Methodology for Clustering High-Resolution Spatiotemporal Solar Resource Data

Dan Getman, Anthony Lopez, Trieu Mai, and Mark Dyson
National Renewable Energy Laboratory

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Technical Report
NREL/TP-6A20-63148
September 2015

Contract No. DE-AC36-08GO28308

NREL
NATIONAL RENEWABLE ENERGY LABORATORY

Methodology for Clustering High-Resolution Spatiotemporal Solar Resource Data

Dan Getman, Anthony Lopez, Trieu Mai, and Mark Dyson
National Renewable Energy Laboratory

Prepared under Task No. SA12.0381

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Technical Report
NREL/TP-6A20-63148
September 2015

Contract No. DE-AC36-08GO28308

NOTICE

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Acknowledgments

We thank Easan Drury, Clayton Barrows, Kelly Eurek, and Elaine Hale for valuable comments and input, and Mike Meshek for editing assistance. This work was supported by the U.S. Department of Energy under Contract No. DE-AC36-08GO28308 with the National Renewable Energy Laboratory. Funding provided by the Office of Energy Efficiency and Renewable Energy and the Office of Electricity Delivery and Energy Reliability. We thank Sam Baldwin, Seungwook Ma, and Gil Bindewald of the above offices for supporting this work. Lastly, we would like to thank our reviewing managers, Nate Blair, Ann Brennan, and Dave Mooney. The opinions represented in this article are the authors' own and do not reflect the view of the U.S. Department of Energy or the U.S. Government. Any and all errors are the responsibility of the authors.

List of Acronyms

DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DNI	Direct Normal Incident
DTW	Dynamic Time Warping
GHI	Global Horizontal Incident
LOWESS	Locally Weighted Scatterplot Smoothing
NARR	North American Regional Reanalysis
NREL	National Renewable Energy Laboratory
NSRDB	National Solar Radiation Data Base
PV	Photovoltaic
TEPPC	Transmission Expansion Planning Policy Committee
SSW	Sum of Squared Within
WECC	Western Electricity Coordinating Council

Executive Summary

In this report, we introduce a methodology to achieve multiple levels of spatial resolution reduction of solar resource data, with minimal impact on data variability, for use in energy systems modeling. The selection of an appropriate clustering algorithm, parameter selection including cluster size, methods of temporal data segmentation, and methods of cluster evaluation are explored in the context of a repeatable process. In describing this process, we illustrate the steps in creating a reduced resolution, but still viable, dataset to support energy systems modeling, e.g. capacity expansion or production cost modeling. This process is demonstrated through the use of a solar resource dataset; however, the methods are applicable to other resource data represented through spatiotemporal grids, including wind data.

In addition to energy modeling, the techniques demonstrated in this paper can be used in a novel top-down approach to assess renewable resources within many other contexts that leverage variability in resource data but require reduction in spatial resolution to accommodate modeling or computing constraints.

Table of Contents

List of Figures

List of Tables

1 Introduction

Solar and wind renewable resources are driven by weather and other atmospheric conditions that span wide-ranging spatial and temporal scales. For this reason, they have intrinsic variability and uncertainty, meaning the amount of resource available changes from one moment to the next and the changes cannot be perfectly forecast. Energy or electricity generated from solar and wind resources are commonly referred to as variable generation. The variability occurs for any given location, and it is correlated between locations with the amount of correlation depending on distance, topography, and other conditions. The existing research to quantify renewable resource variability and correlation has relied on an atmospheric science approach with numerical weather models.

Advances in numerical weather modeling have yielded valuable datasets reaching very high temporal and spatial resolutions that can present billions of estimates over hundreds of thousands of locations [1,2,3,4]. These datasets have been used to assess site- or region-specific resource quality to inform technology developers, system planners, and energy researchers [5,6]. In one use of these data, energy models require input data to inform the characteristics of the different energy sources and technologies represented in the models. With the recent growth in renewable, and especially wind and solar, deployment in many countries and renewable energy's potential to contribute a more sizeable fraction of total energy supply, accurate representation of these data sets in energy models, especially electricity-sector focused models, has become increasingly important. Moreover, the increase in renewable penetration in the electric sector has motivated modelers to develop better techniques for handling resource variability and uncertainty [7], including greater temporal resolution.

One class of energy models of particular focus for the methods introduced in this paper are capacity expansion models for electricity systems, which are used to develop future scenarios of electricity supply and demand for a geographic region of interest [8,9,10]. The geographic structures of these models are often designed to consider a wide range of spatial boundaries, including political, institutional, and electrical ones. Typically, the model structures do not use detailed or quantitative methods to account for the regional characteristics of renewable resources. Instead, capacity expansion models characterize renewable resources based on the arbitrary—from the perspective of the renewable resources—regions designed for other purposes or rely on engineering judgments about renewable resources. The combination of limited data availability, computational limits, and historically low renewable penetration levels have justified modelers' use of this practice.

In this paper, we introduce how clustering methods can be applied to solar resource data. In particular, we demonstrate clustering techniques on regional hourly profiles of solar photovoltaic (PV) plant performance for a sample region. We evaluate two clustering methods and a wide range of cluster sizes as well as four methods of segmentation. Using two different metrics, we attempt to measure the quality of the clusters. The application of numerical clustering techniques to regionalize multi-attribute spatial datasets is not a new concept in the literature [11,12]. However, these methods have not been translated to spatiotemporal datasets, such as PV performance profiles.

