Reducing Energy Time Series for Energy System Models via Self-Organizing Maps

This is a post-peer-review, pre-copyedit version of an article available at www.degruyter.com, DOI: 10.1515/itit-2019-0025

Hasan Ümitcan Yilmaz, Edouard Fouché, Thomas Dengiz, Lucas Krauß, Dogan Keles, Wolf Fichtner

Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Abstract: The recent development of renewable energy sources (RES) challenges energy systems and opens many new research questions. Energy System Models (ESM) are important tools to study these problems. However, including RES into ESM strongly increases the model complexity, because one needs to model the fluctuant, weather-dependent electricity production from RES with a high level of granularity. This leads to long execution times. To deal with this issue, our objective is to reduce the input time series of ESM without losing their energy-related key characteristics, such as weather-dependent fluctuations in production or peak demands. This task is challenging, because of the variety and high-dimensionality of the data. We describe a carefully engineered data-processing pipeline to reduce energy time series. We use Self-Organizing Maps, a specific kind of neural network, to select "representative days". We show that our approach outperforms the existing ones with respect to the quality of ESM results, and leads to a significant reduction of ESM execution times.

ACM CCS:
Computing methodologies → Cluster analysis; Applied computing → Forecasting

Keywords: Energy System Models, Self-Organizing Maps, Data Reduction, Forecasting

1 Introduction

Renewable energy sources (RES) will soon have the largest share of electricity production in many countries. These sources, such as wind and photovoltaics (PV), are weather-dependent – a challenge for energy systems and markets. To study the development of energy systems, so-called Energy System Models (ESM) have been proposed. In recent years, their focus is on understanding the impact of fluctuating electricity production [1].

In a nutshell, an ESM simulates parts of an energy system – or an entire one – in order to study various research questions. Thereby, a broad spectrum of techniques are used for modeling. Mathematical optimization is often used to determine the minimum system costs under various constraints (e.g., political or environmental) for a given geographical area and a pre-defined time horizon, which can range from days to decades, and the area may be arbitrarily large. An ESM outputs a solution in terms of financial costs, greenhouse gas emissions, use of natural resources, energy efficiency, etc.

ESMs can be complex, and they do not scale well with high-resolution data [2]. On the one hand, large input data tends to lead to long execution times, i.e., hours to days. On the other hand, naive data reduction, e.g., via aggregation, reduces the quality of model results by much. This is because energy-related key characteristics, such as fluctuations or peak demands, are lost.

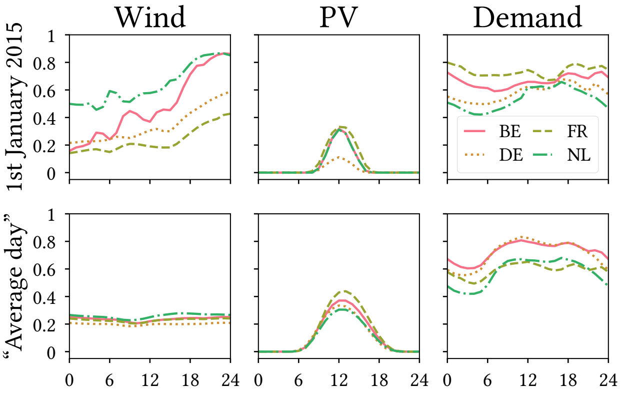

Figure 1: Wind, PV and Electricity Demand Profiles

Figure 1 graphs the normalized electricity production from wind and PV and the demand in four countries on the first day of 2015. PV shows a Gaussian-like profile, of varying mean and amplitude. Wind in turn is rather irregular and fluctuant. The lower part shows the "average day" for that year. Averaging days together – known as the "typical day" approach [3, 2] – is a common way to reduce the input data for ESM. By comparing the wind profiles, one can see that averaging loses much information: One can no longer observe any fluctuation in electricity production from wind. This kind of reduction jeopardizes the output of ESMs, since the RES data is not realistic anymore. However, one still needs to reduce the volume of data to keep ESMs solvable within acceptable execution times.

Consequently, reducing the execution time of ESMs by reducing energy data without destroying the energy-related key characteristics of the data is challenging. Several methods to reduce energy data have been recently proposed [4, 5, 6]. They all target at selecting days which are "representative" of the data.

However, when extending the analysis to larger scales, e.g., to the European level, or to more fine-granular data, we find that existing methods are ineffective.

For example, let us assume that one wants to select a day from a year, when the data consists of three energy time series from four countries, with one measurement per hour. The problem is equivalent to selecting a point with $3*4*24 = 288$ dimensions from a set of 365 points. In other words, the problem is "high-dimensional".

