Cluster Analysis and its Application in the National Resources Inventory

Jianqiang Wang, Jean Opsomer
Department of Statistics, Iowa State University

Abstract

Cluster analysis is a popular statistical tool, which can help researchers explore the structure of multidimensional data, find special groups in populations and seek associations between individual units. It can also be applied to detect unusual points in data. The National Resources Inventory is a longitudinal survey of natural resources information on nonfederal land in the US. One possible problem encountered during NRI data collection and processing is the existence of unusual observations and outliers. These observations need to be identified and evaluated for correctness, in order to ensure the quality of the NRI data. An exploratory study is conducted to investigate the use of clustering approaches for outlier detection in NRI. The performance of different hierarchical clustering methods is compared regarding their ability to isolate artificially constructed outliers.

Keywords: hierarchical agglomerative clustering, outlier detection, survey data collection, National Resources Inventory.

1 Introduction

Cluster analysis is a popular statistical unsupervised learning tool. We partition the data hierarchically or into a specified number of subgroups to optimize a certain objective function. It can be used to partition a population into homogenous subpopulations or classify some objects into similar groups. In this paper, we will examine the performance of different cluster analysis tools in forming valid clusters and separating suspicious points.

The focus of this paper is primarily on hierarchical agglomerative clustering and its applications. In section 2, we will give a description of the National Resources Inventory project, and in section 3, we will talk about defining distance measures, different clustering criteria, and choosing the number of clusters. Then in Section 4, we will explore the National Resources Inventory data, and present our data analysis results. Conclusions will be given in the last section.

2 National Resources Inventory

The National Resources Inventory (NRI) is a long-term survey of natural resources information on nonfederal land in the US, which covers over 75 percent of its total land area. It is conducted by the U.S. Department of Agriculture (USDA) Natural Resources Conservation Services (NRCS), in corporation with the Center for Survey Statistics and Methodology (CSSM) at Iowa State University.

A detailed description of the NRI sample design can be found in Nusser and Goebel (1997). The 1997 NRI survey was based on approximately 300,000 primary sampling units (PSU) and about 800,000 sampling points. The survey was conducted every 5 years before 1997, and annually through a partially overlapping subsampling design since 2000. The most common PSU is a 0.5 mile × 0.5 mile square and in the second stage, we usually take 3 sample points within each PSU. There are some deviations from this standard procedure in practice. The basic sampling design of NRI is a longitudinal stratified two-stage area sample. Variables are collected at point level and PSU-level. The variables for each sampling point include general information of the point (like state, PSU, hydrological unit), indicator variables whether it was sampled in a specific year, broad use and land use, variables concerning USLE (Universal Soil Loss equation), variables concerning WEQ (Wind Erosion equation), conservation practice indicators, and so forth. Not all variables are applicable to each point.

As the NRI data come from a complicated large survey with many variables, it is quite possible to have some outliers resulting from data collection and processing. There can be outliers that have extreme values on a univariate variable, which is not very common because most of them are found in the data editing step. There are also outliers that each univariate value is reasonable, but the combination is rare or not reasonable at all. We are going to come up with some procedure to identify those suspicious points for further examination.

3 Methodology

Hierarchical agglomerative clustering starts with N clusters and merges clusters sequentially until one cluster is left. A dendrogram of nested sequences of clusters is then produced which graphically shows the association between points and how close two clusters are.

3.1 Distance-based hierarchical clustering

The first step in hierarchical agglomerative clustering is to define pairwise distances (dissimilarities) among data points. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ be a point in p-dim space, we define variable-wise dissimilarity first and then combine the variable-wise dissimilarities using a weighted average to form an overall dissimilarity measure.

At each step of iteration, decisions will be made on which two clusters shall be merged based on a certain criterion and this will be carried out iteratively until all points have been clustered. The clustering criterion will determine what the clusters look like given the distance matrix. Commonly used clustering criteria include single linkage (minimum distance between two clusters), complete linkage (maximum distance between two clusters), average linkage (average distance) and Ward's sum of squares. Single linkage method tends to pick up long stringlike clusters and complete linkage is sensitive to outliers, so in our analysis only Ward's method and average linkage (AL) are used.

