Clustering of Smart Meter Data for Data Compression and Fast Power Flow Computation

Christoph KATTMANN
University of Stuttgart – Germany
christoph.kattmann@ieh.uni-stuttgart.de

Krzysztof RUDION
University of Stuttgart – Germany
rudion@ieh.uni-stuttgart.de

Stefan TENBOHLEN
University of Stuttgart - Germany
stefan.tenbohlen@ieh.uni-stuttgart.de

ABSTRACT

The amount of data generated by smart meters is a challenge for storage, but also for computation. In order to derive meaningful knowledge from the recorded data, simulations must be able to use it simply and effectively. This paper presents an algorithm that uses agglomerative hierarchical clustering to compress smart meter data and optimally prepares it for use in Monte Carlo simulations, which are widely utilized in the planning and operation of distribution grids today. These simulations are necessary to determine the probability of voltage band violations for various grid expansion options, but are computationally expensive. By finding clusters of recurring data and replacing them by a single entry in a database, the required storage space can be reduced by around 20%, while the computation time for power flow in Monte Carlo simulations can be reduced by up to 40%.

INTRODUCTION

Smart meters installed in distribution grids generate an enormous amount of data that is difficult to store, transmit and compute [1]. Numerous attempts to apply standard data compression algorithms to the acquired data have proven successful [2],[3],[4]. Their focus is however exclusively on the maximal reduction in size and sometimes on required computing power, which makes sense when the main challenge is the weak hardware and the poor connection to a central server. On those servers however, the requirements are different. Compressed data saves disk space, but it adds additional computation time whenever data is processed.

The subject addressed in this paper is a clustering method for load data that can be used for data compression and additionally speeds up steady-state power flow computations. The principle is outlined in Fig 1. The uncompressed load data of smart meters is first augmented using standard power flow computation, which yields a complete datasets about the grid situation, including complex loads and voltages at individual nodes in the grid and line currents. For this, information about the impedances in the grid is necessary. A cluster analysis can now be conducted on the voltage data. For different time horizons, the cluster analysis reveals if there are voltages profiles that appear more often than others in different times and grids. These voltage profiles can be replaced by one model case, eliminating the storage required for every, only slightly different profile. The result is a database of voltage profiles that appeared multiple times. A following Monte Carlo simulation can now use this database and access it whenever a similar load situation appears, reducing the runtime of several Newton-Raphson iterations down to a database lookup. For single power flow computations, the database approach has been presented in [5].

The runtime of these simulations has become a problem. A major challenge for distribution grid operators is the correct evaluation of grid expansion and control options when faced with potential bottlenecks. With the rise of alternatives to new or reinforced lines like transformers with on-load tap-changers in lower-voltage levels, controllable inverters for PV plants or possibly controllable loads it is necessary to find the technically and economically optimal solution to guarantee compliance to power quality regulations. The expected frequency of voltage band violations, an important part of power quality can be estimated using load flow calculations, using assumptions about the statistical distribution of household loads, weather conditions and other influences. In order to derive meaningful

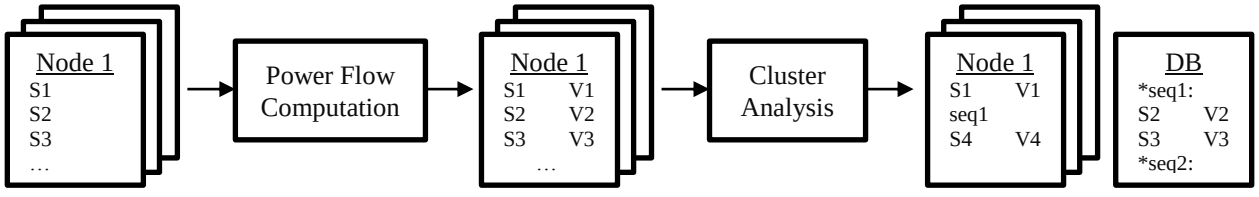

Figure 1: Overview of smart meter data compression on a server

conclusions from these assumptions, thousands and sometimes millions of individual load flow simulations have to be conducted. These Monte Carlo-simulations are the basis for predictions about voltage band compliance and the optimal ways to achieve it.

