SANDIA REPORT
SAND2014-1443
Unlimited Release
Printed February 2014

Clustering Method and Representative Feeder Selection for the California Solar Initiative

Robert J. Broderick, Joseph R. Williams and Karina Munoz-Ramos

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

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.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online order:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2014-1443
Unlimited Release
Printed February 2014

Clustering Method and Representative Feeder Selection for the California Solar Initiative

Robert J. Broderick P.E.
Joseph R. Williams P.E.
Karina Munoz-Ramos

Photovoltaics and Distributed Systems Integration
Military and Energy Systems Analysis
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-1033

Abstract

The screening process for DG interconnection procedures needs to be improved in order to increase the PV deployment level 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 report describes the utility distribution feeder characteristics in California for a large dataset of 8, 163 feeders and summarizes the California feeder population including the range of characteristics identified and most important to hosting capacity. The report describes the set of feeders that are identified for modeling and analysis as well as feeders identified for the control group. The report 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 report presents an approach 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.

ACKNOWLEDGMENTS

CONTENTS

FIGURES

TABLES

NOMENCLATURE

DOE	Department of Energy
SNL	Sandia National Laboratories

1. INTRODUCTION

California has an ambitious goal of increasing its Distributed Generation (DG) by 12,000 MW by the year 2020. A combination of small and large efforts will be required to meet this goal with a significant portion of the DG being Photovoltaic (PV) systems. It is estimated that 8,000 MW of DG will come from “utility-scale” efforts and although these will have a larger impact than the smaller projects, they will be significantly more difficult to implement, more costly and will require more time to implement due to the planning, permitting and construction phases required for larger projects [1]. Although interconnection studies are necessary for planned “utility-scale’ PV systems to mitigate the impact risks of large PV systems, commercial scale PV installations, that are generally less than 2 MW, may not require such rigorous and time consuming interconnection studies.

The typical screening process in use today across the United States utilizes a PV deployment level screen to determine if the amount of aggregated DG (proposed plus installed) exceeds 15% of the peak load on a line section. If this occurs, interconnection studies are 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 [2]. The rationale for the 15% threshold is based on the idea 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. The most recent changes to Rule 21 have implemented an additional screening criteria based on 100% of minimum load that can allow for increased PV deployment.

It has been observed that the existing 15% screen may often be overly conservative and not accurate at determining the full PV hosting capability limit of a particular distribution feeder. In many cases when a PV system is seeking to interconnect, it will fail the 15% screen and therefore require either supplemental and/or full interconnection studies. In many cases these studies do not identify the need for system upgrades and demonstrate that the 15% screen may be overly conservative. There are many examples of circuits in the United States with PV deployment levels above 15% where utility system performance, safety, and reliability have not been affected by crossing this threshold [3].

As part of the ‘*Screening Distribution Feeders: Alternatives to the 15 percent Rule’* project, Sandia National Laboratories (Sandia), National Renewable Energy Laboratory (NREL) and Electric Power Research Institute, Inc. (EPRI) are collaborating to develop new screening methods that utilities can use to quickly and accurately determine the capacity of individual distribution feeders to accept new PV projects without the risk of impacting the grid. A key outcome will be a data-driven, validated approach to determining feeder limits that can simplify interconnection processes and lead to greater PV adoption across the California distribution system. The key Tasks for *Screening Distribution Feeders: Alternatives to the 15 percent Rule’* project are:

Task 1: Document Current Utility Screening Practices and Available Tools
Task 2: Define Distribution Feeder Configurations in California

Task 3: Collect High-Resolution PV Output Data for Use in Feeder Impact Simulation.
Task 4: Complete Detailed Modeling for Selected Feeders in California
Task 5: Run Full Range of High Penetration PV Scenarios on Selected Feeders
Task 6: Develop Practical Screening Method for Handling PV Interconnection Requests
Task 7: Validate the Screening Method Using Site Measurements and Feeder Data

This report describes work that is part of Task 2 of the 'Define Distribution Feeder Configurations in California'. A separate report documenting findings from Task 1 has also been completed and can be found at: http://calsolarresearch.org/images/stories/documents/Sol3_funded_proj_docs/EPRI/CSIRDD_EPRI_UtilityDGInterconnectionBestPracticesFinalRpt_20130606.pdf.

1.1. Task 2 Objectives

The goal of the '*Screening Distribution Feeders: Alternatives to the 15 percent Rule'* project is to develop a screening method that is applicable to the majority of feeder types that are found among the California utilities. The goal of Task 2 is to determine the range of feeder configurations and to develop a database of feeder characteristics for CA utilities. In order to determine a set of representative feeders to be studied it was necessary to determine the statistical range and distribution of feeder configurations and types and their electrical characteristics. The primary challenge of this Task 2 is to understand the overall statistical feeder population and to cluster the universe of feeders into representative groups and then select specific feeders for study and analysis. The primary outcome of Task 2 is to select feeders to be used in the modeling and analysis described in Tasks 4-5 and developing and validating the proposed screening methodology described in Task 6-7.

1.1.1 Task 2.1

A set of initial feeder-specific characteristics affecting hosting capacity were identified as part of Task 2.1. Data sets from three different utilities in California were received and analyzed. The number of feeders in the database and the feeder characteristic data was dependent on the data availability for each utility. We created Microsoft Excel databases containing distribution feeders and the associated characteristics for the feeders within each utility's service territory, along with user guides for the databases; these files can be found on the California Solar Research website at: http://www.calsolarresearch.org/component/option,com_sobipro/Itemid,0/pid,54/sid,88/.

1.1.2 Task 2.2

Twenty-two feeders from the data received were selected as representative feeders as part of Task 2.2. In order to select feeders representative of a cross-section of the known range of feeder types in California as well as representative of the characteristics known to be important to hosting capacity, the clustering approach described in Chapter 3 of this report was implemented. Out of the full set of 22 feeders chosen for evaluation, a control group of 6 feeders were selected for testing/validating the screening methodology.

1.2 Report Overview

The remainder of this report is divided into five main chapters. *Chapter Two* describes the feeder data received from all three participating utilities and discusses similarities and key differences between the data. *Chapter Three* gives a brief overview of two common clustering approaches; *Hierarchical* and *K Means* clustering. *Chapter Three* also details the general clustering approach taken in this project to obtain representative feeders for each of the three participating utilities. While the focus of *Chapter Three* is to describe the general clustering approach taken, *Chapter Four* covers specific examples of how the approach was applied. Although the steps are not described in detail for each utility, the intent is to provide the reader with a better understanding of how the clustering approach was applied to all three utilities. Summaries of the initial data review and cleanup and lists of the final clustering variables can be found in *Chapter Four*. Tables summarizing cluster means and the selected representative feeders for each utility can also be found in *Chapter Four*. *Chapter Five* describes the process for selecting the final list of twenty-two feeders representative of all three utilities. A list of the final twenty-two feeders is also given in *Chapter Five*. Finally, *Chapter six* discusses the important conclusions from this report. *Appendices* provide additional information on the feeder data request, clustering variables for Utility 1 & 2 and the mean values of the clusters for Utility 1 & 3.

2. DESCRIPTION OF FEEDER DATA

One goal for Task 2.1 was to identify feeder characteristics that represent the known range of feeder types in California and are known to affect hosting capacity. The extensive experience that Sandia, NREL and EPRI have with distribution impact studies and the familiarity with utility databases was utilized to compose a set of feeder characteristics that would describe the variation in the feeder population. The set of feeder characteristics were also selected based on their likelihood to affect hosting capacity of the feeder for voltage, thermal and protection impacts. The groups of characteristics identified include:

1. Nominal voltage level (e.g., 4kV, 13kV, 25kV, etc.)
2. Feeder length and main conductor type
3. Three-phase vs. single-phase feeder length
4. Voltage regulation schemes (load tap changes, feeder regulators, switched capacitor banks)
5. Load mix (residential, commercial, industrial)
6. Load shape (peak, minimum load, seasonality)
7. Existing DG and PV deployment levels (kW)
8. Utility operational practices (e.g. use of conservation voltage reduction schemes)
9. System protection devices

These characteristics were used to create a data request list that was sent to the utilities involved in the project to obtain data that was measurable and readily available for all the utilities feeders.

2.1 Feeder Data

Data for more than 8,000 feeders were received from three different utilities in California. Data for 3195 feeders was received from Utility 1, data for 4192 feeders was received from Utility 2 and data for 776 feeders was received from Utility 3. It is important to note that data received for the number of feeders and for the characteristics for each feeder was dependent on availability and ease of retrieval, therefore, data received differed for all utilities. Data requested by EPRI can be found in Appendix A. Table 1 lists the data received from all three utilities. Variable names reflect those provided by each utility regardless of data requested.

Table 1. Data received for all three utilities.

Utility 1	**Utility 2**	**Utility 3**
EPRI Feeder ID	EPRI Feeder ID	EPRI Feeder ID
EPRI Substation ID	EPRI Substation ID	EPRI Substation ID
Nominal voltage, kV	Nominal voltage, kV	Nominal voltage, kV
Total 3- Phase circuit miles	Total 3- Phase circuit miles	Total 3- Phase circuit miles
Total 3- Phase Overhead circuit miles	Total 3- Phase Overhead circuit miles	Total 3- Phase Overhead circuit miles
Total 2- Phase and 1-Phase circuit miles	Total 2- Phase and 1-Phase circuit miles	Total 2- Phase and 1-Phase circuit miles
Total 2- Phase and 1- Phase Overhead miles	Total 2- Phase and 1- Phase Overhead miles	Total 2- Phase and 1- Phase Overhead miles
Number of line voltage regulators, #	Number of line voltage regulators, #	Number of line voltage regulators, #
Number of switched/fixed capacitor banks, #	Number of switched/fixed capacitor banks, #	Number of switched/fixed capacitor banks, #
-	Number of feeder tie points, #	Number of feeder tie points, #
Transformer Count	Connected service transformer capacity, kVA	Connected service transformer capacity, kVA
Summer KW	Recorded peak net load	Feeder peak load, kW
-	Feeder Peak Load Date	Feeder Peak Load Date
-	Feeder Peak Load Time	Feeder Peak Load Time
Winter KW	-	Feeder minimum load (can be estimated), kW
-	Residential, %(Energy-July)	Residential, %
-	Commercial, % (Energy-July)	Commercial, %
-	Industrial, % (Energy-July)	Industrial, %
-	Agricultural, % (Energy-July)	-
Total Customers	Total Customer Count	-
Com Customers	Commercial Customers Count	-
Dom Customers	Domestic Customers Count	-
-	Idle Customers Count	-
Industrial Customers	Industrial Customers Count	-
Other Customers	Other Customers Count	-
Agricultural Customers	-	-
Other Data	**Other Data**	**Other Data**
Supervisory Control and Data Acquisition (SCADA) Breaker	-	Main 3-Phase conductor
Boosters	-	Substation Load Tap Changer, Yes or No
Fuses	-	Load Tap Changer set points, target/total bandwidth
Reclosers	-	Distance between source and voltage regulators (not inline with other regulators stations) Substation to Regulator Station (mi)

Sectionalizers	-	Distance between voltage regulators (inline with other regulator stations) Regulator Station to Regulator Station (mi)
Stepdowns	-	Voltage Regulator set points, target/total bandwidth
Switches	-	Conservation voltage reduction feeder?, Yes or No
Interrupters	-	Archived load data at feeder level, data rate or N/A if not measured
Summer KVA Capability	-	Archived load at station level, data rate or N/A if not measured
Winter KVA Capability	-	-
DG	**DG**	**DG**
Number of DG systems	Total amount of DG on circuit, kW	Total amount of DG on circuit, kW
Number of PV systems	Existing PV capacity installed, kW	Does feeder contain utility-owned PV, Yes or No
kW DG	Largest PV system installed, kW	Solar irradiance monitoring or data?, Yes or No
kW PV (including wind)	PV Capacity with Utility Owned Generation	-
Sum of Non-PV kW	Utility Owned Generation PV, kW	-
# of PV 0-20 kW	# Utility Owned Generation PV Sites	-
# of PV 20-200 kW	-	-
# of PV > 200 kW	-	-

The characteristics of the feeder data received varied significantly within a utility and even more among the three utilities. For example, feeders within Utility 1 were significantly longer than those in Utility 2 and Utility 3. Another example of wide variation among utilities is the use of voltage regulators. While Utility 1 had voltage regulators on almost 30% of its feeders, Utility 2 had voltage regulators on less than 1% of its feeders. This can be attributed to the use of different operational strategies for voltage regulation.

