Clustering Methodology for Classifying Distribution Feeders

Robert J. Broderick and Joseph R. Williams

Sandia National Laboratories, Albuquerque, New Mexico, 87185, USA

***Abstract* — The screening process for DG interconnection procedures needs to be improved in order to increase the penetration of PV systems on the distribution grid. A significant improvement in the current screening process could be achieved by finding a method to classify the feeders in a utility service territory and determine the sensitivity of particular groups of distribution feeders to the impacts of high PV deployment levels. This paper presents a method for separating a utility's distribution feeders into unique clusters using the k-means clustering algorithm. An approach for determining the feeder variables of interest for use in a clustering algorithm is also described. The Cubic Clustering Criterion is used as a quality metric for determining the optimum number of clusters in a large dataset of over 3000 feeders from western utilities. An approach is illustrated for choosing the feeder variables to be utilized in the clustering process and a method is identified for determining the optimal number of representative clusters..**

***Index Terms* — clustering, distribution feeder, cubic clustering criterion, principal components**

I. Introduction

This paper demonstrates a clustering methodology for classifying distribution feeders**.** The methodology groups utility feeders into specific groups of representative feeders. The representative feeders can be used develop a more accurate screening criteria and to help identify those feeders that are more likely to have issues with PV integration. One of the goals for improving screening criteria is to speed up the interconnection process for low risk feeders. This effort is being done in partnership with the Electric Power Research Institute, Inc. for the California Solar Initiative Project to screen distribution feeders to develop alternatives to the 15% of peak load penetration Rule.

The objective of this project is to develop new methods for quickly and accurately determining the capacity of individual feeders to accept new PV projects in order to streamline the interconnection process. The data shown in this paper was provided by a participating utility and consisted of over 3000 distribution feeders. The goal was to identify seven potential representative feeders from the dataset for which detailed models could be created to evaluate high PV penetration scenarios.

II. The 15% Penetration Threshold

The 15% threshold refers to the current practice for screening DG systems. When the amount of aggregated DG exceeds 15% of the peak load on a line section, supplemental studies are then required to determine if system impacts might arise due to the new interconnection request. This practice was first implemented in 1999 through the California Public Utilities Commission (CPUC) Rule 21, and later adapted in the FERC SGIP and remains the current standard in the United States for interconnection procedures [1]. The rationale for the 15% threshold is based on the principal that unintentional islanding, voltage deviations, protection miscoordination, and other potential negative impacts are negligible as long as the DG on the line remains less than the minimum load. The 15% of peak load was intended as a conservative proxy for the minimum load on the circuit.

It has been observed that the existing 15% screen may be conservative and not an accurate way to determine the PV hosting capability limit of a particular distribution feeder. In many cases during supplemental studies required for interconnecting PV, even when penetration is substantially higher than 15%, the review does not identify any necessary system upgrades. There are many examples of circuits in the United States with PV penetration levels above 15% where system performance, safety, and reliability have not been affected by crossing this threshold [2].

III. Clustering Analysis

The purpose of clustering analysis for this project is to place feeders into groups, driven by feeder properties, such that feeders in a given cluster tend to be similar to each other, and dissimilar from feeders in other clusters [3], [4]. Two common methods for clustering are the hierarchical and partitional approach. Hierarchical clustering begins by creating a cluster for each individual element, and combining clusters until the desired grouping is achieved. Partitional algorithms work in the opposite direction by starting with a single cluster and dividing into the desired number of clusters. Classifying feeders based on the hierarchical algorithm was demonstrated in the PNNL Taxonomy Final Report [5]. A well-known and widely used partitional clustering method is the k-means algorithm [6]. For this project, the authors chose the k-means method due to its advantages with working with larger datasets.

A. Principal Components Analysis

Principal Component Analysis (PCA) is a multivariate projection method designed to extract and display the systematic variation in a data set [7]. For our study of distribution feeders where we are examining up to 12

variables, it can reduce the complexity of the variation by projecting the dataset into a lower dimensional space.

The PCA transformation will create a number of principal components equal to the number of variables in the original data matrix. The components are uncorrelated, and ordered so that the first few retain most of the variation present in all of the original variables, i.e. the 1st principal component will account for the most variability, then the 2nd, and so on. Using the 1st and 2nd principal components a biplot can be used to visualize the dominant aspects of variation in the dataset. Figure 1 shows an example biplot of the feeders in the dataset that have been projected to a 2-dimensional space.