The algorithm has been tested with one-phase systems, although an adaptation to three-phase-systems is possible. If the accuracy of three-phase simulations is required however, the inaccuracies introduced by this method are probably also inacceptable.

Related Work

Cluster analysis on smart meter data has been examined before, albeit with different goals. Chicco et al. used classification to separate electricity customers and to identify typical and recurring consumption behavior [6]. Farinaccio et al. used pattern recognition to find information about the end-uses from the total energy consumption characteristics [7]. Ford and Siraj used the same approach with a focus on data security concerns [8]. The important topic of load forecasting is examined by Alzate and Sinn in [9]. They use kernel spectral clustering and wavelet feature extraction on time series of Irish electricity consumers. While cluster analysis has been recognized as a valuable tool in energy data analysis, its importance will grow with the advent of large-scale metering infrastructures.

Input Data

Smart meters usually deliver data about the electricity consumption of households with high temporal resolution. It is usually given as mean power consumption per time interval, often specified to 15 minutes. In order to use smart meter data not just for billing and statistics but also for technical analysis, the voltage is sometimes also recorded, but rarely transmitted and used. The transmission is usually set to take place in fixed time intervals. For this application, it is not necessary to obtain live data, and there are no requirements for refresh rate.

Smart Meter Databases

Relational databases are usually the first choice of storage container because of compatibility and ease of use. This makes sense when the data is only used in relation to the customer. For a statistical analysis about customer behavior and technical analysis of power flow and voltages however, the computational overhead of data retrieval in relational databases is a hindrance. These computations can be more effectively conducted with a simpler storage scheme like lists of profile values. The connection to specific customers is irrelevant to the algorithm. Additionally, by separating the load profiles from other information about the households, privacy concerns can be minimized.

Power Flow Computation in Distribution Grids

In the transmission grid, power flow computations are normally used to determine the direction and the amount of power that flows through the meshed network. In distribution grids however, the direction is usually obvious, and the voltages at the nodes are more critical. Also, reactive power and reactance characteristics are different, so simple optimizations like fast decoupled load flow do not work. The goal of power flow computation in low-voltage grids, to which smart meters are usually connected, is the evaluation of conformity to voltage regulations and line capacity utilization. Also, the non-deterministic nature of consumers and producers like photovoltaic plants makes it necessary to not only consider individual load situations, but deal with loads of a probabilistic nature. One approach is probabilistic power flow [10], which aims to compute the entire power flow using probabilistic distributions instead of single variables. When the power flow computations also contain controlled elements like photovoltaic plants with voltage control or tap-change transformers however, this is not possible due to the limitations of calculations with probabilistic variables.

In these cases, Monte Carlo simulations are used. These simulations are computationally expensive, as thousands and sometimes millions of individual power flow situations have to be individually evaluated in order to get a converged, valid result. Household loads in these situations need to be modeled as realistically as possible. It makes sense to use measured profiles and randomly assign them to households, but installations with special consumption patterns need to be modeled separately.

Benefits of Cluster Analysis for Power Flow Computation

It is inevitable that during Monte Carlo simulations, many power flow situations occur over and over again. Especially low-load situations which appear for instance at night don't deviate much. When data with repeated sections has to be stored, data compression algorithms can be very effective.

The basic idea of most data compression algorithms is based on the exploitation of unnecessary redundancy. Applied to load data, this amounts to the search for recurring segments in the loads characteristics, in order to save these segments only once and reference them whenever they reappear. Instead of looking for regularities only in the load characteristics, one can also use the corresponding voltages as source data and exploit the regularities in those. This can lead to reductions that were not possible when exclusively using loads as the data source.

The successful application is based on the construction of

a database, a simple table of clusters, which can then also be used as a shortcut during power flow computations. For segments in the cluster table, no power flow computations have to be performed as the results for entire parts of load patterns can be pulled directly from the database.

CLUSTER ANALYSIS

For the application of the database for power flow solutions, a cluster algorithm has to be chosen first. There are multiple variations suitable for different tasks.