Energy models represent one application of the techniques presented in this paper. For example, energy modelers can cluster high spatiotemporal resolution renewable resource data into a smaller and computationally manageable set of regions. Mai et al. [13][1] describe the Resource Planning Model, a capacity expansion model for regional power systems in the Western United States, which relies on the clustering techniques to model transmission interconnection options for solar resources. The application of clustering techniques to renewable resource data also enables a measure of the level of approximation through the metrics we present.

In addition to energy modeling, the techniques demonstrated in this paper can be used in a novel top-down approach to assess renewable resources. For example, while renewable resource behavior is intrinsically driven by atmospheric and ground conditions, the clustering techniques are blind to these factors, but they might be useful in assessing their effects. Patterns driven by ground features and weather can be revealed through cluster results. While we do not comprehensively evaluate these here, the methods introduced can be applied in future work.

[1] This forthcoming paper builds on the model described in Mai et al. [10].

2 Methodology

This research outlines a process for selecting the most appropriate statistical clustering method and data segmentation strategy for clustering simulated solar PV performance data.

The state of Colorado in the United States is chosen as a sample region to develop and evaluate the methodology. However, the techniques presented here are applicable to any region where gridded solar resource data are available. The variability of solar resource results in a complex solar capacity factor[2] dataset, which presents a challenge in the determination of the most appropriate combination of data segmentation and clustering method. For example, Figure 1 shows the normalized PV capacity factor for three locations in Colorado over a three-day period. The output profiles demonstrate the variability of PV output in each location and between locations. Clustering techniques can be used to systematically evaluate the temporal correlations between locations and group sites accordingly.

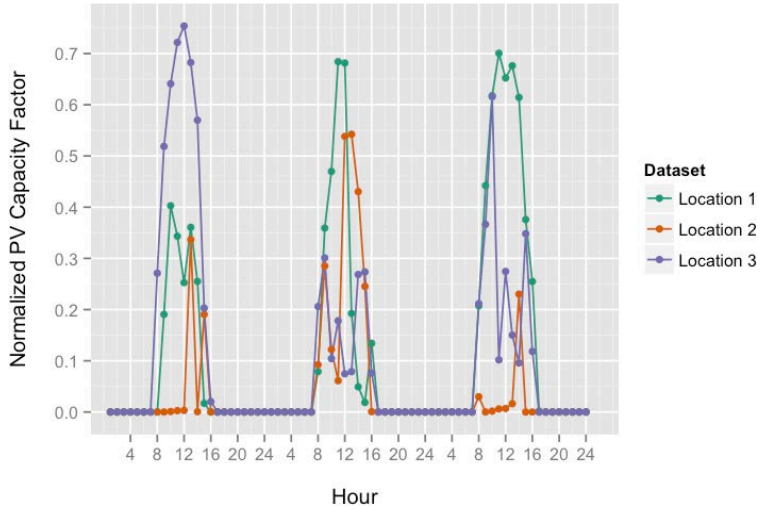

Figure 1. PV output in terms of capacity factor over a three-day period for three locations in Colorado

While clustering can be applied for time-series data over any amount of time or any resolution, we focus on one-year hourly profiles. To reduce computation time, we further segment the annual hourly PV performance data using four different methods. This produces four individual performance datasets on which to cluster. We apply two clustering methods, k-means [14] and Max-P [15], with a range of required clusters (20–140) to each segmented dataset. Finally, we test the quality of the various cluster techniques, cluster sizes, and data segmentations using a simple sum of squares and a cross validation within each cluster.

As a result of the clustering, the PV performance data are reduced from about 3,000 time-series vectors (each with 8,760 data points) to between 20 and 140 time-series vectors (again, each with 8,760 data points).

[2] Capacity factor is a unit-less representation of the system output performance relative to its nameplate capacity over a defined duration (e.g., instantaneously, over an hour, or over a year). We use hourly data in the analysis presented.

2.1 Source Data

We use solar radiation and weather data from the National Solar Radiation Data Base (NSRDB) [3]. The solar data are hourly and on a 10 x 10 km^2 grid spanning the contiguous United States and Hawaii. The gridded solar product from the NSRDB spans 12 years from 1998 to 2009, but for the present analysis, we use only data from 2006.[3] The solar data are measured using geostationary satellites, which calculate the direct normal incident (DNI) and global horizontal incident (GHI) radiation by measuring the sunlight that is reflected or scattered in the atmosphere. The weather data come from the North American Regional Reanalysis (NARR) and are coarser in time (three-hour) and space (40 x 40 km^2). The NARR weather data are spatially joined to the solar data using a nearest neighbor method and linearly interpolated to match the solar hourly time step.

We calculate hourly PV generation using the PVWatts model,[4] which estimates the electricity production of a grid-connected PV system based on system characteristics (e.g., tilt, derate, orientation), solar radiation, and weather data.[5] We convert the hourly generation output from PVWatts into hourly capacity factors (Figure 1).

