Energy Strategy Reviews 25 (2019) 75–85

Contents lists available at ScienceDirect

Energy Strategy Reviews

journal homepage: www.elsevier.com/locate/esr

Impact of the choice of regions on energy system models

Kais Siala*, Mohammad Youssef Mahfouz

Technical University of Munich, Chair of Renewable and Sustainable Energy Systems, Germany

ARTICLE INFO

Keywords:
Spatial clustering
Energy system model
Expansion planning
Optimization
GIS

ABSTRACT

Many existing energy system models rely on input data available at country-level, or at the level of administrative divisions. However, there is usually no correlation between the distribution of data such as solar radiation, wind speed, and electrical load on one hand, and the administrative divisions on the other hand. The goal of the research is to measure the impact of the shape of model regions on the results of system optimization models. A novel clustering methodology for high-resolution data is presented and applied to define new regions for an energy system model which optimizes expansion planning and unit commitment. The new model regions take into account the bottlenecks in the transmission system and their effect on the expansion of renewable energy sources. We compare the obtained energy mixes, new capacities, and curtailment levels against a model using administrative divisions. The results show discrepancies between the models in the case of a high share of variable renewable energy, and quantify the impact of the distribution of load, wind and solar resources on energy system models. Possible applications of the new model regions are discussed to emphasize their utility for modelers and policy-makers.

1. Introduction

The electricity system in Europe is increasingly relying on decentralized generation from variable renewable energy sources. The conventional power plants that used to be conveniently located next to load centers are being replaced with wind and solar power plants in areas with high renewable potential. This trend is set to continue if the decarbonization targets are to be achieved without relying heavily on nuclear and hydro power generation.

The increased reliance on geographically distributed, time-dependent renewable energy generation requires a deep understanding of the spatial and temporal distribution of wind and solar resources, and their eventual correlation with load patterns [1]. This is helpful to determine the system flexibility requirements, such as the need for new transmission lines or for storage devices. Thanks to improvements in the processing power, it is now possible to solve multi-regional optimization problems with various wind and solar potentials. The advantage of using a high number of model regions has been quantified in previous studies [2]. However, the definitions of the model regions has been mostly dictated by the political boundaries of countries and their administrative subdivisions, especially if the models span over many states or support policy-makers on an international level [2,3]. The lack of diversity in the shape of model regions might be partly due to the fact that the input data of energy system models are usually available on the level of administrative divisions, and are seldom readily available to be used in another spatial configuration.

One way of achieving a higher flexibility in the definition of regions is to first obtain or generate high resolution data using geographic information systems (GIS), and them cluster them into categories or groups, depending on certain trends in the data set. The generation of high resolution data is a critical research topic that has been addressed by previous studies. In this analysis, we assume that high resolution data is readily available and we focus therefore on the clustering step.

Since Lloyd [4] published the *k-means* algorithm as a clustering technique, other variants have been developed to accommodate a wide range of user requirements, as explained by Jain and Dubes [5]. For spatial clustering, it is possible to use such algorithms and add contiguity constraints to them, as done by Adam et al. [6] in their analysis of the pan-European electricity system for 2050. But as data analysis developed and GIS became more common, new algorithms specific to spatial clustering emerged, such as the *max-p regions* algorithm by Duque et al. [7]. Despite these developments, the ability of spatial clustering algorithms like *max-p regions* is still limited to small data sets with hundreds of data points. The novelty of our work resides in the clustering of high resolution spatial data (tested with $\sim 10^8$ data points) into contiguous regions that can be used in energy system optimization models. We apply the method in the case of Europe to compare the results of the models in 2015 and in 2050 under the constraint of a 95%

* Corresponding author.
E-mail address: kais.siala@tum.de (K. Siala).

https://doi.org/10.1016/j.esr.2019.100362
Received 4 October 2018; Received in revised form 11 March 2019; Accepted 7 May 2019
Available online 05 June 2019

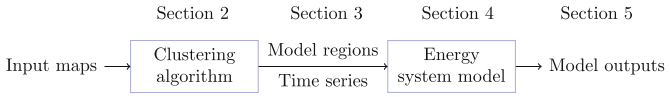

Fig. 1. Paper workflow.

CO_2 emissions reduction.

Hence, we structured our analysis in the following way. First, we go through the clustering method used to obtain the new regions in Section 2. The obtained clusters are described qualitatively and quantitatively