Cluster Algorithm

The chosen algorithm is an agglomerative hierarchical cluster analysis algorithm. It first defines every data point as its own cluster and then merges clusters based on their distance. The broadly used k-means algorithm is not applicable because it requires knowledge of the amount of clusters.

Two parameters of the hierarchical clustering algorithm have to be chosen:

Distance function:

The *distance function* or *metric* is a function that defines a scalar distance between elements. Here, the elements are points in high-dimensional space, one dimension for every voltage value included. The simple and intuitive Euclidian metric

$$D_E = \left(\sum_{i=1}^{n} \left| \underline{U}_{1i} - U_{2i} \right|^2 \right)^{1/2}$$

where n is the number of voltage values in the data point, and $\underline{U}_{1i}$ and $\underline{U}_{2i}$ are the voltages (active and reactive components). The result is a scalar value that defines the similarity between two data points that each represent a time series of voltage data in an entire grid.

It is possible to modify this metric, for instance by putting weights on active and reactive components, but the simple metric proved sufficient in tests.

Cluster distance:

In order to obtain a meaningful set of clusters, a maximal distance has to be defined for two points to be in the same cluster. This can be achieved by building a table of distances between all data points or alternatively by using an optimized method like SLINK [11].

The optimal distance to be chosen here is not intrinsically defined by the problem, but has to be empirically discovered and tuned. One way to visualize the data points and have an overview of the clusters is a method called Sammon mapping [12]. This is a projection method that aims to map the multidimensional data onto two dimensions which can be easily plotted and interpreted.

The Sammon mapping of two-hour segments which is shown in Fig. 2 shows several distinct clusters in the data that comprises about 1000 distinct days of real load data. Although not the entire dataset can be assigned to one of the clusters, some of the clusters show enough cohesion to be merged into one representative dataset. This process is repeated for different timeframes. Clusters, i.e. recurring voltage characteristics can be found for up to 8-hour segments.

Automatic methods for valid cluster distance detection can be constructed, but the maximal acceptable distance might differ between simulations. Ultimately, the

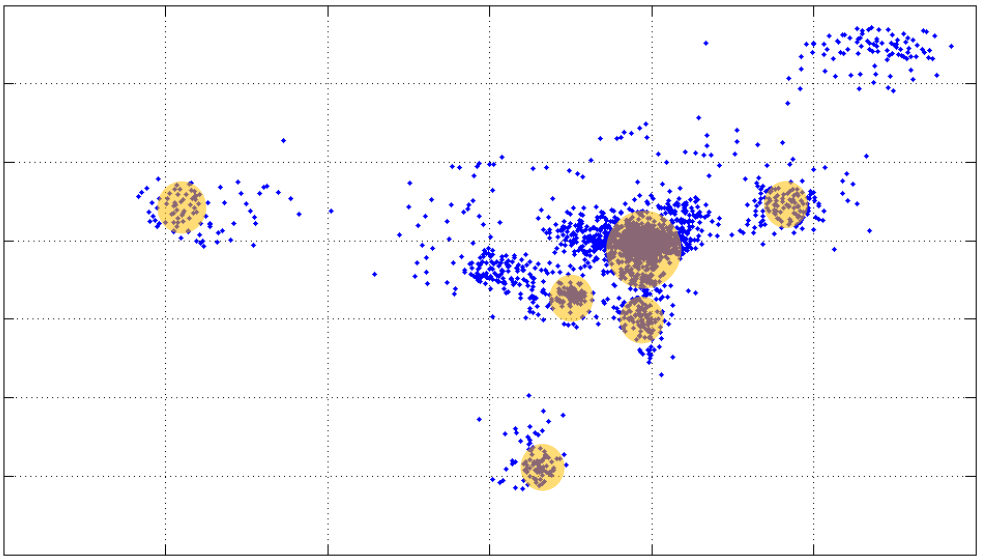

Figure 2: Sammon mapping of voltage segments. Each point represents one two-hour segment. As Sammon mapping aims to preserve the distances between the high-dimensional data into a two-dimensional space, the exact values of the axes have no meaning and are left out

distance is an arbitrary choice and has to be balanced based on the desired compression rate and performance gain, and the acceptable loss in accuracy.