2.1.1 Key Differences

Figure 1 shows the distribution of feeder kV for all three utilities. Utility 3 provided data only for 12 kV feeders which is the voltage class for the majority of their feeders. The majority of feeders for Utility 1 and Utility 2 also had a nominal voltage of 12 kV. The data set for Utility 1 contained a few feeders with a nominal voltage greater than 33 kV.

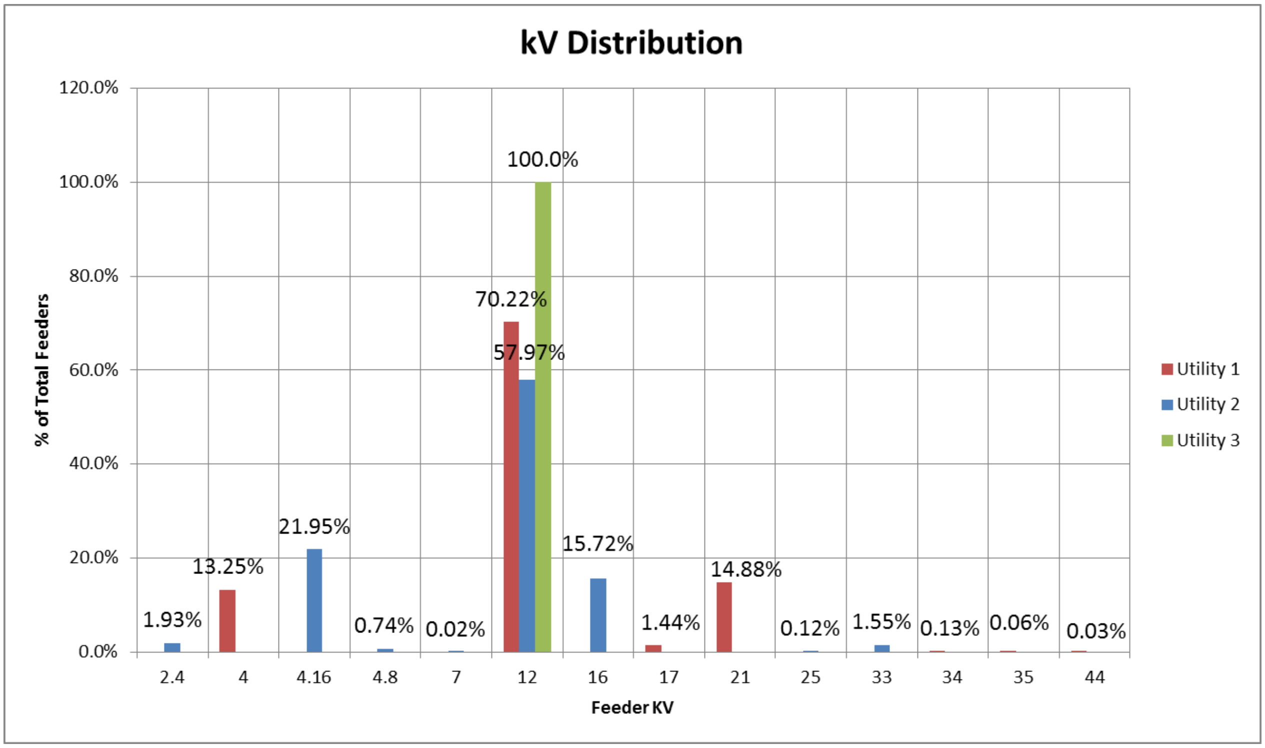

Figure 1. Voltage (kV) distribution for all three utilities.

Figure 2 shows the 3-phase length distribution for all three utilities. The 3-phase length distribution is defined as the sum of all 3-phase sections within the feeder. The majority of 3-phase feeders for Utility 2 and Utility 3 were less than 20 miles in length and all of them were less than 80 miles in length. Utility 1 had longer feeders with several over 80 miles in length.

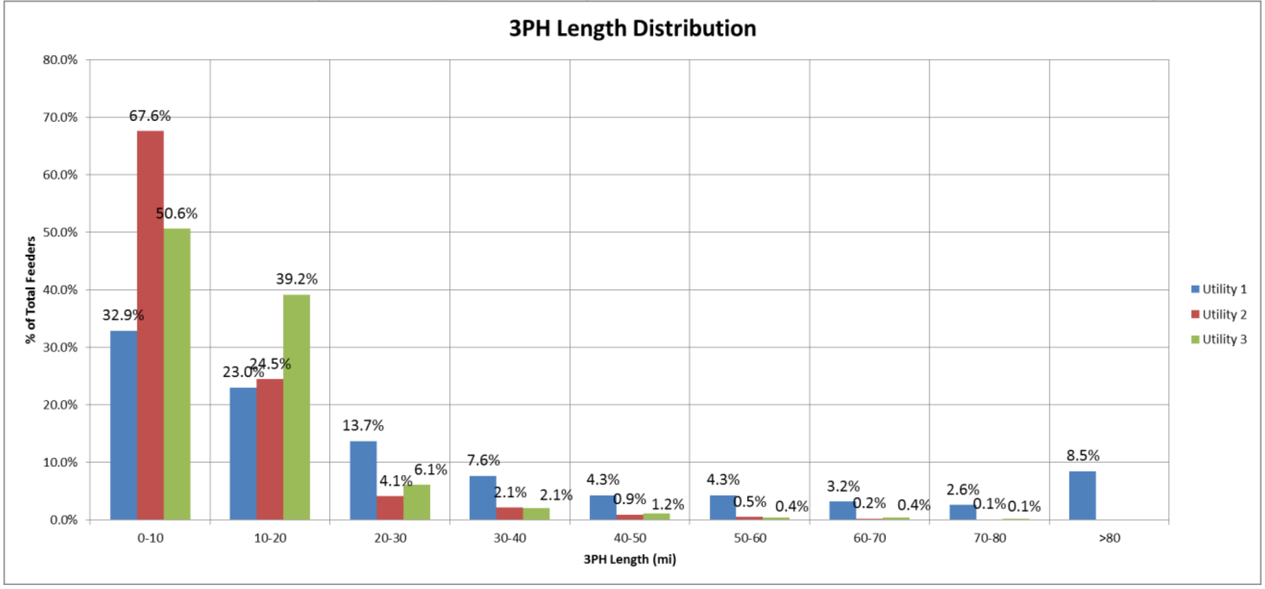

Figure 2. Three-phase length distribution for all three utilities.

Figure 3 shows the voltage regulators distribution for all three utilities. As stated previously, less than 10% of the feeders within Utility 2 and Utility 3 had voltage regulators while almost 30% of the feeders within Utility 1 had at least one voltage regulator and as many as 27 voltage regulators. For comparison, feeders within Utility 2 had at most 3 voltage regulators and feeders within Utility 3 had at most 6 voltage regulators.

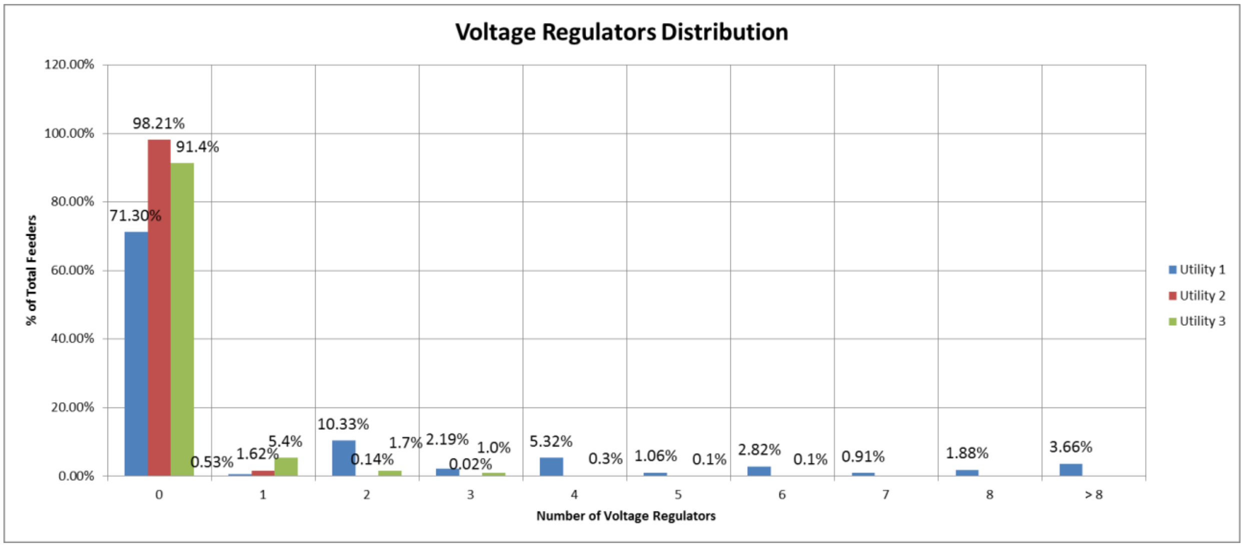

Figure 3. Voltage regulators distribution for all three utilities.

Figure 4 shows the distribution for number of capacitors for all three utilities. Most feeders had at least one capacitor. Feeders within Utility 1 had the greatest number of capacitors while Utility 3 feeders had the least. Almost 20% of feeders within Utility 3 had no capacitors.