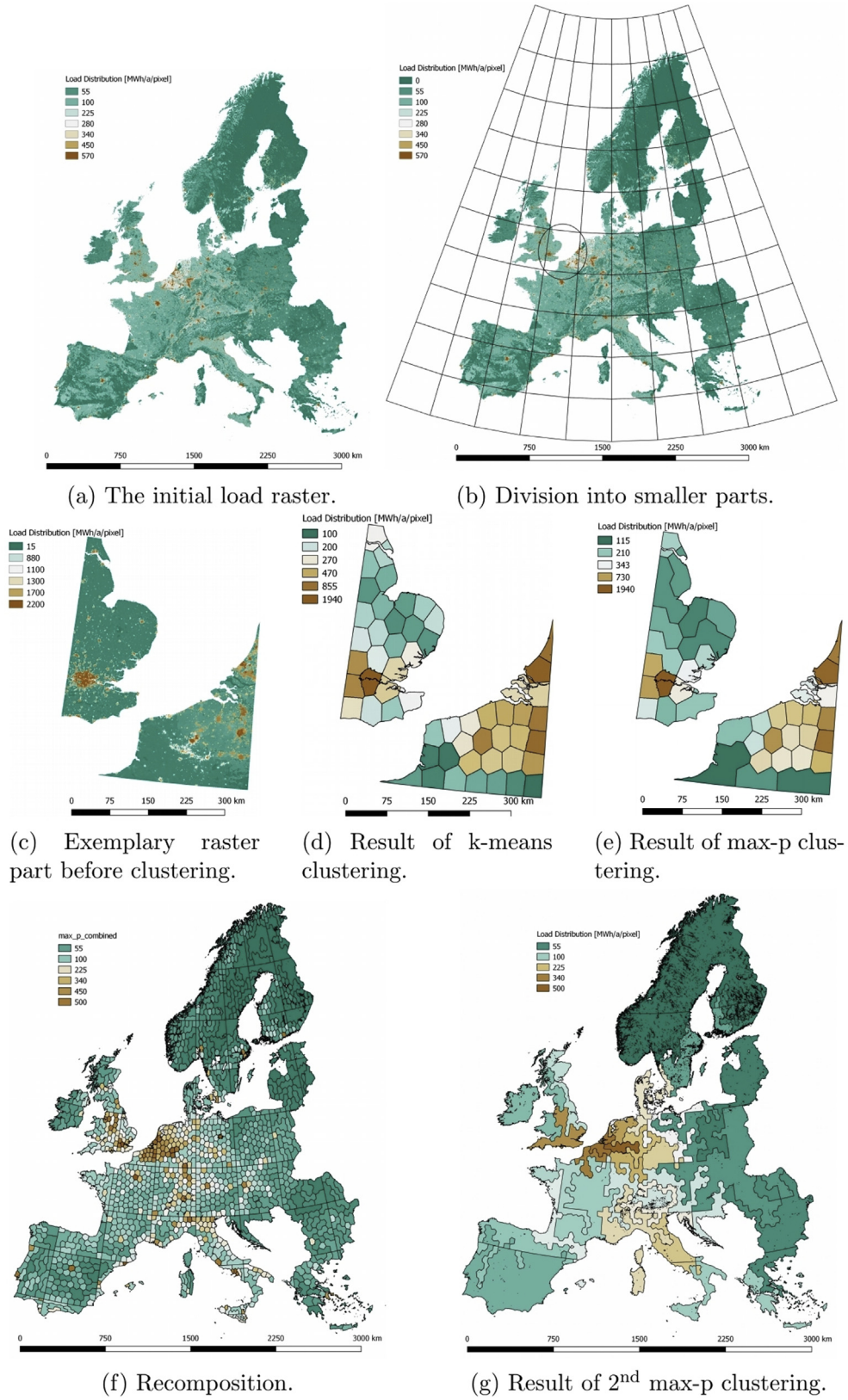

Fig. 2. Clustering methodology steps as applied to the load raster map.

in Section 3. Then, in Section 4, we introduce *urbs*, the open-source model framework used for the optimization of the energy system of Europe. The input data and the assumptions for the models are described in the same section. Finally, we analyze the results of the optimization in Section 5, followed by a conclusion and a discussion of possible applications. An overview of the paper workflow is shown in Fig. 1.

2. Spatial clustering

The goal of the spatial clustering is to create contiguous clusters with maximum data homogeneity within each one of them. These clusters are used afterwards as model regions in an energy system optimization model.

Since Fischer [8] laid the groundwork for the taxonomy of spatial clustering problems, several algorithms were developed to either decrease the running time or to optimize the clustering process. One of the main problems with the existing spatial clustering algorithms is their inability to handle data with huge resolution. Therefore, a new approach is adopted in this paper, which applies two existing algorithms (*k-means++* and *max-p regions*) in a three-step process. A short description of the process is provided below. The inputs and outputs of the clustering are also presented for the case of Europe.

2.1. Clustering method

The clustering method introduced in this section is a multi-stage process involving the *k-means++* and *max-p regions* algorithms. A formal description of the process is provided in Appendix A. The code is also available as open source [9].

The core of *k-means++* is the standard *k-means* algorithm developed by Lloyd [4] which assigns each data point to the nearest centroid of k initial clusters. Arthur and Vassilvitskii [10] enhanced the *k-means* algorithm by choosing the initial centroids such that they are as far away from each other as possible. The algorithm is fast and capable of handling a big amount of data, but it does not produce spatially contiguous regions. If a contiguity constraint is added in the cost function of the distance minimization, the algorithm generates compact clusters with the shape of Voronoi polygons with the respective centroids at their centers. Enforcing a strict contiguity constraint would lessen the importance of the data homogeneity within each cluster. Hence, *k-means++* cannot be applied solely to achieve the desired outcome of the study.

The spatial contiguity and the data homogeneity within clusters are, on the other hand, the strengths of the *max-p regions* clustering algorithm [11]. The algorithm is one of the first mixed-integer programming (MIP) spatial clustering algorithms developed for the p-regions problem [7]. It clusters data of n regions into p contiguous clusters where every cluster satisfies a minimum threshold value, such as the minimum solar energy potential that should exist in every cluster. The information lost due to clustering is minimal because the algorithm produces the maximum number of clusters that can be achieved in regard to the set threshold value. That number is unknown, but it increases if the threshold value decreases. With proper understanding of the data variance, the number of output clusters can be known to be within a certain range. The contiguous shapes of the clusters depend only on the input data and are not necessarily compact. Moreover, it is well implemented in python in the pySAL library [12] and in GeoDa [13], which makes it user-friendly. The main drawback for *max-p regions* is the inability to handle huge data. According to Duque et al. [7], the max-p regions problem algorithm has computational complexity of $(n-1)n^2 + \frac{n^2-n}{2}$ variables and $3n + (n-1)n^2 + n\frac{n^2-n}{2}$ constraints, where n is the number of data points to be clustered. This corresponds to a complexity of $O(n^3)$ according to the Bachmann—Landau notation. Hence, if the number of areas increases, the problem becomes intractable [7]. Therefore, *max-p regions* cannot handle our input data without decreasing its resolution drastically.

Our approach combines the strengths of both algorithms to achieve the desired outcome:

1. Starting with rasters of $\sim 10^8$ data points, we apply the "divide and conquer" principle to split the data into 100 equally-sized rasters that can be processed in parallel (see Fig. 2a and b).
2. We then cluster the data of each raster part (at most $\lesssim 1.27{\cdot}10^6$ numerical data points) using *k-means++*, as exemplified in Fig. 2c and d. In order to determine the number of clusters k_i for every map tile a_i, we first search for a reference area a_r which has the highest product of relative standard deviation σ_r/σ_{max} and relative size (number of valid data points) n_r/n_{max}. The elbow method [14] was used to determine the number of clusters k_r for a_r. Second, for each tile a_i, the number of clusters k_i was calculated using the following expression:

$$k_i = k_r \cdot \left(0.7\frac{n_i}{n_{max}} + 0.3\frac{\sigma_i}{\sigma_{max}}\right). \quad (1)$$

   We end up with $\lesssim 100$ clusters for each map part. This step is fast ($\sim$ 3 h for all tiles) and leads to the largest data compression (down to 1 cluster for $\sim$ 15000 data points in the case of load).
3. The output of *k-means++* is a raster for each tile, which we polygonize into a shapefile.
4. For each tile a_i, we run the *max-p* algorithm. We define the threshold of the *max-p* algorithm as a function of relative standard deviation and relative number of valid data points, as described in the following equation:

$$thr_i = A \cdot e^{-B\left(\frac{n_i}{n_{max}} + C\frac{\sigma_i}{\sigma_{max}}\right)}, \quad (2)$$

   where thr_i is the threshold for tile a_i and A, B, and C are three global parameters for the complete data set. We determine A, B, and C through regression so that:
   - if a tile has the smallest n_i and the smallest σ_i, it will be split into at most two clusters;
   - if a tile has the largest n_i and the largest σ_i, it will have a very low threshold so that it may retain all of its parts from *k-means++*;
   - if a tile has the smallest size but the highest standard deviation, its threshold will be average.

   Despite a limited data compression (the total number of clusters is almost halved), this step lasts $\sim$ 3 h for all the tiles.
5. After applying the *max-p* algorithm to every tile a_i, we merge all the tiles together again to get the full map of Europe (see Fig. 2f). The total number of clusters at this stage is $\lesssim 1800$ in all three maps, which can be clustered at the next step.
6. *max-p* algorithm is applied on the whole map to obtain the final map shown in Fig. 2g. In this study, we chose a target number of 28 regions to match the number of countries in Europe that are usually used as model regions. In order to obtain exactly 28 regions at the end, we varied the threshold through trial and error. The algorithm required 5–8 h for the clustering of $\sim$ 1800 data points.