The x and y axis of the biplot do not represent any particular unit and are bi-products of the PCA transformation. The first step in the PCA transformation is the scaling of the data. Variables often have substantially different numerical ranges, for example voltage regulators might range from 0 to 8, and 3-Phase miles might range from 0 to 200. Since PCA uses a maximum variance projection method scaling is necessary to prevent variables like 3-Phase miles from dominating over voltage regulators. Unit Variance Scaling is done on each variable by dividing each value by the standard deviation. Each variable will then have equal (unit) variance. Each variable also goes through a mean centering transformation where the average for each variable is calculated and then subtracted from each value. This helps in improving the interpretability of the model.

Each point on the biplot in Figure 1 represents a single distribution feeder. We can see areas of high density where the feeders will tend to share similar characteristics. Points on the biplot that are isolated represent feeders that are more unique not sharing similar characteristics with other feeders, these represent outliers within the dataset.

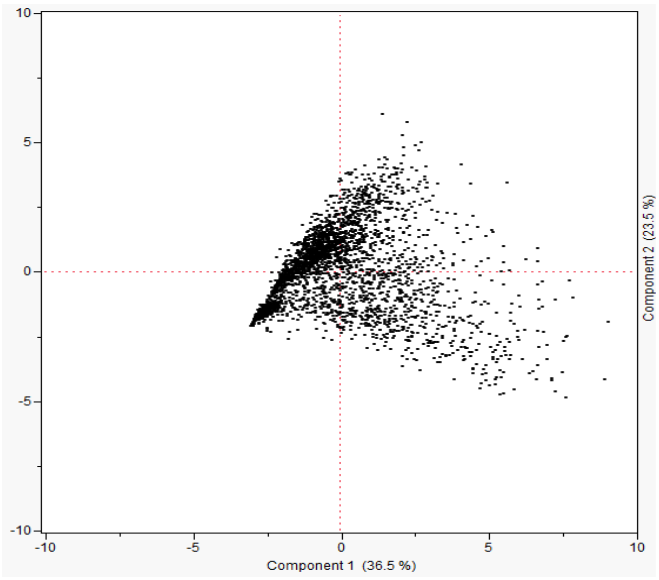

Fig. 1. Biplot of Feeders in Dataset.

Figure 2 shows a vector for each variable used in the PCA transformation. The vector plot shows that feeders in the upper right quadrant will tend to have larger kV, larger summer kVA capacity, a larger customer count, and a larger industrial customer count.

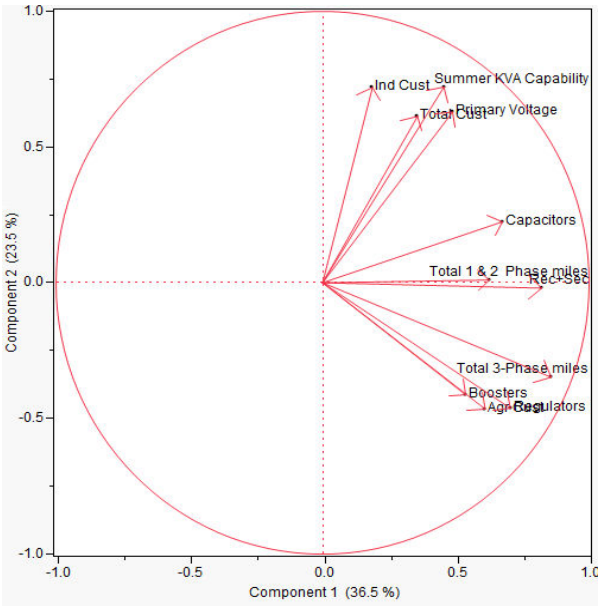

Fig. 2. Vector Plot of PCA Transformation.

The PCA transformation is the basis for the k-mean clustering algorithm. Feeders are separated/grouped based on the proximity of their principal components. Figure 3 shows a graphical representation of how the k-means algorithm would create 3 clusters.

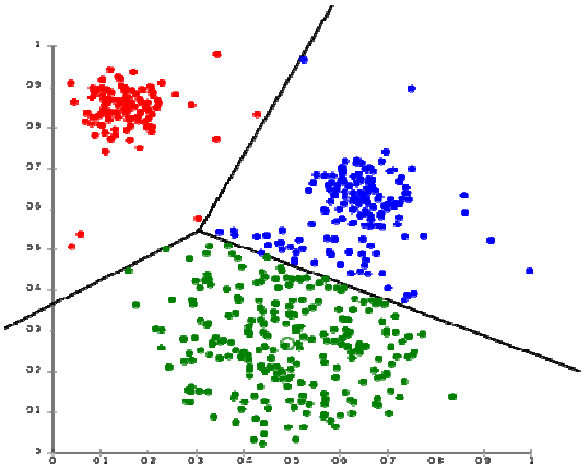

Fig. 3. K-Means Algorithm for 3 clusters.

