Title
Exploratory Clusters of Student Technology Participation with Multivariate Regression Trees

Author
Skipper, Peter Tyrel

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Exploratory Clusters of Student Technology Participation with Multivariate Regression Trees

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science in Statistics

by

Peter Tyrel Skipper

2014

ABSTRACT OF THE THESIS

Exploratory Clusters of Student Technology Participation with Multivariate Regression Trees

by

Peter Tyrel Skipper

Master of Science in Statistics

University of California, Los Angeles, 2014

Professor Frederic R. Paik-Schoenberg, Chair

Classroom practices in regards to technology use may have a significant impact (positive or negative) on the effectiveness of a curriculum. This paper looks at temporal frequency of technology use in the context of a high school statistics curriculum, and generates exploratory clusters of that usage with multivariate regression trees. It examines both Euclidean distance and two versions of Kullback-Leibler divergence, ultimately discovering that Euclidean clusters are more robust to outliers and have lower cross-validated error.

The thesis of Peter Tyrel Skipper is approved.

_____

Robert L. Gould

_____

Hongquan Xu

_____

Nicolas Christou

_____

Frederic R. Paik-Schoenberg, Committee Chair

University of California, Los Angeles

2014

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 MOTIVATION

Mobilize (MZ) is a collaboration between Los Angeles Unified School District (LAUSD) and the University of California at Los Angeles (UCLA). MZ designs data-science curriculum at the secondary level. That curriculum involves participatory sensing campaigns, in which students collect data about themselves via smartphones on a variety of topics (e.g. snacking habits, trash and recycling in their neighborhood, media placement in their neighborhood). The provided data can be measured repeatedly and delved into on more than a surface level (think "What type of food did I eat today, and how healthy was it?" as opposed to "What is my favorite color?"). The data is shared with classmates and analyzed using the open-source R statistical software package. The curriculum thus provides a pragmatic framework to develop skills in algebra, biology, computer science and statistics, one that is grounded in issues affecting the students' own communities.

Of course, all of this requires that students actually collect the data. Many of the statistical and computational aspects of the curriculum are new to teachers, and as such the course creators are not certain that teachers are strictly adhering to those aspects of the program. As a result, how students actually collect the data is of interest. Particular classroom processes may have a positive (or negative) effect on instruction, and this information would be useful for future implementations of the curriculum.

