Application of Cluster Analysis on Antenna Factor Measurements Data

[1]M. Kollár, [2]K. Kováč

[1]Seibersdorf Laboratories GmbH, RF-Engineering, A-2444 Seibersdorf Austria, Email matej.kollar@seibersdorf-laboratories.at

[2]Department of Measurement, Faculty of Electrical Engineering and Information Technology, Ilkovičova 3, 81219 Bratislava, Slovak Republic

Abstract. *Clustering is a method of partitioning set of observations into subsets (clusters) in a way that observations within individual subsets are similar in some sense. Antenna calibration is a method for obtaining antenna factor for specific antenna. Antenna factor is then used to determine the actual radiation emission level in various measurements where the antenna is involved. This paper deals with cluster analysis of antenna factor data for various antenna models and different calibration methods and parameters. As the cluster analysis is extremely data specific, related problems are discussed. The application of several clustering methods to the antenna factor data has been explored. The evaluation of the obtained results is presented.*

Keywords: Antenna Calibration, Antenna Factor, Data Clustering

1. Introduction

Calibration of the antennas is a very important type of measurement which enables today industry to perform various radiation emission and immunity tests on electronic devices. The purpose of the antenna calibration procedure is to obtain antenna factor data that can be used for correction of the received radiation emission level. When calibrating an antenna, there are plenty of effects which influence resulting antenna factor e.g. manufacturing properties, errors caused by a measurement system, errors originating from the antenna setup, and measurement errors inherent in the method. All these form the actual appearance of the calibration data. When various measurements are displayed in one chart they tend to create natural groups (clusters), see Fig.1. Being able to recognize these groups, one may extract valuable information, about the data generation process. It is possible to say, if the data measured with a specific antenna diverts from the others of the same model, the antenna or measurement is likely to be invalid and further verification has to be conducted.

2. Subject and Methods

Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval, and bioinformatics. Antenna factor data can be viewed as values of a function of frequency, and thus our problem can be seen as functional clustering. For this purpose we have selected two well-known clustering approaches.

1. Hierarchical clustering finds successive clusters using previously established clusters. These algorithms are either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

2. Partitional algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering.

An important step in clustering is to select an appropriate distance/dissimilarity measure for the given problem (e.g., Euclidean, City-block, Chebyshev). The choice of a particular dissimilarity measure may significantly affect the actual shapes of the resulting clusters.

Antenna calibration data are characteristic by non-uniform sampling in the frequency domain. This is caused by different frequency steps used for different frequency ranges. E.g. in the range between 1 and 100MHz the data are acquired with 1MHz step. On the other hand, in the range between 100MHZ and 200MHz the frequency step is 5MHz. In our experiments we consider the calibration curves as vectors and use the Euclidean distance that has proved to be a proper distance measure. It progressively places a greater weight on measurement points which are further apart, and thus it appropriately determines dissimilar measurements.

Next, the agglomerative hierarchical and k-means++ clustering algorithms were applied to our data. In the first step of the agglomerative hierarchical clustering, all the elements are considered as individual clusters. Afterwards, they are sequentially paired according to the selected distance measure and linkage criterion [1], until there are no clusters to be merged. In our analysis, the complete linkage has been used.