Existing state-of-the-art methods, relying on distance-based clustering, fall prey to the so-called "curse of dimensionality" [7], and the selection of days is only marginally better than random.

Our main contribution is proposing a carefully engineered data-processing pipeline to reduce high-dimensional energy time series. It leads to superior results compared to the state of the art. We also show that Self-Organizing Maps (SOM) [8], a prominent type of neural network, helps in our specific context. SOM gives way to good ESM results with a high reduction in execution times.

2 Related Work

Clustering algorithms have been widely used to reduce time series for ESMs. Heuberger et al. [9] use k-means to reduce the size of a raw data set for their generation expansion model, and Green et al. [10] use it to cluster electricity demand and wind generation data. They conclude that reducing the input data to 10 days (10 clusters) yields a good trade-off between execution time and accuracy. Nahmmacher et al. [4] use Ward's hierarchical clustering for a long-term model of the European electricity system. They claim that 6 representative days are sufficient for good model results. ElNozahy et al. [5] combines principal component analysis with k-means, fuzzy c-means and hierarchical clustering. Based on their own internal validity index, k-means is the best. Pfenninger [3] and Kotzur et al. [6] compare heuristic day selection, down-sampling and clustering methods for the selection of representative days. Their results suggest that the best method heavily depends on the input data and the model constraints.

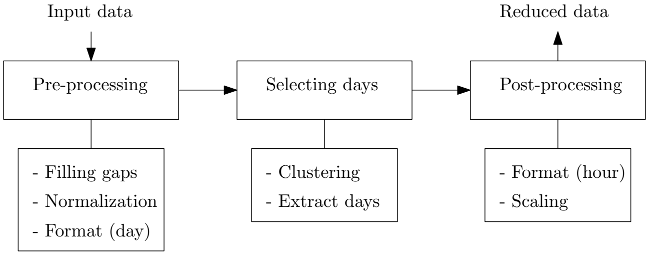

Figure 2: Schema of our pipeline

While averaging is outperformed by all other methods, no approach yields the best results for all problems. A crucial drawback of the methods used so far is that they are ineffective with high-dimensional data. The reason is that the sparsity of the data increases exponentially with increasing number of dimensions such that, for metrics based on all coordinates, e.g., the Euclidean distance, the relative differences in the distances between each pairs of points become dramatically smaller and smaller [11, 12]. In this context, SOM seems to be a particularly promising approach. Bação et al. [13] compare k-means to SOM and conclude that SOM is superior as it leads to a better exploration of the search space. Mangiameli et al. [14] show that it yields higher accuracy than hierarchical clustering. SOM was applied in Astel et al. [15] to cluster a large data set of chemical indicators and in Park et al. [16] to select representative species for ecological communities.

In comparison, this study is – to our knowledge – the first to propose a generic pipeline to reduce energy time series, and to consider SOM as a means to reduce reducing energy data.

3 Pipeline for Data Reduction

This section describes our pipeline for reducing the input time series for ESMs. Section 3.1 describes the preprocessing of input data by filling the gaps, normalizing and changing the format. Then, Section 3.2 explains how we use this prepared data as input for clustering and then select the representatives from the resulting clusters. Finally, Section 3.3 presents the post-processing of data. The reduced data can then be used as input for ESMs. Figure 2 serves as a summary. The pipeline relies on a given algorithm to cluster the data. However, our approach is "generic", i.e., it is independent from which algorithm and underlying data set one uses.

Let $C = \{C_1, \ldots, C_m\}$ be a set of m countries and $S = \{S_1, \ldots, S_n\}$ a set a n time series types (e.g., "electricity production from PV", "electricity demand"). In our case, the input data consists of a set of time series at hourly resolution from $C \times S$, as in Figure 3.

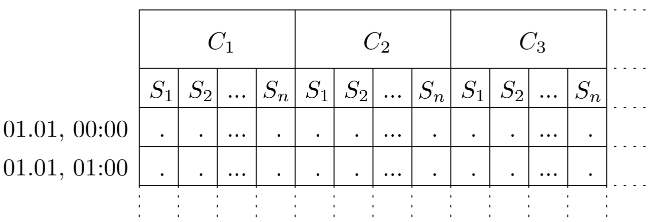
C_1
C_2
C_3
S_1 S_2 ... S_n S_1 S_2 ... S_n S_1 S_2 ... S_n
01.01, 00:00
01.01, 01:00

Figure 3: Data in hourly format

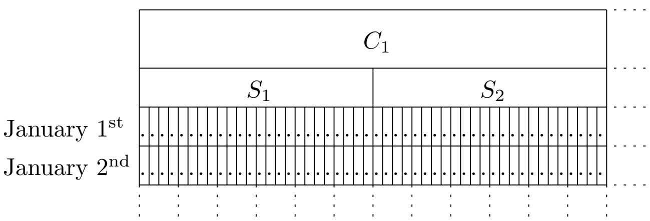

Figure 4: Data in daily format

3.1 Pre-Processing

Filling Gaps: Real-world data typically has missing values – our case is no exception. We replace missing values, i.e., "gaps", in time series by linear interpolation between both ends of the gap.