2.2. Clustering input

As mentioned before, the method used in this analysis is applicable on raster data sets with a high resolution, which cannot be clustered with standard methods. In the following we will use three rasters of Europe in 15 arcsec resolution: one for the wind potential (expressed in kWh/kW_p), one for the photovoltaic (PV) potential (in kWh/kW_p), and one for the load density (in MWh/pixel/a) [15]. The rasters for the solar and wind potentials and their corresponding model regions after the clustering are displayed in Fig. 3. For the load distribution map, please refer to Fig. 2a and g.

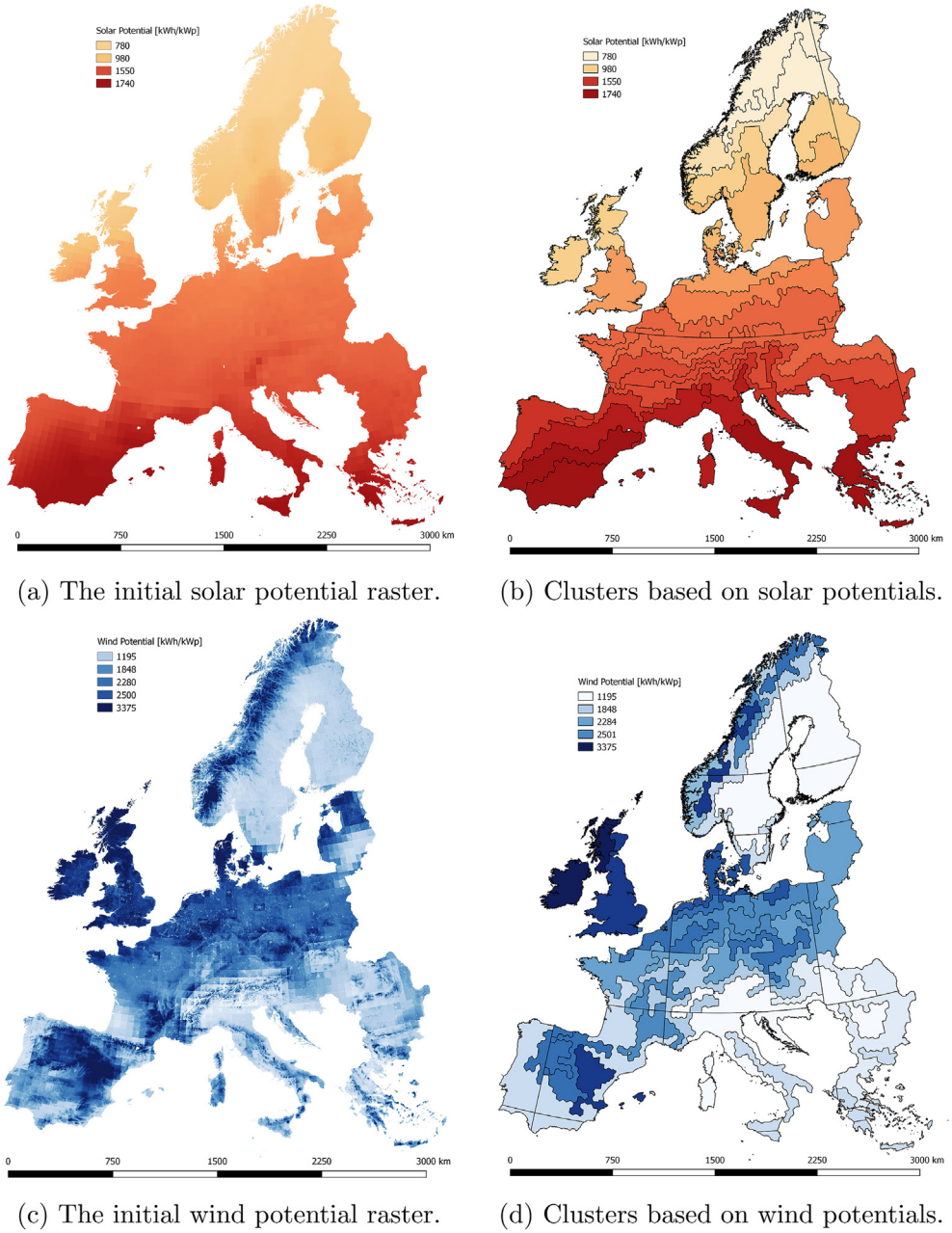

Fig. 3. Input rasters and output regions produced by the clustering process.

3. Results of the spatial clustering

Before using the clusters to create energy system models, we analyze them qualitatively and quantitatively by describing them and comparing them to each other and to regions delimited by national borders.

3.1. Qualitative description

The outputs of the clustering process have different characteristics that might impact the results of the energy system optimization. The *Load* map is characterized by small, stretched out regions in the densely populated areas of Central Europe, between Northern Italy and the Netherlands, and by large regions in the periphery (Scandinavia, Iberian Peninsula, Eastern Europe). The *Solar* map is mostly made of equally sized regions that have the shape of horizontal bands, since the solar potential highly correlates with the latitude. Regarding the *Wind* map, we observe that the shapes of the clusters are affected by the elevation (Northern Spain, Norway) and by the distance to the shore, with many distinct clusters along the coast and others that are mostly continental.

3.2. Information loss

Clustering is used in this study to summarize the large amount of information contained in the high resolution maps into homogeneous groups. The compression of data leads inevitably to loss of information at the different stages of the algorithm, as displayed in Table 1. The largest data compression occurs after running *k-means++*. The compression ratios at that stage are affected by the maximum number of clusters k_r, which was determined using the elbow method for the largest and most diverse region, and by the sizes and standard deviation of the tiles. Choosing a higher k_r will lead to more clusters. Despite the high ratios, the information loss at this stage is not critical for the solar and wind potential maps, because the full-load hours do not variate a

Table 1
Evolution of the number of valid data points over the clustering process.

Map	Input	*k-means* + +	*max-p* 1	*max-p* 2
Solar	~ $36.7 \cdot 10^6$	3490	1853	28
Wind	~ $36.7 \cdot 10^6$	2792	1780	28
Load	~ $36.7 \cdot 10^6$	2347	1340	28

lot within a small radius, as reflected in the low coefficients of variation (below 0.04 for the solar potential map). However, for the load map, small urban settlements with high load densities are dissolved with their surrounding areas, and the coefficients of variation may exceed 20 in northern Scandinavia. Increasing the number of *k-means* clusters (above the optimum number derived from the elbow method) might preserve the contrast between urban and rural areas until the next stage, but it would lead to longer computation times for the *max-p* algorithm running on all tiles.