2.2 Temporal Resolution Reduction Methods

One of the main challenges of working with resource data—in general and in energy models—is the large volume of data especially for uses that require high spatial and temporal resolution. The volume of data can adversely impact computational runtime for clustering algorithms. To examine the possibility of reducing data volume, we reduce the data size in four ways. Each method attempts to retain the variability of the capacity factors at key time spans. In Section 4, implications of different segmentation methods are presented. The four data reduction methods include:

- Daytime (DT) Only: We subset the hourly capacity factors by extracting only daylight hours (9 a.m. – 4 p.m. local standard), as PV output is negligible for all regions during non-daylight hours.
- Peak Load Day: We subset the hourly capacity factors by the 24 hours in which the peak load occurs.[6] Historical (2006) electricity load profiles are from the Western Electricity Coordinating Council (WECC) Transmission Expansion Planning Policy Committee (TEPPC) 2020 report [16]. The peak load day was identified by summing the WECC TEPPC load profiles in Colorado. We select these hours because of the importance of peak demand to utility planning. Clustering over this period of time may be preferable over annual profiles for analyses or models that are focused on resource adequacy and peak capacity needs.
- Peak Load Week: We apply the same method as for the Peak Load Day segmentation except the full week (168 hours) in which the peak load occurs is used.

[3] Solar data from 2006 are chosen as they overlapped the load and wind data used in the Resource Planning Model [10, 13].
[4] http://www.nrel.gov/rredc/pvwatts/
[5] The nominal PV system configuration used in PVWatts to develop hourly profiles includes the following specifications: 20-MW system size; 0.85 derate factor; 1-axis tracking; 0-degree tilt.
[6] The Peak Load Day and Peak Load Week methods also sample only the daytime hours and therefore include fewer than 24 and 168 hours, respectively.

- Wavelet Transformation: A discrete wavelet transformation was created from the daytime hours dataset to determine whether clustering on the wavelets would provide similar or better results than modeling the entire hourly dataset. Wavelet decomposition is one method of looking at the variability within specific temporal frequencies within a time series [17]. Wavelet decomposition is beneficial in time series analysis because it provides a view into the variability at specific frequencies through an estimate of the energy at that frequency while retaining all original data [17]. We use the default Daubechies S8 symmlet transform. In doing so, we reduce the data used in the clustering process from 8,760 values per data point to 11.

2.3 Clustering Methods

Aggregation of spatial data is often a necessary part of spatial data analysis. Compatibility of datasets within a single analysis and spatial or temporal resolution constraints are two common reasons for this. Data aggregation can be accomplished in numerous ways, but a significant component in an aggregation process is the decision whether to aggregate based on predefined regions or discover regions using the relationships in the data themselves. The latter case is referred to as clustering and it is the method used in this analysis. Clustering is essentially a process in which similar areas are categorized into regions that have some type of internal consistency and are different from surrounding regions.

There are numerous methodologies for categorizing data and this research is not intended to completely evaluate or compare these methods. For the purpose of this research, two clustering methods are used and the results compared.

First, traditional k-means clustering [14] is selected to provide a baseline or benchmark classification to use in comparing other classification methods. The k-means algorithm is a distance-based clustering algorithm. The number of desired number of clusters, the parameter k, must be specified as an input parameter for this algorithm. Given this input parameter, k centroids are selected at random. Each data point is assigned to the centroid with the shortest distance, the centroids are recalculated based on these assignments, and the process is repeated until there are no changes in centroid assignment. On completion, data points that have the same centroid assignment are considered to be in the same cluster. The k-means algorithm is not explicitly spatial and so the parameters for latitude and longitude for each data point are included in the vector used for classification. Although k-means does result in regional clusters using this strategy, many resulting clusters are not contiguous and are spread out across the geographic space of the region (see Figure 1 for examples of non-contiguous regions from k-means clustering). A secondary process would be needed to finalize or refine the regions to remove these potentially undesirable traits.[7]

The second method explored is the max-p clustering algorithm [15]. Unlike k-means clustering, the max-p algorithm does not require the user to predefine a number of clusters. Instead, the algorithm is designed to find the optimal combination of clusters based on retaining spatial homogeneity while ensuring that the regions remain within some predefined, spatially constrained, threshold [15]. As a result, the algorithm returns spatial contiguous regions. To ensure a direct comparison between k-means and max-p, a threshold is calculated such that the

[7] For example, connectedness of regions might be required under some model structures.

total number of clusters approximately matched k. This is accomplished using a simple calculation in which the total number of data points is divided by the desired number of clusters. All of the data points are given an additional attribute with a value of 1, and the sum of this parameter for each cluster is used as the maximum threshold for max-p. In this way, we consistently generate datasets with a known number of clusters using both k-means and max-p.

2.4 Evaluation Process

Traditionally, determining the best clustering process is hampered by selecting the "best" number of clusters or, in the case of max-p, the limiting value determining the size of each cluster. Finding this optimum number is less of a concern in our analysis than is setting the resolution of the data to a specific value, and hence setting the number of clusters, is a desired outcome of the process. For this reason, the process of evaluating combinations of data and method is based on identifying the best performers within two ranges of desired resolutions for each region. These ranges are 20–40 clusters for the low-resolution version of the dataset and 120–140 clusters for the high-resolution version of the dataset.