Normalization: To compare values between countries, we normalize the input time series. We divide the electricity production time series by the installed capacities, and the electricity demand by the maximal peak demand in the related country. This way, the time series are scaled to $[0, 1]$.

Format (day): We format the data, such that each instance (i.e., row) stands for an entire day. An instance contains all values of all time series for all countries. Figure 4 illustrates this format. We use it in the next pipeline step for clustering. Here, the dimensionality of the data increases inevitably. In our case, the number of dimensions is multiplied by a factor of 24, i.e., the number of values per day.

3.2 Selecting Days

Clustering: We cluster the pre-processed data. We assume that the output of the clustering simply is a set of n cluster centers in the input space. Our main innovation here is to use Self-Organizing Maps, which we describe in detail in Appendix 7.2.

Extracting Days: As in [4, 6], we select the closest day to each cluster center as a "representative". Then, we weight each representative using the number of days contained in their respective clusters. Representatives from larger clusters have larger weights in the modeling step, as they account for a larger share of the data.

3.3 Post-Processing

Format (hour): We format the data back to its original shape (Figure 3). Thus, the data does not have any missing values, it is normalized and much smaller than the raw data.

Scaling: We scale the reduced data so that the sums of each original time series equal the sums of each reduced series weighted by the number of days for each representative. Skipping this step may lead to biased ESM results. For example, if the total PV electricity production in the reduced data exceeds the one in the original time series, the ESM may favor investments in PV.

4 Evaluation

We evaluate the benefits of our method w.r.t. different baselines (random and "typical days" approaches) and clustering algorithms:

- **Random**: Select n days randomly from input data.
- **Typical days (1)**: Divide the input data into n (quasi-)equal sets of consecutive days, and average them together.
- **Typical days (2)**: Average week days, Saturdays and Sundays for each season, producing 12 representative days.
- **HC-DTW**: Hierarchical Clustering based on Dynamic Time Warping distance, as in [17]. We choose the level with n clusters from the resulting hierarchical cluster structure and select the days closest to the cluster centers.
- **Ward**: Ward's hierarchical clustering algorithm [18]. We proceed similarly as with HC-DTW. The only difference is the distance measure used for clustering.
- **K-means**: The well-known partitioning clustering algorithm, which was used in many other studies [19].
- **SOM**: We adapt SOM as a clustering algorithm by considering each neuron in the grid after training as a cluster center.

All in all, each method outputs a reduced number of days from the input data. The only difference is how they choose these days. Random, k-means and SOM are not deterministic. Therefore, their results in our experiments are averaged over 10 executions.

4.1 Evaluation Methodology

We use the PERSEUS-EU [20, 21, 22] model as an exemplary ESM, which we review in Appendix 7.1. ESMs use in general similar approaches [1]. Therefore, we expect the results to be comparable with other ESMs.

In our case, the input time series are the electricity production profiles from RES (wind onshore and PV) and the electricity demand of four neighbouring countries: Germany, France, Belgium and the Netherlands. We obtain the data for 2015 from ENTSOE [23].

We follow the so-called "greenfield" [24] investment approach, i.e., we assume that there are no existing power plants, and that ESM can invest optimally in a power plant portfolio. The ESM is executed with three different configurations:

- **Configuration 1:** Consider a time horizon of 40 years, under the constraint that RES must represent at least 30% of the total electricity production in the first period, then must increase linearly until 80%.
- **Configuration 2:** Consider a time horizon of 20 years, under the constraint that RES must represent at least 30% of the total electricity production in the first period, then must increase linearly until 50%.
- **Configuration 3:** Consider a time horizon of 5 years, under the constraint that investment in renewables are not allowed. In this configuration, we use the so-called "net demand" (the total demand minus the renewable electricity production) as an input for data reduction and the ESM.

To evaluate the quality of a data reduction approach, we compare the output of the ESM model with the reduced data *Red* and the full input data (reference case *Ref*). As an "error" metric, we use the relative deviation of the ESM production mix, and the deviation is calculated for each country c, each model period p and each power plant technology t in the output:

$$\text{Error} = \frac{\sum_c \sum_p \sum_t \left| Red_{c,p,t} - Ref_{c,p,t} \right|}{\sum_c \sum_p \sum_t Ref_{c,p,t}} \tag{1}$$

4.2 Results

We evaluate against different numbers of days, namely, we run each approach for 4, 9, 16, 25, 36, 49, 64, 81 and 100 representative days. The experiment consists of the execution of hundreds of model instances and after 100 days the ESM execution time becomes too long. Therefore, we analyze up to 100 days to limit the computation time required for the experiment.