2 DATA

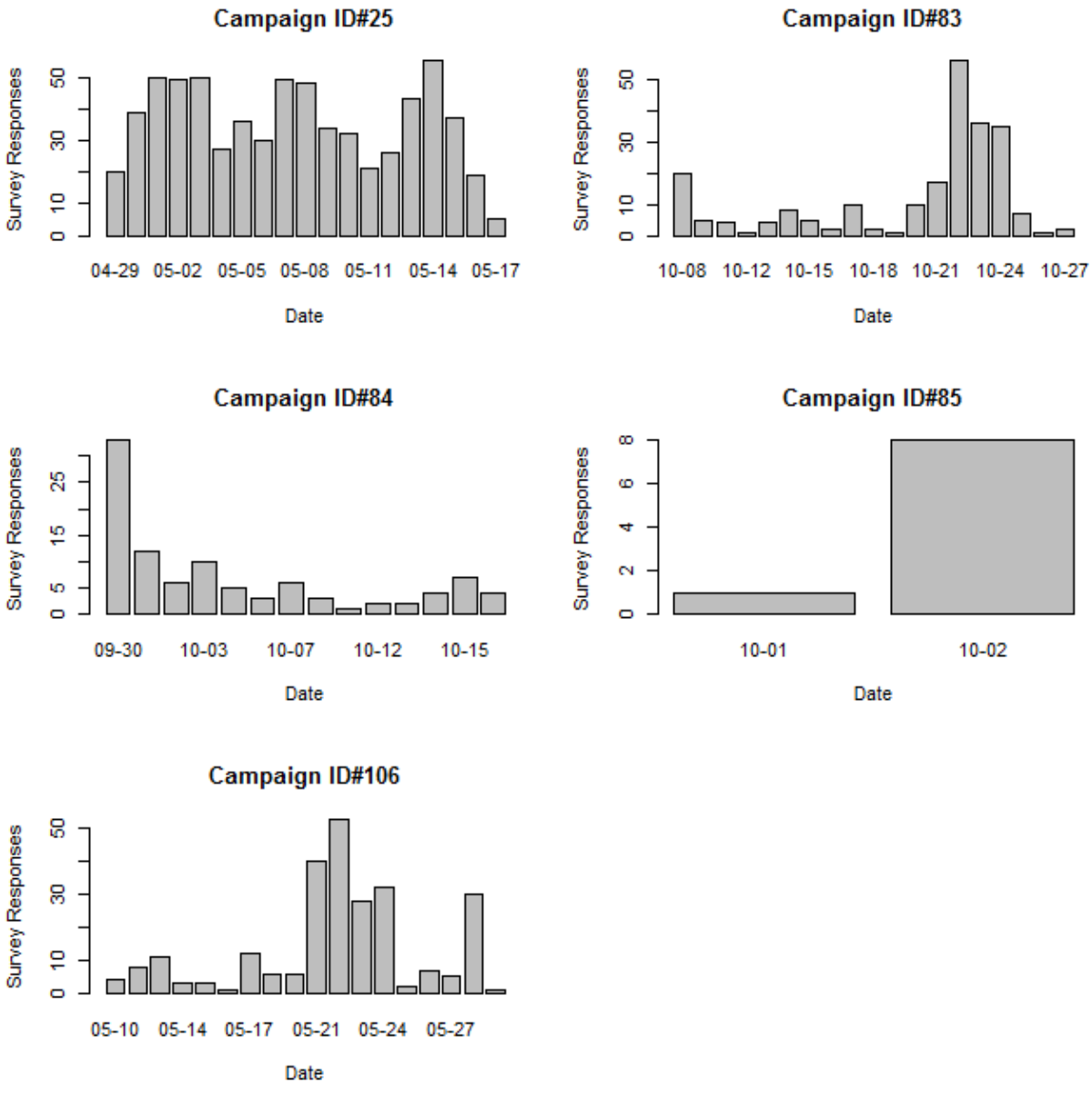
Table 1 summarizes the volume of survey responses by general subject area. A "response" occurs when a student contributes data to the shared dataset, via a smartphone (or on a computer).

Table 1: Number of Campaigns and Average Responses per Campaign

Subject	Total Campaigns	Average Responses per Campaign
Exploring Computer Science (ECS)	15	364.06
Algebra 1	35	116.51
Biology	4	77.0

Student responses are collected in a campaign. One campaign represents a single teacher's implementation of the provided curriculum. For example, a biology campaign might ask students to collect data about the types of trash they throw away, answering survey questions on the phones about frequency, size, type, etc. Fifty-four campaigns are considered here, executed from spring 2012 to spring 2013. Twenty-one teachers participated in the examined campaigns, most of whom executed a couple of them. Campaigns in the dataset range in length from a single day to almost two months. The average campaign lasts 17.5 days. Student responses are distributed across those days in myriad ways. Figure 1 shows a sample of five of the campaigns. Some of them, like number 25, appear to have a cyclical pattern. Others, such as number 84, have significant interest initially, but responses peter out as the campaign continues. Still others, like number 85, are over almost as quickly as they begin. Of course, these anecdotal comparisons may or may not hold in a mathematical sense. In the following section, methods to determine rigorous clusters will be discussed.

Figure 1: Bar Plots of Five of the Campaigns



3 METHODS

3.1 MULTIVARIATE REGRESSION TREES

Classification trees attempt to predict a categorical response via a decision tree. A decision tree is a graphical technique, recursively dividing the dependent variable into a number of groups (or classifications); the impact of the explanatory variables is described at each split of the tree. Regression trees apply a similar procedure when the response variable is numeric. Breiman et al (1984) combined the two techniques into Classification and Regression Trees, or CART.

De'ath (2002) suggested an extension to CART with a ***multivariate*** response, *Multivariate Regression Trees*. MRT begins with a parent node containing all of the observed data. The first explanatory variable is examined, and all possible splits of the observations along that single variable are considered. MRT inspects the within-group sums of squares on the response, looking for the split that minimizes SS_{group1} and SS_{group2} (or alternatively, maximizes the **between-group** sums of square). The process is then repeated for the other explanatory variables. Whichever split results in the absolute minimum sum of squares is chosen, resulting in two child node clusters. Splits are recursively examined for each of the child nodes individually, and the single best split (from ALL the nodes) is chosen for the next branching of the tree.