Average linkage defines the distance between two clusters as the average distance between points in the two clusters. It is not as extreme as single linkage or complete linkage and tends to form clusters of equal variance. Kamvar et al. (2002) showed that the probabilistic model is similar to an equal-variance configuration if squared distances are used. It is not invariant to monotone increasing transformations $h(\cdot)$ if the transformation function h is not linear. However, single linkage and complete linkage are both invariant to monotone increasing transformations, which has been used by some people as an argument in favor of single or complete linkage methods. The average linkage criterion is defined as:

$$d_{AL}\left(U,V\right)=\frac{1}{|U||V|}\sum_{\substack{i\in U\\ j\in V}}D\left(\mathbf{x}_i,\mathbf{x}_j\right), \qquad (1)$$

where U,V stands for two clusters and $D\left(\mathbf{x}_i,\mathbf{x}_j\right)$ represents the distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$.

Ward's method (Ward, 1963) uses the sum of squared errors as the clustering criterion and its statistical interpretation is straightforward. The distance is defined as the increase in error sum of squares after combining two clusters, and it corresponds to a multivariate normal model with a spherical variance-covariance matrix.

$$d_{WD}\left(U,V\right)=SSE\left(U\cap V\right)-SSE\left(U\right)-SSE\left(V\right), \quad (2)$$

where $SSE\left(U\right)=\sum_{\mathbf{x}_i\in U}\left(\mathbf{x}_i-m_U\right)'\left(\mathbf{x}_i-m_U\right)$ and m_U is the sample mean vector of the cluster. $SSE\left(U\cap V\right)$ and $SSE\left(V\right)$ are defined similarly.

3.2 Choose the number of clusters

In applications where we do not have a pre-specified number of clusters but are interested in the optimal clustering structure of the data, we face the task of determining the correct number of clusters. Numerous procedures have been proposed for determining the number of clusters (Jain and Dubes 1988), and in Milligan and Cooper (1985), 30 procedures were compared in a Monte Carlo simulation. Their artificial datasets contained either 2, 3, 4, or 5 distinct nonoverlapping clusters, and the performances of stopping rules (criteria for deciding the number of clusters) were compared so as to identify the best performing criteria. The two best criteria in Milligan and Cooper's study are the Calinski and Harabasz index (Calinski and Harabasz 1974), the Duda and Hart criterion (Duda and Hart 1973). It should be noted that their sample size was really small, only 50 points in each simulation, so some methods that would do well in large samples did a poor job in their study, like the Likelihood Ratio. We will describe two of the top performers briefly and demonstrate how they perform in the NRI data.

The Calinski and Harabasz index (referred to as CH value) starts from defining within-group and between-group deviation matrices. The total deviation matrix is defined as

$$\mathbf{T}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(\mathbf{x}_{gi}-\bar{\mathbf{x}}\right)\left(\ \mathbf{x}_{gi}-\bar{\mathbf{x}}\right)',$$

where G is the total number of clusters, and the size of cluster g is n_g, $g=1,2,...,G$. $\bar{\mathbf{x}}$ is the overall mean vector.

The within group deviation matrix is defined as

$$\mathbf{W}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(\mathbf{x}_{gi}-\bar{\mathbf{x}}_g\right)\left(\ \mathbf{x}_{gi}-\bar{\mathbf{x}}_g\right)',$$

where $\bar{\mathbf{x}}_g$ is the mean vector of cluster g.