Figure 5 shows that the error decreases with increasing number of days for each configuration and algorithm, and that SOM is consistently better. Next, we see that the results from the "Typical Days" approaches are not much better than Random.

In Configuration 2, the error is slightly lower for every approach, except for Random, compared to Configuration 1, due to the lower complexity of the ESM.

In Configuration 3, the error is much lower overall, because the ESM is much less complex. Here, the input data is different and has fewer dimensions (96 instead of 288), thus the effects of the curse of dimensionality are not that distinct. Therefore, k-means and Wards also perform well.

Next, we show that our results hold for different SOM configurations. We vary between an hexagonal (H) and rectangular (R) topology, between a "gaussian" (G) and

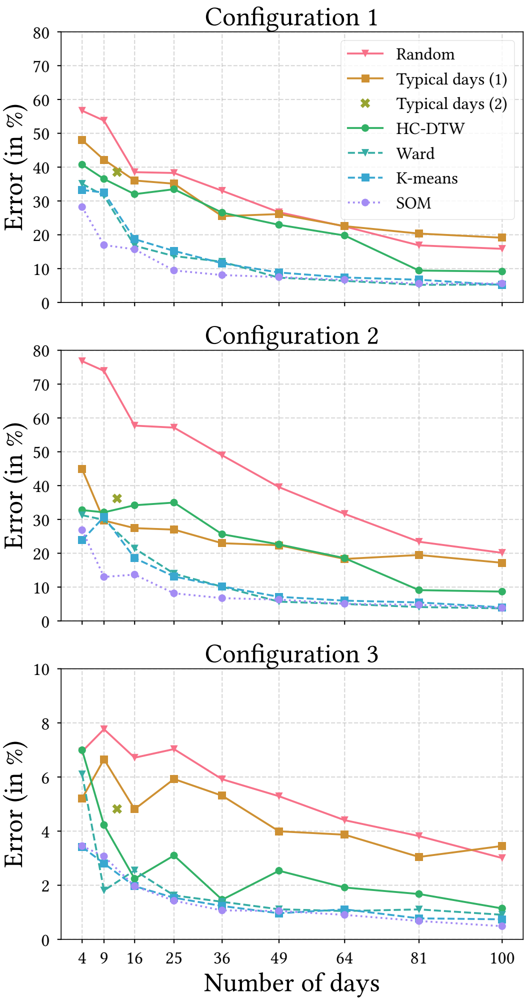

Figure 5: Error in the ESM results with Configuration 1,2,3

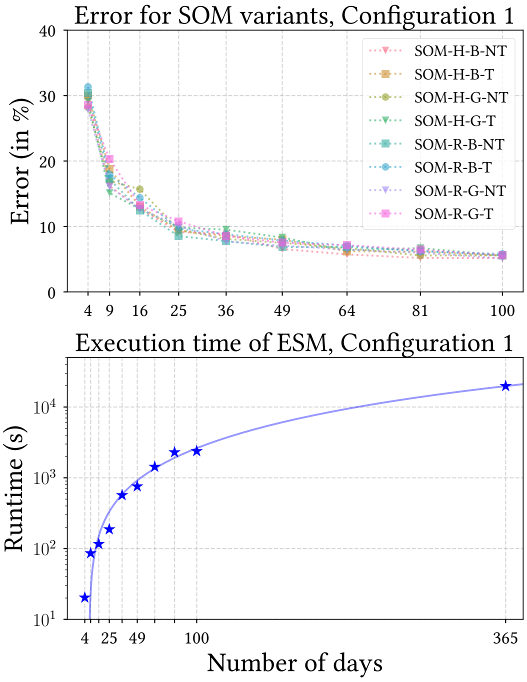

Figure 6: Error for variants of SOM and execution time

“bubble” (B) neighborhood function, and between a torroidal (T) and non-torroidal grid (NT). This accounts for 8 SOM parameter combinations in total. In Figure 6, we see that each combination performs well, and that the difference between them is not very large. This means that our results hold quite independently from SOM’s (hyper-)parameters.

Finally, Figure 6 graphs the benefits of reducing energy time series on the execution time of the ESM. For Configuration 1, selecting 25 days leads to less than 10% error in the output for an execution time two orders of magnitude shorter, compared to the reference. After 64 days, the error in the ESM results reduces very slowly. We ran our experiments on a server with 72 cores (Intel® Xeon® Processor E5-2697 2.30GHz) and 300GB RAM. Each ESM instance runs with 10 parallel threads.