In the next sections we will discuss how to choose the variables to use for clustering and how to find the optimal number of clusters.

B. Determining the Clustering Variables

We determined which variables to use from the dataset by looking for which were of interest and relevant to the projects goal of investigating interconnection impacts as well as choosing variables such that the optimum number of clusters is more accurately attained. Primary voltage along with data relating to the length of the feeder can have significant impact on the ability to integrate PV systems into the circuit so these variables were chosen to study. Variable related to regulating voltage on the line (regulators, capacitors, boosters) are important for classifying feeders and were added to the list of potential variables. The types of loads on the feeder are also of interest and data relating to the type of customers (domestic, commercial, industrial, agricultural) on the circuit were examined. Variables dealing with circuit protection (fuses, reclosers, etc.) are not tied directly to the topology of the feeder so this category of variables was not examined, however, protection issues will be captured in the analysis of the final feeders selected. Data was available on the winter and summer peak loading on the feeder, but these variables were not chosen directly, but rather it was determined that the ratio of summer peak to winter peak would be of interest because the ratio is effective at identifying seasonal circuits. Lastly the KVA capability of the feeder was added to the list of variables to be studied.

The optimum number of clusters is more accurately achieved when the variables chosen are independent of each other. In order to achieve optimal clustering the correlation of the chosen variables was determined and were compared as shown in the heat map in Figure 4 below.

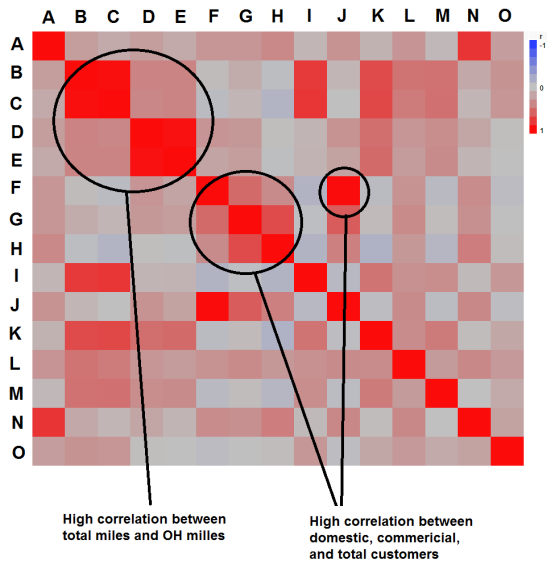

Fig. 4. Correlation of feeder variables.

TABLE I
VARIABLE LABELS FOR HEAT MAP IN FIGURE 4

Label	A	B	C	D	E	F	G	H
Variable Name	Primary Voltage	Total 3-Phase Miles	OH 3-Phase Miles	Total 1&2 Phase Miles	OH 1&2 Phase Miles	Domestic Customers	Commercial Customers	Industrial Customers

Label	I	J	K	L	M	N	O
Variable Name	Agricultural Customers	Total Customers	Regulators	Capacitors	Boosters	Summer KVA Capability	Summer / Winter Peak Ratio

Blocks of dark red on the heat map represent a high correlation between two variables. Variables such as 3 phase over-head (OH) miles, 1&2 phase OH miles, domestic customers, and commercial customers were eliminated from the clustering list due to their high correlation with other parameters within the variable list. The final list of chosen variables is shown in Table 2.

TABLE II
VARIABLES CHOSEN FOR CLUSTER ANALYSIS

Primary Voltage	**Regulators**	**Industrial Customers**	**Ratio of Summer Peak to Winter Peak**
Total 3-Phase Miles	**Capacitors**	**Agricultural Customers**	**Summer KVA Capability**
Total 1&2 Phase Miles	**Boosters**	**Total Customers**	

By reducing the total number of variables the optimum number of clusters will more easily be achieved which will be shown later. Also by reducing the number of variables related to customer data, this category will not dominate the clustering selection by watering down other important variables such as primary voltage and circuit length. This represents a good list of variables for clustering because it is not over burdensome, but yet still captures key topology issues of a distribution feeder.

The list of variables from Table 2 will not necessarily be the same for each utility, and is dependent on the data available as well as the operation of the utility's distribution system. For example line voltage regulators, whose presence can have a significant impact on a feeder's ability to manage distributed PV, may not always be a good cluster variable as in some California utilities they are not a common occurrence. Other variables, such as the conductor type (4/0, 336.4kcmil, etc.) may make an excellent candidate as a clustering variable, but may not always be available within the dataset.