Automated cluster evaluation is challenging when optimal boundaries are not known. If there is no ground truth, many different combinations of clusters can provide similar results. In our analysis, two tests are selected to evaluate clusters from each clustering method. A goal of this research is the reduction of spatial resolution in resource datasets used in modeling. As some representative value will be calculated for each cluster and used in further analysis, it is critical that the data within each cluster be as representative of the whole cluster as possible. Because of this, both tests are selected on the premise that they would provide some measure of consistency within each cluster.

The first test involves calculating the sum of squared within (SSW) each cluster. This calculation simply sums the squared differences of each observation from the overall mean within the cluster, where smaller values representing more similar resource within a cluster. The resulting values provide a representation of the variance within the cluster.

The second test involves estimating the capability to use a subset of the data points within a given cluster to estimate the values of other data points within the same cluster. This cross-validation test is intended to determine how accurately the data points within a single cluster could be used to predict a randomly selected data point within the same cluster. The first step in the process is to reduce the total number of points by 10% through random selection. Then, 10% of the points in the cluster are selected one at a time and a linear regression is performed using all of the remaining points (independent variables) in the cluster to model each model each select point (dependent variable), resulting in an adjusted R^2 value. The adjusted R^2 value gives us an indication of the amount of the variance that can be explained by the model, and hence the level of accuracy with which the remaining values in the cluster can estimate the missing values.

In an effort to put into perspective the differences between the results of this testing for each run, the resulting statistics are compared to determine whether the differences were statistically significant. As the data are not found to be normal, the Wilcoxon rank-sum test is used in this test.

3 Results

When considering the results of these evaluations, it is first important to ask whether differences between the results are significant. In this case, the results of both tests for all datasets are compared with each other using the Wilcoxon rank-sum test. Specifically, the set of SSW results from one run are compared to the set of SSW results from all other runs to determine whether the differences between the result sets are significant. The same process was performed for the Adjusted R^2 values. Datasets with a high p value from this test, in this case $p >= 0.05$, are considered to have failed to reject the null hypothesis that the data are from the same population, and they are therefore statistically equivalent. The combinations of runs that fail to reject the null hypotheses are outlined as rows in Table 1. Each row indicates a set of results that are deemed statistically equivalent to each other using this test.

Table 1. Datasets Determined to be from the Same Population Using the Wilcoxon Rank-sum Test

Statistically Equivalent SSW Results (for low-resolution cluster sets)				
	Peak Load Week / K-Means	Wavelet / Max-P		
Statistically Equivalent SSW Results (for high-resolution cluster sets)				
	None			
Statistically Equivalent Median Adjusted R^2 Results (for low-resolution cluster sets)				
	All DT Hours / K-Means	Peak Load Day / K-Means	Peak Load Week / K-Means	
	Peak Load Day / K-Means	Peak Load Day / Max-P	Peak Load Week / K-Means	
	Peak Load Day / Max-P	Peak Load Week / K-Means		
	Peak Load Week / Max-P	Wavelet / Max-P		
Statistically Equivalent Median Adjusted R^2 Results (for high-resolution cluster sets)				
	All DT Hours / K-Means	Peak Load Day / Max-P	Peak Load Week / Max-P	Wavelet / Max-P
	Peak Load Day / K-Means	Peak Load Week / K-Means		
	Peak Load Day / Max-P	Peak Load Week / Max-P	Wavelet / Max-P	
	Peak Load Week / Max-P	Wavelet / Max-P		

[a] Each row in the table represents a set of results that are deemed statistically equivalent to each other using this test.

Figure 2 shows example clusters generated through the two algorithms (k-means and max-p) evaluated and for two resolution levels. The k-means algorithm produces some clusters with disconnected data points. This could be resolved by applying another regionalization algorithm to the k-means results; however, that would result in far more clusters for each run. The max-p algorithm spatially constrains the clusters and therefore has no disconnected clusters in the results. In both cases, the algorithms produce clusters with simpler edge complexity in the eastern part of the state and clusters that have more edge complexity in the mountainous western part of the state. This is likely due to the high variability of the data in the western part of the state.