5 Conclusion

In this paper, we have described a data-processing pipeline to reduce energy time series for energy system models. Our results show that this pipeline yields better results than common heuristics used in the field. Next, Self-Organizing Maps as a clustering algorithm yields better results than any other approach. We hypothesize that the topology-preserving property of SOM [8] allows to overcome the effects of the curse of dimensionality in this particular setting more effectively. Our data reduction approach leads to very good ESM results for a much shorter runtime, allowing the use of ESMs for larger scale energy system analyses which were intractable so far.

6 Discussion and Future Work

Traditionally, the input data for ESM must also include the extreme cases affecting the grid, such as times of very high electricity demand or very low solar and wind power production, because energy systems need to satisfy the demand at all times. Still, with clustering, there is no guarantee that those cases – which one can see as “outlier” from a data perspective – are part of the reduced series. Thus, in the future, it would be interesting to analyze the clustering results to understand why our results are good, and if necessary, extend them to become aware of those outliers. In addition, one could assess the impact of our approach on the other ESM outputs and could analyze the results in more detail, e.g., with respect to investment in individual power plant technologies. Furthermore, one could use other ESMs to confirm our results and the validity of our approach. Lastly one could consider other approaches to deal with high dimensionality, e.g., dimensionality reduction methods such as principal component analysis.

7 APPENDIX

7.1 Exemplary ESM

PERSEUS-EU [20, 21, 22] is an optimization model for the electricity sector of 28 European countries with a multi-periodic linear optimization approach. PERSEUS-EU is used in particular for analyzing the impact of changing framework conditions caused by political or environmental reasons, with the objective to minimize total system costs under a set of technical, ecological and political constraints. Examples of important cost parameters are fuel costs for electricity generation, variable and fixed operating costs of power plants as well as fixed capital costs of new generation units.

The main decision variables of the optimization model are the production level of the existing and new capacities, investment in new capacities and energy exchanges between neighboring countries. In addition to future capacity and production mix, the model outputs – among others – details on primary energy mix, cross border exchanges, emissions in each country and marginal costs of electricity generation.

The model structure relies on a directed graph, where the nodes are connected to each other through energy flows. At the system nodes, several energy conversion technologies (e.g. power plants) are available. The source node provides fuel imports to the graph, while sink node contains the energy demand that is to be served through the inflows to this node. Exchange flows represent the electricity exchange between the system nodes (e.g. between European countries). A main restriction is the flow balance for each node.

To reduce the model complexity, generally, periods with duration of five years are selected and each 5 years is represented by a characteristic year. In addition, each characteristic year has an intra-year time resolution with several model time slices. The reduced intra-yearly time structure must represent the whole year.

For our experiments, we simplify the PERSEUS-EU model to limit the required computation time, since they involve the executions of hundreds of model instances. Still, we model the core restrictions of an ESM via the following equations:

7.1.1 Objective function

$$
\begin{aligned}
min \sum_{y\in Y} (\frac{1}{1+r})^{y} \cdot \\
\left(\begin{aligned} &\sum_{so\in SO}\sum_{no\in NO}\sum_{ec\in EC} FL_{so,no,ec,y}\cdot c^{fuel}_{so,no,ec,y} \\ &+\sum_{u\in U}(K_{u,y}\cdot c^{fix}_{u,y} + K^{new}_{u,y}\cdot c^{inv}_{u,y}) \\ &+\sum_{pc\in PC}\begin{pmatrix} PL_{pc,y}\cdot c^{var}_{pc,y} \\ +\sum_{t\in T} LV_{pc,y,t-1,t} \quad \cdot c^{lv}_{pc}\end{pmatrix} \end{aligned} \right)
\end{aligned} \quad (2)
$$

7.1.2 Energy balance restrictions

$$\sum_{no' \in NO} FL_{no',no,el,y,t} + \sum_{pc \in PC_{no}} PL_{pc,y,t} = \sum_{no' \in NO} \frac{FL_{no,no',el,y,t}}{\eta_{no,no',el,y}} \quad (3)$$

$$\forall no \in NO^{sys}, \forall y \in Y, \forall t \in T$$

$$\sum_{no' \in NO} FL_{no',no,ec,y,t} + \sum_{pc \in PC_{no}} PL_{pc,y,t} \cdot \lambda^{prod}_{pc,ec} = \sum_{no' \in NO} \frac{FL_{no,no',ec,y,t}}{\eta_{no,no',ec,y}} + \sum_{pc \in PC_{no}} \frac{PL_{pc,y,t} \cdot \lambda^{cons}_{pc,ec}}{\eta_{pc,y}} \quad (4)$$

$$\forall no \in NO^{sys}, \ \forall ec \in EC, \ \forall t \in T, \ \forall y \in Y$$