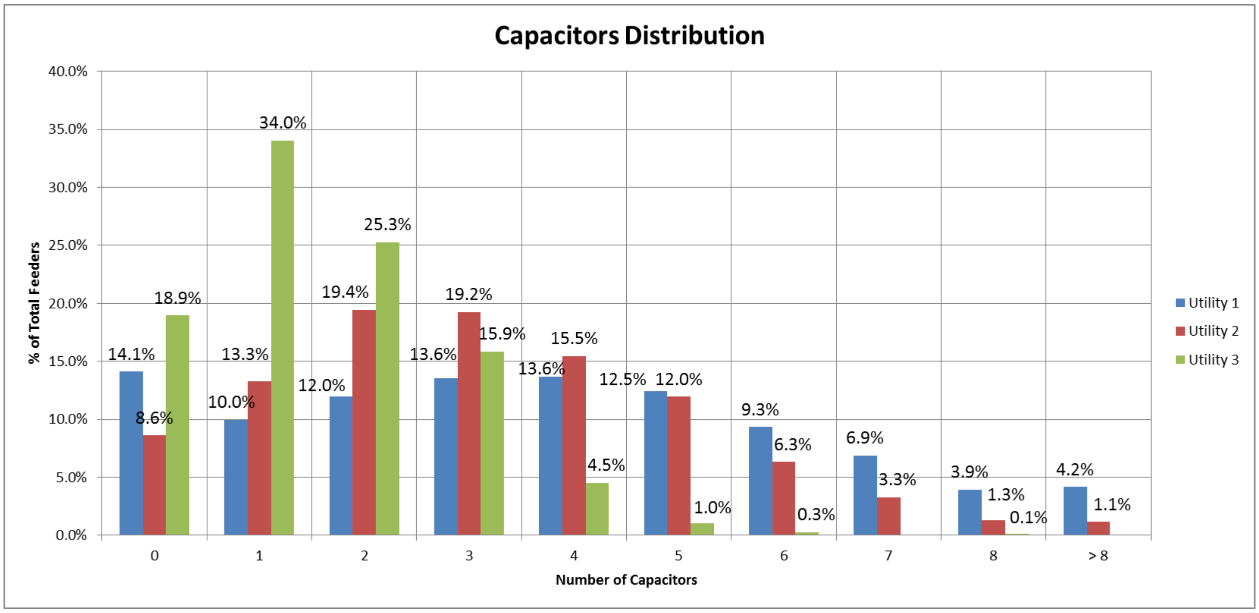

Figure 4. Number of capacitors for all three utilities.

Figure 5 shows the peak load distribution for all three utilities. Almost 60% of the feeders within Utility 1 and about 50% of the feeders within Utility 2 have a peak load below 7.5 MW. Almost 70% of the feeders within Utility 3 have a peak load between 5-10 MW and more than 90% have a peak load below 10 MW.

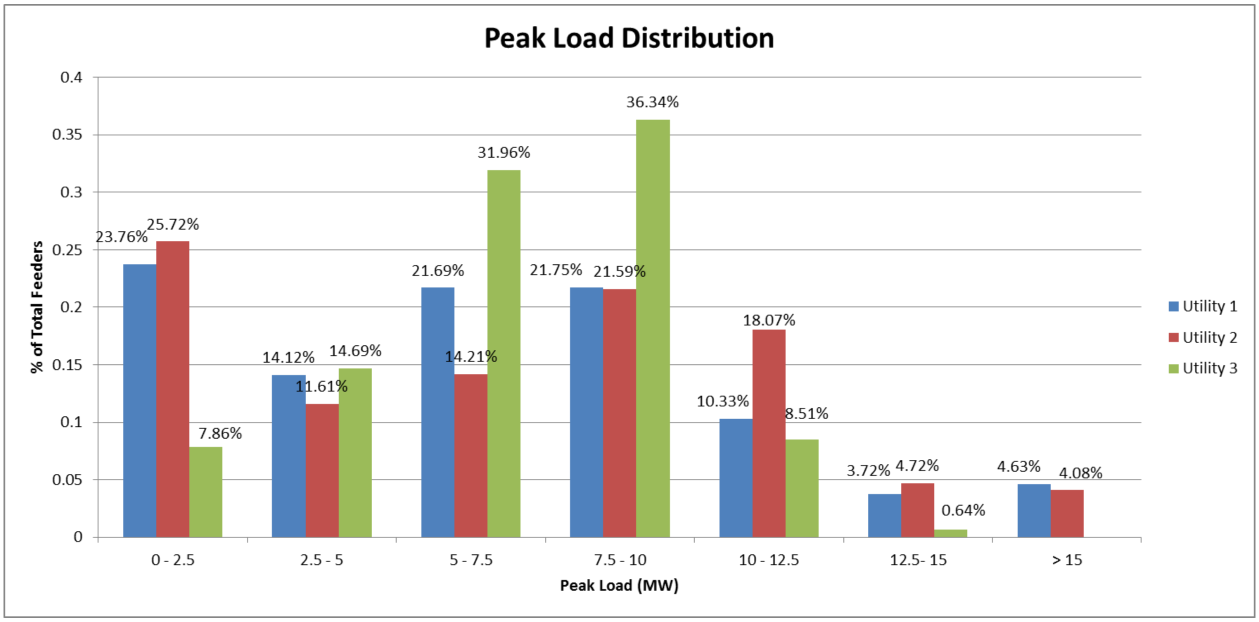

Figure 5. Peak load distribution for all three utilities.

Figure 6 shows the distribution of the connected service transformer rating for Utility 2 and Utility 3. Information on connected service transformer rating was not received for Utility 1. While there seem to be large numbers of smaller transformers, the overall shape of the distribution appears to be similar to that of a normal distribution. Almost 20% of the transformers within Utility 2 are less than 2,500 kVA. Utility 3 has a higher percentage of larger (> 15,000 kVA) transformer.

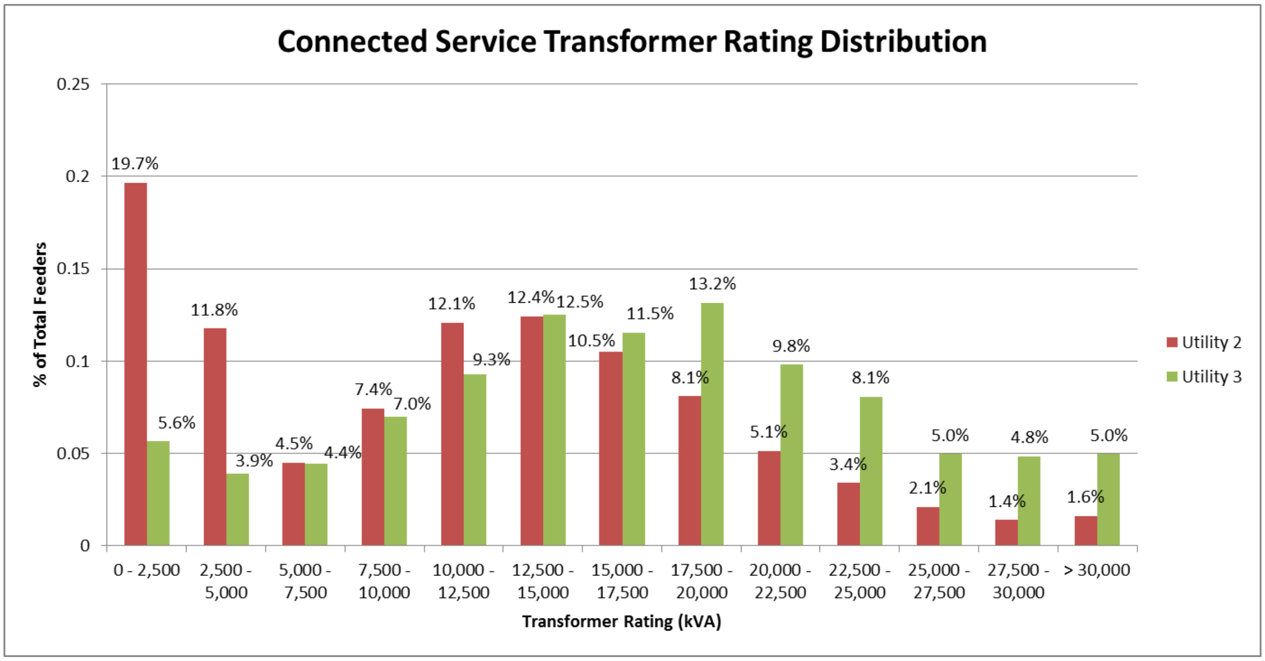

Figure 6. Connected service transformer rating distribution for Utility 2 and Utility 3.

It is clear that there are significant differences between feeders not only within a utility but also among the utilities as well. Therefore, it is important when developing the new screening methods to understand the broad range of differences present within utilities and how these differences can impact the methodologies proposed. Furthermore, when selecting representative feeders for developing and validating the screening methods, it is important to select a set that is representative of all utilities involved.

The significant differences between feeders not only within a utility but also among the utilities will impact the hosting capacity of these feeders. The wide variation in key feeder characteristics such as voltage class and length of the feeder will often result in a wide variation of voltage and thermal impacts that are key elements of the hosting capacity calculation for each feeder. The number of voltage regulators and capacitors also provides an indication of how complex the voltage and power factor control scheme is on a feeder and this variation in complexity can result in wide variations in hosting capacity. Finally the peak load distribution and connected service transformer rating provide information on how heavily loaded the feeder is and how likely reverse power flow will be on the feeder. The occurrence of reverse power on a feeder can often limit the hosting capacity of the feeder.

3. OVERVIEW OF CLUSTERING APPROACH

The purpose of the clustering analysis is to place feeders into groups, distinguished by feeder properties, such that feeders in a given cluster are similar to each other, and dissimilar from feeders in other clusters [4], [5]. There are several approaches used for clustering. Two commonly used clustering algorithms are *Hierarchical* and *K Means* [6].

3.1 Hierarchical Clustering

Hierarchical clustering can be either agglomerative or divisive. Agglomerative hierarchical clustering begins by creating a cluster for each individual element. The clusters with the shortest distance between them, i.e. most similar, are then merged to form a single cluster. This process is repeated until the desired numbers of clusters have been formed. Divisive hierarchical clustering begins with all elements in a single cluster. The cluster is then are divided into sub-clusters (criteria for doing so varies). This process is repeated until all elements end up in their own cluster. The two main benefits of hierarchical clustering are; 1) flexibility in selecting the number of clusters and 2) visualization of the clusters and the distance between clusters through a resulting dendrogram. One drawback of hierarchical clustering is that it requires a similarity matrix between all elements. For large data sets, generally greater than 200 elements, using a hierarchical approach can be time consuming. Classifying feeders based on the hierarchical algorithm was demonstrated in the PNNL Taxonomy Final Report [7].

3.2 K Means Clustering

A well-known and widely used partitional clustering method is the *K Means* algorithm [8]. The goal of the K Means algorithm is to minimize the within-cluster sum of squares of distances between elements through an iterative approach. Unlike Hierarchical methods, when using K-means for clustering, the number of clusters, *k*, needs to be defined a priori and selecting different initial clusters can results in different final clustering results.

