Accuracy of Clustering as a Method to Group Distribution Feeders by PV Hosting Capacity

Robert J. Broderick, Karina Munoz-Ramos, and Matthew J. Reno
Distributed Solar Grid Integration
Sandia National Laboratories
Albuquerque, NM, USA

***Abstract*— This paper examines the accuracy of clustering techniques for predicting hosting capacity. Hosting capacity results for 214 study feeders were used to predict a range of hosting capacities for an addition 7929 feeders using clustering techniques. Several methods were explored to try to improve the accuracy for predicting hosting capacity, including increasing the number of clusters, selecting variables that are highly correlated to hosting capacity for clustering, and weighting highly correlated clustering variables. The average normalized interquartile range (ANIQR) is used to compare the accuracy of several clustering methods for predicting hosting capacity.**

***Index Terms*—clustering methods, distributed power generation, power distribution, principal component analysis**

I. Introduction

As photovoltaic (PV) energy penetration continues to increase, utilities are becoming concerned about the impact that these systems will have on the distribution system. The analysis required to quantify the impact of PV systems on the distribution system can be time consuming and costly. Gathering the data and creating feeder models that accurately predict the behavior of a distribution feeder is a difficult and tedious process. Recently, data clustering techniques for grouping distribution feeders with similar characteristics have been proposed for simplifying PV interconnection studies [1], [3-7]. However, the accuracy of these clustering methods to predict hosting capacity on distribution feeders from a given cluster has not been evaluated. This work explores the accuracy of clustering techniques for predicting hosting capacity and discusses possible limitations of these techniques.

II. Background

A k-means clustering methodology was employed in [1] with the objective of developing a more accurate screening criteria for PV interconnection. Correlation maps on variables of interest were used to select the set of variables used for clustering, principal components analysis (PCA) was used to transform the data along with unit variance scaling of the data set to obtain equal weighting among the selected clustering variables and the Cubic Clustering Criterion (CCC) [2] was used to determine the number of clusters. Variables used for clustering consisted of variables related to feeder topology, feeder voltage control, feeder load data and customer data. Over 8,000 feeders from three different utilities were clustered. Clustering was performed separately on each individual utility. Eleven variables were used for clustering feeders from Utility 1 and Utility 3 and twelve variables were used for clustering feeders from Utility 2. Seven, ten and five clusters were obtained for Utility 1, Utility 2 and Utility 3 respectively.

Classifying feeders based on the hierarchical algorithm was first demonstrated in the PNNL Taxonomy Final Report [3]. Hierarchical clustering using Ward's method was used to cluster 13,007 low voltage (LV) feeders in [4]. Three statistical parameters including Semi-Partial R^2, Pseudo-F statistic and a Pseudo-T^2 Test were used to determine the optimal number of clusters for the data set. Seven variables were used for clustering including variables related to feeder topology, feeder load data and customer data. In the end nine clusters were selected to represent the 280 HV feeders and ten clusters were selected to represent the 13,007 LV feeders. In [5] a clustering algorithm was developed and used to cluster 1295 Arizona Public Service (APS) distribution feeders. Techniques used include, using a k-mediods algorithm for clustering, using the error ellipse method for outlier identification and removal and using the Calinski Index to select the optimal number of clusters. Sixteen variables were used for clustering and consisted of variables related to feeder topology, feeder voltage control, feeder load data and customer data. Nine clusters were used to cluster the 1295 feeders.

Four clustering approaches including hierchical, k-mediods++, improved k-means++ and Gaussian Mixture Model are employed and compared in [6]. The 232 feeders were partitioned into two data sets, one with PV and one without PV. Clustering was performed on both data sets individually. Four indices were used to determine the optimal number of clusters for each data set. The four indices include Variance Ratio Criterion (VRC), Similarity Matrix Indicator (SMI), Global Silhouette Coefficient (GS) and Average Silhouette Coefficient (AvgSC). Nineteen variables were used for clustering and consisted of elements related to customer

data, feeder topology, feeder load data and PV data. Eleven clusters in total were obtained to represent the 232 distribution feeders.

Several other clustering approaches have been proposed [7]. However, the accuracy of these clustering methods for predicting PV hosting capacity has not been explored. This paper utilizes results, including hosting capacity, for 214 feeders that have been thoroughly analyzed and attempts to determine a hosting capacity range for more than 7,900 additional feeders by utilizing clustering techniques. Metrics for quantifying the accuracy of the clustering techniques for predicting hosting capacity are developed.