One important measure of a tree's "fit" to the data is *relative error* (RE). RE is defined as the ratio of the sum of the within-group SS from each node, to the global SS of the complete data. Thus, RE is a measure of the variation in the data still unexplained by the tree. While useful, an obvious problem arises from sole dependence on this metric. The splits could logically continue until each data point occupies its own cluster. This would minimize the within-

group sums of square, but it would hardly be informative for whatever purpose the researcher had in mind. To avoid this trivial solution, De'ath proposes "pruning" the tree (i.e. removing branches) by resampling and cross-validating. Data are randomly assigned into a pre-defined number of groups, all roughly equal in size. Each of the groups is excluded from the analysis in turn, and a tree is constructed from the remaining data (called the "training" data). The observations from the excluded group (called the "test" group) are individually assigned to the terminal nodes in the tree (the "leaves") following the splitting rules previously derived from the training data. A distance is computed between the centroid of that leaf and the test observation. This distance can be summed over all test groups via the following formula for cross-validated relative error (CVRE):

$$CVRE = \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} \left(y_{ij(k)} - \hat{y}_{j(k)}\right)^2}{\sum_{i=1}^{n} \sum_{j=1}^{p} \left(y_{ij} - \bar{y}_j\right)^2}$$

In the above formula, $y_{ij(k)}$ is a single observation from test set *k,* $\hat{y}_{j(k)}$ is the predicted value of variable *j* for that observation (the centroid of its predicted leaf), and $\bar{y}_j$ is the overall mean for variable *j*. This is now a ratio of the errors summed across all *K* test groups compared to the total variance of the data. Perfect predictor(s) would thus have a CVRE of 0, with values closer to 1 representing a poorer set of independent variables. It is possible for the ratio to extend above 1, if the splitting rules in the training group force the test group into clusters which are not the nearest, or to which they are very distant from the centroids.

Appropriate tree size can be derived from calculating CVRE for successive numbers of total leaves (1, 2, 3, etc.). CVRE should decrease at first for most datasets. At some point, however, a larger tree will result in a flattened or perhaps even increased CVRE, as progressively more complicated splitting rules force test observations into groups that they are

(in an overall sense, across all parameters in an individual y observation) more distant from the centroid.

CVRE is subject to some variability. For each partitioning of the tree, data are randomly assigned to the *K* test groups, resulting in slightly different error estimates. For small numbers of *K*, it is feasible to create estimates of the standard error of CVRE. Express the divisions of the regression tree as a function of θ, where θ is the size of the tree (i.e. the number of nodes) and $f_\theta(y_i)$ returns the predicted value (i.e. the centroid of the corresponding leaf) for observation y_i in that particular tree. The particular estimate will depend on which test group is excluded from the training set, so the exclusion of test set *k* from the tree algorithm is expressed as f_θ^{-k}, and the error from test set *k* as $e_k(\theta) = \sum_{i \epsilon k}(y_i - f_\theta^{-k}(y_i))^2$. Then the CVRE for test set *k* of tree size θ is $CVRE_k(\theta) = \frac{e_k(\theta)}{\sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij}-\bar{y}_j)^2}$ and a sample standard deviation of the set $CVRE_1(\theta), \ldots CVRE_K(\theta)$ is $\mathrm{SD}(\theta) = \sqrt{var(CVRE_1, \ldots CVRE_K)}$. An estimate of the standard error of CVRE for tree size θ is thus $\mathrm{SE}(\theta) = SD(\theta)/\sqrt{K}$.

The simplest choice the researcher can make is to choose tree size θ that minimizes CVRE. However, due to the variability of this statistic, individual iterations of tree formation may result in different recommended tree sizes. Borcard et al suggest repetition of the procedure 100-500 times, and selecting the most frequent recommended tree size (Borcard, 2011). Breiman et al recommend a further adjustment, choosing the smallest tree within one standard error of the minimum (Breiman, 1984). This has the advantage of retaining a tree with small CVRE while further reducing the likelihood of over-fitting to the training data. The method has been popularized in many studies, and is often referred to colloquially as the "1se" rule. Lane takes issue with its application to functional response data, and asserts that using 1se results in over-pruning when the response is a probability distribution (Lane, 2012). Because the data vectors are similar in structure to a pdf (see *Feature Selection* below), both standards

(minimum and 1se) are applied in order to compare their predictive powers (see the *Results* section below).