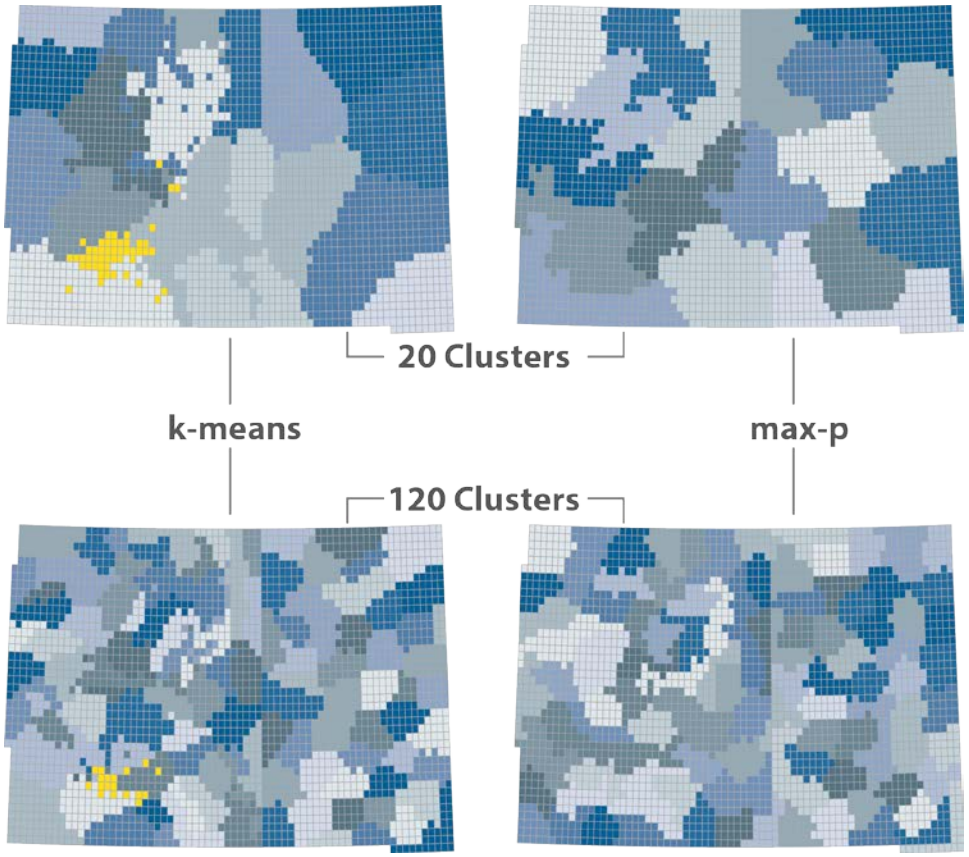

Figure 2. Examples of cluster regions for k-means and max-p at 20 and 120 clusters each

Disconnected data points within a single cluster are highlighted for one sample cluster (yellow) in both k-means result sets.

In Figures 3 through 9, the results of the two tests are shown in box-and-scatter plots for each of the datasets at both the low and high-resolution clusters. Results in the figures are standardized using a simple 0-1 scaling (X - min X)/(max X - min X), where X represents the metric presented in each figure.

Figures 3 and 4 clearly show that the variance of the SSW values is less in the higher-resolution clusters than it is in the lower-resolution clusters for all datasets. Also, the datasets with more attributes on which to cluster (All DT Hours with 2,920 attributes and Peak Load Week with 55 attributes) generally perform better with both algorithms. The exception to this observation is the performance of the wavelet dataset (10 attributes) with the max-p algorithm, which performs second best of the datasets. Additionally, the difference in performance between k-means and max-p is accentuated in the lower-resolution clusters, where dataset and algorithm selection seem to have a much greater impact on the results. Figure 6 demonstrates that the relative differences between the datasets is fairly constant for this statistic and that the performance gets exponentially better as the number of clusters increases, but only to about 100 clusters, where performance improvement becomes more linear. This is expected as the variability of the clusters decreases significantly as the number of data points within each cluster decreases.

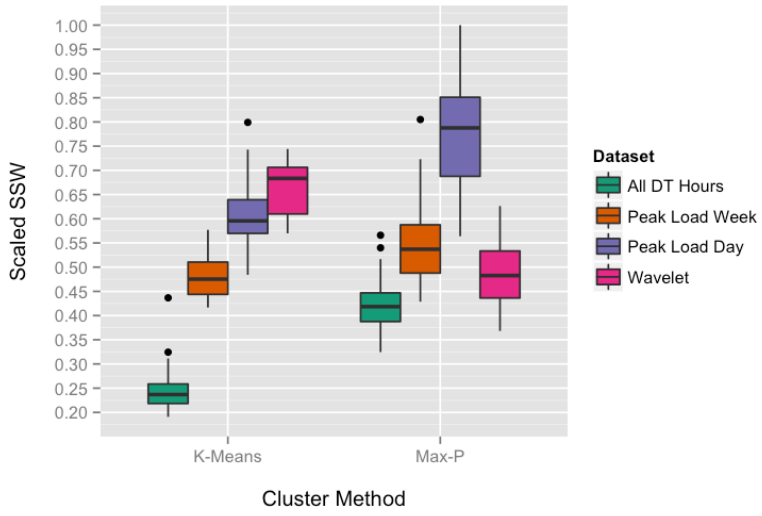

Figure 3. All estimates of the SSW for each dataset and clustering method in the low-resolution clusters

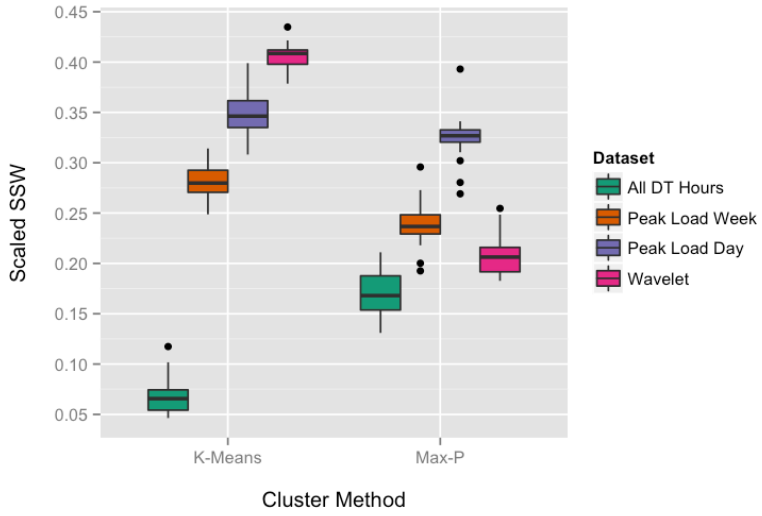

Figure 4. All estimates of the SSW for each dataset and clustering method in the high-resolution clusters