The between group deviation matrix is defined as,

$$\mathbf{B}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}n_g\left(\bar{\mathbf{x}}_g-\bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}_g-\bar{\mathbf{x}}\right)'.$$

The multivariate decomposition is similar to the univariate sum of squares decompostion,

$$\mathbf{T}=\mathbf{B}+\mathbf{W}.$$

Given the above deviation matrices, the Calinski and Harabasz (1974) index is defined as

$$C\left(G\right)=\frac{\mathrm{tr}(\mathbf{B})}{G-1}\Big/\frac{\mathrm{tr}(\mathbf{W})}{n-G}. \qquad (3)$$

It actually involves all the clusters in the current step, not just the two clusters about to merge and we need a group membership for all points to actually calculate this criterion. The univariate version of this criterion is the F-statistic in ANOVA.

The second criterion on their ranking list is the Duda and Hart (1973) criterion (referred to as DH criterion), which only involves the two clusters we are merging.

Suppose we have G clusters now and the optimal way to get $G+1$ clusters is to split cluster g of size n_g into clusters g_1 and g_2. Then, we define,

$$\mathbf{W}_g=\sum_{i=1}^{n_g}\left(\mathbf{x}_{gi}-\bar{\mathbf{x}}_g\right)\left(\ \mathbf{x}_{gi}-\bar{\mathbf{x}}_g\right)',$$

and

$$\mathbf{W}_{gj}=\sum_{i=1}^{n_{g1}}\left(\mathbf{x}_{gj,i}-\bar{\mathbf{x}}_{gj}\right)\left(\ \mathbf{x}_{gj,i}-\bar{\mathbf{x}}_{gj}\right)',$$

for $j = 1, 2$.

The $\text{tr}(\mathbf{W}_g)$ is the within-cluster sum of squared distances between objects and centroid of cluster g, and $\text{tr}(\mathbf{W}_{g1})$, $\text{tr}(\mathbf{W}_{g2})$ are the within-cluster sums of squared distances between objects and centroid in clusters g_1 and g_2.

The DH criterion is defined as follows,

$$\begin{aligned} L(g) &= \left\{1 - \frac{\text{tr}(\mathbf{W}_{g1}) + \text{tr}(\mathbf{W}_{g2})}{\text{tr}(\mathbf{W}_g)} - \frac{2}{\pi p}\right\} \\ &\quad \times \left\{\frac{n_g p}{2[1 - 8/(\pi^2 p)]}\right\}^{\frac{1}{2}}. \end{aligned}$$

It is a local criterion which only involves the two clusters being merged at the current step. The null hypothesis that the cluster is homogeneous should be rejected if the calculated value is larger than the critical value from a standard normal distribution. This criterion depends on the number of dimensions, and cluster size.

A more recent method was introduced by Tibshirani et al (2001) for estimating the number of clusters. This method includes defining a proper measure of within-cluster dispersion and comparing it with its expectation under an appropriate reference null distribution. It is computationally intensive, as it requires drawing samples of size n from the reference distribution repeatedly. In this method, the definition of a within-cluster dispersion measure and choice of null distribution are both very important, and we have different choices of the latter. The reference null distribution is usually chosen as unimodal distribution or uniform distribution with convex support.

Their method starts with an initial partitioning of the dataset from any other algorithm, and say we have G clusters labeled $C_1, C_2, ...C_G$ of sizes $n_1, n_2, ...n_G$. We define a measure of within-cluster dispersion as

$$W_G = \sum_{g=1}^{G} \frac{1}{2n_g} \sum_{i \in C_g} \sum_{j \in C_g} D(\mathbf{x}_i, \mathbf{x}_j).$$

The "gap statistic" is defined as the difference between the logarithm of the finite sample expectation of W_G and W_G itself.

$$\text{Gap}(G) = \text{E}_n^* \log W_G - \log W_G.$$

Note that $\log W_G$ and $\text{E}_n^* \log W_G$ are both decreasing when we have more partitions and the clusters become more homogeneous. Here, $\text{E}_n^* \log W_G$ can be obtained by repeatedly drawing a sample of size n from the reference distribution, applying the same clustering rule until we have G clusters and calculating the mean of $\log W_G$ from these samples. Suppose the true number of clusters is G^*. When $G \leq G^*$, we would expect a sharp increase in the gap statistic when G is increasing because $\log W_G$ is decreasing faster than its expectation under the null distribution as a result of splitting loose clusters into several tight clusters; as G increases and $G \geq G^*$, we would expect $\log W_G$ to decrease at a slower rate because the improvement by partitioning a tight cluster will not be that much.