Time Intervals

The number of clusters that can intuitively be identified is largest for two-hour segments. This might be different for bigger sets of input data. It is not necessary for the algorithm to work that the time periods are similar in the times of day they really occurred. On the contrary, looking at an individual clusters, the data points stem from many different times of day and grids. Again, a bigger sample of input data could yield different results.

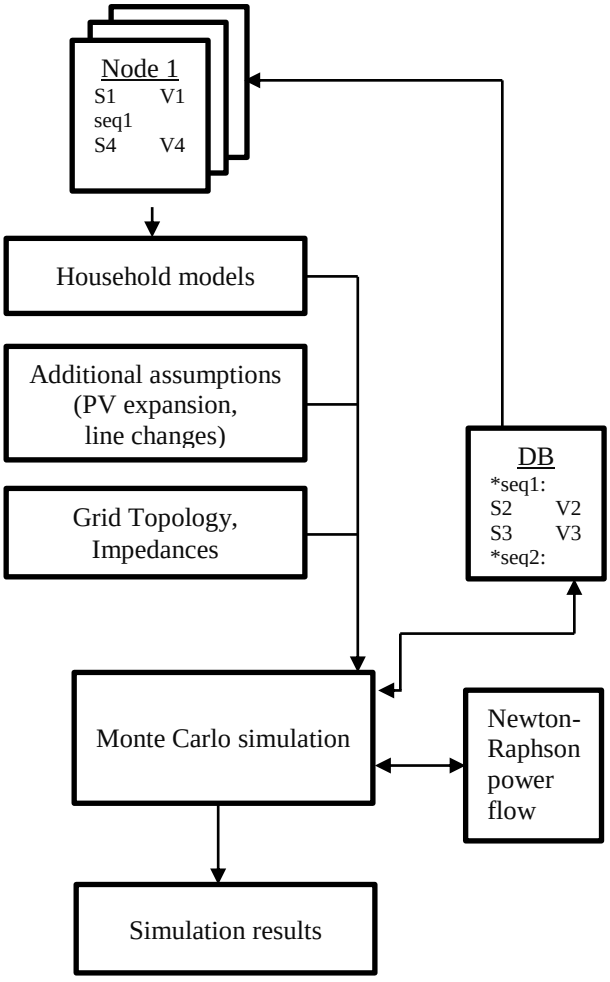

Figure 3: Utilization of database in Monte Carlo simulations

APPLICATION

Available Input Data

The smart meter data available to the authors was recorded over a time period of a year in around 80 households in Germany. As there are many faulty recordings, the data had to be inspected first and implausible data like zero consumption was removed. The algorithm works on bigger timeframes of data, so periodical errors in the data make it unusable. The load data was then compiled into a matrix that separates all load data from additional information about the households.
The interval of load data was uniformly chosen and, if necessary, transformed to be 15 minutes.

Observable Clusters

Using Sammon mapping, some strong clustering can be observed, as pictured in figure 2. The clusters get more diluted with bigger time intervals up to about eight hours, where no obvious cohesion can be identified any more. This means that there are no similar time intervals of eight hours or more in the dataset.

Database Access

During the Monte Carlo simulation, the database built by the clustering algorithm is used to accelerate the power flow computations. The process is outlined in Fig. 3. For every sequence of load data, the database is checked for similar entries, which can be directly applied as the solution. Sequences that cannot be found in the database have to be computed using conventional Newton-Raphson power flow.