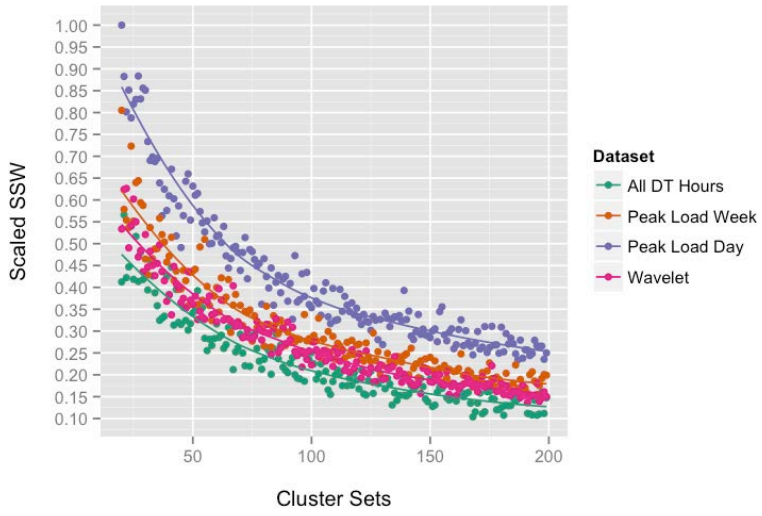

Figure 5. All estimates of the SSW for each dataset and clustering method at all cluster levels with locally weighted scatterplot smoothing (LOWESS) shown as a line for each dataset

The results of the cross-validation tests (Figures 6 through 9) tell a slightly different story.[8] The difference in the performance of each of the datasets at the two resolutions is less pronounced with this statistic. Also, performance in this test is better at lower resolutions than it is at higher resolutions. This is expected, as the lower-resolution clusters will have more points with which to build a model, which will result in a better-adjusted R^2 value.

The max-p clusters perform better than the same datasets using k-means in all cases. The full dataset performance—although it is still better in the max-p clusters—is much closer to the other datasets in performance. The wavelet dataset performs poorly with the k-means algorithm and close to the level of the full dataset when using max-p at both resolution levels. Also, the results for both algorithms are highly correlated between datasets, according to the Wilcoxon results (Table 1). As with the SSW values, the relationship between the performance of each of the datasets is shown to be consistent for all cluster sizes.

Because the cross validation is run on each cluster within each clustered dataset—between 20 and 199 times depending on the dataset resolution—we are able to look at the percentiles of the result of the test. Figure 8 shows the values of the cross validation for both max-p and k-means for all runs. This figure highlights the tighter spread of both the values themselves and the range of the 5th and 95th percentiles for the max-p algorithm.

[8] Higher R-squared values in Figures 6 through 9 indicate better cluster quality, whereas the opposite is true for Figures 3 through 5, in which lower SSW values reflect better clustering.

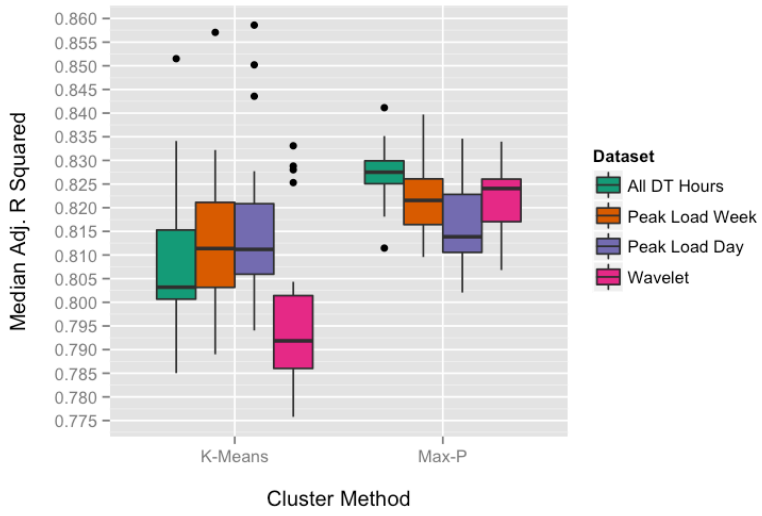

Figure 6. All estimates of the median adjusted R^2 for each dataset and clustering method in the low-resolution clusters