3.2 JACCARD COEFFICIENT

Consider two possible partitions (P_1 and P_2) of a dataset X into clusters. When all pairs of points in X are examined, those points must fall into one of four categories:

- The two points belong to the same group in P_1 and they belong to the same group in P_2. Designate the total of all these pairs ***a***.
- The two points belong to the same group in P_1, but they belong to different groups in P_2. Designate the total of all these pairs ***b***.
- The two points belong to different groups in P_1, but they belong to the same group in P_2. Designate the total of all these pairs ***c***.
- The two points belong to different groups in P_1 and they belong to different groups in P_2. Designate the total of all these pairs ***d***.

If X has N total observations, then the total number of pairs $T = N(N-1)/2 = a + b + c + d$. Paul Jaccard proposed an index for the similarity of two partitions $J = a / (a + b + c)$ (Halkidi, 2001). Jaccard's similarity coefficient ranges [0, 1]. It holds the significant advantage that it ignores pairs that are dissimilar in both groups, rather than artificially inflating the similarity of two clustering schemes because they both put pairs of points into different groups. Indeed, in any dataset with moderately large numbers of clusters the outcome in which a particular pair of points is split into different groups is likely to be quite common.

4 CLUSTERING STRUCTURE

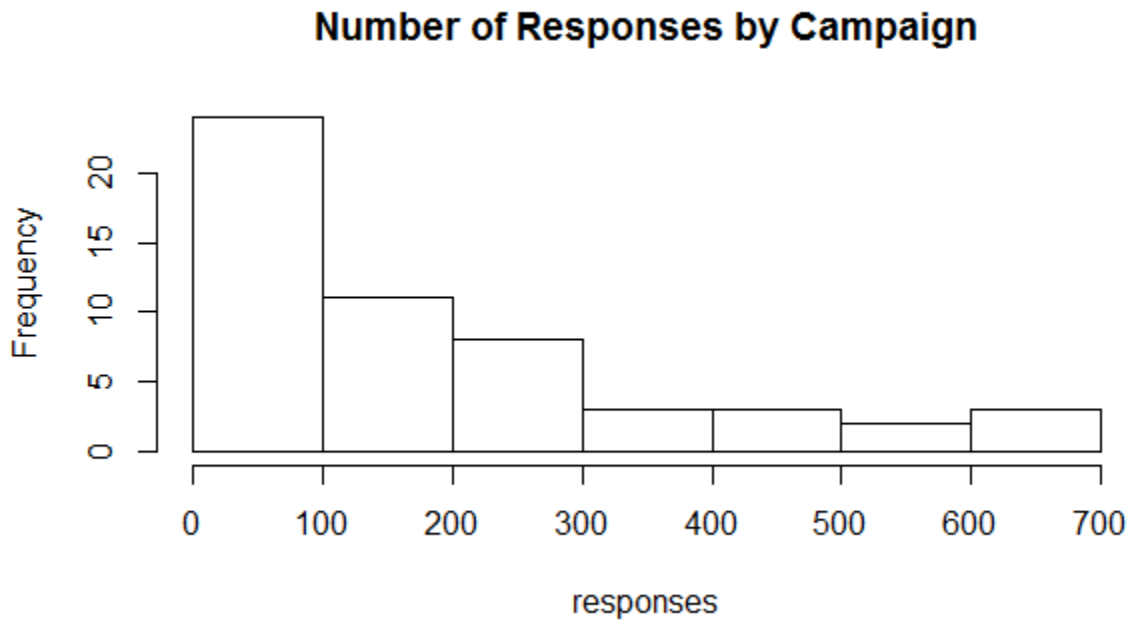
4.1 FEATURE SELECTION

The first question to answer when measuring the similarity between two data points is, *according to what*? In other words, what information will be collected about the data in order to frame comparisons? Rather than type of interaction, this study is interested specifically with the temporal frequency of technology use. How often the students are interacting with the smartphones in the context of the curriculum needs to be quantified.