In Tibshirani et al (2001), two reference distributions were presented, both of which are multivariate uniform distributions.

1. Generate reference datasets from a uniform distribution over the range of each variable. This method respects the range of each univariate variable, and is generally simple to implement.

2. Generate reference data uniformly over a hyperbox aligned with each principal component of the data. To be specific, suppose X is $n \times p$ matrix of observations, and assume a singular value decomposition $X - \bar{X} = UDV^T$. Then we let $Z = \left(X - \bar{X}\right) V$, and we uniformly draw samples from the range of Z, and back-transform to get simulations of X. In applications where we have both categorical and continous variables, we can draw replicates of categorical variables from their range and treat continous variables in this way to form a reference dataset.

The second method is also referred to as GapPC (Tibshirani et al 2001), where PC stands for principal component. The use of uniform distribution is under the null hypothesis that the data are sampled from a p-dimensional uniform distribution. An alternative null hypothesis is the unimodalilty hypothesis in which the data are considered to be a random sample from a multivariate normal distribution. The unimodal model usually gives a high probability of rejecting the null hypothesis of $G = 1$ if the data are sampled from a distribution with a lower kurtosis than normal, and methods based on uniformity are generally conservative, leading to fewer rejections of the null hypothesis (Sarle 1983). A reference distribution corresponding to unimodality can be constructed by taking the mean vector and variance-covariance matrix of the orginal data (or some robust measures of centroid and dispersion) and form a multivariate normal distribution.

The whole procedure for estimating the number of clusters is as follows:

1. Cluster the observed data using any of the methods described above, calculate the within cluster dispersion W_G for $G = 1, 2, ..., G_M$, where G_M is the maximum number of clusters that has been pre-specified by the analyst.

2. Generate B reference data sets, using any of the reference distributions above, and cluster each data set in the same way as the observed data. Suppose we get with-in cluster dispersion measures $W_{Gb}^*, b = 1, 2, ..., B; G = 1, 2, ..., G_M$. Compute the gap statistics

$$\text{Gap}(G) = \frac{1}{B} \sum_{b=1}^{B} \log W_{Gb}^* - logW_G$$

3. Let $\bar{W}_G^* = \frac{1}{B}\sum_{b=1}^{B}\log W_{Gb}^*$, compute the standard deviation

$$sd_G = \left[\frac{1}{B}\sum_{b=1}^{B}\left\{\log W_{Gb}^* - \bar{W}_G^*\right\}^2\right]^{\frac{1}{2}}$$

and define $s_G = sd_G\sqrt{1+1/B}$. Finally choose the number of clusters via

$$\hat{G} = \min_G\{G : \text{Gap}(G) \geq \text{Gap}(G+1) - s_{G+1}\} \quad (4)$$

Here, we define $\hat{G}$ in this way to make sure there is a significant drop of "Gap statistics" from having $\hat{G}$ clusters to $\hat{G}+1$ clusters.

The approach above is easy to implement, but can be computationally exhaustive, especially when the sample size n and number of reference sets B are increasing.

4 Data Analysis

4.1 Data description and exploration

Our data set is taken from the 2003 NRI survey conducted in Kansas. We included data from the collection years 1997, 2000 and 2003, obtained from the core sample and those observed in 2002. There are two major categories of points in our data: points that have soil erosion and points that do not. The points that have soil loss are sampled from cultivated cropland, noncultivated cropland, pastureland or land that has enrolled in the Conservation Reserve Program (CRP). The other points are from rangeland, forest land, minor land, urban and built-up land, rural transportation, water areas. Some points switch from cropland to urban land or vice versa, and these will be treated as from the second category as soil erosion is not applicable in certain years. The first category will be referred to as USLE data and the second as NONUSLE data.

