day- and night-time kWh usage. It can be easily seen that this methods does not classify vacancy very well. Even cases with a low ratio can, judging by the path of the kWh usages, very well be from non-vacant households.

Volatility measure

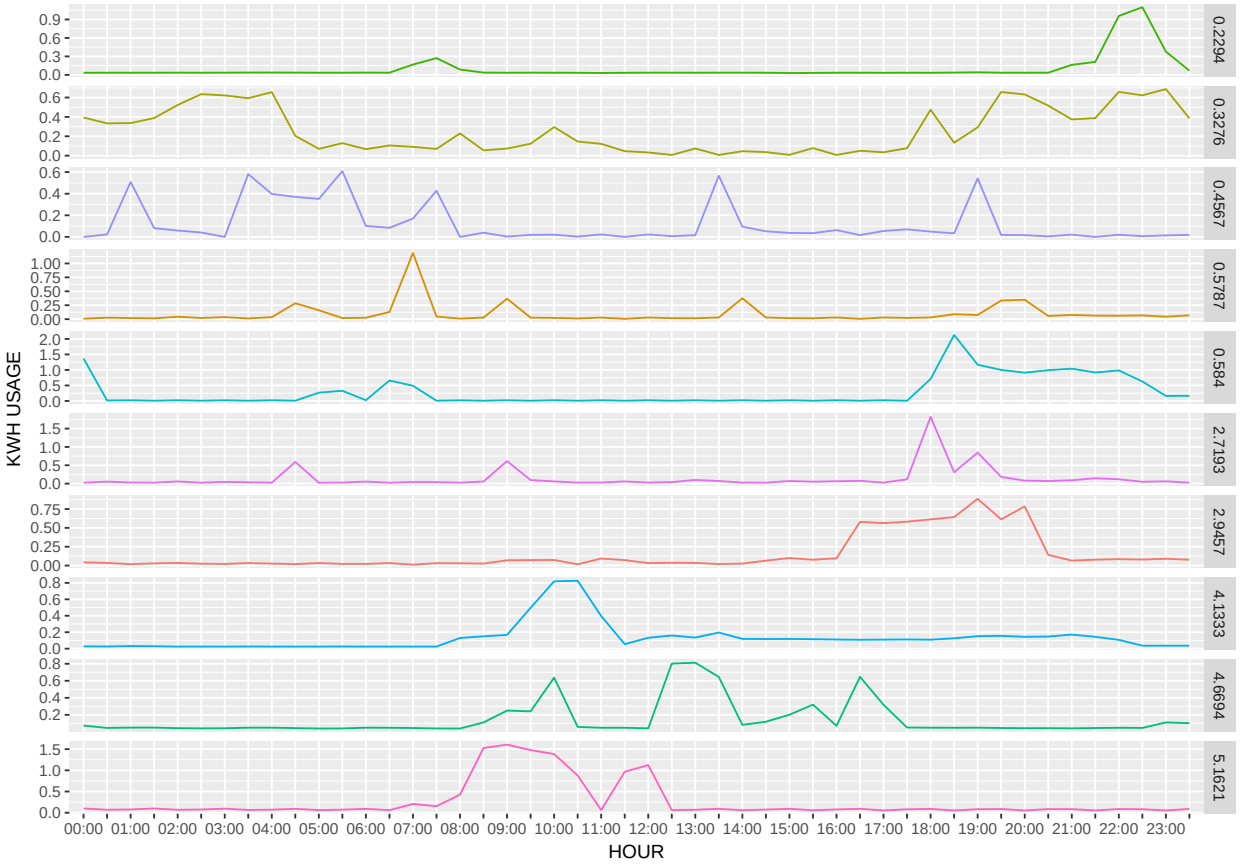

Figure 3: Examples using time day- and night-time approach.

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 vacant households. A high volatility would result from frequent and large variations in kWh usage which furthermore could point to a non-vacant household.

Figure 4. shows the KWH usage of specific days and households with the right bars showing the estimated volatility. This methods seems to perform poor as well, since a high volatility does not necessarily coincide with a non-vacant household.