K-means is a partitional algorithm that starts with all elements in a single cluster and divides the initial cluster into the desired number of clusters. The goal is to minimize within cluster distances. The K-means process begins by randomly selecting *k* elements, termed "means", from the data set. K clusters are then created by associating every observation with the nearest mean. Next, the mean element is replaced by the centroid of each cluster, and element assignments are repeated. Figure 7 illustrates the process. In **Step 1**, three mean elements are selected. **Step 2** shows the region corresponding to each mean, consisting of all points closer to that mean than to any other mean. **Step 3** shows how the centroid of each of the *k* clusters becomes the new mean and after step 2 and 3 are repeated **Step 4** shows the final converged clusters.

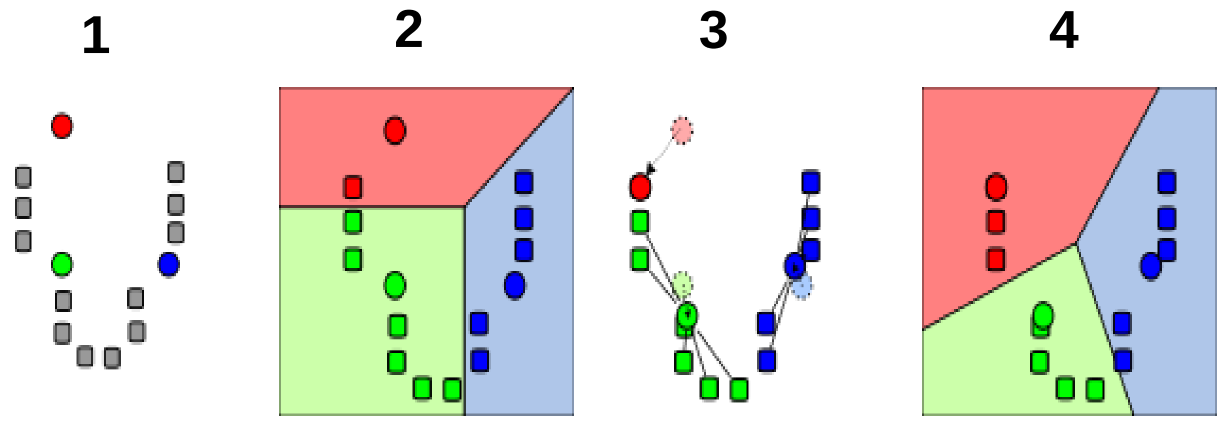

Figure 7. Example of K-means partitional algorithm. Source: Wikimedia Commons

One of the main advantages of the K-means algorithm is its quick convergence for large data sets (greater than 200 elements) making it more popular than hierarchical clustering approaches. For this project, K-means, specifically the Expectation-maximization algorithm (used by SAS JMP[1]) and known for its ability to accommodate clusters of variable size much better than the original K-means algorithm) was used.

3.3 Clustering Approach

The following section describes the general steps taken during the clustering approach. The process outlined was followed for all three utilities. For several of these steps, a statistical analysis program called SAS JMP was used.

3.3.1 Initial Data Review and Cleanup

Although the same data request was sent to all utilities, data received differed due to availability and ease of retrieval. Once received, the data went through an initial review process. The review process consisted of the following steps:

- *Histogram generation.* Histograms were generated to understand the distribution of all variables of interest (e.g. Nominal voltage, total circuit miles, number of capacitors, etc.). Histogram plots can be found in the *EPRI RD&D3 Feeder Database and User's Guide*[2] for each utility.

[1] More information on SAS JMP can be found at http://www.jmp.com
[2] http://www.calsolarresearch.org/component/option,com_sobipro/Itemid,0/pid,54/sid,88/.

- *Data clarification.* Often times it was necessary to get clarification from the utility as to how certain data was defined, calculated, etc. A list of questions was formulated and sent to the utility during this initial data review process.
- *Outlier identification.* Histograms and filtering were two methods that were used to identify feeders that were obvious outliers. Depending on the circumstance, these feeders were sometimes excluded from the clustering approach.
- *Boundary definition.* Some initial boundaries were defined during this initial review process. For example, feeders with a length of less than 0.1 miles were excluded from the clustering approach due to their scarcity and irrelevance.
- *Data anomaly documentation.* Some data anomalies were found during the initial data review process. These anomalies are captured in the *EPRI RD&D3 Feeder Database and User's Guide* for each utility.
- *Data set preparation.* Before using JMP program for clustering, the data had to be prepared. Formatting consisted of filling in all blank columns with null values, converting *Yes* or *No* columns to binary, etc.

3.3.2 Selecting Variables for Clustering

Initial variables were selected based on the impact they might have on differentiating feeder types and on DG hosting capacity. The initial variables varied among utilities as needed to account for differences in availability of data from each utility. These initial variables were analyzed using a correlation map, similar to the one shown in Figure 8, to show the degree of correlation among all variables. Blocks of dark red on the heat map represent a high correlation between two variables. The dark red diagonal is expected since these blocks show correlation of each variable with itself.

Because the optimum number of clusters is more accurately achieved when the chosen variables are independent of each other, pairs of highly correlated variables were examined more closely to determine if it was beneficial to remove one of the variables before clustering. The degree of correlation was used to develop a list of candidate pairs for evaluation. For example, *Total 3-Phase Circuit Miles* and *Total 3-Phase Overhead Circuit Miles* had a strong positive correlation, shown by the circle in the upper left of the figure. Therefore one of these two variables was picked and only *Total 3-Phase Circuit Miles* was used for clustering in all cases. Other correlations which resulted in omission of other variables are shown by the other circles in the figure. Once an optimal clustering was obtained, no further variables were removed.

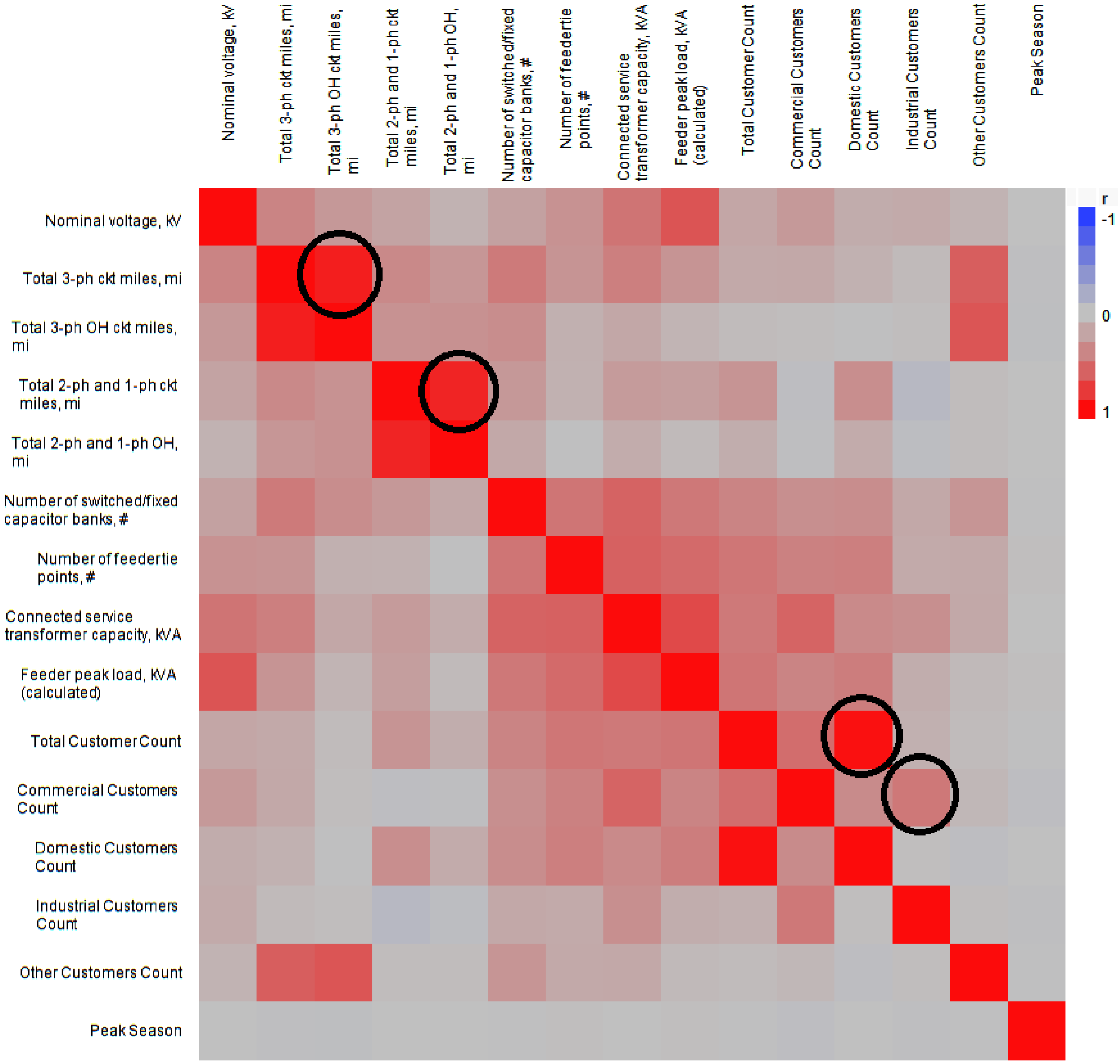

Figure 8. Example of correlation map.

3.3.3 Removing Outliers

Feeders labeled as outliers are those that are not representative of the overall data set. K-means clustering algorithms can be very sensitive to outliers, especially if the initial cluster means are chosen based on the outliers, which is often the case since many algorithms start by choosing initial cluster means as far apart from each other as possible. [10]. Therefore, when using K-means as the clustering technique, removing outliers can help improve convergence speed and will make the clustering more reliable. Outliers in the dataset were identified as follows: *Distance*, a multivariate calculation that is a measure of how similar a particular feeder is to its closest neighbor, was used as the basis for outlier removal. Although two feeders may share similar characteristics (and thus have a small distance between them) they themselves may be unique among the dataset. Therefore, rather than basing outlier removal on the distance between a feeder and its closest neighbor, a distance measure from each feeder to its twelve closest neighbors was computed and these distances were used to compute an average distance. If the average distance was above a certain threshold (different for all utilities) the feeder was considered an outlier and was removed from the clustering process.

3.3.4 Selecting the Number of Clusters

K-means algorithms require the number of clusters to be specified in advance, so choosing the optimal number of clusters can be one of the most difficult tasks. One popular approach for determining the optimal number of clusters is to use Cubic Clustering Criterion (CCC). CCC is a quality metric that quantifies the dissimilarity between a set of *k* clusters and the *k* clusters that would result for uniformly random data. [10]. CCC was used for this project to determine the number of clusters for each of the three data sets.

3.3.4.1 Cubic Clustering Criterion

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 accepted 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 clarify that the objective is not to find the number of clusters that gives the highest CCC value. As can be seen from Figure 9, as the number of clusters increases the CCC will also increase reaching a maximum when each data point lies within its own cluster, resulting in a minimized within-cluster sum of squares. Statistical analysis was performed using the SAS JMP software tool to calculate the CCC value for each cluster number.