III. Methodology

A. Clustering Algorithm (k-means)

For this project, a k-means clustering approach was used. Variables used for clustering were selected based on the impact they might have on differentiating feeder types and on DG hosting capacity. Because the optimum number of clusters is more accurtely achieved when the chosen variables are independent of each other, the intial variables were analyzed using a correlation map and pairs of highly correlated variables were examined more closely to determine if it was beneficial to remove one of the variables before clustering. K-means clustering algorithms can be very sensitive to outliers [2] and therefore, feeders that were considered outliers were removed from the data set. The sensitivity of K-means to outliers is shown in Fig. 1 where the choice of the number of outliers significantly changes the solution for the optimal number of clusters from 8 to 11 clusters.

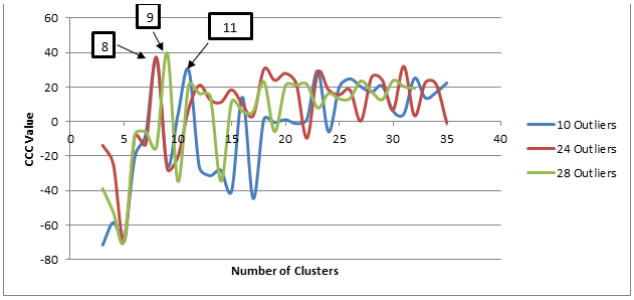

Fig. 1. Cluster solution dependence on outlier selection.

The most difficult problem in cluster analysis is how to determine the optimal number of clusters. A quality metric for determining the optimum number of clusters is based on the CCC. The optimum number of clusters can be derived from a CCC value based on minimizing the within-cluster sum of squares. Although not a mathematical law and more of a rule of thumb that has been validated in the statistical community, the optimal number of clusters can be determined by plotting the CCC value against the number of clusters and finding a local maximum after the CCC rises above 2 and before it drops below 2. Fig. 1 shows three solutions of 8, 9 and 11 optimal clusters using this rule. It is important to note that you are not necessarily looking for the highest CCC value as this will be achieved when a cluster is created for each individual element which is not representative of optimal clustering. Statistical analysis was performed using the SAS JMP software tool to calculate the CCC value for each cluster number.

B. Analyzed Study Feeders

In order to characterize the accuracy of clustering feeders to determine hosting capacity, 214 feeders have been selected and analyzed for their hosting capacity. The distribution systems are from various utilities around the United States. The majority of feeders also included a year of substation SCADA measurements from the utility. The power systems models include the full details about voltage regulator settings, capacitor switching controls, and up to 6000 buses per distribution system.

Fig. 2 shows characteristics of the total set of 8143 feeders as well as the 214 study feeders. About 67% of all feeders and 65% of the study feeders are 12 kV feeders. Almost 88% of all feeders have no regulators as compared to 80% of the study feeders. Overall there are no major differences between the characteristics of the collection of 8143 feeders and the 214 study feeders.

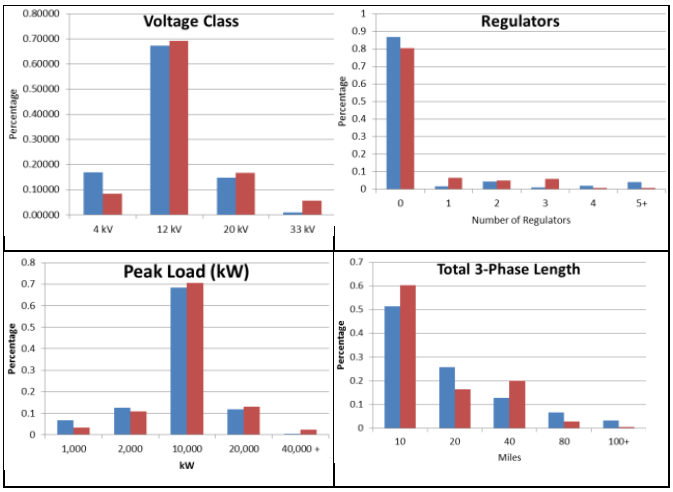

Fig. 2. Feeder characteristics for full set of feeders (Blue) and 214 study feeders (Red)

C. Feeder PV Hosting Capacity Analysis

Each of the study feeders is analyzed using a detailed hosting capacity analysis. The methodology in [8-9] is used to investigate a large number of potential PV scenarios (combinations of PV size and location) in OpenDSS [10]. On average, there are around 40,000 PV scenarios analyzed per feeder.