The longest campaign in the dataset lasted for fifty-nine days. Campaigns are defined on a vector from day 1 to day 59, with the total number of responses on each day corresponding to each individual "bucket." Campaigns that saw no observations on a particular day simply score a zero for that bucket. Seven of the observed variables (days 41, 47, 53, 54, 55, 56, and 58) were perfectly collinear through all 54 observations, and were thus removed for clustering purposes.

Figure 2 shows a histogram of the number of responses by campaign. The distribution is clearly right-skewed. The median number of responses is 121.5 and the inter-quartile range is 214.25. To prevent skewing of the results from a handful of campaigns with a great number of responses, standardization is necessary. Thus, the campaign vectors are divided by the total number of responses for each campaign to create a proportion of technology use on each of the days in the range [0, 1].

Figure 2: Histogram of Total Responses in the 54 Campaigns



4.2 PROXIMITY MEASURE

How to quantify the "closeness" of one data point to another must also be addressed. After standardization, a logical representation is squared Euclidean distance. For observations x_i and x_k, with j features, the squared Euclidean distance is simply:

$$\sum_j \left(x_{ij} - x_{kj}\right)^2$$

The metric has the advantage of straightforwardness, and that it weights each of the fifty-nine days equally. It is also the standard for calculation of the CVRE in R's statistical package, *mvpart*.

The Kullback-Leibler divergence is a measure of the distance between two probability distributions, and can be thought of as the cost when one pdf is used to approximate another.

The formula for the divergence of distribution *p* from distribution *q* is $D(P|Q) \equiv \sum_j p(j)\log(\frac{p(j)}{q(j)})$. Lennert-Cody et al show that Kullback-Leibler divergence can be effectively used to compare binned frequency distributions (Lennert-Cody, 2010). Lennert-Cody defines the average of a set of distributions P_1 to P_n, $\bar{P} = \frac{1}{n}\sum_{i=1}^{n} P_i$ and calculates the divergence of a node *m* as $KLD_m = \sum_{i \in m} D(P_i|\bar{P}_m)$.

One drawback of Kullback-Leibler divergence is that it is not a symmetric measure. Lennert-Cody accommodates for this by constricting the divergence measure to always be an observation's distance from an individual observation to some centroid, as defined above. As a result, distances between individual observations are never calculated. In most clustering applications a more robust, symmetric measure of distance is preferred. Lee uses the Kullback-Leibler *average*, a symmetric distance measure:

$$KLD_{avg} = \frac{1}{2}(D(P|Q) + D(Q|P))$$

Findings suggest that Kullback-Leibler provides more accurate clustering of time-series forecast data than traditional Euclidean metrics (Lee). For this study, as interest pertains only to the *relative* distance between observations, the average is dropped and the Kullback-Leibler distance is defined as $KL_{dist} = D(P|Q) + D(Q|P)$. Using this metric, the technology participation data is clustered and compared to the results of traditional Euclidean measures (see *Results* below).

One additional caveat is necessary. Because Kullback-Leibler divergence uses a ratio of probabilities, $\frac{p(i)}{q(i)}$, the metric explodes whenever the two distributions are not symmetrically zero at *i*. Lennert-Cody accommodates for this by very carefully choosing binning sizes to create non-zero frequencies. This additional constraint is undesirable in the current context, in