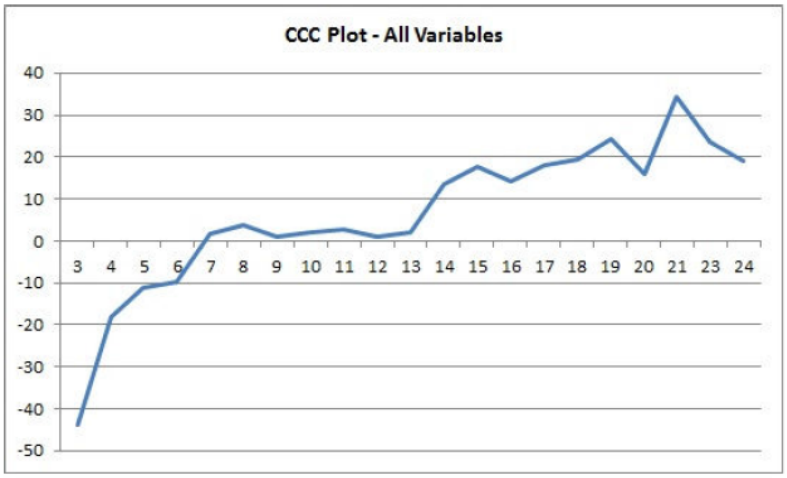

Figure 9. CCC plot with 24 clusters.

As discussed previously, the variables selected for clustering were down selected from a larger set by excluding certain highly correlated variables. The down-select process for the variables to be used in the clustering algorithm helps with selecting the optimum number of clusters for a given data set as shown below in Figure 10. Two example CCC plots are shown below in Figure 10. In Figure 10A all of the original variables were used in the clustering algorithm and to compute the CCC value. As the number of clusters increases, there is a continual rise in the CCC value with no definitive peaks up until 22 clusters. The CCC value never drops back below 2. Figure 10B shows a CCC plot using the down-selected variables based on correlation. There is a definitive peak occurring at 12 clusters, followed by a drop in the CCC value that goes below 2. This indicates that the ideal number of clusters for this data set is 12.

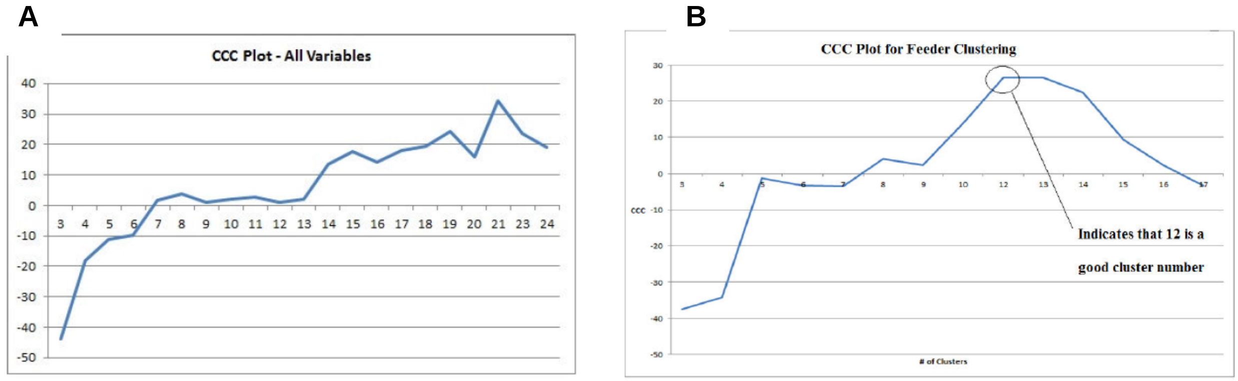

Figure 10:A) CCC plot with original variables; and B) CCC plot with down-selected variables.

3.3.4.2 Cluster Grouping Decision

The JMP software package was used to perform the K-means clustering approach to identify an optimal number of clusters by using the CCC. The resulting clusters were reviewed and clusters that had similar characteristics were evaluated and the clusters that best captured the similar characteristics were retained while the other redundant clusters were eliminated. This was done to help minimize the number of representative feeders for each utility given the project objectives and limitations.

3.3.5 Feeder Selection

Figure 11 shows an example of a biplot for the elements within a single cluster, where the multiple data dimensions have been reduced using Principle Component Analysis (PCA) to the two dominant aspects of variation. The '*90% radius'* depicted in Figure 11 represents how tightly grouped the feeders are within the cluster and is the length of the radius from the cluster's center that captures 90% of the elements within the given cluster. Feeder selection from within the cluster was accomplished by sorting the feeders based on their distance from the center mean and selecting feeders that were closest to the center of the cluster, and therefore highly representative of the cluster. Other important parameters used to make final feeder selection included significant PV system presence and the existence of feeder SCADA data. These parameters are critical for developing the accurate feeder models needed for analysis.

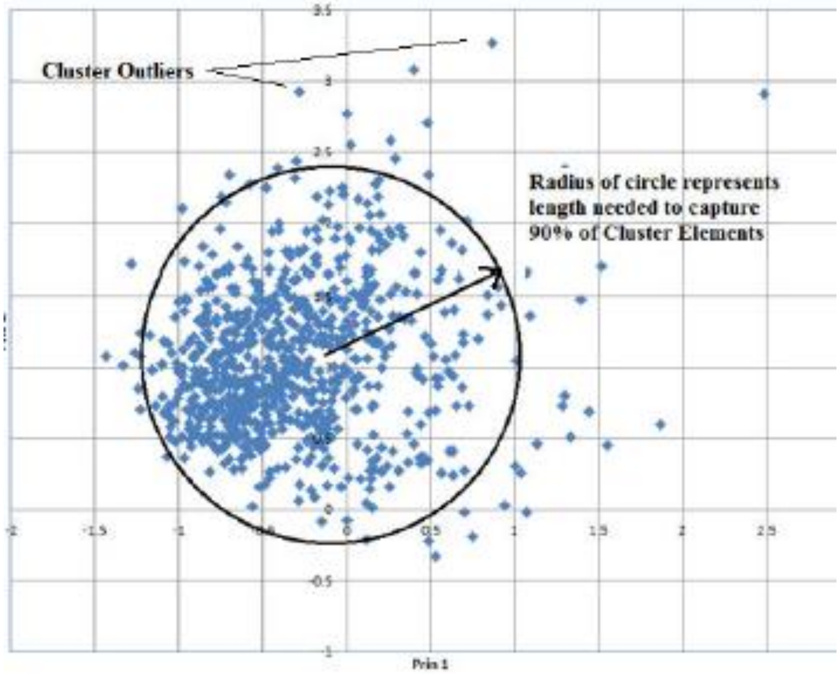

Figure 11. Example of cluster biplot.

4. CLUSTERING RESULTS FOR THE THREE UTILITIES

This chapter describes the steps taken as part of the clustering approach for each of the three utilities. Each section summarizes in general how the steps were applied to each utility. In addition, detailed examples of steps taken for a single utility are used to better illustrate the approach.

4.1. Initial Data Review and Cleanup

The initial data review and cleanup process differed for all utilities. Utilities provided data based on availability and ease of retrieval; therefore, there was significant variation in what was received from the three utilities.

The data received was used to generate histograms that helped visualize the distribution of variables of interest. The histograms often highlighted issues that needed further investigation. For example, the '*Current Carrying Capacity*' histogram for Utility 3 validated the assumption that a standard conductor was used regardless of the expected peak load. Therefore, '*Current Carrying Capacity'* was removed as a clustering variable since it was not representative of the feeder characteristics.

For all three utility data-sets, feeders that were missing critical information (e.g. Nominal voltage, Total 3-PH Circuit Miles, etc.) were excluded from the clustering approach. In addition, feeders with '*Total Circuit Miles*' less than 0.1 miles were also excluded from the clustering approach for all three utilities. Table 2 summarizes the steps for initial data cleanup for all three utilities.

Table 2. Summary of initial data cleanup for each utility.

Utility	Initial Data Cleanup
1	• Removed 2 feeders with blank kV • Removed 96 feeders with '*Total Circuit Miles (mi)*' less than 0.1
2	• Removed 12 feeders with '*Total Circuit Miles (mi)*' equal to 0 • Combined feeders with 4.16 kV and 4.8 kV into a single 4 kV group • Removed 85 feeders with kV equal to 2.4 kV, 7 kV and 25 kV due to the relatively low number of feeders represented by these voltages and to reduce the number of voltage levels for optimal clustering.
3	• Removed 8 feeders with total length of less than 0.1 miles • Translated the 'Main 3-PH Conductor' data to 'Current Carrying Capacity (AMPS)' • Translated the 'Feeder Peak Load month/time' data to a value between 0 and 24 • Feeders listing multiple climate zones were modified to include only the first climate zone listed • Data for switched/fixed capacitor banks was modified to reflect total number of switched and fixed capacitor banks (data was supplied in the format X / Y where X is the # of switched cap banks and Y is the # of fixed cap banks)

4.2. Selecting Variables for Clustering

The following initial clustering variables were common to all utilities:

- Nominal Voltage (kV)
- Total 3-PH Circuit Miles (mi)
- Total 3-PH Overhead Circuit Miles (mi)
- Total 2-PH and 1-PH Circuit Miles (mi)
- Total 2-PH and 1-PH Overhead Circuit Miles (mi)
- Number of Switched/Fixed Capacitor Banks (#)

Correlation maps for all three utilities showed high correlations between 'Total 3-PH Circuit Miles' and 'Total 3-PH Overhead Circuit Miles' and between 'Total 2-PH and 1-PH Circuit Miles' and 'Total 2-PH and 1-PH Overhead Circuit Miles.' Both 'Total 3-PH Overhead Circuit Miles' and 'Total 2-PH and 1-PH Overhead Circuit Miles' were excluded from the final clustering analysis.

Figure 12 shows the correlation map between the fifteen initial clustering variables chosen for Utility 3. The first point highlights the high correlation between Total Miles (3-PH, 2-PH and 1-PH) and Overhead Miles (3-PH, 2-PH and 1-PH) as discussed above. The second point shows a high correlation between the '*Number of line voltage regulators*' and Total/Total Overhead Miles (3-PH, 2-PH and 1-PH). The third point shows a medium-high correlation between '*Connected Service Transformer Capacity (kVA*)' and '*Total 3-PH Circuit Miles*'. In the fourth point a medium-high correlation is shown between '*Feeder Peak Load (kW)*' and '*Connected Service Transformer Capacity (kVA)*'. The fifth point shows a negative correlation between '*Commercial %*' and '*Residential %*.'