For each PV scenario, a series of simulations is performed to determine if that particular scenario would cause issues on the distribution system. The simulations include a range of load values that occur during daytime hours throughout the year, a range of feeder states as far as regulation equipment taps and switching capacitor states, and simulation of extreme PV output ramps. Steady-state voltage violations are determined using ANSI C84.1, thermal violations are defined by the component's amp rating, and temporary voltage violations are determined using the ITIC (CBEMA) curve.

Using the detailed simulation results, the feeder hosting capacity is defined as the maximum PV size that can be placed anywhere on the feeder without causing issues.

IV. Validation/Accuracy/Analysis of Previous Clustering Algorithm

A. Clustering All California Feeders and Study Feeders

The data set used for clustering consisted of a total of 8,143 feeders with feeder characteristics for 7,929 California feeders and 214 feeders analyzed to determine the hosting capacity of the feeder. The clustering was performed using 8 feeder characteristic variables with a double weighting on feeder primary voltage. The other feeder variables were: Total 3-Phase miles, Total 1 & 2 Phase miles, Residential Customer %, Regulator #, Capacitor # and feeder peak load. These variables are typically easy for utilities to determine and were available for each of the 8,143 feeders.

B. Clustering Results

The clustering of a total of 8,143 feeders resulted in a solution of 8 clusters as shown by the colored regions in the biplot of Fig. 3 with study feeders shown by different markers. The study feeders are widely distributed across the clusters with the largest concentration in cluster 7, a 12KV cluster.

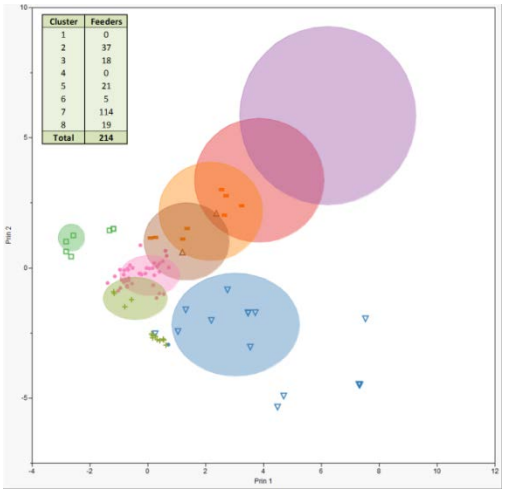

Fig. 3. Biplot of the 8 cluster solution for 8,143 feeders with study feeders shown by markers

The range of hosting capacities for each cluster is shown using boxplots in Fig. 4. Boxplots are useful for identifying outliers and for comparing distributions. The blue box is the interquartile range (IQR) and it represents the values between the 75th percentile and the 25th percentile covering the middle 50% of the data. The median for the cluster is shown by the red line and outliers are shown by the red "+" markers. The whiskers extend out to capture all values that are less than third quartile +1.5 IQR and greater than first quartile -1.5 IQR. Any data not included between the whiskers is plotted as an outlier.

The number of study feeder in each cluster is shown in the upper portion of the figure. Clusters with four or more feeders were plotted and analyzed. The minimum number of study feeders per cluster was set at four to ensure sufficient data to define a meaningful range of hosting capacities per cluster. The most populated cluster is cluster 7 with 114 study feeders. It is a 12-13.8 kV cluster with a range of hosting capacities from 0.2 to 4.3 MW excluding outliers. The box height is 1.5 MW which is the range of hosting capacity values for 50% of the study feeders in the cluster. The cluster with the greatest range of hosting capacities is cluster 2 with 37 study feeders. It is a 19.8kV to 34.5kV cluster with a range of hosting capacities from 0.3 to 10.2 MW excluding outliers. The box height is 4.6 MW which is the range of hosting capacity values for 50% of the study feeders in the cluster.