Variables	Types	USLE	NONUSLE
bu1997	categorical	×	×
bu2000	categorical	×	×
bu2003	categorical	×	×
lu1997	ordinal	×	×
lu2000	ordinal	×	×
lu2003	ordinal	×	×
cfact2003	continous	×	
pfact2003	continous	×	
slopenlen2003	continous	×	
slope2003	continous	×	
usleloss1997	continous	×	
usleloss2000	continous	×	
usleloss2003	continous	×	

Table 1: Variables in our NRI dataset, their types and whether they exist in USLE/NONUSLE data.

The variables of interest are broad use (bu), land use (lu), C factor (cfact), P factor (pfact), slope length (slopelen), slope (slope), USLE loss (usleloss) in the years 1997, 2000, 2003. For C factor (cfact), P factor (pfact), slope length (slopelen), slope (slope), we only observe in 2003, as they do not change much across years. The original data has 13 dimensions, and the variables as well as their types are listed in Table 1. Bu is a categorical variable with 12 categories, explained in Table 2. Lu is a finer categorization of bu and it is treated as an ordinal variable at the moment. The C factor is a USLE cover and management factor and P factor is a USLE support practice factor. USLELOSS is calculated from Universal Soil Loss Equation. Variables other than bu and lu are treated as continous variables. The USLE data have all 13 variables in Table 1, but generally, NONUSLE data only have broad use (bu) and land use (lu).

Value of Broad Use	Point Type
1	Cultivated cropland
2	Noncultivated cropland
3	Pastureland
4	Rangeland
5	Forest land
6	Minor land
7	Urban and built-up land
8	Rural transportation
9	Small water areas
10	Large water areas
11	Federal land
12	Conservation Reserve Program (CRP)

Table 2: Broad Use Categories.

As is seen from Table 1, NONUSLE points usually do not have all the variable values, we redefine "not applicable" as some reasonable values and work with the complete data.

4.2 Cluster analysis on original NRI data

The distance matrix is defined with regard to the type of each variable. Average linkage, Ward's method and rescaled Ward's method are used to create dendrograms based on the distance matrix.

We use three methods to choose the number of clusters, CH values, DH criterion and GapPC under uniform hypothesis. The maximum number of clusters is set to be $G_M = 50$. Table 3 summarizes the estimated number of clusters .

After a careful look at the three clusters generated by Ward's method, we can tell that the major classification variable is land use, as its variance dominates other variables. So we standardize each variable before using Ward's method, which is referred to as rescaled Ward's method.

	CH	DH	GapPC
Avg. Linkage	2	2(8)	4
Ward's	3	2(10)	8
Rescaled Ward's	2(5)	2(5, 28)	6

Table 3: Estimated number of clusters from original NRI data, numbers in parentheses are alternative values (see text).

To evaluate the performance of the three clustering methods in forming valid clusters, we take the USLE/NONUSLE classification as the truth, and tabulate the misclassification of each clustering approach at $G = 2$ clusters as shown in Tables 4-6. We can see that under Ward's method, there is a huge misclassification (around 30%). But if we standardize the variables and use a rescaled Ward's method, we can significantly decrease the number of misclassifications. But a close examination into the two clusters generated by rescaled Ward's method tells us that CRP points are merged into the same cluster as NONUSLE points, as broad use is treated as a continuous variable. Average linkage has the fewest misclassifications among all three approaches.

	USLE data	NONUSLE data
cluster 1	2143	22
cluster 2	0	2705

Table 4: Description of the two clusters generated from average linkage method, using CH (also DH) criterion.

	USLE data	NONUSLE data
cluster 1	2143	1488
cluster 2	0	1239

Table 5: Description of the two clusters generated from Ward's method, using DH criterion.