Currently, the first run of the *max-p* algorithm has the lowest data compression ratios. It is possible to alter the threshold values by setting different constraints for the regression that determines the parameters *A*, *B*, and *C*. Higher thresholds lead to less clusters for each tile, which can speed up the second run of *max-p* for the whole map. However, the quality of the final maps will be lower in this case. Hence, we chose the constraints so that we obtain as many clusters after the first run of the *max-p* algorithm as we can process in a reasonable amount of time (below 8 h).

3.3. Homogeneity and contrast

By design, the final solar and wind clusters should be overall more homogeneous with respect to the solar and wind potentials than individual countries. This is for instance visible in Fig. 4, which depicts the PV full-load hours of a sample of data points for each region (blue lines, sorted by the median of the regions) and for all Europe (black line in the background). The median values for each region, whether a country (top) or a solar cluster (bottom), are shown in red. Fig. 4 shows that the new clusters have smaller value ranges, are of similar sizes, and their medians lie almost always on the black curve of the sorted FLH values in Europe. Hence, the clustering method splits Europe in homogeneous regions of equal sizes, that can be represented by their median values without sacrificing too much accuracy. This is not the case for the map using country borders, where countries with very good locations for solar projects are misrepresented by their low median values.

In all three maps, the coefficients of variation are usually smaller than for the map of countries (with respect to the data type chosen for the clustering), both for the extremes and on average. It varies between 1.14 and 3.56 for the load clusters (countries: 1.23–6.05), between 0.01 and 0.09 for the solar clusters (countries: 0.01–0.15), and between 0.11 and 0.50 for the wind clusters (countries: 0.12–0.59). The improved homogeneity may have a positive impact on the robustness of the modeling of energy systems with high shares of solar and wind power supply. In fact, model regions are usually represented by one or a few points with particular FLH values, for which time series are generated. Therefore, it is crucial that these points (best, upper 10%, median, etc.) summarize the quality of the region without distortion.

In addition to homogeneity, higher contrast between the regions may be desired for some applications. By clustering the load map so that the regions satisfy a minimum total load, we created clusters of high load densities in Central Europe and low load density on the periphery. The load densities are plotted against the area of the regions in the case of countries (crosses) and load clusters (filled circles) in Fig. 5. Compared to countries, the load clusters have more distinguishable regions on the extremes (large area and low density, or small

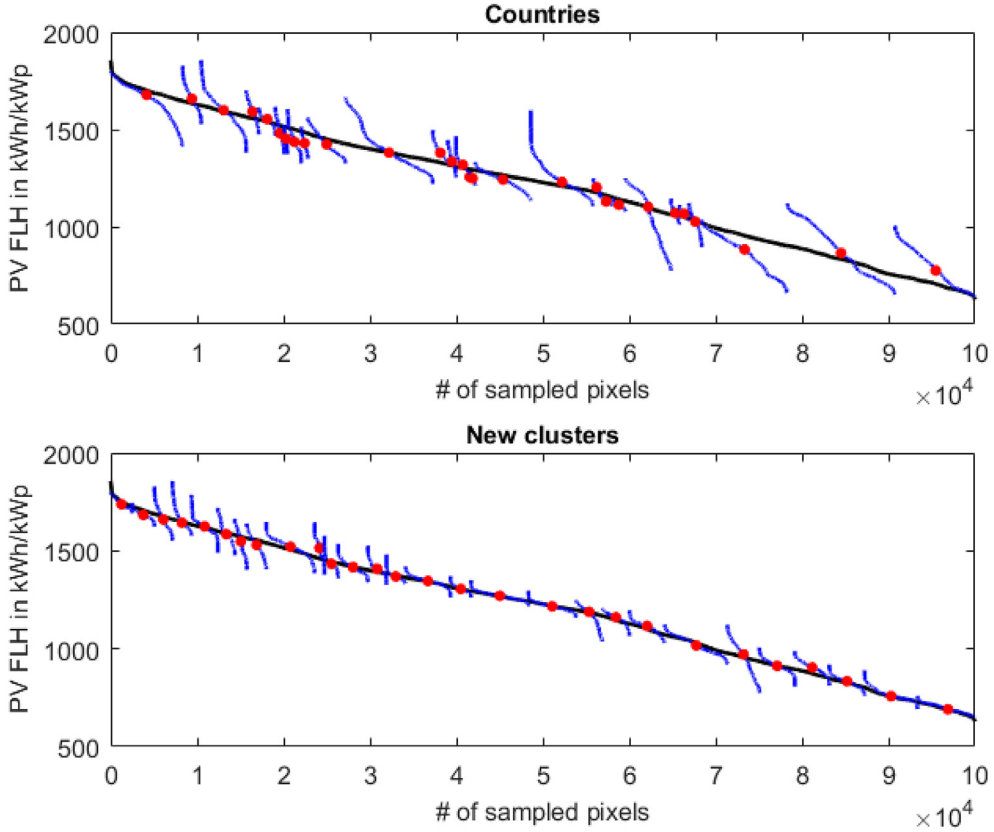

Fig. 4. Solar FLH values within regions (blue lines), sorted in descending order by their median value (red circles), plotted against the sorted values for Europe (black line in the background). The top figure corresponds to the countries, the bottom one to the solar clusters.

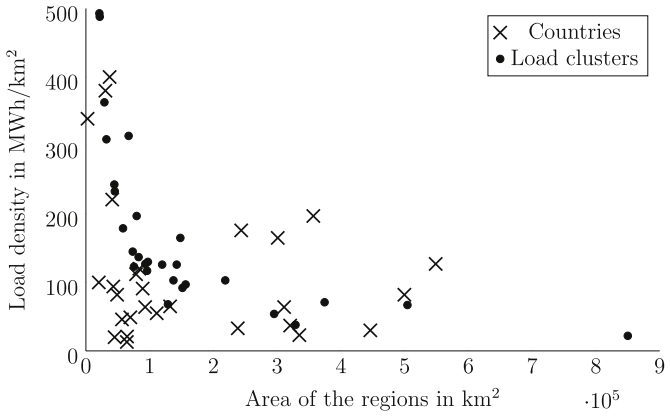

Fig. 5. The load density in relation to the area of each region in the case of countries (crosses) and load clusters (filled circles).

area with high density), which can be useful for studies that investigate the energy system of high-demand regions (e.g. cities) surrounded by load-demand regions (e.g. countryside).

Whether these topological differences have a measurable impact on energy system optimizations can be determined by running an experiment using energy system models. The procedure of the experiment used in this study is explained in the next section.

4. Energy system optimization

In order to quantify the impact of the choice of regions on energy system optimizations, we build models using the regions of Figs. 2g, 3b and 3d. We first describe the modeling framework, then provide information about the data and the assumptions used in this experiment.