The normalized box height of each cluster is calculated by dividing the IQR by the median hosting capacity of the cluster. The variation across the whole set of clusters can be measured by taking the average of all the normalized cluster box heights for clusters with 4 or more feeders.

$$ANIQR = \frac{1}{n}\sum_{1}^{n}((P75 - P25) \div P50) \quad (1)$$

Where n equals the cluster number and P equals the percentile. The box heights or IQR represents the central variation in the cluster hosting capacity distribution. The normalized IQR removes the bias of large hosting capacity feeders and taking the average ensures that we are improving the overall cluster solution. The ANIQR for this set of clusters was 109%, meaning that the average variation in hosting capacity for each cluster was 109% of the mean hosting capacity of that cluster.

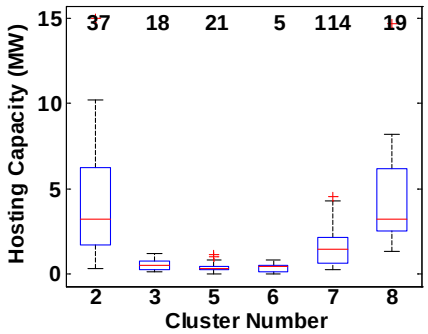

Fig. 4. Boxplot of hosting capacity variation per cluster for 8 cluster solution for 8,143 feeders using 8 c luster variables.

C. Clustering Accuracy to Predict Hosting Capacity

Ideally the clustering would provide a tight range of hosting capacities for each cluster, but the results show that the range of hosting capacities varies widely depending on the cluster with some box plots showing a narrow range and others a very broad range. The variation in hosting capacity within the cluster is not dependent on the location within the cluster. The source of the hosting capacity variation in Fig. 2 was investigated by looking at the hosting capacity violation type per feeder in each cluster. Fig. 5 uses the marker shape to denote the type of issue that first violated on each study feeders which determined the hosting capacity of the feeder. Cluster 6 for example, has 12 kV feeders that all have hosting capacity limitations caused by over-voltage conditions. The relative size of the marker indicates the relative feeder peak load of each feeder. The large variation in the violation types within each cluster illustrates the difficulty of capturing the key characteristics that drive the hosting capacity value with clustering.

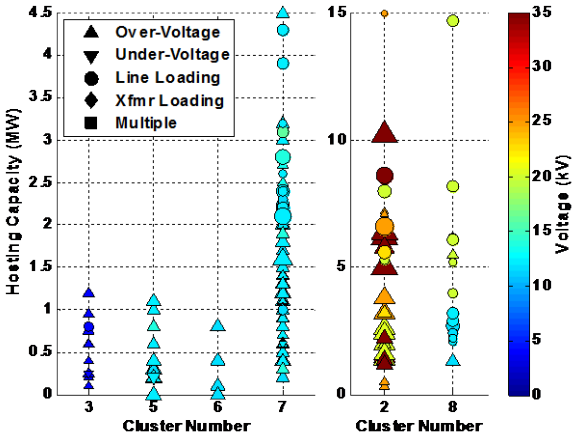

Fig. 5. Hosting capacity violation type shown by marker shape for each feeder in the clusters.

V. Variations in Clustering Methodology

A. Dependence of Accuracy on Number of Clusters

One possible solution to improve the accuracy of hosting capacity per cluster is to increase the number of clusters. Fig. 6 shows a 16 cluster solution and Fig. 7 shows a 32 cluster solution using the same variables as the 8 cluster solution discussed earlier. Clusters with four or more feeders were plotted and analyzed.

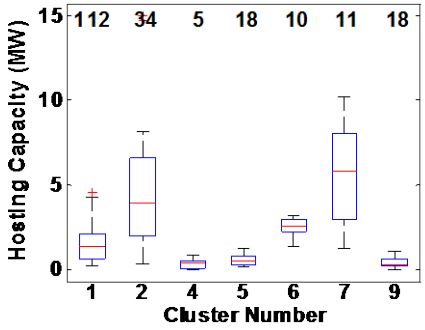

Fig. 6. Boxplot of hosting capacity variation per cluster for 16 cluster solution for 8,143 feeders using 8 cluster variables.