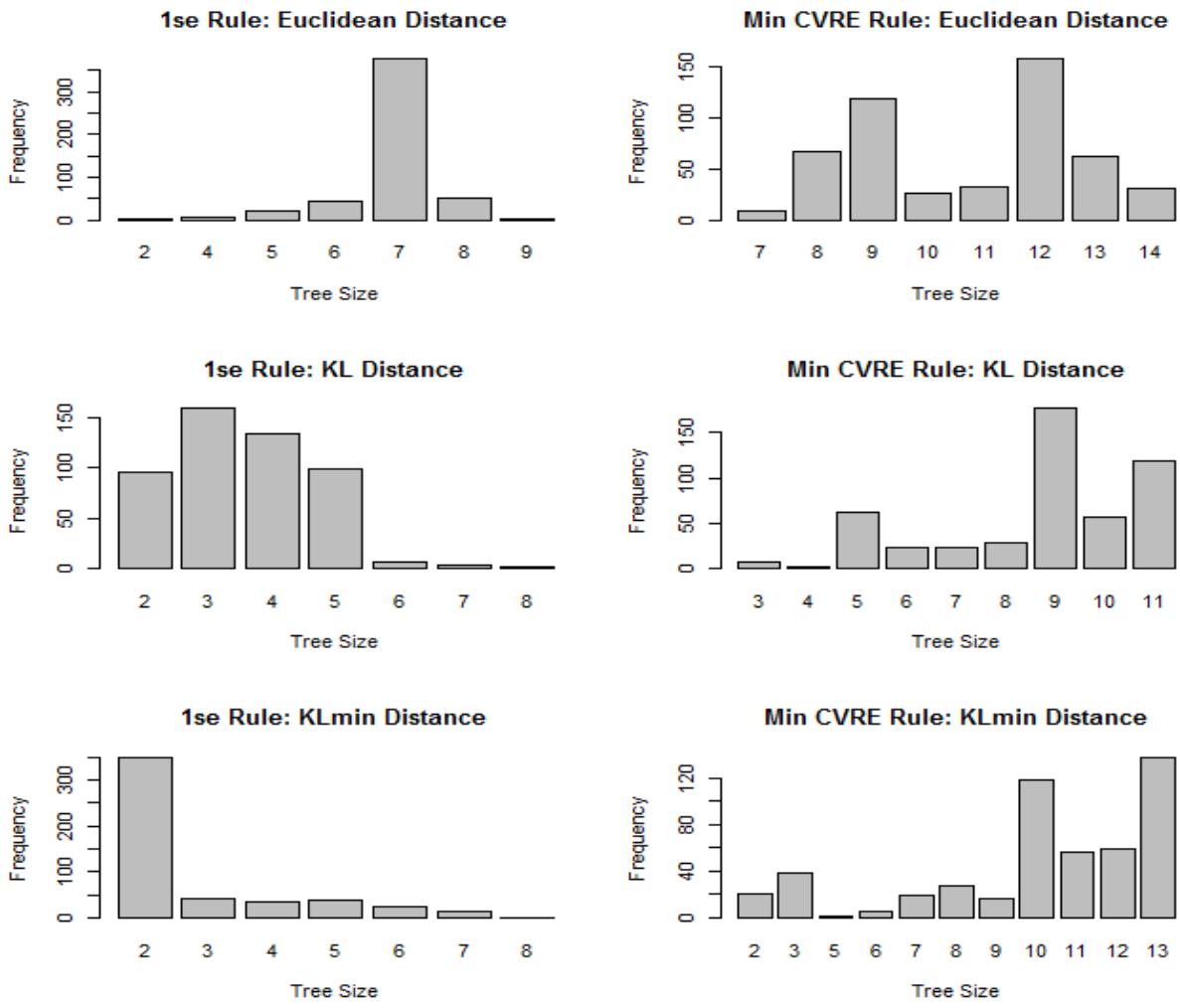
which day-to-day use of the technology is a natural and meaningful descriptor, even though it results in some "zero days." Two remedies are examined. (1) For all $y_{ij} = 0$, simply add one "dummy" response to the *jth* category of observation *i* before normalizing the vector. In campaigns with large numbers of responses (>100), this should have minimal effect on the shape of the distributed responses. (2) For all $y_{ij} = 0$, add 0.0001 to the *jth* category of observation *i* before normalizing the vector. This allows campaigns with fewer observations to avoid being overwhelmed by the dummy responses. This second procedure of negligibly altering the observations will be referred to as KL_{min}.

5 RESULTS

5.1 OPTIMUM TREE SIZE

The default 10-fold cross-validation was used throughout to choose the appropriate multivariate regression tree size. Figure 3 displays bar plots for the chosen tree size according to the different standards for CVRE and distance metrics discussed above, after 500 repetitions each.