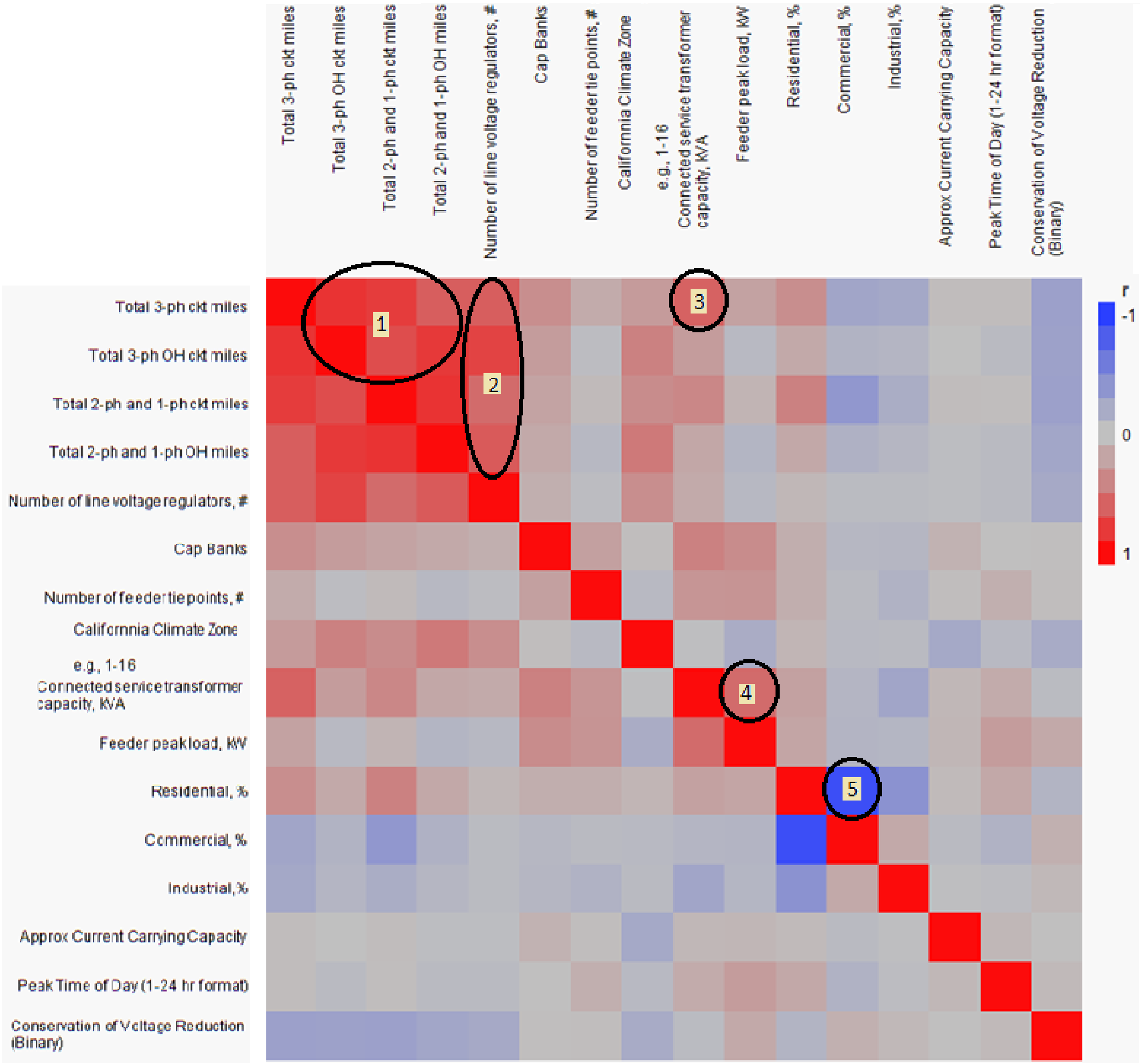

Figure 12. Correlation map for Utility 3.

After close examination of the correlation map and several team discussions, the following eleven variables were used for Utility 3 clustering.

1. Total 3-PH Circuit Miles (mi)
2. Total 1-PH and 2-PH Circuit Miles (mi)
3. Number of Line Voltage Regulators (#)
4. Fixed and Switched Capacitors Banks (#)
5. Number of Feeder Tie Points (#)
6. Connected Service Transformer Capacity (kVA)
7. Feeder Peak Load (kW)
8. Residential (%)

9. Commercial (%)
10. Industrial (%)
11. Feeder Peak Load Time (#)

Initial and final clustering variables for the other Utilities 1 and 2 can be found in Appendix B.

4.3. Removing Outliers

The approach described in section 3.3.3 was used to remove outliers from the clustering analysis. Using the *Declutter* tool in the JMP software, distance plots were generated for each utility. Figure 13 is a distance plot for utility 2 showing the average distance for each feeder and its twelve closest neighbors. The distance plot was used to decide the average distance threshold for removing outliers. The distance threshold for each utility was determined by looking at the distance plot and visually determining where the separation was most apparent. The red line in the figure shows the threshold that was chosen for Utility 2. The darker points represent feeders that were excluded from the cluster analysis because they were above the distance threshold for this utility. Thresholds used for detecting outliers for each utility are given in Table 3, along with the number of feeders that were labeled as outliers and where therefore, excluded from the cluster analysis.

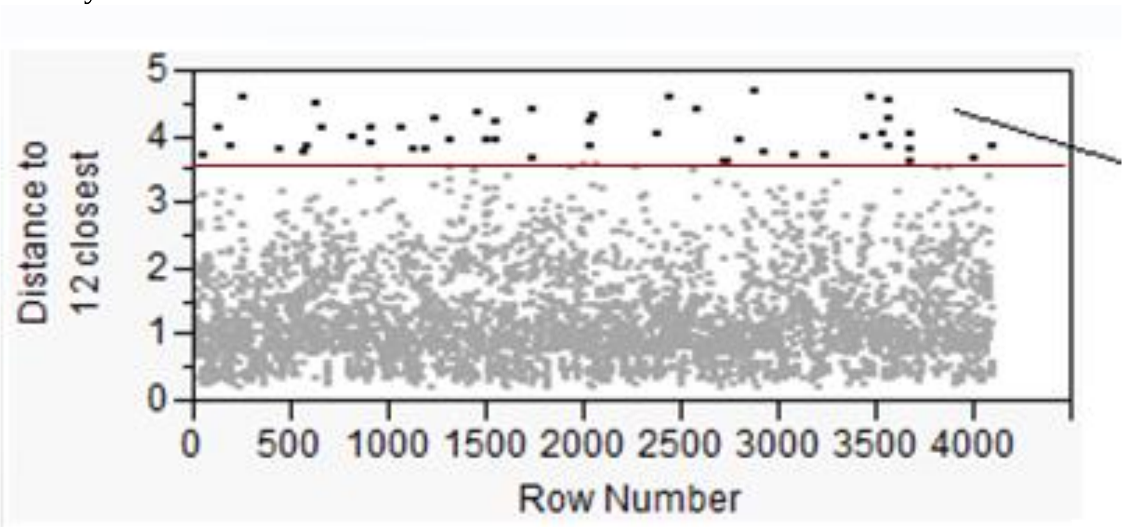

Figure 13. Distance plot (for 12 closest feeders) for Utility 2.

Table 3. Distance threshold for outlier detection & number of excluded feeders for each utility.

Utility	Distance Threshold for Outliers	Number of Feeders Excluded
1	> 4.0	47
2	> 3.3	49
3	> 4.6	17

4.4. Selecting the Number of Clusters

The approach outlined in 3.3.4 was used to determine the number of clusters for each utility. K Means clustering was performed using a varying K, where K represents the number of clusters. For each given K, JMP calculated and returned a CCC value. As stated previously, a local maximum that comes after a rise above a CCC value of 2 and that comes before a drop below a CCC value of 2 indicates a good selection for the number of clusters adequate for a given data-set. Figure 14 shows '*Number of Clusters K'* vs. '*CCC value*' for Utility 2. In this case, it is apparent that for the given data-set, having K equal to eight clusters would be optimal. Eight clusters also make sense keeping in mind the project objectives and limitation of selecting no more than twenty representative feeders across all three utilities.

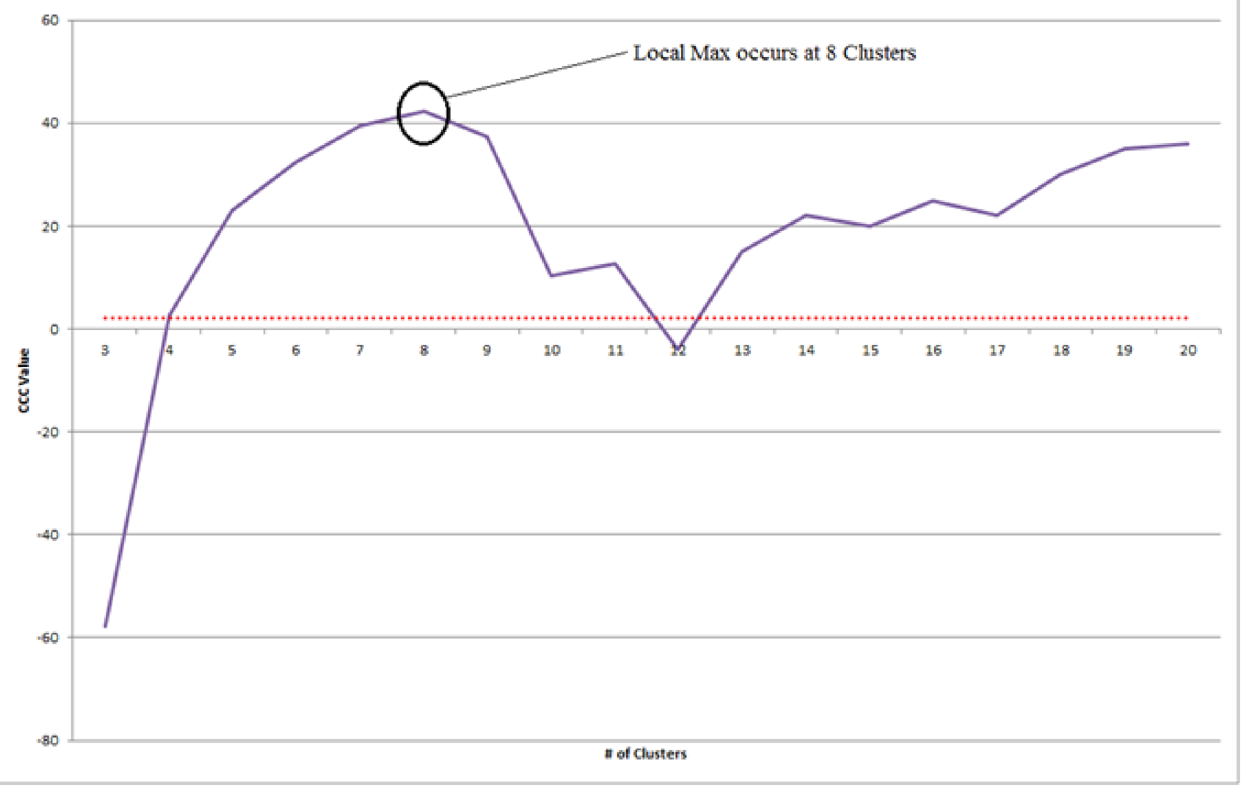

Figure 14. Number of cluster vs. CCC value for Utility 1.