C. Determining the Optimum Number of Clusters

The most difficult problem in cluster analysis is how to determine the number of clusters. A quality metric for determining the optimum number of clusters is based on the

Cubic Clustering Criterion (CCC) [8]. 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. 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 does 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.

In Section B we discussed the process for selecting the variables to be used in the clustering algorithm. The down-select process for the variables to be used in the clustering algorithm helps finding the optimum number of clusters. An example CCC plot is shown below in Figure 5 in which all the original variables were used in the clustering algorithm. We see that there is a continual rise in the CCC value with no definitive peaks occurring until 22 clusters, and the CCC value does not drop back below 2.

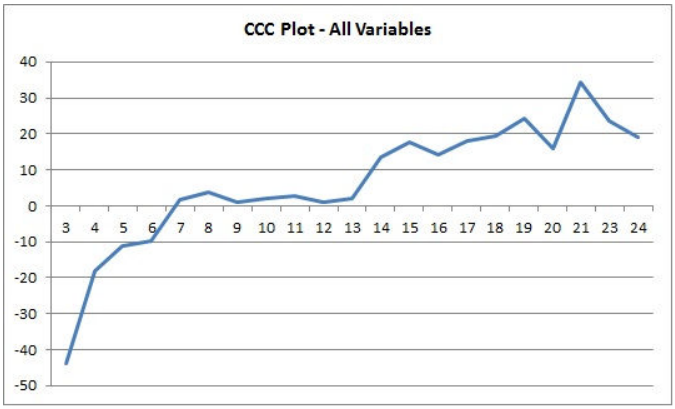

Fig. 5. CCC Plot with all variables.

By following the method presented in section B for selecting the clustering variables we can more accurately find the optimum number of clusters. Figure 6 shows a CCC plot when clustering around the variables from Table 2. We see a definitive peak occurring at 12 clusters, followed by a drop in the CCC value below 2. A comparison of the CCC plot for clustering with the final variables from Table 2 to the CCC plot of the original variables shows how a reduced variable list improves the clustering process.

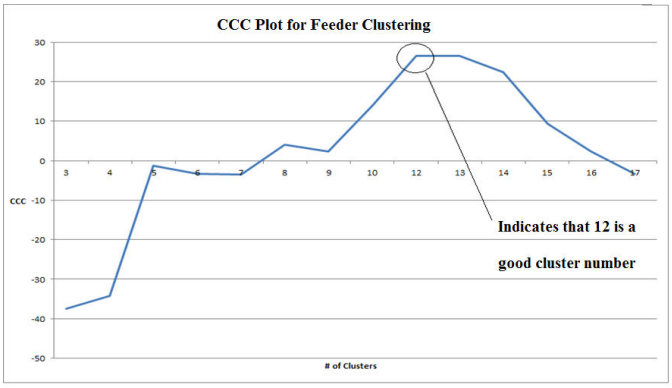

Fig. 6. CCC Plot with reduced variable list.

Based on the CCC plot in Figure 6 a local maximum that met the clustering criteria was achieved at 12 clusters and this value was chosen as an optimum number of clusters.

IV. Selection of Feeders

Once each feeder within the dataset was classified into one of the 12 clusters the next step was to identify the key clusters for further study. Table 3 below shows the mean values of certain variables for each cluster; that is for each column listed the values shown are the average of all the feeders within that particular cluster. This is why some of the values in the primary voltage column do not fall exactly within a standard distribution voltage kV level, as they are often the average of various voltage levels.

TABLE III
Cluster Means

Cluster	Feeder Count	Primary Voltage	Total 3-Phase miles	Regulators	Capacitors	Total Cust	90% Radius
1	59	12.00	171.79	10.42	5.98	2267.68	60.73
2	390	20.75	32.71	0.60	4.78	2920.27	39.83
3	779	12.00	24.76	0.31	5.01	2874.60	11.88
4	619	12.02	18.02	0.36	2.20	833.35	11.39
5	13	19.00	173.10	11.92	7.62	3588.92	106.74
6	28	13.32	62.59	1.86	2.82	786.39	139.55
7	41	20.80	239.80	6.32	6.51	2524.59	89.63
8	29	12.97	314.80	7.59	7.83	1781.03	110.43
9	390	4.00	6.26	0.12	1.56	923.26	3.20
10	164	12.27	85.93	3.86	4.68	2144.50	32.90
11	111	12.18	222.20	6.32	6.37	1037.86	41.97
12	294	12.02	117.92	3.09	4.88	860.91	20.62

The far right column '90% radius' value represents how tightly grouped the feeders are within the cluster. The mean values represent the center of the cluster and the 90% radius

column is the length of the radius from that center that captures 90% of the elements within the given cluster. Figure 7 below gives an example of a biplot for the elements within a single cluster.