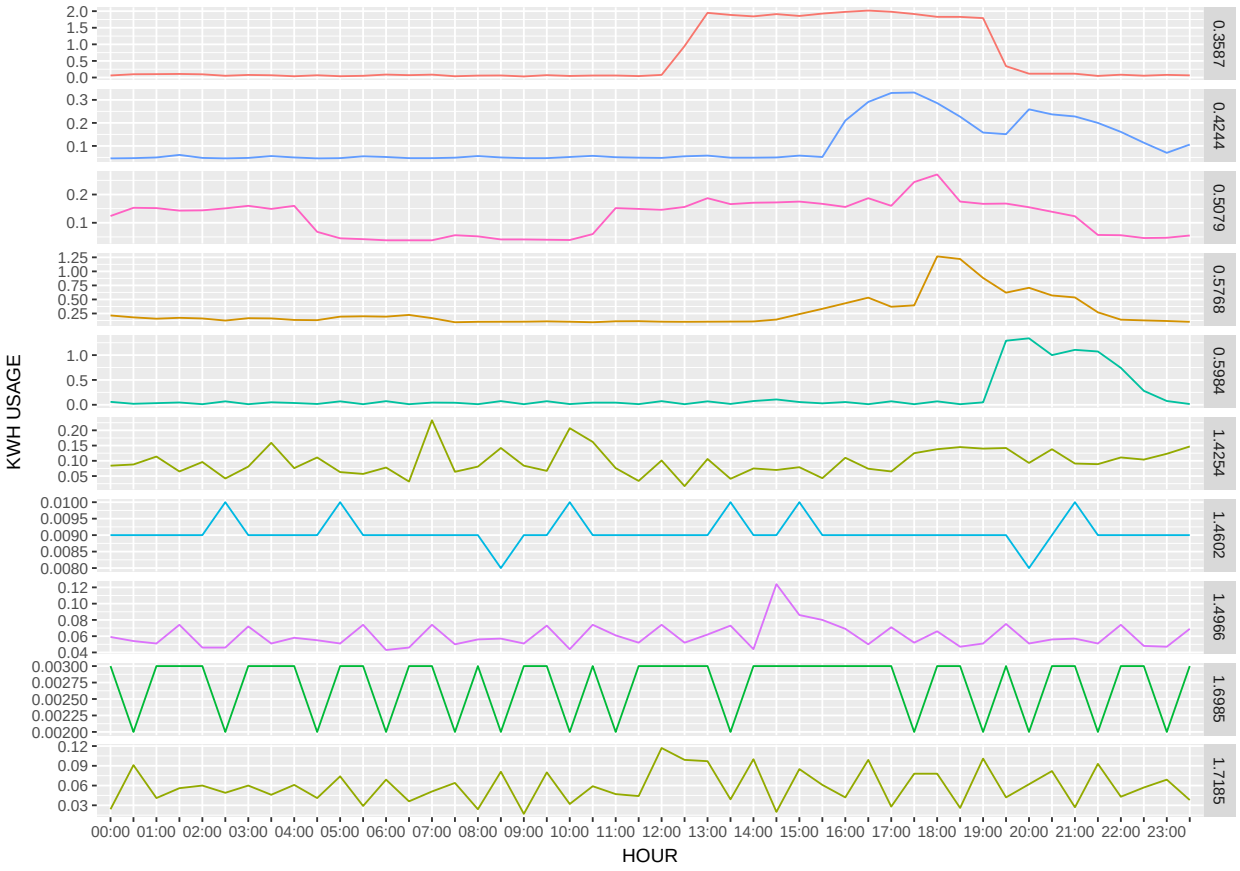

Figure 4: Examples using volatility approach.

Cell-wise outlier detection

Another approach in order to classify vacant 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`. The main idea of this method is, given a data matrix $X \in \mathbb{R}^{n \times p}$, to use univariate outlier detection methods on each column as well es the correlation between every column to detect potential outlying cells. Some 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 on the total kWh usage per household and day, 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 lists 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 skewed heavily to the right, as displayed by Figure 5, which shows the density estimates of kWh usage per day. Since this outlier detection method method internally applies, onto each column, robust estimates for location and scale, which expect the data to be approximately normal, the Box-Cox transformation, see Box and Cox (1964), was initially applied to each column. Because zeros can occur in the data we choose to use the two-parametric 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 groups were labeled *low*,*mid* and *high* and contained of households with 0-2 occupants, 3-4 occupants, and 5+ occupants.