4.1. Expansion planning models with urbs

We use the open-source modeling framework *urbs* to generate the models for our analysis. The created models co-optimize capacity expansion as well as hourly dispatch of generation, transmission, and storage from a social planner perspective. The optimization goal is to minimize the costs of expanding and operating the energy system including the annualized capital costs, the fuel costs, as well as other fix and variable operational costs. Major inputs are the hourly time series for the load and the capacity factors of renewable energy sources. Other important input data include the existing infrastructure (grid, power plants, storage) which can eventually be built in the models. Techno-economic parameters cover costs for investments, maintenance, fuels, and emissions. Finally, restrictions can be set such as maximal capacities for grid/generation expansion or a limit on the CO_2 emissions.

Each model solves a linear optimization problem that is written in Pyomo using gurobi. Major outputs include the installed capacities (generation, grid, storage) and the hourly operation of the system. The models also provide the emissions, the costs, and the marginal costs at each region. The source code for *urbs* and an extensive description can be found on GitHub [16].

4.2. Data and assumptions

The main assumptions and data sources used to conduct this analysis are described below. The data pre-processing is automated in order to ensure a uniform model generation process.

Geographic coverage. The analysis covers the 28 countries of the European Union, excluding Malta and Cyprus and adding Norway and Switzerland.

Representation of time. We use the second weeks of January, April, July, and October to represent the full year of 2015. The time resolution is 1 h.

Model regions. We use four models based on: country borders (later referred to as *Countries*), regions with homogeneous wind potential in terms of full-load hours (*Wind*), regions with homogeneous solar photovoltaic potential (*Solar*), and regions with similar total electricity demand (*Load*).

Load time series. Hourly time series for each country [17] are disaggregated into sectoral load time series based on typical sectoral load profiles [18]. Sectors are distributed geographically based on land use types [19]. Load is aggregated again for the new model regions. The same load time series are used in 2015 and 2050.

Renewable time series. Wind and solar hourly capacity factor time series are generated by combining MERRA-2 radiation, temperature and wind speed data [20] with maps of land use, elevation, and protected areas. Among the suitable locations in each region, we pick the time series of the pixel with the median full-load hours for onshore wind and solar PV, and at the top 10% for offshore. We use a uniform correction factor lower than 1 for each technology in order to match the wind and solar generation of Europe in 2015 approximately. Identical time series were used in 2050. A new version of the code for generating the time series and the input rasters for wind and solar is available open source [21].

Commodity prices. We assume the same prices in 2015 and 2050. The prices are constant over the year and respect the merit order of power plants observed in 2015. For more information, see the inputs files for the *urbs* models [15].

Conventional power plants. Power plants in operation in 2015 [22] were allocated to the regions based on their coordinates, then aggregated based on their types. In 2050, none are still existing. New capacities for nuclear are not allowed to exceed the levels of 2015. Missing power plant characteristics were filled with own assumptions.

Renewable power plants. Installed capacities in each country in 2015 were collected from IRENA [23] then distributed geographically based on technical potential maps (wind and solar), land use types (others), and a randomness factor. There are no pre-existing capacities in 2050. New capacities for hydro and biomass are not allowed to exceed 2015 levels.

Transmission lines. Transmission lines are extracted from GridKit [24]. Based on their lengths and voltage levels, a transmission capacity is assigned to them. Only connections between different regions are considered.

Economic parameters. Investment, fix and variable costs for most power plant types are derived from ETRI [25].

CO_2 emissions. There are no limits for 2015. For 2050, we assume a 95% reduction compared to 2015. No CO_2 certificate price are used.

5. Results of the optimization

In this section, we discuss the results of the energy system optimization for 2015 and for 2050 using different regions within Europe.

We first compare the results of all models (*Countries*, *Wind*, *Solar*, *Load*) to ENTSO-E statistics for the year 2015. Despite using only four weeks instead of a full year, the models manage to match the energy mix of Europe with minor discrepancies, as shown in Fig. 6. The models overestimate the share of nuclear power, but the error is compensated by an underestimation of coal, gas, and other mixed fuels. Almost no energy is curtailed (less than 0.1% in all models). Comparing the model *Countries* to the three others, we observe negligible differences in the solar PV generation (relative error between 1% and 5%). However, the wind generation is overestimated in all three models, with a relative error ranging from 5% for *Solar* to 10% for *Wind*. The reason for this lies in the choice of the representative time slices and in the time series for renewable generation. First, although the model *Countries* was calibrated to match the energy mix of ENTSO-E using a full-year optimization,[1] this calibration is lost when running a four-week optimization. The time series happen to be representative for the solar generation in

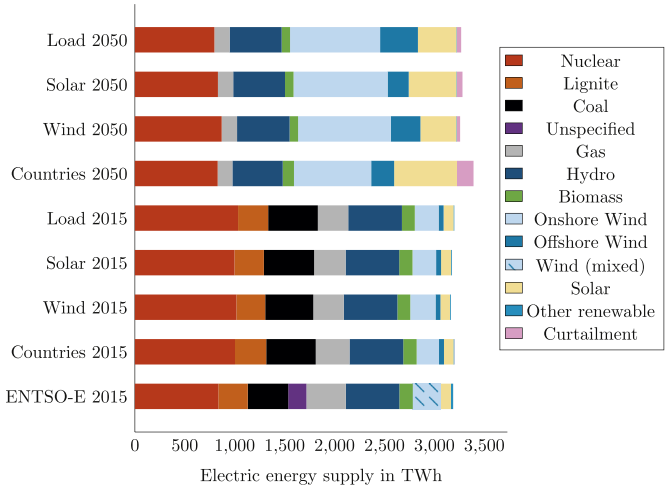

Fig. 6. Electric energy supply in Europe in TWh according to the *urbs* model and to ENTSO-E statistics of 2015.

that time frame, but they are not for wind generation. Second, as mentioned in section 4.2, the renewable generation time series were determined for a pixel at the median locations for onshore wind and solar PV, and at the top 10% for offshore wind. Since the shape of the regions are not the same in the models, the locations of the medians and the top 10% are also different, which leads to different time series.

The discrepancies between the models widen for the year 2050. Due to the stringent CO_2 emissions constraint, lignite and coal power plants disappear from the power mixes. The only CO_2 emitting technology which is still allowed is gas, due to its lower specific emissions. The combined share of hydro power and biomass remains almost constant in all four cases. On the other hand, the shares of solar PV, onshore and offshore wind, nuclear, and the amount of curtailment vary considerably between the models.

In the model *Countries*, the ratio of solar to wind is 1:1.6, the lowest in all four models. It is also coupled with a higher amount of curtailment (166 TWh compared to 31 TWh for *Wind*, the lowest value). All in all, a positive correlation exists between the amount of curtailment and the ratio of solar to wind. Also, the higher the capacity of solar PV, the bigger is the installed capacity of battery storage, as shown in Fig. 7. One possible explanation for this could be the low wind FLH values of countries in Southern Europe, which have high FLH of solar PV. Yet this explanation is not sufficient, because an even stronger effect would have been visible in the model *Solar*. However, we observe that the cost-optimal solution for the model regions in *Solar* is to build less PV and more onshore wind than in *Countries*. This counter-intuitive result could be explained with the shape of the regions: the stretched horizontal bands of the solar cluster have more transmission line connections to their neighbors than the compact model regions in *Countries*. With less transmission bottlenecks, the model *Solar* can integrate wind and solar generation better in the North-South direction.