$$PL_{pc,y} = \sum_{t \in T} PL_{pc,y,t} \qquad \forall pc \in PC, \ \forall y \in Y \quad (5)$$

$$FL_{no',no,el,y} = \sum_{t \in T} FL_{no',no,el,y,t} \quad (6)$$

$$\forall no', no \in NO^{sys}, \ \forall y \in Y$$

7.1.3 Capacity restriction

$$K_{u,y} \cdot av_{u,y,t} \geq \sum_{pc \in PC_u} PL_{pc,y,t} \quad (7)$$

$$\forall u \in U, \ \forall y \in Y, \ \forall t \in T$$

$$K_{u,y} = \kappa_{u,y} + \sum_{y'=y-\tau_u}^{y} K^{new}_{u,y'} \qquad \forall u \in U, \ \forall y \in Y \quad (8)$$

7.1.4 Load variation restriction

$$LV_{pc,y,t-1,t} = \left| \frac{PL_{pc,y,t}}{h_t} - \frac{PL_{pc,y,t-1}}{h_{t-1}} \right| \cdot tr_{t-1,t} \quad (9)$$

$$\forall pc \in PC$$

7.1.5 Expansion targets for RES

$$\sum_{pc \in PC^{re}} PL_{pc,y} = \alpha_y \cdot \sum_{pc \in PC} PL_{pc,y} \qquad \forall y \in Y \quad (10)$$

7.1.6 List of Symbols

$ec \in EC$	Energy carriers
$EC^{re} \subset EC$	Renewable energy carriers
$el \in EC$	Electricity
$no \in NO^{sys}$	System Nodes
$pc \in PC$	Production Processes
$PC^{re} \subset PC$	Production processes from RES
$si \in SI \in NO$	Sinks of the graph structure
$so \in SO \in NO$	Sources of the graph structure
$t \in T$	Time slices
$u \in U$	Units
$y \in Y$	Years
α_y	Electricity production target of RES in year y
$\beta_{ec,no,y}$	Capacity expansion target of node n for energy carrier ec in y
$\eta_{n,n',ec}$	Flow efficiency of the flow between n and n' for ec
$\eta_{pc,y}$	Efficiency of process pc in year y
$\lambda_{pc,ec}$	Share of produced or consumed energy carrier ec of process pc
τ_u	Physical lifetime of unit u
$av_{u,y,t}$	Availability of u in y at time t
$c^{fix}_{u,y}$	Fixed annual operation costs of unit u in year y
$c^{fuel}_{ec,y}$	Fuel costs of ec in year y
$c^{inv}_{u,y}$	Annuitised investment expenditures for commissioning u in year y
c^{lv}_{pc}	Load variation costs of process pc
$c^{var}_{pc,y}$	Variable operating costs of pc in y
h_t	Number of hours in time slice t
$k^{exist}_{u,y}$	Initial capacity of unit u in year y
r	Discount rate of future cash flows
sc_u	Secured capacity of unit u
sf	Security factor for security of supply
$tr_{t-1,t}$	Number of transitions between time slices $t-1$ and t
$FL_{no,no',ec,y,t}$	Flow level of energy carrier ec between no' and no at time t in y
$FL_{no,no',ec,y}$	Flow level between no' and no in y
$K^{new}_{u,y}$	Newly installed capacity of u in y
$K_{u,y}$	Capacity of unit u in year y
$LV_{pc,y,t-1,t}$	Load variation of process pc between time $t-1$ and t in year y
$PL_{pc,y,t}$	Production level of pc in y at time t
$PL_{pc,y}$	Production level of pc in year y

7.2 Self-Organizing Maps

Self-Organising Maps (SOM) are a specific type of artificial neural networks traditionally used for dimensionality reduction and visualization of high-dimensional data [8]. It is a projection of a set of n-dimensional points to a two-dimensional neuron grid. The SOM is a $p \times q$ neuron matrix, associating each neuron to a n-dimensional weight vector $w_{ij}, i \in \{1, \ldots, p\}, j \in \{1, \ldots, q\}$. Figure 7 illustrates the architecture of the SOM. Training the