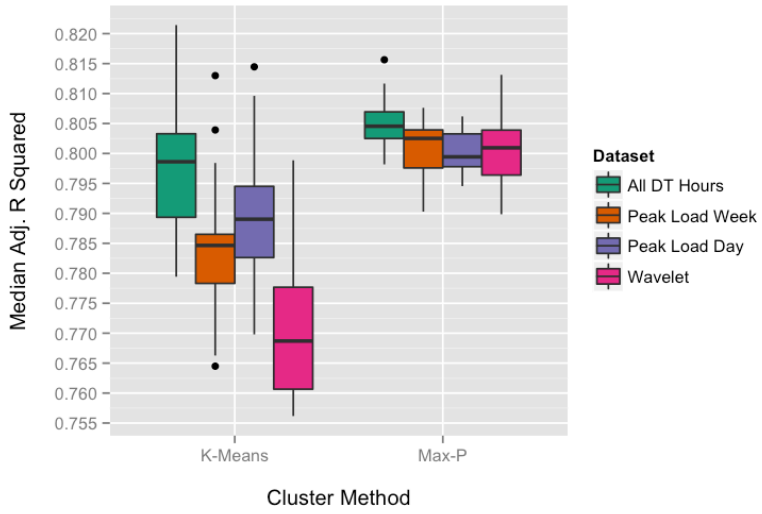

Figure 7. All estimates of the median adjusted R^2 for each dataset and clustering method in the high-resolution clusters

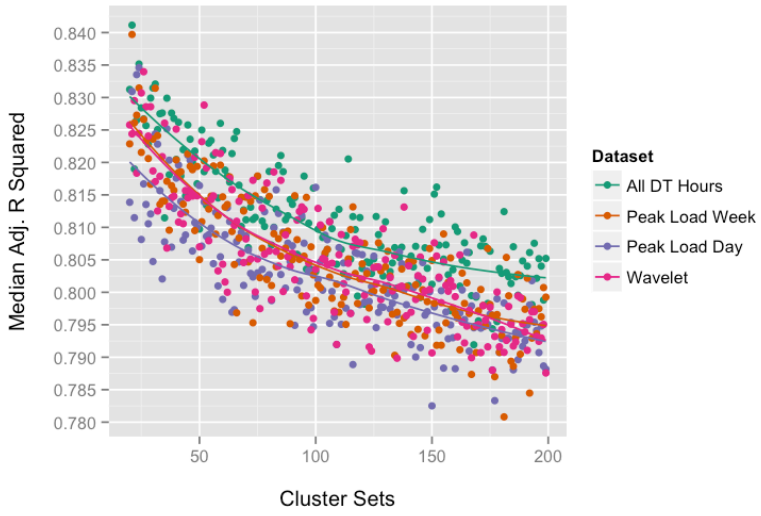

Figure 8. All estimates of the median adjusted R^2 for each dataset and clustering method at all cluster levels with LOWESS (locally weighted scatterplot smoothing) shown as a line for each dataset

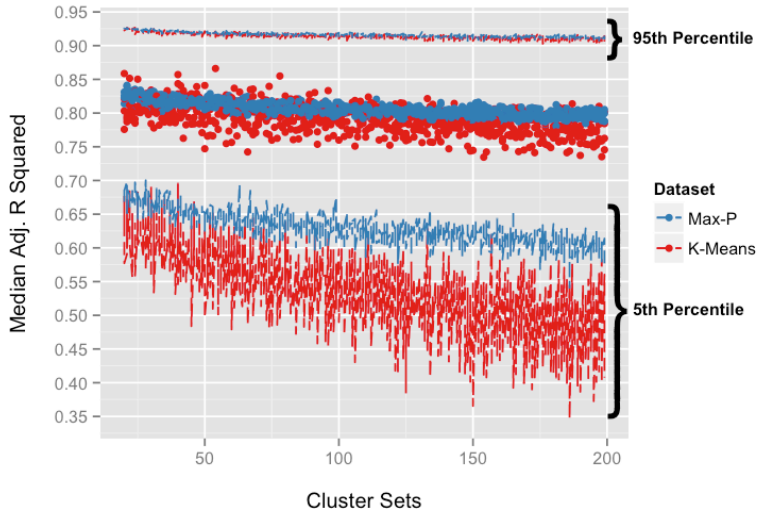

Figure 9. All estimates of the median adjusted R^2 at all cluster levels including the 5th and 95th by cluster method

4 Conclusions

The purpose of this research is twofold: outline a process for determining the best clustering method and dataset segmentation type, and outline the process for selecting values for a specific modeling application that could be expanded to include a variety of datasets and statistical tests for cluster performance. In the context of these goals, several conclusions can be drawn from our results. Upon comparing the results of the SSW and cross-validation tests, we can see creating a cluster with low internal variability does not necessarily provide a cluster that can reliably estimate missing values or generate typical values for the cluster. In a comparison of the algorithms, max-p seems better able to use the information in the full dataset, and significantly better able to use the information available in the wavelet dataset, to estimate missing values. Lastly, although both algorithms produce results that are consistent across all datasets regardless of the resolution of the clusters, attention should be paid when determining the best combination of data reduction method and clustering method at lower resolutions.

Given our results, it would seem that the max-p algorithm applied to the full dataset would give the best results. This combination provides spatially contiguous clusters, better performance when extracting a reasonable approximation of the data represented within a cluster, and more consistent results over many runs. Reducing the variability within clusters, as measured with the SSW estimates, is important but only in so far as doing so impacts the ability to extract a reasonable representation of the data in the cluster.