As expected, the model *Solar* has a higher share of solar PV than *Wind*, which has on the other hand a higher share of onshore and offshore wind, combined. This is most probably due to the existence of very favorable locations with high full-load hours of solar PV (in the case of *Solar*) and onshore wind (in the case of *Wind*). The model *Load* is characterized by the highest share of offshore wind generation (379 TWh). Here again, the shape of the regions and their geography provide a possible explanation. Many of the regions with high load density lie in Northern Europe, close to the North Sea, an area with very good conditions for offshore wind power plants due to its shallowness and high wind speeds. Hence, offshore wind power plants located in that area operate with high capacity factors and at competitive costs to cover a high proportion of the electricity demand of the model regions of *Load*.

Last but not least, Fig. 7 shows no investment in transmission lines and a little investment in battery storage, when compared to the large investments in solar PV and wind. This is due to the way the transmission grid is simplified: There are no transmission constraints within each region, and interconnections between model regions limit the amount of electricity that can be traded between them. With only 28 model regions, none of which are designed to reflect transmission bottlenecks, the limits for electricity transport are usually high and renewable power generation seems to be easy to integrate with no grid expansion nor large amounts of storage.

The discrepancy between the models is not limited to the aggregated capacities and energy production values, but affects the geographic distribution of the power plants as well. Although the *urbs* models do not deliver exact locations for the new power plants, it is possible to draw approximate geographic distribution maps based on the potential map and a randomness factor. In Fig. 8 we plot possible distributions of onshorewind power plants based on the results of the four models, as an example.

The darker the color in Fig. 8, the more agreement there is between the models that wind power plants should be built in the area. This seems to be the case for the coastal areas in Northern Europe and Norway, the United Kingdom, Estonia and Latvia. There is also a consensus to avoid Northern Italy, Southern Spain, Bulgaria and Central Sweden. Despite the differences between the models, these results are considered robust.

6. Conclusion and discussion

In this paper, we argue that creating energy system models with different shapes for the regions has several advantages. The focus in the first part was on the clustering methodology that we used to obtain regions with homogeneous characteristics out of high resolution data. The method is scalable and we were able to apply it on data sets with roughly 10^8 data points to obtain 28 regions. We explained how the number of clusters at the end of each stage of the algorithm can be varied, and which impacts this would have on the quality of the results and on the computation time. For the sake of clarity, the clustering was conducted using one single characteristic at a time (either similar total electricity demand, or wind potential, or solar potential). However, it is possible to use two or more parameters to obtain regions with a combination of characteristics. This could be done in a future study.

The regions created through clustering have inherent properties that are crucial for certain analyses. For instance, the model based on load density clusters is suitable for studying the interplay of cities and countryside, particularly in the power sector. Using regions with different renewable potentials ensures their homogeneity, which is usually implicitly assumed in energy system models that rely on one or a few time series per region. We showed that this assumption does not hold for models using countries as model region, and that clusters based on wind or solar full-load hours have lower coefficients of variation. Homogeneous regions are adequately represented through single time series, which leads to more robust results in energy system optimizations.

Using countries as model regions is appropriate in many situations, such as the analysis of policies on a national level. However, energy system modelers should be able to assess the magnitude of the errors caused by the choice of the regions. This study provided an example of an electricity system optimization for Europe for 2015 and 2050. Whereas the errors in 2015 were minimal, due to the limited share of renewable power supply, they were higher for 2050, causing discrepancies 5–10% in the energy mix. This example showed some

[1] A full-year optimization was conducted using the model *Countries* for the purpose of calibration, but its results are not discussed in this paper, because the focus lies in the relative differences between the models.

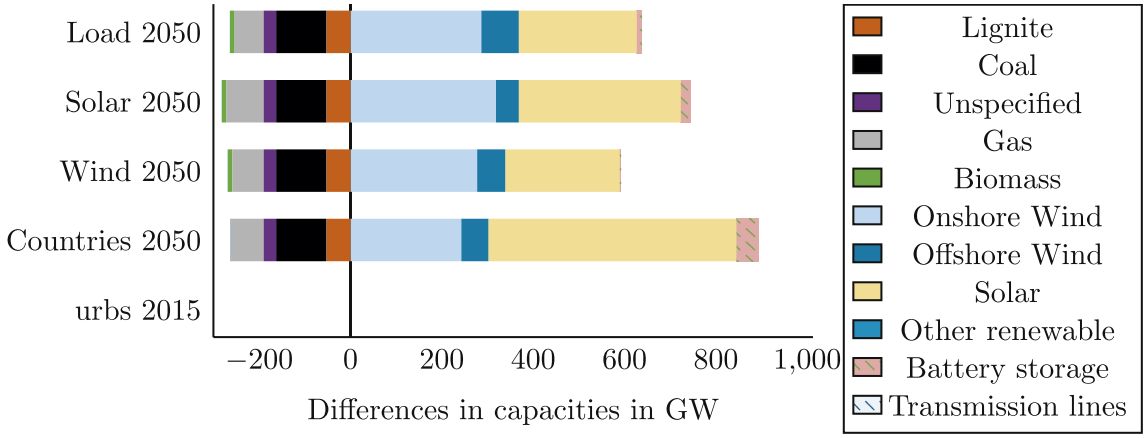

Fig. 7. Differences in capacities of power plants, storage units, and transmission lines in GW compared to the capacities in 2015.

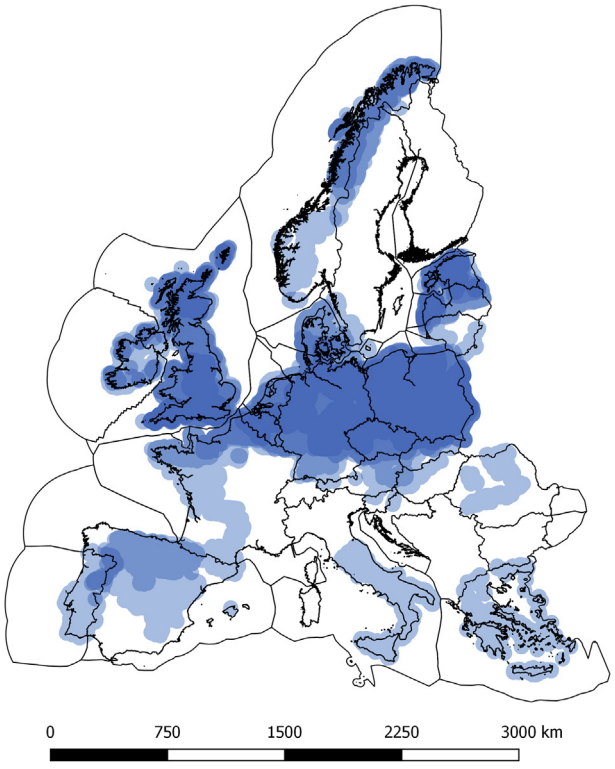

Fig. 8. Possible geographic distribution of new wind capacities in 2050. Each shade of blue corresponds to the distribution of one of the models, so that dark blues areas are those identified by all models as most likely project locations.