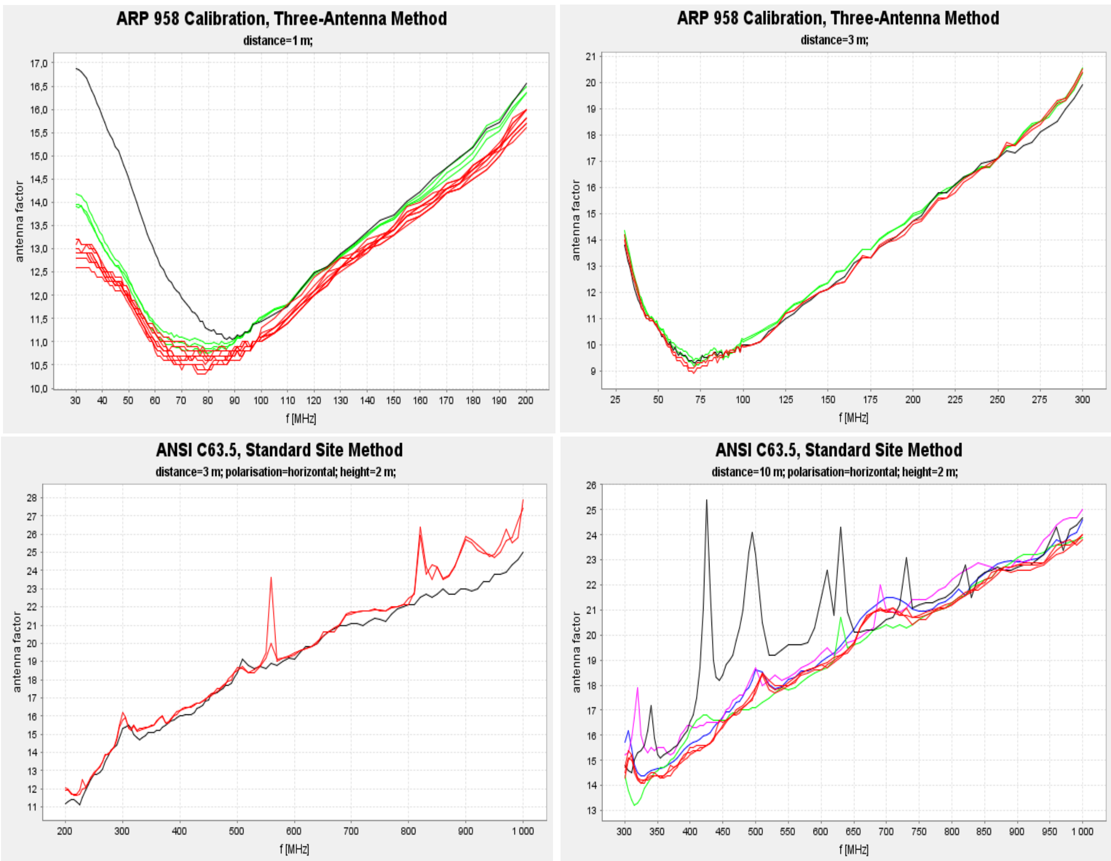

Fig. 1. Result of the cluster analysis applied to the antenna factor measurements. Individual clusters are drawn in different colours.

At the beginning of k-means clustering, a set of randomly chosen data points are set as *k* initial cluster centers (centroids). In the next step, all data points are assigned to the nearest centroids. Afterwards, new cluster centers are computed. This process is repeated until cluster centers arrive at the stable positions or maximum number of iterations is reached. This method does not guarantee that the final clusters always contain the same elements and therefore the k-means++ [2] algorithm was used. This modification of the original algorithm determines the initial cluster centers in such a way that previous property is satisfied.

Data for clustering were selected from a set of calibrations performed according to ANSI C63.5 [3] and ARP 958 [4]. In Fig.1 some of the clustering results for different measurement distances (1m, 3m, 10m) are depicted. One can see that the given set of data contains correct calibration data, as well as false antenna calibrations. Using the Euclidean distance measure, there was no difference in performance between both clustering methods.

The crucial property of the both clustering algorithms is that they require the desired number of clusters is defined prior to the execution of the algorithm. If the number of clusters is not explicitly known, one needs to estimate this number empirically from the data. This is a problem on its own for which a number of techniques can be employed. Some of them are as follows: entropy-based partitioning of dendrogram [5], information theoretic approach [6], Silhouette [7], *v*-fold cross-validation [8]. The common problem of these methods is that they are more suited for large datasets which is not the case of the antenna calibration data. Furthermore, some of them involve algebraic matrix operations, which fail due to singularities and violated assumptions when applied to unsuitable data.

The principle of determining the number of clusters in arbitrary dataset is to define a criterion function that measures the clustering quality of any partition of the data. Then the problem is to find such a partition of data that minimizes this criterion function [9]. If the dataset to be clustered contains a reasonable number of elements, it is possible to run clustering algorithm iteratively for various numbers of clusters starting with one to a size of the dataset and stopping the algorithm when the minimum value of the criterion function is reached.