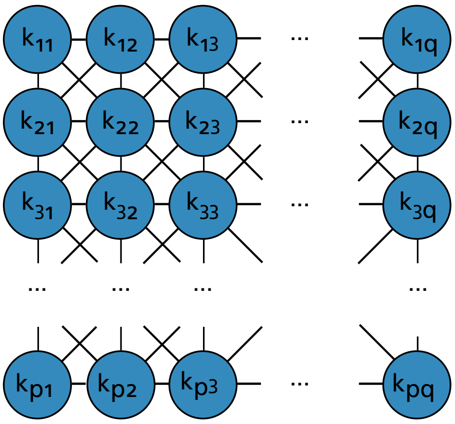

Figure 7: Architecture of a Self-Organising Map

SOM happens in an iterative way. Let $w_{ij}(t)$ denote the weight vectors after iteration t, $w_{ij}(0)$ the initial weight vectors, $\alpha(t)$ a learning rate decreasing with t and $h_{ij}(t)$ a neighborhood function instantiated as a smoothing kernel whose width decreases with t. At each iteration, we randomly select an object x_t and update the weight vector of each neuron as follows:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)h_{ij}(t)(x_t - w_{ij}(t)) \quad (11)$$

until the model has converged. Then, for our data reduction approach, we select as a representative the closest day to each neuron. We consider only square grids, i.e., $p = q$. This way, by setting $p = 3, 4, 5, \dots$ we obtain $n = 9, 16, 25, \dots$ representatives.

Acknowledgement

This work was partially supported by the DFG Research Training Group 2153: "Energy Status Data - Informatics Methods for Its Collection, Analyses and Exploitation" and the BEAM-ME Project founded by the Federal Ministry for Economic Affairs and Energy.

Literature

[1] "A review of computer tools for analysing the integration of renewable energy into various energy systems," *Applied Energy*, vol. 87, no. 4, pp. 1059 – 1082, 2010.

[2] K. Poncelet, E. Delarue, D. Six, J. Duerinck, and W. D'haeseleer, "Impact of the level of temporal and operational detail in energy-system planning models," *Applied Energy*, vol. 162, pp. 631–643, 2016.

[3] S. Pfenninger, "Dealing with multiple decades of hourly wind and pv time series in energy models: A comparison of methods to reduce time resolution and the planning implications of inter-annual variability," *Applied Energy*, vol. 197, pp. 1–13, 2017.

[4] P. Nahmmacher, E. Schmid, L. Hirth, and B. Knopf, "Carpe diem: A novel approach to select representative days for long-term power system modeling," *Energy*, vol. 112, pp. 430–442, 2016.

[5] M. S. ElNozahy, M. M. A. Salama, and R. Seethapathy, "A probabilistic load modelling approach using clustering algorithms," in *IEEE Power and Energy Society general meeting (PES), 2013.* Piscataway, NJ: IEEE, 2013, pp. 1–5.

[6] L. Kotzur, P. Markewitz, M. Robinius, and D. Stolten, "Impact of different time series aggregation methods on optimal energy system design," *Renewable Energy*, vol. 117, pp. 474–487, 2018.

[7] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*, R. E. Bellman, Ed. Princeton University Press, 1961.

[8] T. Kohonen, *Self-Organizing Maps*, ser. Springer Series in Information Sciences. Berlin, Germany: Springer Berlin Heidelberg, 1995.

[9] C. F. Heuberger, I. Staffell, N. Shah, and N. M. Dowell, "A systems approach to quantifying the value of power generation and energy storage technologies in future electricity networks," *Computers & Chemical Engineering*, vol. 107, pp. 247–256, 2017.

[10] R. Green, I. Staffell, and N. Vasilakos, "Divide and conquer? *k*-means clustering of demand data allows rapid and accurate simulations of the british electricity system," *IEEE Transactions on Engineering Management*, vol. 61, no. 2, pp. 251–260, 2014.

[11] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *Database Theory - ICDT 1999*, ser. Lecture Notes in Computer Science, G. Goos, J. Hartmanis, J. van Leeuwen, C. Beeri, and P. Buneman, Eds. Berlin, Heidelberg: Springer Berlin / Heidelberg, 1999, vol. 1540, pp. 217–235.

[12] P. Indyk and R. Motwani, "Approximate nearest neighbors," in *Proceedings of the Thirtieth annual Symposium on the Theory of Computing*, J. Vitter, Ed. New York, NY: Assoc. for Computing Machinery, 1998, pp. 604–613.

[13] F. Bação, V. Lobo, and M. Painho, "Self-organizing maps as substitutes for k-means clustering," in *Computational science - ICCS 2005*, ser. Lecture Notes in Computer Science, G. D. van Albada, J. J. Dongarra, P. M. A. Sloot, and V. S. Sunderam, Eds. Berlin [u.a.]: Springer, 2005, vol. 3516, pp. 476–483.

[14] P. Mangiameli, S. K. Chen, and D. West, "A comparison of som neural network and hierarchical clustering methods," *European Journal of Operational Research*, vol. 93, no. 2, pp. 402–417, 1996.

[15] A. Astel, S. Tsakovski, P. Barbieri, and V. Simeonov, "Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets," *Water research*, vol. 41, no. 19, pp. 4566–4578, 2007.