Figure 3: Bar Plots for "1se" Rule and Minimum CVRE Rule, by Distance Metric

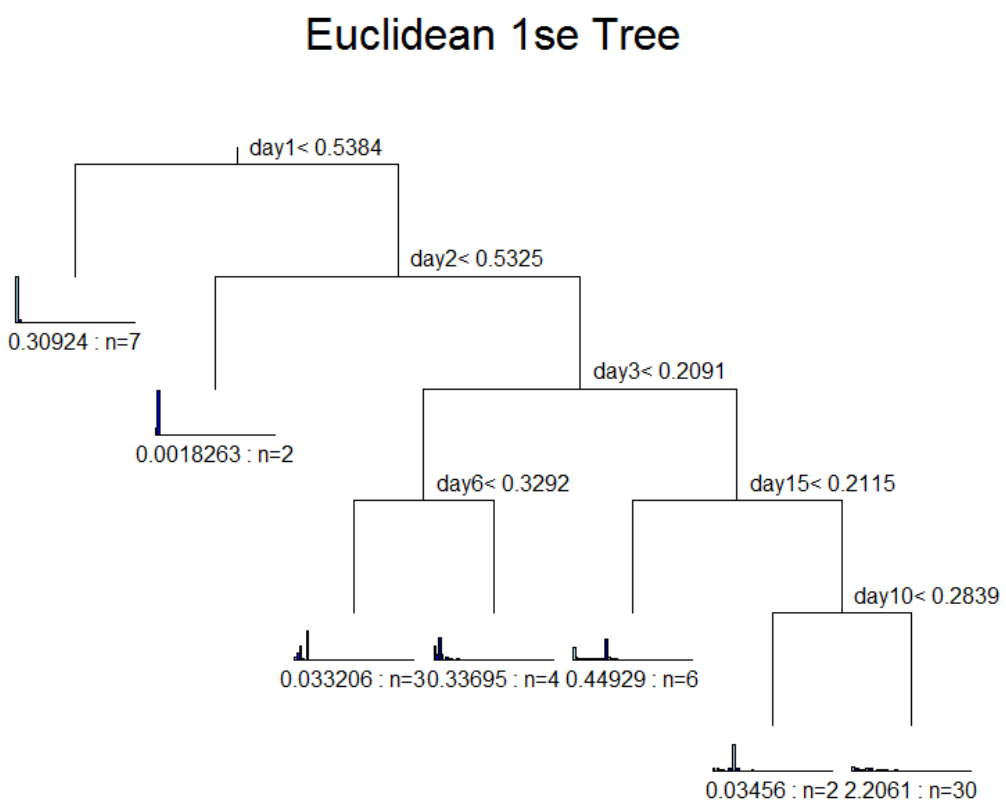


The modes for optimal tree size are twelve and seven (Euclidean), nine and three (KL distance), and thirteen and two (KL_{min} distance). As expected, choosing the minimum CVRE results in larger trees for all of the distance measures. The Kullback-Leibler distances generally exhibit greater variability, which can be an indication of a larger standard error of the CVRE (see *Evaluation of Trees* below). Furthermore, the difference between KL_{min}'s minimum CVRE and 1se tree is very large (thirteen compared to two), indicating that the information gained by adding nodes to that tree is minimal (a lot of complexity must be added to gain a nominal increase in predictive power). In the next section, the trees themselves will be examined in greater detail.

5.2 REGRESSION TREE OUTPUT

Figure 4 shows the Euclidean 1se rule tree.

Figure 4: Euclidean Tree Following 1se Rule



The most important splits happen on early days, with the first partition separating campaigns with more than 53% of their responses on day 1 from the rest of the group. This type of campaign indicates significant early enthusiasm for the technology, followed by a significant decline in participation. Leaves of the tree indicate the number of campaigns in each split, and the within-group sum of squares. Figure 5 shows the fuller Euclidean tree with minimum CVRE. The splits are retained and expounded on, creating smaller clusters on the right side of the tree.

The Kullback-Leibler 1se tree is displayed in Figure 6. The tree is much smaller and (unsurprisingly) the group sizes are much larger. The splits, however, are quite different. The KL tree splits on day 17 (the average campaign length for the 54 campaigns was 17.5 days), creating clusters out of the short and long campaigns first, rather than the initial participation rates. Because the bins are all non-zero (see Proximity Measure above), a wider range of possible splits is available. Figure 7 expands the splits into the minimum CVRE tree for Kullback-Leibler.