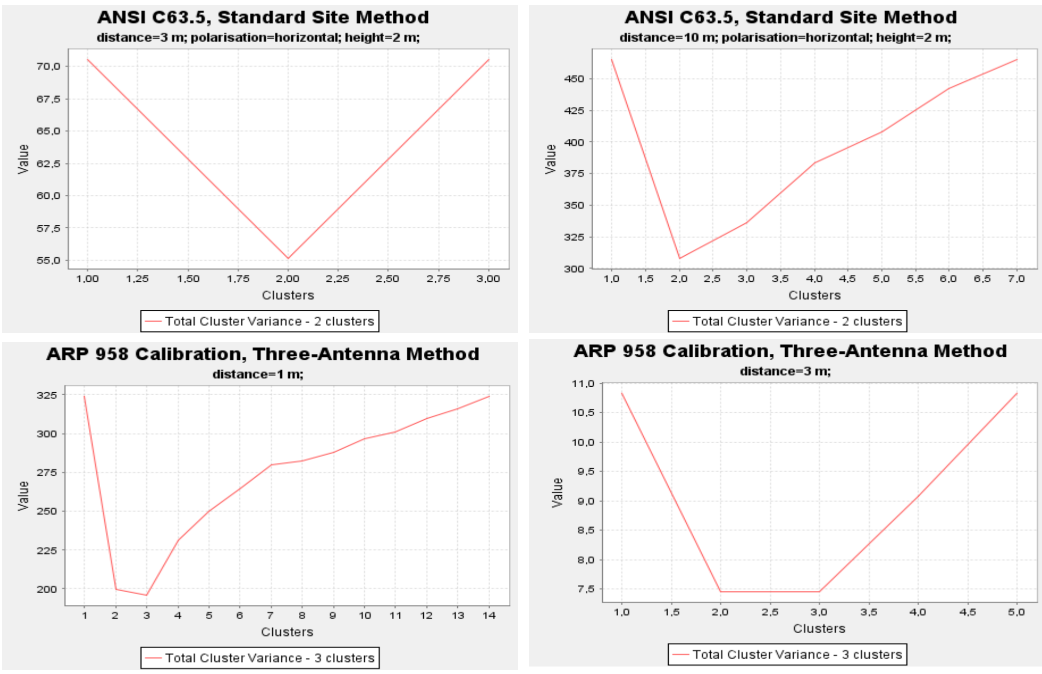

Fig. 2. The actual shapes of the total cluster variance criterion functions applied to different datasets. The minimum gives the estimate on the number of clusters.

3. Results

Examining various criterion functions, the total cluster variance S_T delivered the most acceptable results. In our case the total cluster variance is defined by the sum of between cluster variance S_B and within cluster variance S_W.

$$S_T = S_B + S_W$$

The resulting criterion functions, calculated for the dataset are depicted in Fig.1, are shown in Fig.2. In the case of ANSI C63.5 10m distance dataset only 2 instead 5 clusters were determined.

4. Discussion

The application of the agglomerative hierarchical, as well as the k-means clustering was successful for partitioning of the measured data of the antenna factor. In determination of the number of clusters in the dataset, we used the total variance criterion function in combination with the agglomerative hierarchical clustering.

In general, the cluster analysis appeared to be a complex application-specific problem. The standard clustering algorithms performed well on the antenna factor measurements. One can expect that the same situation may apply also to other similar data. In determination of the number of clusters, a difficult question has arisen, what is the real number of clusters? Usually it is being resolved by an expert opinion which explicitly assigns particular data elements to predefined groups. Although a promising automatic method for determination of the number of clusters for antenna factor data has been identified, it has to be verified on larger datasets.

References

[1] "The CLUSTER Procedure: Clustering Methods". SAS/STAT 9.2 Users Guide. SAS Institute.

[2] Arthur, D. and Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding". Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms: 1027–1035.

[3] ANSI C63.5-1988; American National Standard for Electromagnetic Compatibility - Radiated Emission Measurements in Electromagnetic Interference (EMI) Control - Calibration of Antennas (9 kHz to 40 Ghz).

[4] SAE ARP 958; SAE Aerospace Recommended Practice a Standard Method for Checkout and Calibration of Electromagnetic Interference Measurement Antennas

[5] Casado et al. (1997). “An objective method for partitioning dendrograms based on entropy parameters”. Plant Ecology Volume 131, Number 2, Springer Netherlands, pp. 193-197.

[6] Catherine A. Sugar and Gareth M. James (2003). "Finding the number of clusters in a data set: An information theoretic approach". Journal of the American Statistical Association 98 (January), pp. 750–763.

[7] Peter J. Rousseuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65.

[8] Geisser, Seymour (1993). Predictive Inference. New York: Chapman and Hall.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification,. New York: John Wiley & Sons, 2001, Unsupervised Learning and Clustering, pp. 29-37a