Table 4 shows the resulting cluster means for Utility 2 and Table 5 describes the characteristics of each of the eight clusters for Utility 2. In Table 5 it is apparent that the characteristics of cluster 4 are well captured in clusters 1 and 8 and therefore, cluster 4 was eliminated leaving Utility 2 with a total of seven clusters. Cluster means for Utility 1 and Utility 3 can be found in Appendix C.

Table 4. Resulting cluster means for utility 2.

Cluster	Feeder count	Nominal, voltage, kV	Total 3-ph ckt miles, mi	Total 2-ph and 1-ph ckt miles, mi	Number of switched/fixed capacitor banks, #	Number of feeder tie points, #	Connected service transformer capacity, kVA	Feeder peak load, kVA (calculated)	Total customer count	Industrial customer count	Other customer count	Peak season	90% Radius
1	70	12.57	31.86	38.92	5.89	4.97	19094	6707	1536	1.74	32.64	2.97	57
2	63	33	19.14	0.14	0.06	2.14	3969	12177	8	0.73	1.71	2.84	43.2
3	174	14.05	9.06	2.13	4.33	7.38	21134	7607	1489	29.99	12.83	2.89	44.1
4*	130	12.83	25.17	7.12	4.32	6.28	12711	4977	829	2.27	74.06	2.79	45.7
5	950	4	2.9	2.21	1.8	2.61	2199	1609	665	0.51	2.43	2.56	9.8
6	1184	12.66	6.87	3.3	2.4	4.75	10426	5502	685	194	6.09	2.87	21.5
7	1404	12.89	11.1	9.49	4.37	7.67	16730	8960	1875	2.01	11.12	2.9	26.4
8	73	12.49	37.57	3.87	5.41	8.12	15351	5752	796	2.1	265.7	2.82	69

* Cluster 4 was eliminated due to the representation of its characteristics in other clusters.

Table 5. Cluster characteristics for utility 2.

Cluster	Characteristics
1	Long feeders, Long 1&2 phase
2	33 kV feeders
3	Medium length, high load feeders with heavy industrial
4	Long agricultural feeders
5	4 kV feeders
6	Short feeders
7	Medium length, high load feeders
8	Long 3 phase agricultural feeders

4.5. Selection of Representative Feeders

Representative feeder selection for all utilities was based on a combination of distance to the cluster mean and existing PV capacity. The first feeder selected for each cluster was simply the feeder with the shortest distance to the cluster mean. The second feeder for a given cluster was selected based on distance (closest to the cluster mean) and the PV Capacity (greater than 100 kW). PV capacity was not used earlier in the clustering process, but was used as a down select criterion to improve the likely hood of having feeders with a significant amount of PV systems. The third feeder selected for a given cluster was also based on distance and PV capacity (greater than 1000 kW). Table 6 shows initial feeder selection for the final seven clusters for Utility 2.

Table 6. Initial feeder selection for utility 2.

EPRI Feeder #	Nominal Voltage	Total 3-ph miles	Total 2-ph and 1-ph miles	Cap Banks	Feeder Tie Points	kVA Capacity	Peak kVA	Res %	Com %	Ind %	Agr %	Total Cust	Largest PV system, kW	PV Capacity w/ UOG	Distance
Cluster 1 Mear	12.57	31.9	98.92	5.89	4.97	1904	6707					1536			
1494	12	39.5	37.73	5	8	19604	7233	78	16	0	7	1845	6	35	5.82
1242	12	33.5	33.73	4	3	15070	5046	69	21	0	10	1352	84	220	6.68
3366	12	36.7	27.20	5	4	15997	4292	62	9	0	30	1182	994	1367	11.75
Cluster 2 Mear	33.00	19.1	0.14	0.06	2.14	3969	12177					8			
52	33	22.9	0.39	0	2	300	13270	97	30	0	0	15	0	0	1.70
3999	33	29.8	0.00	0	5	10050	15524	0	0	98	2	6	895	1793	11.42
3358	33	17.2	0.00	0	3	20000	8735	0	23	72	0	7	895	1793	13.51
Cluster 3 Mear	14.05	9.1	2.13	4.33	7.38	21134	7607					1489			
3618	12	8.0	1.20	4	6	20926	6797	10	37	53	0	1603	0	0	5.46
2802	16	8.0	1.10	4	9	20558	5474	29	55	16	1	1348	43	181	5.94
228	12	8.2	0.40	6	4	24315	7400	5	66	28	1	650	1000	1999	9.98
Cluster 5 Mear	4.00	2.9	2.21	1.80	2.61	2199	1609					665			
2480	4	2.8	2.55	2	3	2280	1685	91	9	0	0	711	0	0	0.71
2543	4	2.9	2.16	2	2	3218	1913	68	32	0	0	713	997	1012	0.83
3655	4	2.5	2.58	1	4	2038	2427	92	8	0	0	664	997	1039	1.42
Cluster 6 Mear	12.66	6.8	3.30	2.40	4.75	10426	5502					685			
545	12	6.2	4.51	3	5	11583	6085	20	30	50	0	531	8	19	1.13
149	12	6.5	3.22	3	5	12770	5343	13	15	67	6	477	996	2026	1.35
4122	12	7.2	7.70	2	4	7900	5160	93	0	7	0	670	998	1041	2.33
Cluster 7 Mear	12.89	11.1	9.49	4.37	7.67	16730	8960					1875			
498	12	10.1	9.26	4	8	17100	10167	54	36	7	4	1829	10	27	1.63
420	12	10.2	10.52	4	7	16255	9717	51	33	12	4	1962	508	890	1.64
2649	12	12.2	11.53	4	8	18810	8888	46	46	8	1	2144	994	1031	1.71
Cluster 8 Mear	12.49	37.6	3.87	5.41	8.12	16351	5752					796			
2921	12	37.1	1.29	6	7	14023	7946	6	4	0	90	411	23	24	2.99
1206	12	42.1	0.82	5	6	19187	7116	2	8	2	89	349	498	536	6.86
1787	12	49.5	1.21	6	10	18250	6048	5	3	6	87	495	1000	1104	10.30

In addition to the feeders selected based on shortest distance to cluster mean and existing PV capacity, several other feeders were selected because they contained Utility Owned Generation (UOG) Photovoltaics and had feeder SCADA data. Table 7 shows feeders (highlighted in gray) that were selected for clusters 6, 7 and 8 due to the criteria just described.

Table 7. Initial feeder selection for Utility 2 feeders with utility owned PV.

EPRI Feeder #	Nominal Voltage	Total 3-ph miles	Total 2-ph and 1-ph miles	Cap Banks	Feeder Tie Points	kVA Capacity	Peak kVA	Res %	Com %	Ind %	Agr %	Total Cust	Largest PV system, kW	PV Capacity w/ UOG	Distance	UOG PV, kW	#UOG PV Sites
Cluster 6 Mear	12.66	6.81	3.30	2.40	4.75	10426	5502					685					
545	12	6.24	4.51	3	5	11583	6085	20	30	50	0	531	8	19	1.13		
149	12	6.51	3.22	3	5	12770	5343	13	15	66	6	477	996	2026	1.35		
4122	12	7.22	7.70	2	4	7900	5160	93	0	7	0	670	998	1041	2.33		
172	12	3.25	0.07	3	6	11885	7025	0	8	92	0	9	0	2500	3.95	2500	1
1225	12	10.73	0.03	2	5	17885	6399	0	50	50	0	87	0	2500	5.18	2500	2
83	12	8.44	0.71	1	2	15970	7494	0	6	94	0	136	174	5348	6.23	5000	2
2159	12	9.59	0.23	4	10	13627	5739	0	52	48	0	74	0	5000	8.76	5000	2
Cluster 7 Mear	12.89	11.10	9.49	4.37	7.67	16730	8960					1875					
498	12	10.11	9.26	4	8	17100	10167	53	36	7	4	1829	10	27	1.63		
420	12	10.18	10.52	4	7	16255	9717	51	33	12	4	1962	508	890	1.64		
2649	12	12.21	11.53	4	8	18810	8888	45	46	8	1	2144	994	1031	1.71		
281	12	8.70	8.70	5	12	18923	8538	55	11	34	0	1880	603	1648	5.04	1000	1
1774	12	14.60	10.69	2	5	22779	9169	55	40	5	0	2176	7	5511	7.43	5500	4
2037	12	11.96	2.07	3	9	21631	6839	15	65	20	0	597	625	4645	10.34	3500	1
29	12	16.15	0.16	5	10	19440	8559	0	64	32	4	242	113	4726	14.06	4500	2
3278	12	24.68	14.63	6	10	20556	10557	54	34	0	12	1841	16	6100	15.59	6000	1
27	12	14.71	1.99	2	10	27569	8040	12	39	46	3	329	264	7050	18.22	6500	3
407	12	13.36	0.10	5	9	35205	10439	0	69	31	0	158	352	1989	30.07	1000	1
Cluster 8 Mear	12.49	37.57	3.87	5.41	8.12	16351	5752					796					
2921	12	37.10	1.29	6	7	14023	7946	6	4	0	90	411	23	24	2.99		
1206	12	42.13	0.82	5	6	19187	7116	2	7	2	89	349	498	536	6.86		
1787	12	49.52	1.21	6	10	18250	6048	4	3	6	87	495	1000	1104	10.30		
2151	12	39.74	2.16	4	6	15689	2989	7	8	0	85	604	2	5002	58.31	5000	1

5. SUMMARY OF RESULTS

There were twenty two feeders selected from those identified in the clustering. The final set includes 16 feeders for detailed analysis and development of the screening methodology, while 6 feeders are for validation of the methodology. There have been two ‘bonus’ feeders added in addition to the initially planned twenty. These ‘bonus’ feeders have been previously analyzed under another EPRI project but also fit well into the specified clusters. Leveraging this previous work will improve the results from the CSI project.

The intent was to select sixteen feeders that would represent the range of differences seen within and among utilities to develop a screening method that would be widely applicable. Ideally, one feeder would have been chosen from each identified cluster to represent all of the three participating utilities appropriately, however, there were more than 16 total clusters. The clusters were examined to identify similar primary characteristics and reduced the clusters in which feeders were chosen for detailed analysis.

One of the primary characteristics to manually reduce clusters was nominal voltage. The utility feeder clusters represent a wide range of voltage classes from 4 kV to 33 kV. The majority of the feeders fall into the 12 kV class, therefore, the majority of the feeders chosen for analysis were selected from that kV class. Three clusters were used to represent outlying voltage classes such as 4 kV and 33 kV. The 12 kV voltage class feeders have a more detailed representation of the range in feeder lengths and characteristics. From each of the remaining 17 unique clusters, one or more feeders were chosen for detailed analysis.

The specific feeder chosen from each cluster had

1) a relatively high customer count
2) SCADA measurement data
3) a relatively low number of small residential PV

Customer count is important to the Distributed PV (DPV) analysis in that the amount of small residential PV analyzed is dependent on the potential customers. Measurement data is necessary to accurately build and validate the feeder model. Large utility owned PV is beneficial for the analysis since the location and output data is more easily included in the electrical model. Many small residential systems are more difficult to place and accurately account for when decoupling from the load and overall feeder measurement data.