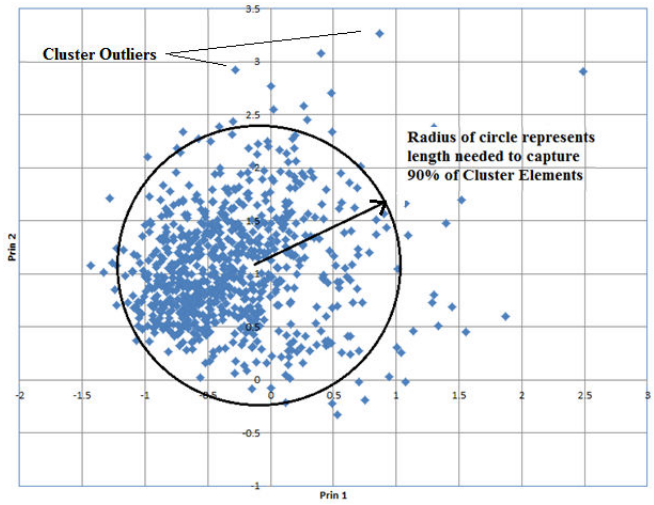

Fig. 7. Cluster Biplot.

One of the objectives of the project is to identify 20 feeders representing the range of distribution feeder types within the California grid. Since we are limited to a total of 20 feeders for further study it is necessary to be selective on what feeders to choose. Due to this limitation on the number of feeders to study we examined each cluster to see if there was any grouping that could be eliminated as a cluster of interest. Clusters 5, 6, and 8 were eliminated because they are low population clusters that are not tightly grouped. Clusters 3 and 4 were similar so only one was chosen for further study, and Cluster 3 was picked because of its higher customer count. Lastly cluster 10 was eliminated because the other 12kV class clusters (1, 3, 11, 12) did a good job of capturing the range of values for the feeder variables. Selection of the actual feeders from within the cluster was done by sorting the feeders by their distance from the mean, and choosing those feeders that are closest to the center of the cluster. This method of feeders selection will give those that are best representative of the given cluster. Different feeders may need to be selected based on the existence of SCADA equipment and PV monitoring equipment present on the feeder. Once the potential feeders are identified the next step is to begin creating detailed models of the chosen feeders for analysis on PV screening methods.

III. Conclusion

This paper outlined the method for using the k-means clustering methodology for classifying distribution feeders into a sub-group and the use of the Cubic Clustering Criterion for determining the optimum number of clusters. The significance of this work is that it demonstrates a method to classify distribution feeders into a specific subgroup. This can have significant impact on the screening process for interconnection request of PV systems on the distribution grid. Through modeling and analysis a utility could determine which sub-group of feeders is more or less sensitive to the effects an interconnecting PV system might have on that particular feeder. This could lead to a more streamlined approach to interconnection procedures to avoid unnecessary interconnection studies, cost, and delays.

Acknowledgement

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] *Updating Interconnection Screens for PV System Integration,* Technical Paper. NREL/TP-5500-54063, January 2012 [Online]. Available: http://energy.sandia.gov/wp/wp-content/gallery/uploads/Updating_Interconnection_PV_Systems_Integration.pdf

[2] M. Braun et al, "Is the distribution grid ready to accept large-scale photovoltaic deployment? State of the art, progress, and future prospects." *26th EU PVSEC*, Hamburg, Germany 2011, [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/pip.1204/pdf

[3] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[4] J.A. Hartigan, *Clustering Algorithms*. New York: Wiley, 1990.

[5] *Modern Grid Initiative Distribution Taxonomy Final Report*, November 2008 [Online]. Available: www.gridlabd.org/models/feeders/taxonomy_of_prototypical_feeders.pdf

[6] J. MacQueen, "Some methods for classification and analysis of multi-variate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probab.,* L.M.L. Cam and J. Neyman, Eds. Berkeley, CA: Univ. California Press, 1967, vol. I.

[7] I.T. Jolliffe, *Principal Component Analysis, 2nd Edition*, New York, New York: Springer-Verlag, 2002.

[8] *SAS Technical Report A-108 Cubic Clustering Criterion,* 1983. [Online]. Available: http://support.sas.com/documentation/onlinedoc/v82/techreport_a108.pdf