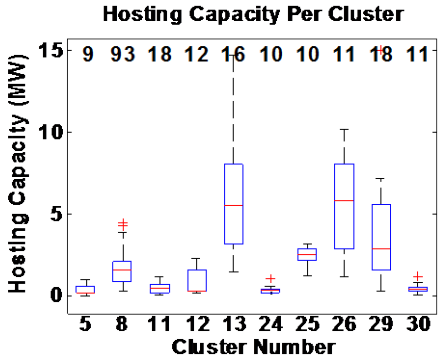

Fig. 7. Boxplot of hosting capacity variation per cluster for 32 cluster solution for 8,143 feeders using 8 cluster variables.

The ANIQR did not change materially: the 16 cluster solution had an ANIQR of 102% showing a slight improvement from the 109% for the 8 cluster solution and the 32 cluster solution had an ANIQR of 129% showing degradation. The box height for the most populated cluster did not change for the 16 cluster solution and decreased to 1.2 MW for the 32 cluster solution. The conclusion is that increasing the number of clusters does not particularly change the overall accuracy of the clustering solution.

B. Correlation of Clustering Variables to Hosting Capacity

Another possible solution to improve the accuracy of hosting capacity per cluster is to increase the weighting of key clustering variables based on their correlation factor (CF) with hosting capacity. Table I shows the correlation coefficients between the clustering variables used earlier and the hosting capacities of the 214 feeders. Using the correlation coefficients the relative weighting factors shown in the table were implemented.

TABLE I: SEVEN CFs & WEIGHTING FACTORS

Variable	Correlation Factor (|CF|)	Weighting
Primary Voltage (kV)	0.60	4X
Total 3-Phase Conductor (miles)	0.33	2X
Total 1 & 2 Phase Conductor (miles)	0.10	1X
Residential Customers (%)	0.33	1X
Number of Regulators	0.28	1X
Number of Capacitors	0.19	1X
Feeder Peak Load (kW)	0.31	2X

Table II below shows ANIQR for each of the three cluster solutions. A significant improvement occurred in both the 8 and 32 cluster solutions, but the 16 cluster solution got worse.

TABLE II: ANIQR FOR CLUSTERING SOLUTIONS

Cluster Solution with Weightings	ANIQR
8	94%
16	106%
32	77%

Although the weighting of the initial variables did improve the ANIQR in two cases the range in hosting capacity variation is still very high even in the best case at 77%.

The persistence of a wide variation in hosting capacities in each cluster indicates that the clustering variables chosen may not be correlated enough with hosting capacity to exactly predict it and perhaps adding more highly correlated variables will reduce the variation in each cluster and more accurately predict the hosting capacity.

C. Addition of New Clustering Variables

There are currently a fairly limited number of variables of feeder characteristics available for clustering. In the future, new feeder characteristic variables could become available, and should be studied to determine which are the most important for predicting hosting capacity. This could include things like calculating the X/R ratio at key buses on the circuit. The set of feeder characteristics for the 214 study feeders is much more comprehensive set than is available for clustering of the 7,929 California feeders. The data set includes many important characteristics such as feeder impedance values, short circuit current capability, X/R ratios

of various buses, etc. that were not in the California feeder characteristic data set. Table III shows the ranked order of correlation factors for the study feeder characteristics and the hosting capacity of the study feeders

TABLE III.CFs & WEIGHTING FACTORS FOR 214 FDRs

Variable	Correlation Factor (|CF|)	Weighting Factor
Feeder Voltage (kV)	0.69	2X
Minimum Short-Circuit Current	0.67	1X
Impedance (3-phase buses at feeder voltage) Min X/R	0.66	1X
Service Transformers Median Size (kVA)	0.52	1X
Density (kW/sq-km)	0.48	1X
Min X/R to VREG	0.45	1X
3 phase conductor rating- Lowest	0.44	1X
Max R to VREG	0.42	1X
Impedance (3-phase buses at feeder voltage) Max R	0.41	1X
3 phase conductor rating- Weighted Average	0.39	1X
Daytime Peak (MW)	0.36	1X
Avg X/R to VREG	0.33	1X
Percent Residential Customers	0.31	1X

A clustering analysis was performed only on the 214 feeder studies. Fig. 8 shows the boxplot for the most optimal clustering for the 16 cluster solution. The ANIQR for the 16 cluster solution was 76%, the lowest value found. Although this is slight improvement, the average variation in hosting capacity of 76% is still very high and demonstrates that the clustering method has limited accuracy, even when the new highly-correlated feeder characteristics are utilized. There are too many interconnected pieces to group feeders into precise ranges of hosting capacity. For example,