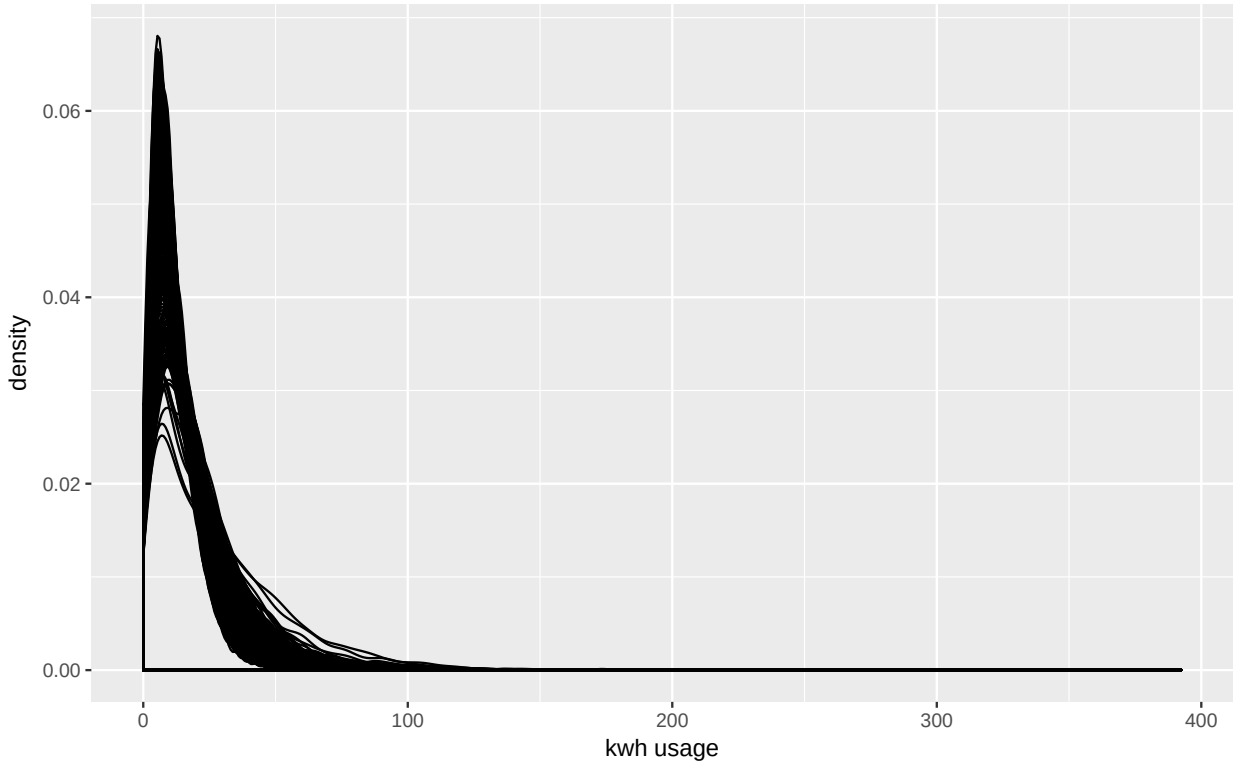

Figure 5: Density estimates on kwh usage for each day.

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

Figure 6. shows the total kWh usage per day for some of the households with red dots showing bottom outliers e.q. vacant households and yellow dots upper outliers, which are not of interest in this context but are displayed for the sake of completeness. It seems, that the methods perform reasonably well, especially when flagging, what seems to be, vacancy due to long holidays. Nevertheless it can't be said what really is or is not a vacancy, which leaves questions about a goodness of classification, e.q. false-positive- or false-negative-rate, unanswered.

Random forest

The final approach consists of a random forest model that was trained on a labeled subset. The regression variables of the model consists of previously listed classifiers as well as additional information. In detail the variables we used per household were estimates concerning kWh usage per day and quarter consisting of:

- Variance

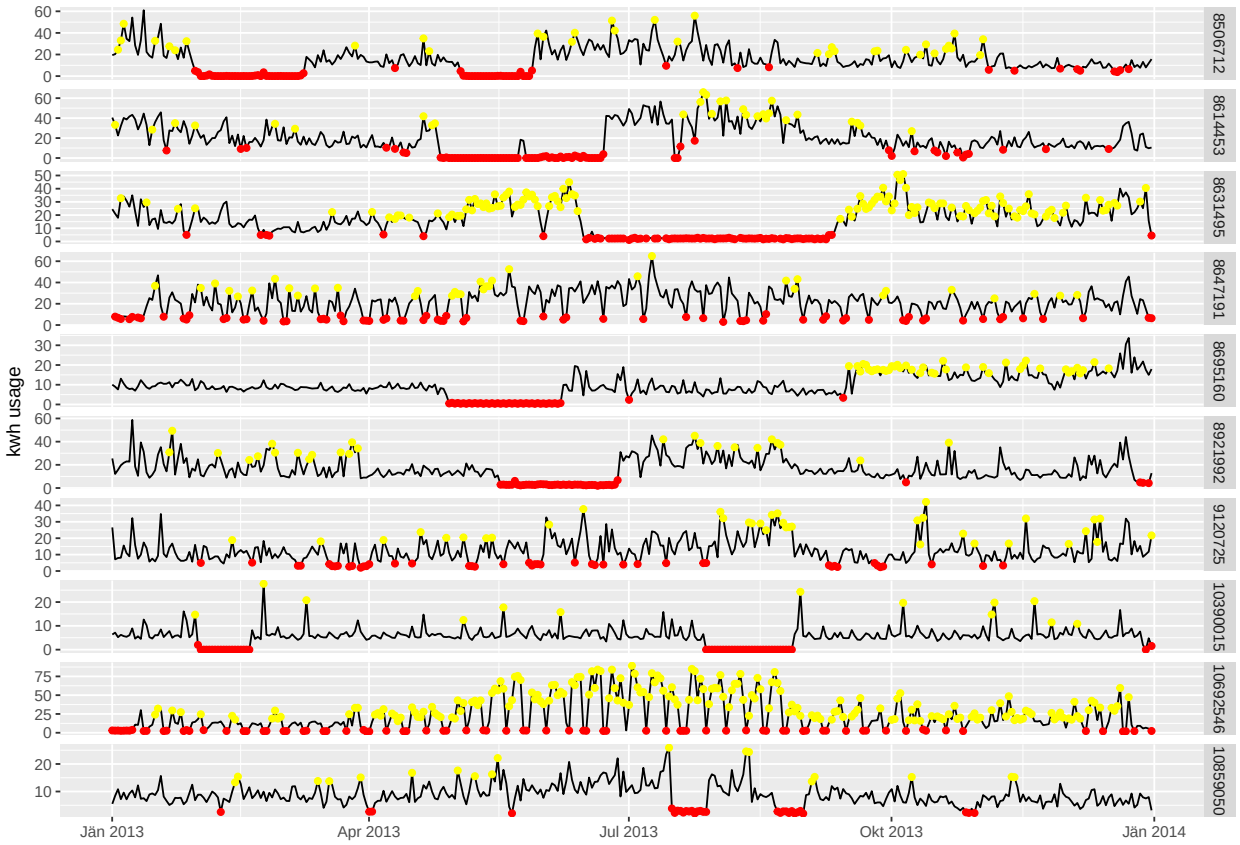

Figure 6: Results of cell wise outlier detection on daily kwh usage for some of the households. Red dots display lower and yellow dots upper potential outliers.

- Range
- Mean
- Median
- Standard deviation of first differences of the scaled kWh usage
- Share between mean day- and night-time kWh usage
- Number of occupants in each households grouped as in the cell wise outlier detection method

The labeled subset was created using the cell-wise outlier detection method and than handpicking 75 households for which this outlier detection method seemingly produced reliable results. For the calculation of the random forest model the function `ranger` from the R-package which carries the same name.

The goodness of fit for the random forest, on the training data, was the following:

		predicted	
		non-vacant	vacant
true	non-vacant	24210	135
	vacant	358	2010

About 85% of the assumably vacant cases were successfully detected with a false discovery rate of about 6%.

Concluding remarks

The outlier detection method as well es the random forest approach seem to produce, at least, more reliable results. Nevertheless the quality of these classifications can not truly be measured without a pre-labeled data set. If a pre-labeled data set is provided more fine-tuning steps can be taken, like:

- Using other measures like variance or range on the cell-wise outlier detection approach.
- Using other machine learning algorithms, like boosting or support vector machines.
  - Without a labeled data set, the results would not gain more insight into the problem of classification as was displayed by the random forest model.
- Using cluster algorithms on time series to identify specific patterns for vacant households.

References

Box, G. E. P., and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B (Methodological)* 26 (2). [Royal Statistical Society, Wiley]: 211–52.

Rousseeuw, P. J., and W. V. den Bossche. 2016. "Detecting Deviating Data Cells." *ArXiv*. https://arxiv.org/abs/1601.07251.