advantages of using other model regions for a complementary analysis. For modelers, the variation of the regions works as a sensitivity analysis. It puts the robustness of the model results to the test and helps visualize their actual impacts on the geography of the energy system. It can also reveal the weaknesses of the model assumptions. In our analysis, this was particularly the case of the misrepresentativity of some renewable time series and of the lack of transmission line congestion within regions. The clustering algorithm can be used to solve the former problem, by creating homogeneous areas (either independently or within political or administrative divisions), for which representative time series can be generated. As of the modeling of transmission lines, all our cluster models and the one based on national borders neglected bottlenecks within the regions, which is not a valid assumption in many instances. We recommend clustering transmission networks based on line contingencies, yet we were not able to achieve this due to the lack of topologically valid open grid data for Europe.

For policy-makers, trying various shapes of regions for the same study can be a powerful argument in a time characterized by a lack of acceptance for infrastructure projects, such as transmission lines and onshore wind power plants. It shows willingness to search for alternatives, and overcomes certain modeling weaknesses (e.g. abstraction, unsubstantiated assumptions, etc.) through a visually appealing sensitivity analysis.

Acknowledgements

The authors acknowledge the financial support by the Federal Ministry for Economic Affairs and Energy of Germany (BMWi) in the project “4NEMO - Research Network for the Development of New Methods in Energy System Modeling; research project: 0324008A”.

Appendix A. Pseudocodes

Pseudocode 1: Main Code

D: High resolution data raster
m: Number of tiles in the vertical dimension
n: Number of tiles in the horizontal dimension
F: Output shapefile
α: Range of clustering values for elbow method
L: Limit of accepted distance in `Call-k-means++`

A= `Split` (D, m, n), set of split areas
k_r n_{max} n_{min} σ_{max} σ_{min} = `Elbow Method` (A, α, L)
for *every* $a_i \in A$ **do**
  φ_i = `Call-k-means++` $(a_i, k_r, \sigma_{max}, n_{max}, L)$
E = `Call-max-p`$_1$ $(n_{min}, n_{max}, \sigma_{min}, \sigma_{max}, \varphi)$
F = `Call-max-p`$_2$ (E, thr_2, N_2)

Pseudocode 2: Split

Inputs: D, m, n
Read number of rows of input raster r_D and the number of columns c_D
$r = r_D/m$, number of rows in each tile
$c = c_D/n$, number of columns in each tile
A is empty
for $v = 1 \cdots m$ **do**
  for $h = 1 \cdots n$ **do**
    Cut the data set D between $c \cdot (h-1)$ and $c \cdot h$ in the horizontal dimension, and $r \cdot (v-1)$ and $r \cdot v$ in the vertical dimension
    Save sub-data set in a_i
    Add a_i to A
Outputs: $A = a_1, a_2, \cdots, a_i, \cdots, a_n$

Pseudocode 3: ElbowMethod

Inputs: A, α, L

e_h: Inertia of `k-means++` clusters.

for *every* $a_i \in A$ **do**
  Calculate standard deviation σ_i and number of non-null points n_i of every area a_i
  Calculate the product X_i of σ_i and n_i of every area a_i
Pick area a_r with maximum value of X_i
Save n_{max}, the maximum value of number of non-null points in all areas
Save n_{min}, the maximum value of number of non-null points in all areas
Save σ_{max}, the maximum value of standard deviation in all areas
Save σ_{min}, the maximum value of standard deviation in all areas

G is an empty plot
for *every element* h *of* α **do**
  $\sim, e_h =$ `k-means++`(a_r, h, L), Euclidian square distance from `k-means++` output
  Plot the point with the coordinates (h, e_h) in graph G
Apply the elbow method to G as described in [14]
Choose the point P_α at which the gain of adding more clusters drastically decreases
$k_r =$ x-coordinate of P_α, number of clusters of reference area a_r

Outputs: k_r, n_{max}, n_{min}, σ_{max}, σ_{min}

Pseudocode 4: Call-k-means++

Inputs: a_i, k_r, σ_{max}, n_{max}, L

$k_i = k_r \cdot (0.7\frac{n_i}{n_{max}} + 0.3\frac{\sigma_i}{\sigma_{max}})$, Number of clusters for area a_i
$\varphi_i, \sim =$ `k-means++`$(k_i, a_i, L, \sigma_{max}, n_{max})$ as described in [10], clustered data set for area a_i

Output: φ_i

Pseudocode 5: Call-max-p_1

Inputs: n_{min}, n_{max}, σ_{min}, σ_{max}, φ

A, B, C: Regression constants used in calculating thr_1

Islands: all φ_i which do not have common borders with other areas
thr_1: Threshold value for max-p_1
N_1: Neighbor matrix for max-p_1
d: one data feature in a_i picked according to the clustering analysis priority (Solar, Wind, etc.)
Create N_1 matrix

if *islands* $\neq 0$ **then**
  Assign every island to the nearest neighboring area

Calculate the regression equation constants A, B, C.

$$A \cdot e^{-B(\frac{n_{min}}{n_{max}}+C\frac{\sigma_{min}}{\sigma_{max}})} = 0.5$$

$$A \cdot e^{-B(\frac{n_{min}}{n_{max}}+C\frac{\sigma_{max}}{\sigma_{max}})} = 0.1$$

$$A \cdot e^{-B(\frac{n_{max}}{n_{max}}+C\frac{\sigma_{max}}{\sigma_{max}})} = 0.01$$

Calculate the threshold value thr_1 for area a_i.

$$thr_1 = A \cdot e^{-B(\frac{n_i}{n_{max}}+C\frac{\sigma_i}{\sigma_{max}})}$$

for $h = 1 \cdots n$ **do**
  e_i: `max-p1` output as described in [7] $e_i = \texttt{max-p}_1(\varphi_i, thr_1, N_1)$
Merge e_i zones to get E
E: Map of all areas A after first step of max-p

Output: $E = e_1, e_2, \cdots e_i, \cdots, e_n$

Pseudocode 6: Call-max-p_2

Inputs: E

Islands: clusters in E which do not have common borders with other clusters
d: one data feature picked from E according to the clustering analysis priority (Solar, Wind, etc.)
thr_2: Threshold value for max-p_2
N_2: Neighbor matrix for max-p_2

Create N_2 matrix
if *islands* $\neq 0$ **then**
  Assign every island to the nearest neighboring area

$$thr_2 = \frac{\sum_{d \in E} d}{40}$$

$F = \texttt{max-p}_2(E, thr_2, N_2)$

Output: F

References

[1] S. Pfenninger, A. Hawkes, J. Keirstead, Energy systems modeling for twenty-first century energy challenges, Renew. Sustain. Energy Rev. 33 (2014) 74–86 ISSN 1364-0321 https://doi.org/10.1016/j.rser.2014.02.003.