So far, we have 3 or 4 estimated number of clusters for each clustering criterion and it is hard to say which number is optimal. Generally speaking, CH method usually gives small values, which represent the overall structure of data, and DH values reveals more information about the local structure of data. GapPC gives a series of values that satisfy (4), and they correspond to the number of clusters where there is a large departure of within-cluster dispersion from the reference distribution. If we compare the number of misclassifications as shown in Tables 4, 5 and 6, we can see there is a serious misclassification in Ward's method (according to the a priori classification of USLE/NONUSLE points), and rescaling as well as average linkage method has fewer misclassifications. Additionally, the major classification variable is land use in Ward's method, and after standardizing the variables, land use is no longer dominating other variables in classification. Average linkage method takes the type of each variable into consideration, and does well in forming valid clusters.

	USLE data	NONUSLE data
cluster 1	1819	19
cluster 2	324	2708

Table 6: Description of the two clusters generated from rescaled Ward's method, using CH (also DH) criterion.

4.3 Cluster analysis on contaminated NRI data

Now we want to examine our method's ability of isolating outliers. We create 12 artificial outliers, with illegal univariate variables (category 1), mismatch of broad use and land use (category 2), extreme USLELOSS value relative to broad use (category 3) and temporal inconsistency(category 4), with 3 outliers in each category. Outliers from categories 1 and 2 are impossible in a proper dataset, the outliers from category 3 are possible but very unlikely, and outliers from category 4 are more likely to occur but are suspicious and need careful examination.

We add all 12 outliers into NRI data, do hierarchical clustering, choose the number of clusters and see how many of them can be isolated.

We use CH value, DH criterion and GapPC to choose the number of clusters, where the results are summarized in Table 7. From Table 7, we can conclude that average linkage method has done a better job of isolating unusual point from clean data than Ward's method, where the latter treats all variables as continous variables. GapPC gives a sequence of solutions to equation (4), which we only take the smallest one. So we can see from table 7 that we can not identify many outliers from the number of clusters suggested by GapPC.

		CH	DH			GapPC
Avg. Linkage	# of clusters	2	2	10	18	4
	# of outliers	0	0	5	7	0
Ward's	# of clusters	3	2	11	24	8
	# of outliers	0	0	0	0	0
Rescaled Ward's	# of clusters	2	5	2	32	3
	# of outliers	0	0	0	7	0

Table 7: Estimated number of clusters from NRI data with 12 outliers; We have only one estimate for CH criterion and GapPC, respectively; for DH criterion, we have 3 estimates corresponding to the local modes on the plot of DH value; the number of outliers are how many outliers we can isolate at that stage of clustering

Now, we add these artificial outliers into the data one at a time, use average linkage, usual Ward's method and rescaled Ward's method to do the clustering, and we are interested in when each artificial outlier is separated under each distance measure. Here, 'being separated' means being classified as a singleton. An outlier will be separated from other points eventually, but a good clustering method should isolate it at relatively fewer clusters. Table 8 summarizes the results.

We can see from Table 8 that the first two outliers with illegal broad use category are isolated very early using average linkage method, but not separated even with 100 clusters using other mehtods. The third outlier has negative USLELOSS values, but it is not identified even with 100 clusters in any of the three methods. The explanation is that a great many points have a USLELOSS value of or close to 0, and a deviation of 1 is not serious enough for this point to outly from the majority. But we will not worry about this in our project, because negative USLELOSS are dealt with in data editing. Average linkage method is still doing a better job in isolating points with illegal combination of broad use and land use, and when it comes to points with extremely large land use, Ward's method could isolate them in fewer than 100 clusters, but it takes more clusters than the other two methods. The third category of outliers is mismatch of broad use and USLELOSS, we can see that outlier number 8 is isolated earlier than number 7, which is consistent with the fact that it has a more extreme USLELOSS. The last category of outliers are temporally inconsistent, for which average linkage method is still doing a better job than the other two methods, and Ward's method still can not isolate any outlier even with 100 clusters.