[16] Y.-S. Park, J. Tison, S. Lek, J.-L. Giraudel, M. Coste, and F. Delmas, "Application of a self-organizing map to select representative species in multivariate analysis: A case study determining diatom distribution patterns across france," *Ecological Informatics*, vol. 1, no. 3, pp. 247–257, 2006.

[17] Y.-J. Weng and Z.-Y. Zhu, "Time series clustering based on shape dynamic time warping using cloud models," in *2003 International Conference on Machine Learning and Cybernetics.* Piscataway, NJ: IEEE, 2003, pp. 236–241.

[18] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, p. 236, 1963.

[19] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics.* Berkeley, Calif.: University of California Press, 1967, pp. 281–297.

[20] H. U. Heinrichs, *Analyse der langfristigen Auswirkungen von Elektromobilität auf das deutsche Energiesystem im europäischen Energieverbund.* Karlsruhe, Germany: KIT Scientific Publishing, 2014.

[21] S. Babrowski, P. Jochem, and W. Fichtner, "How to model the cycling ability of thermal units in power systems," *Energy*, vol. 103, pp. 397 – 409, 2016.

[22] S. Babrowski, H. Heinrichs, P. Jochem, and W. Fichtner, "Load shift potential of electric vehicles in europe," *Journal of Power Sources*, vol. 255, pp. 283 – 293, 2014.

[23] ENTSOE, "Central collection and publication of electricity generation, transportation and consumption data and information for the pan-european market." visited 2019-02-10.

[24] M. Geidl, P. Favre-Perrod, B. Klöckl, and G. Koeppel, "A greenfield approach for future power systems," *41st International Conference on Large High Voltage Electric Systems 2006, CIGRE 2006*, 01 2006.

Dipl.-Inf. Hasan Ümitcan Yilmaz is a Ph.D. candidate at the chair of Energy Economics of the Karlsruhe Institute of Technology (KIT). He holds a Diplom degree in Computer Science from KIT. His main research topics include energy system modeling and the decarbonization of the European energy system.

Address: Karlsruhe Institue of Technology (KIT), Germany,
E-Mail: hasan.yilmaz@kit.edu

M.Sc. Edouard Fouché is a Ph.D. candidate in Data Mining at the Karlsruhe Institute of Technology (KIT), under the supervision of Prof. Dr.-Ing. Klemens Böhm. He holds a master's degree in Computer Science from KIT and a master's degree in Engineering from ESIEE Paris. His research focuses on Correlation Analysis, Bandit Algorithms, Outlier Detection and Clustering.

Address: Karlsruhe Institue of Technology (KIT), Germany,
E-Mail: edouard.fouche@kit.edu

M.Sc. Thomas Dengiz is a Ph.D. candidate at the chair of Energy Economics of the Karlsruhe Institute of Technology (KIT). He holds a master's degree in Industrial Engineering from KIT. His main research topics include the design of algorithms for smart grids and analyzing the flexibility of electric heating devices in buildings.

Address: Karlsruhe Institue of Technology (KIT), Germany,
E-Mail: thomas.dengiz@kit.edu

B.Sc. Lucas Krauß is a M.Sc. candidate in Computer Science at the Technische Universität Berlin. He is interested in scalable and online machine learning.

Address: Karlsruhe Institue of Technology (KIT), Germany,
E-Mail: lucas.krauss@campus.tu-berlin.de

Dr. Dogan Keles graduated as industrial engineer from the Karlsruhe Institute of Technology (KIT) in 2006 and received his doctoral degree at the Economics Department of KIT in 2013 with summa cum laude. During his work at the KIT he analysed uncertainties in energy markets and developed methods to evaluate energy investments. During his research visit to University of California Berkeley in 2011, he worked on stochastic modeling. During his Senior Research Fellowship at Durham University in 2019, he carried out research on the effect of RES on electricity markets and designing systems with large shares of renewables. Currently, Dogan Keles is head of the research group "energy markets and energy system analysis" at the Institute of Industrial Production (IIP) at the KIT and works on different projects on the design of energy markets, evaluation of energy technologies (under uncertainty) and modelling of energy systems. His studies resulted in different publications in highly ranked journals and peer-reviewed conference proceedings.

Address: Karlsruhe Institue of Technology (KIT), Germany,
E-Mail: dogan.keles@kit.edu

Prof. Dr. Wolf Fichtner is Director of the Institute for Industrial Production (IIP) and the French-German Institute for Environmental Research (DFIU) of the Karlsruhe Institute of Technology (KIT). His main research topics include Energy Systems Analysis and Energy Modelling.

Address: Karlsruhe Institue of Technology (KIT), Germany,
E-Mail: wolf.fichtner@kit.edu