Performance

The performance of the algorithm can be defined by the compression rate and the speed-up of Monte Carlo simulations. As model Monte Carlo simulation, a simple voltage distribution evaluation is chosen, which calculates the probabilities of voltages in certain intervals depending on the assumed household models fed by the measured data.
The compression rate of the algorithm depends on the maximal tolerated voltage deviation in the clusters. A rate of 20% can be achieved with errors of < 1 p.u.
The gain in computation speed depends heavily on the input data chosen for the Monte Carlo simulation. Optimally, the algorithm replaces nearly every Newton-Raphson computation by a database lookup, reducing the computation time for a two-hour segment comprised of eight individual computations from around 100 ms to around 0.2 ms, a factor of 500. On average, the conducted simulations ran around 40% faster on a standard PC.
The overhead of the cluster analysis is not negligible. The construction of the reusable database built from 1000 days of consumer load data took around an hour. However, these computations can be continuously performed when the smart meter data arrives.

CONCLUSION

An algorithm that compresses smart meter data based on the voltages in the grid has been presented. It uses hierarchical agglomerative clustering for voltage data to identify redundant data points in different time intervals. The strongest clusters were identified for two-hour-periods, where the amount of data to be stored can be reduced by about 20%, while speeding up power flow computations using the data by about 40% on average. Bigger time intervals promise an even bigger increase in compression and computation speed, but greater inaccuracies and bigger overhead have to be accepted.

ACKNOWLEDGMENTS

The authors would like to thank the Ministry of Science, Research, and the Arts Baden Württemberg in Stuttgart for their support.

REFERENCES

[1] IBM White Paper, 2012, *"Managing big data for smart grids and smart meters"*, http://www-935.ibm.com/services/multimedia/Managing_big_data_for_smart_grids_and_smart_meters.pdf.

[2] Ringwelski, M.; Renner, C.; Reinhardt, A.; Weigel, A.; Turau, V., "The Hitchhiker's guide to choosing the compression algorithm for your smart meter data," *Energy Conference and Exhibition (ENERGYCON), 2012 IEEE International* , vol., no., pp.935,940, 9-12 Sept. 2012.

[3] Tse, Norman C.F.; Chan, John Y.C.; Lai, L L, "Development of a smart metering scheme for building smart grid system," *Advances in Power System Control, Operation and Management (APSCOM 2009), 8th International Conference on* , vol., no., pp.1,5, 8-11 Nov. 2009

[4] Unterweger, A.; Engel, D., "Resumable Load Data Compression in Smart Grids," *Smart Grid, IEEE Transactions on* , vol.PP, no.99, pp.1,1 doi: 10.1109/TSG.2014.2364686

[5] Kattmann, C.; Abdel-Majeed, A.; Tenbohlen, S., "Database-assisted load flow simulation for low voltage grids using a model reduction approach," *PES General Meeting, Conference & Exposition, 2014 IEEE* , vol., no., pp.1,5, 27-31 July 2014 doi: 10.1109/PESGM.2014.6939202

[6] G. Chicco, R. Napoli, and F. Piglione. "Comparisons Among Clustering Techniques for Electricity Customer Classification", *Power*, 21(2):933-940, 2006.

[7] L. Farinaccio and R. Zmeureanu. "Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses, *Energy and Buildings*, 30:245-259, 1999.

[8] Ford, V.; Siraj, A., "Clustering of smart meter data for disaggregation," *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE* , vol., no., pp.507,510, 3-5 Dec. 2013 doi: 10.1109/GlobalSIP.2013.6736926

[9] Alzate, C.; Sinn, M., "Improved Electricity Load Forecasting via Kernel Spectral Clustering of Smart Meters," *Data Mining (ICDM), 2013 IEEE 13th International Conference on* , vol., no., pp.943,948, 7-10 Dec. 2013 doi: 10.1109/ICDM.2013.144

[10] Giraldo, J.S.; Castrillon, J.A.; Mauricio, G.E.; Castro, C.A., "Efficient probabilistic power flow for weakly-meshed distribution networks," *Transmission & Distribution Conference and Exposition - Latin America (PES T&D-LA), 2014 IEEE PES* , vol., no., pp.1,6, 10-13 Sept. 2014 doi: 10.1109/TDC-LA.2014.6955231

[11] R. Sibson: SLINK: an optimally efficient algorithm for the single-link cluster method, In: *The Computer Journal*. 16, Nr. 1, British Computer Society, 1973, S. 30-34

[12] Sammon JW (1969). A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers* 18: 401–409.