I also add all 12 outliers into NRI data and see if they can be isolated in separate clusters containing only the outliers in fewer than 100 clusters, which is obviously not the case. Then I delete outlier number 3 and add other 11 outliers into NRI data because outlier number 3 will not be isolated in less than 100 clusters in any clustering method. Average linkage method can isolate all 11 outliers when we form 24 clusters but the other two methods can not isolate them even with 100 clusters. Overall, average linkage method is doing a much better job than the other two methods in isolating unusual points, and rescaled Ward's method performs better than the usual Ward's method, especially when there is temporal inconsistency or mismatch of broad use and soil erosion. Ward's method, which is seriously affected by univariates whose variances dominate other variables, is inferior to the other two clustering methods.

5 Conclusion

In this project, we have used three clustering methods, average linkage, Ward's method and rescaled Ward's method to form clusters on original NRI data and the validity of the clusters has been assessed. We have also created different kinds of artificial outliers and added them into the NRI data to see if our cluster analysis methods can isolate them or not.

In summary, Ward's method treats all variables as if they were continous variables, which is not proper in our application. Another problem with Ward's method is that its classification variables are those with large variances, and the dissimilarity between points heavily relies on their values.

Rescaling the variables before using Ward's method is helpful in forming easily interpretable clusters and isolating outliers. It avoids the fact that variables with dominating variance will be the primary classification variable.

Average linkage is better in forming reasonable clusters and isolating unusual points than the previous two methods. Pairwise distance among points should be defined in a careful way, taking consideration of relative importance of different variables.

CH value is a global criterion and it tells us about the main structure of the data. However, if we are not interested in the overall structure but some special minor groups, DH criterion gives more guidance in exploring them.

Gap statistic considers the decrease of within cluster dispersion measure when we have more clusters, which is compared with the same measure under reference null distribution. It gives us some insight concerning the local structure of the data and how it 'should' perform if our data has no classification structure. It gives a sequence of numbers when applied to the data, and when we are faced with data that contain unknown outliers, we can use these estimated numbers of clusters and see if suspicious points are isolated into small clusters.

A next step would be to look at the whole NRI dataset and see if our approach can be extended to the whole dataset and how it is going to work. Additionally, theoretical work is needed to show if we can do inference on the population based on the clustering structure of sample data from a complex survey design.

Acknowledgement

This research was supported in part by the USDA Natural Resources Conservation Service cooperative agreement NRCS-68-3A75-4-122.

References

Calinski, T. and J. Harabasz (1974). A dendrite method for cluster analysis. *Commun. Stat 3*, 1–27.

Duda, R. and P. Hart (1973). *Pattern Classification and Scene Analysis.* New York, NY:Wiley.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning.* Springer.

Jain, A. and R. Dubes (1988). *Algorithms for Clustering Data.* Prentice-Hall.

Kamvar, S., D. Klein, and C. Manning (2002). Interpreting and extending classical agglomerative clustering algorithms using a model-based approach. In

Outlier type	ID	Avg Linkage	Ward's	Rescaled Ward's
Illegal univariate variables	1	5	≥ 100	≥ 100
	2	4	≥ 100	≥ 100
	3	≥ 100	≥ 100	≥ 100
Mismatch of bu and lu	4	11	85	55
	5	16	≥ 100	≥ 100
	6	9	84	30
Extreme USLELOSS	7	23	≥ 100	33
	8	11	≥ 100	30
	9	18	≥ 100	≥ 100
Temporal inconsistency	10	3	≥ 100	28
	11	4	≥ 100	57
	12	5	≥ 100	13
	1-12	≥ 100	≥ 100	≥ 100
	all but 3	24	≥ 100	≥ 100

Table 8: The number of clusters that are needed to isolate each artificial outlier.

Proceedings of the International Conference on Machine Learning (ICML), pp. 283–290.

Milligan, G. W. and M. C. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50*, 159–179.

Nusser, S. M. and Goebel, J. J. (1997) The National Resources Inventory: A long-term multi-resource monitoring programme *Environmental and Ecological Statistics 4*, 181-204

Sarle, W. (1983). Cubic clustering criterion. Technical report, A-108, SAS Institute, Inc.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 63*(2), 411–423.

Ward, Joe H., J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*, 236–244.