[2] B.A. Frew, M.Z. Jacobson, Temporal and spatial tradeoffs in power system modeling with assumptions about storage: an application of the POWER model, Energy 117 (2016) 198–213, https://doi.org/10.1016/j.energy.2016.10.074 ISSN 0360-5442.

[3] S. Gago Da Camara Simoes, W. Nus, P. Ruiz Castello, A. Sgobbi, D. Radu, P. Bolat, C. Thiel, E. Peteves, The JRC-EU-TIMES Model - Assessing the Long-Term Role of the SET Plan Energy Technologies, EUR - Scientific and Technical Research Reports, European Commission Joint Research Centre, Institute for Energy and Transport, 2013, https://doi.org/10.2790/97596 https://doi.org/10.2790/97596.

[4] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (2) (1982) 129–137, https://doi.org/10.1109/TIT.1982.1056489 ISSN 0018-9448.

[5] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988 ISBN 0-13-022278-X.

[6] K. Adam, M. Müller-Mienack, M. Paun, G. Sanchis, K. Strunz, e-HIGHWAY 2050 — The ENTSO-E facilitated study programme towards a modular development plan on pan-european electricity highways system 2050, in: 2012 IEEE Power and Energy Society General Meeting, ISSN 1932-5517, 1–6, doi:10.1109/PESGM.2012.6344807, 2012.

[7] J.C. Duque, L. Anselin, S.J. Rey, The max-P-regions problem, J. Reg. Sci. 52 (3) (2012) 397–419, https://doi.org/10.1111/j.1467-9787.2011.00743.x ISSN 00224146.

[8] M.M. Fischer, Regional taxonomy, Reg. Sci. Urban Econ. 10 (4) (1980) 503–537, https://doi.org/10.1016/0166-0462(80)90015-0 ISSN 01660462.

[9] M.Y. Mahfouz, W.S. Khan, K. Siala, Geoclustering: v0.1 (Beta), (2019) https://github.com/tum-ens/geoclustering.

[10] D. Arthur, S. Vassilvitskii, K-means + +, The advantages of careful seeding, Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, 2007.

[11] J.C. Duque, M.C. Vélez-Gallego, L.C. Echeverri, On the performance of the subtour elimination constraints approach for the p -regions problem: a computational study, Geogr. Anal. 50 (1) (2018) 32–52, https://doi.org/10.1111/gean.12132 ISSN 00167363.

[12] S. Rey, L. Anselin, P. Amaral, D. Arribas-Bel, M. Wigginton Conway, C. Famer, D.C. Folch, J. Gaboardi, M. Hwang, W. Kang, E. Knaap, M. Kolak, J. Laura, Z. Li, S. Lumnitz, T. Oshan, R. Pahle, C. Schmidt, H. Shao, P. Stephens, S. Wang, L.J. Wolf, pySAL Documentation - Release 1.14.4, (2018) http://pysal.org/.

[13] L. Anselin, S. Ibnu, K. Youngihn, GeoDa: an introduction to spatial data analysis, Geogr. Anal. 38 (1) (2006) 5–22.

[14] R.L. Thorndike, Who belongs in the family? Psychometrika 18 (4) (1953) 267–276, https://doi.org/10.1007/BF02289263 ISSN 0033-3123.

[15] K. Siala, M.Y. Mahfouz, Supplementary materials for the spatial clustering of

Europe, https://doi.org/10.5281/zenodo.1439342, (2018).
[16] J. Dorfner, M. Dorfner, K. Schönleber, S. Candas, smuellr, wyaudi dogauzrek, adeeljsid yunusozsahin, T. Zipperle, S. Herzog, L. Odersky, K. Siala, O. Akca, Urbs: v0.7.3: A Linear Optimisation Model for Distributed Energy Systems, (2018), https://doi.org/10.5281/zenodo.1228851 https://doi.org/10.5281/zenodo.1228851.
[17] ENTSO-E, Monthly hourly load values, https://www.entsoe.eu/data/power-stats/hourly_load/, (2015).
[18] BDEW, Standardlastprofile strom (German),, 2017. https://www.bdew.de/energie/standardlastprofile-strom/.
[19] P.D. Broxton, X. Zeng, D. Sulla-Menashe, P.A. Troch, A global land cover climatology using MODIS data, J. Appl. Meteorol. Climatol. 53 (6) (2014) 1593–1605, https://doi.org/10.1175/JAMC-D-13-0270.1.
[20] R. Gelaro, W. McCarty, M.J. Suárez, R. Todling, A. Molod, L. Takacs, C.A. Randles, A. Darmenov, M.G. Bosilovich, R. Reichle, K. Wargan, L. Coy, R. Cullather, C. Draper, S. Akella, V. Buchard, A. Conaty, A.M. da Silva, W. Gu, G.-K. Kim, R. Koster, R. Lucchesi, D. Merkova, J.E. Nielsen, G. Partyka, S. Pawson, W. Putman, M. Rienecker, S.D. Schubert, M. Sienkiewicz, B. Zhao, The modern-era retrospective analysis for research and applications, version 2 (MERRA-2), J. Clim. 30 (14) (2017) 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1.
[21] K. Siala, H. Houmy, S.A. Huezo Rodriguez, Renewable-Timeseries (Beta), (2019) https://github.com/tum-ens/renewable-timeseries.
[22] F. Hofmann, J. Hörsch, F. Gotzens, FRESNA/powerplantmatching: v0.2 (Version v0.2), Zenodo, (2018), https://doi.org/10.5281/zenodo.1405595 doi:10.5281/zenodo.1405595.
[23] O. Lavagne d'Ortigue, A. Whiteman, S. Elsayed, Renewable Energy Capacity Statistics 2015, Tech. Rep., IRENA, (2016) http://www.irena.org/-/media/Files/IRENA/Agency/Publication/2015/IRENA_RE_Capacity_Statistics_2015.pdf.
[24] B. Wiegmans, GridKit Extract of ENTSO-E Interactive Map, (2016), https://doi.org/10.5281/zenodo.55853 https://doi.org/10.5281/zenodo.55853.
[25] R. Lacal Arantegui, A. Jaeger-Waldau, M. Vellei, B. Sigfusson, D. Magagna, M. Jakubcionis, M. d. M. Perez Fortes, S. Lazarou, J. Giuntoli, E. Weidner Ronnefeld, G. De Marco, A. Spisto, C. Gutierrez Moles, ETRI 2014 - Energy Technology Reference Indicator Projections for 2010-2050, EUR - Scientific and Technical Research Reports, Joint Research Center of the European Union, 2014, https://doi.org/10.2790/057687 http://publications.jrc.ec.europa.eu/repository/handle/JRC92496.