Given that the results from the max-p algorithm are very similar and even statistically equivalent in some cases, the option to swap one of the data subsets for the full dataset is viable if computational or temporal limitations require a smaller input dataset. If data segmentation is needed, the results suggest that the wavelet dataset represents enough of the information in the time series at each point to use in place of the full dataset for high-resolution cluster runs.

We identify several options to expand this research. Initially, several other clustering methods could be added as options, depending on the nature of the datasets being evaluated. Options include DBSCAN [18], self-organizing maps [19], and clustering with mixture models. Additionally, the evaluation statistics presented here are sufficient for general approximation of the runs; however, several enhancements could be made in this area. These include adding additional evaluation methods such as using a dynamic time warping (DTW) value to measure overall difference between the values within a cluster or enhancing the model-based evaluation by testing the capability to predict missing values within the cluster.

Because the data being clustered are spatiotemporal, investigating fuzzy clustering within specific time ranges (e.g., seasonal clustering) and then determining changes in cluster inclusion over time might offer some insight into selection of the best methods. As topography, localized weather, and other external factors can impact renewable energy potential, exploring comparisons of clustering methods with respect to these factors would also be useful. Undertaking focused efforts to understand the performance within mountainous or coastal regions, or exploring the removal of regions that have inherently low potential for renewable energy development due to population or resource restrictions, would be a useful addition to this research. Finally, while we have focused on solar PV profiles, clustering techniques can be applied to other renewable energy resources, such as wind, or other time-dependent data relevant for energy models.

5 References

[1] AWS https://www.awstruepower.com

[2] 3Tier http://www.3tier.com/en/

[3] NSRDB http://rredc.nrel.gov/solar/old_data/nsrdb/

[4] MERRA http://gmao.gsfc.nasa.gov/research/merra/

[5] Lopez, A., Roberts, B., Heimiller, D., Blair, N., Porro, G., 2012. U.S. Renewable Energy Technical Potentials: A GIS-Based Analysis, NREL/TP-6A20-51946. Golden, CO: National Renewable Energy Laboratory.

[6] Eurek, K., Denholm, P., Margolis, R., Mowers, M., 2013. Sensitivity of Utility-Scale Solar Deployment Projections in the SunShot Vision Study to Market and Performance Assumptions. NREL/TP-6A20-55836. Golden, CO: National Renewable Energy Laboratory.

[7] Sullivan, P., Eurek, K., Margols, R. 2014. Advanced Methods for Incorporating Solar Energy Technologies into Electric Sector Capacity-Expansion Models: Literature Review and Analysis. NREL/TP-6A20-61185. Golden, CO: National Renewable Energy Laboratory.

[8] Short, W., Sullivan, P., Mai, T., Mowers, M., Uriarte, C., Blair, N., Heimiller, D., Martinez, A., 2011. Regional Energy Deployment System (ReEDS). NREL/TP-6A20-46534. Golden, CO: National Renewable Energy Laboratory.

[9] Nelson, J., Johnston, J., Mileva, A., Fripp, M., Hoffman, I., Petros-Good, A., Blanco, C., Kammen, D.M., 2012. High-Resolution Modeling of the Western North American Power System Demonstrates Low-Cost and Low-Carbon Futures. *Energy Policy*, 43 (April), pp. 436-447.

[10] Mai, T., Drury, E., Eurek, K., Bodington, N., Lopez, A., Perry, A., 2013. Resource Planning Model: An Integrated Resource Planning and Dispatch Tool for Regional Electric Systems. NREL/TP-6A20-56723. Golden, CO: National Renewable Energy Laboratory.

[11] Openshaw, S., 1973. A Regionalisation Program for Large Data Sets, *Computer Applications*, 3-4, pp. 136–147.

[12] Webster, R., Burrough, P. A., 1972. Computer-Based Soil Mapping of Small Areas from Sample Data II, *European J. of Soil Science*, 23(2), pp. 222–234.

[13] Mai, T.; Barrows, C.; Lopez, A.; Hale, E.; Dyson, M.; Eurek, K. (forthcoming). Implications of Model Structure and Detail for Utility Planning: Scenario Case Studies using the Resource Planning Model.

[14] MacQueen, J., 1967. Some Methods for Classification and Analysis of Multivariate Observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[15] Duque, J.C., Anselin, L., and Rey, S.J., 2012. The max-p-regions problem. *Journal of Regional Science*, 52 (3), pp.397–419.

[16] Western Electricity Coordinating Council (WECC), 2012. Assumptions Matrix for the 2020 Transmission Expansion Planning Policy Committee (TEPPC) Dataset. http://www.wecc.biz/ library/StudyReport/Documents/Assumptions%20Matrix%20for%20the%202020%20TEPPC %20Dataset.pdf

[17] Percival D., Walden A., 2000. Wavelet Methods for Time Series Analysis. Cambridge, MA: MIT Press.

[18] Ester, M., Kriegel H.-P., Sander J., Xu X., 1996. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in: Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, 1996, pp. 226–231.

[19] Kohonen, T., 1982. Self-organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43, pp. 59-69.