The six validation feeders were chosen from the full cluster set. Some validation feeders were chosen from within clusters used for the detailed analysis while others were chosen from those clusters not included in the detailed analysis. The methodology developed from the detailed analysis should apply for all feeders not necessarily in a specific cluster, thus the validation feeders could be attained from any data set. The validation feeders must have large PV systems (preferably utility owned) where measurement data can be attained. These measurements will be utilized to validate the methodology and construct the feeder model for additional methodology validation.

Table 8. Final Feeder Selection

Utility	Feeder No.	Feeder ID	Notes	Cluster
SCE	1	967	Study Feeder	1
SCE	2	3999	Study Feeder	2
SCE	3	2802	Study Feeder	3
SCE	4	2543	Validation Feeder	5
SCE	5	1231	Validation Feeder	6
SCE	6	420	Study Feeder	7
SCE	7	2921	Study Feeder	8
PGE	1	2093	Study Feeder	3
PGE	2	2885	Study Feeder	4
PGE	3	142	Study Feeder	6
PGE	4	281	Validation Feeder	6
PGE	5	888	Study Feeder	8
PGE	6	1354	Study Feeder	9
PGE	7	1140	Validation Feeder	10
SDGE	1	683	Study Feeder	1
SDGE	2	404	Study Feeder	2
SDGE	3	296	Study Feeder	4
SDGE	4	525	Study Feeder	5
SDGE	5	679	Validation Feeder	5
SDGE	6	514	Validation Feeder	1
SDGE	7	631	Study Feeder	2
SDGE	8	440	Study Feeder	N/A

6. CONCLUSIONS

This work demonstrates a method to classify distribution feeders into clusters and to select representative feeders from each cluster. This paper outlined the method for using the K-means clustering methodology for grouping distribution feeders and the use of the Cubic Clustering Criterion for determining the optimum number of clusters. K-means clustering was found to be a very effective and versatile method for clustering.

A key finding of this work is that a relatively small number of initial clusters (5-12) are needed to represent the variation in the feeder characteristics for each utility. This work also demonstrates the fundamental importance of voltage class and feeder length for distinguishing between clusters of feeders which matches the power engineering design criteria for distinguishing between feeders.

Representative feeders were selected from each of the final 17 unique clusters as shown in table 8, but the limitation on the total number of feeders to be analyzed of 22 under this project also limited the opportunity to sample multiple feeders from each cluster to better understand how the cluster variation affects hosting capacity.

Analysis of these representative feeders can have significant impact on the screening process for requests for interconnection 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.

7. REFERENCES

[1] Berkeley Law, "California's Transition to Local Renewable Energy: 12,000 Megawatts by 2020," September 29, 2013, Available at: http://www.law.berkeley.edu/files/Transition_to_Local_Renewable_Energy_February_2012_DRAFT(1).pdf.

[2] Updating Interconnection Screens for PV System Integration, NREL/TP-5500-54063, January 2012 [Online], Available at: http://energy.sandia.gov/wp/wpcontent/gallery/uploads/Updating_Interconnection_PV Systems_Integration.pdf.

[3] 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 at: http://onlinelibrary.wiley.com/doi/10.1002/pip.1204/pdf.

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

[5] J.A. Hartigan, "Clustering Algorithms," New York: Wiley, 1990.

[6] O.A. Abbas, "Comparisons Between Data Clustering Algorithms," in The International Arab Journal of Information Technology, 2008, vol. 5, No. 3, [Online], Available at: http://www.ccis2k.org/iajit/PDF/vol.5,no.3/15-191.pdf.

[7] K. P. Schneider et al, "Modern Grid Initiative Distribution Taxonomy Final Report," November 2008 [Online], Available at: www.gridlabd.org/models/feeders/taxonomy_of_prototypical_feeders.pdf.

[8] 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.

[9] M. Norusis, "IBM SPSS Statistics 19 Statistical Procedures Companion," 2012 [Online], pp. 390, Available at: www.norusis.com/pdf/SPC_v13.pdf .

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

APPENDIX A
Feeder Data Requested

A. Feeder Identifiers;

- Utility ID
- Feeder ID
- Substation ID

B. Feeder Characteristics

- Nominal voltage (kV)
- Type of feeder (Radial or Network)
- Total 3-ph ckt miles (mi)
- Total 3-ph OH ckt miles (mi)
- Main 3-ph conductor (e.g. "336 kcm OH")
- Total 2-ph and 1-ph ckt miles (mi)
- Total 2-ph and 1-ph OH miles (mi)
- Substation LTC (Yes or No)
- LTC set points, target / bandwidth
- Number of line voltage regulators (#)
- Distance between voltage regulators
- VR set points, target / bandwidth
- Number of switched/fixed cap banks (#)
- Number of feeder tie points (#)
- Short circuit capacity at 3-ph node farthest from station (MVA)
- Conservation voltage reduction feeder? (Yes or No)
- Station latitude
- Station longitude
- California Climate Zone (e.g., 1-16)
- Approximate service area (sq mi)
- Connected service transformer capacity (kVA)

C. Load Characteristics

- Feeder peak load (kW)
- Feeder peak load month/time (mo/hr)
- Feeder minimum load (estimated) (kW)
- Feeder minimum load month/time (mo/hr)
- Residential (%)
- Commercial (%)
- Industrial (%)
- Agricultural (%)

D. DG and PV Installed

- Total amount of DG on circuit (kW)
- Existing PV capacity installed (kW)
- Largest PV system installed (kW)
- Does feeder contain utility-owned PV (Yes or No)

E. Measurement Data Available

- (Archived load data at feeder level (data rate or N/A if not measured)
- Archived load at station level (data rate or N/A if not measured)
- Highest possible station/feeder data rate with existing equipment (data rate)
- Archived PV system output data (data rate or N/A if not measured)
- Highest possible PV plant output data rate with existing equipment (data rate)
- Solar irradiance monitoring or data? (Yes or No)

APPENDIX B

Utility 1 & 2 Clustering Variables.

Utility 1	
Initial Clustering Variables	**Final Clustering Variables**
1. Primary Voltage 2. Total 3PH Miles 3. Total 3PH Overhead Miles 4. Total 1&2 PH Miles 5. Total 1&2 PH Overhead Miles 6. Regulators 7. Capacitors 8. Boosters 9. Sectionalizers + Reclosers 10. Domestic Customers 11. Commercial Customers 12. Industrial Customers 13. Agricultural Customers 14. Total Customers 15. Ratio of Summer Peak to Winter Peak 16. Summer Peak kW 17. Summer kVA Capability	1. Primary Voltage 2. Total 3PH Miles 3. Total 1&2 PH Miles 4. Regulators 5. Capacitors 6. Boosters 7. Sectionalizers + Reclosers 8. Industrial Customers 9. Agricultural Customers 10. Total Customers 11. Ratio of Summer Peak to Winter Peak 12. Summer kVA Capability

Utility 2	
Initial Clustering Variables	**Final Clustering Variables**
1. Primary Voltage 2. Total 3PH Miles 3. Total 3PH Overhead Miles 4. Total 1&2 PHMiles 5. Total 1&2 PH Overhead Miles 6. Capacitors 7. Feeder Tie Points 8. Connected Transformer Capacity (kVA) 9. Feeder Peak Load (kVA) (calculated) 10. Total Customers 11. Commercial Customers 12. Domestic Customers 13. Industrial Customers 14. Other Customers (Agr) 15. Peak Season	1. Primary Voltage 2. Total 3PH Miles 3. Total 1&2 Miles 4. Capacitors 5. Feeder Tie Points 6. Connected Transformer Capacity (kVA) 7. Feeder Peak Load (kVA) 8. Total Customers 9. Industrial Customers 10. Other Customers (Agr) 11. Peak Season

APPENDIX C

Utility 1 & 3 Cluster Means.

Utility 1:

Feeder Count	Primary Voltage	Total 3-Phase miles	Total 1 & 2 Phase miles	Ind Cust	Agr Cust	Total Cust	Regulators	Capacitors	Boosters	Reclosers + Sectionalizers	Summer KVA Capability	Summer/Winter	90% Radius
136	12.29	48.41	57.99	30	26	2218	4.54	4.5	0.35	4.94	9650.82	1.14	38.78
735	12	20.31	11.02	62	5	2929	0.3	4.96	0.07	1.97	11309.24	1.29	15.06
114	12.86	132.15	15.78	19	286	1232	6.89	6.82	2.18	5.5	10902.51	1.96	85.54
290	12.02	64.6	6.38	16	108	750	3.27	4.94	0.94	2.88	9540.62	1.81	25.37
94	12	49.6	40.86	35	26	1794	3.09	5.01	3.16	4.32	9481.72	1.2	36.11
214	20.64	41.3	31.09	74	25	3628	1.11	5.72	0.24	4.68	20083.13	1.58	56.56
237	21.01	18.55	10.52	63	6	1713	0.31	3.21	0.08	1.31	19202.85	1.47	29.1
410	4	3.79	2.16	7	0	883	0.11	1.48	0.03	0.19	2497.93	0.94	4.35
749	12.01	10.49	3.57	27	5	704	0.26	2.09	0.07	0.54	9335.52	1.19	13.53
59	13.02	88.78	116.11	31	53	2653	8.61	5.97	1.97	8.22	10683.2	1.29	74.77

Utility 3:

Cluster	Total 3-ph ckt miles	Total 2-ph and 1-ph ckt miles	Number of line voltage regulators, #	Cap Banks	Number of feeder tie points, #	Connected service transformer capacity, kVA	Feeder peak load, kW	Residential, %	Commercial, %	Industrial, %	Hour	Count
1	29.17	34.56	1.58	1.90	1.94	20620.39	5532.90	82.42	17.50	0.07	12.50	31
2	16.04	14.17	0.06	2.03	2.70	21348.99	8523.84	90.62	9.34	0.04	15.01	263
3	2.17	0.02	0.00	1.18	1.05	5779.09	6061.82	0.00	49.70	50.30	14.47	22
4	7.86	4.92	0.02	1.26	1.97	12831.42	6118.02	86.38	13.44	0.17	13.77	283
5	5.81	0.27	0.03	1.31	1.88	13283.80	6111.84	4.62	84.82	3.31	11.78	152

DISTRIBUTION

External distribution (Electronically distributed unless otherwise noted)

1 Jeff Smith
Electric Power Research Institute
942 Corridor Park Boulevard
Knoxville, TN

2 Joe Williams
UK Power Networks
Barton Road
Bury St Edmunds
IP32 7BG

Internal distribution (Electronically distributed unless otherwise noted)

1	MS1033	Robert Broderick	6112 (1 electronic & 1 paper)
1	MS1188	Karina Munoz-Ramos	6114
1	MS0899	Technical Library	9536
1	MS0115	WFO/CRADA Agreements	10769

Sandia National Laboratories