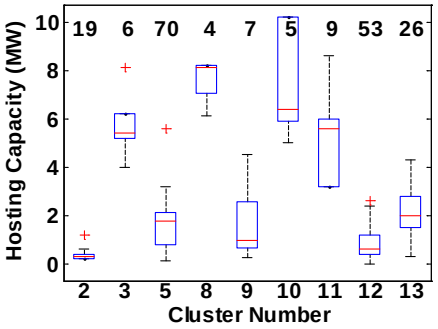

Fig. 8. Boxplot of hosting capacity variation per cluster for 16 cluster solution for 214 study feeders using 14 cluster variables.

line rating or capacitors might not matter if the voltage set point is very high, but for a lower regulator set point, the new hosting capacity variable determination is driven by capacitors or line rating. Clustering is a simple method that, by its nature, is unable to capture the interaction between the different variables. Not only is clustering for PV hosting capacity prediction imprecise due to interrelation of variables, but there are so many unique features of feeders that are difficult to capture. For example, one of the study feeders had a voltage limit on hosting capacity that was driven by a service transformer with a tap setting above nominal resulting in higher voltages than normal. These types of feeder characteristics and operational characteristics of utility feeders will remain difficult to quantify.

VI. CONCLUSIONS

The accuracy of clustering as a method to group distribution feeders into specific ranges of PV hosting capacity has been shown to be relatively inaccurate. Clustering is still useful as it provides good separation between clusters in many cases, but it has its limitations. The best clustering solutions for the various methods explored did not predict the hosting capacity accurately and the best solution had an average hosting capacity variation of 76%. Clustering will never perfectly group feeders such that all unique characteristics match with a single PV hosting capacity for the feeder, but it can provide a rough estimate of the hosting capacity for similar types of feeders.

ACKNOWLEDGMENTS

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000

REFERENCES

[1] Broderick, R.J.; Williams, J.R., Munoz-Ramos, K.; "Clustering method and representative feeder selection for the California Solar Initiative," *Sandia National Laboratories report SAND2014-1443: Albuquerque, New Mexico, February 2014*

[2] SAS Institute Inc., SAS Technical Report A-108 Cubic Clustering Criterion, Cary, NC. 1983. 56pp.

[3] K. P. Schneider, Y. Chen, D. P. Chassin, R. G. Pratt, D. W. Engel, and S. E. Thompson, "Modern Grid Initiative Distribution Taxonomy Final Report," PNNL, Richland, WA, Technical Report PNNL-18035, 2008.

[4] Yingliang Li; Wolfs, P., "Statistical identification of prototypical low voltage distribution feeders in Western Australia," *IEEE Power and Energy Society General Meeting*, pp.1-8, 22-26 July 2012

[5] Cale, J.; Palmintier, B.; Narang, D.; Carroll, K., "Clustering distribution feeders in the Arizona Public Service territory," *IEEE 40th Photovoltaic Specialist Conference (PVSC)*, pp.2076-2081, 8-13 June 2014

[6] Rigoni, V.; Ochoa, L.F.; Chicco, G.; Navarro-Espinosa, A.; Gozel, T., "Representative Residential LV Feeders: A Case Study for the North West of England," *IEEE Transactions on Power Systems*, no.99, pp.1-13

[7] F. Dehghani, M. Dehghani, H. Nezami and M. Saremi, "Distribution Feeder classification based on self-organized maps" (Case Study: Lorestan Province, Iran).

[8] K. Coogan, M. J. Reno, S. Grijalva, and R. J. Broderick, "Locational Dependence of PV Hosting Capacity Correlated with Feeder Load," in IEEE PES Transmission & Distribution Conference & Exposition, Chicago, IL, 2014.

[9] M. J. Reno, K. Coogan, S. Grijalva, R. J. Broderick, and J. E. Quiroz, "PV Interconnection Risk Analysis through Distribution System Impact Signatures and Feeder Zones," IEEE PES General Meeting, 2014.

[10] M. J. Reno and K. Coogan, "Grid Integrated Distributed PV (GridPV) Version 2," Sandia National Labs SAND2014-20141, 2014.