Similar to the KL 1se tree, the KL_{min} 1se tree is most noticeable for its parsimony. The tree makes only a single split, as shown in Figure 8. The minimum CVRE tree, by contrast, has the largest number of nodes in the group (thirteen). Some of these splits are on vanishingly small distinctions (e.g. the split on day 13 for campaigns with less than 0.0000008375 of the total responses). The entire KL_{min} tree for the minimum CVRE rule is shown in Figure 9.

Figure 5: Euclidean Tree Following Minimum CVRE Rule

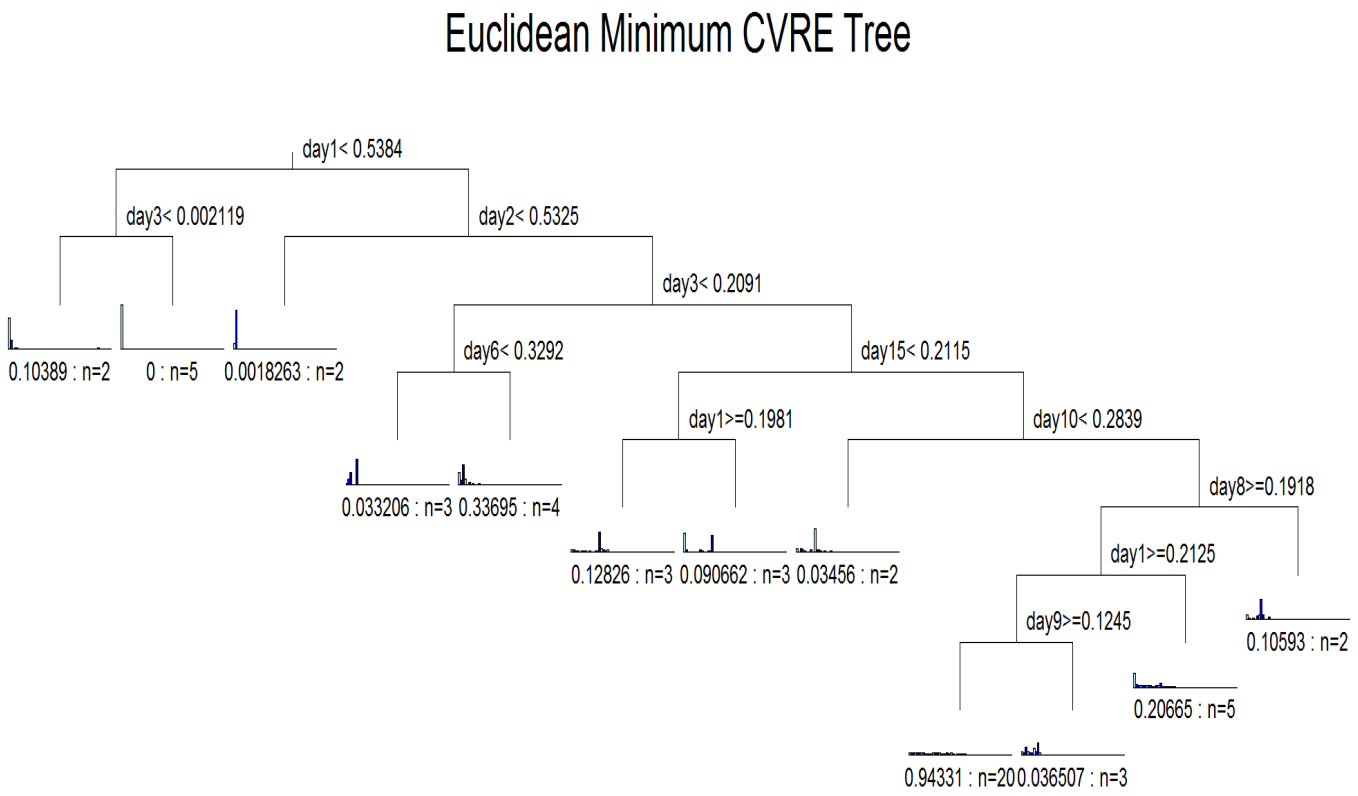


Figure 6: Kullback-Leibler Tree Following 1se Rule

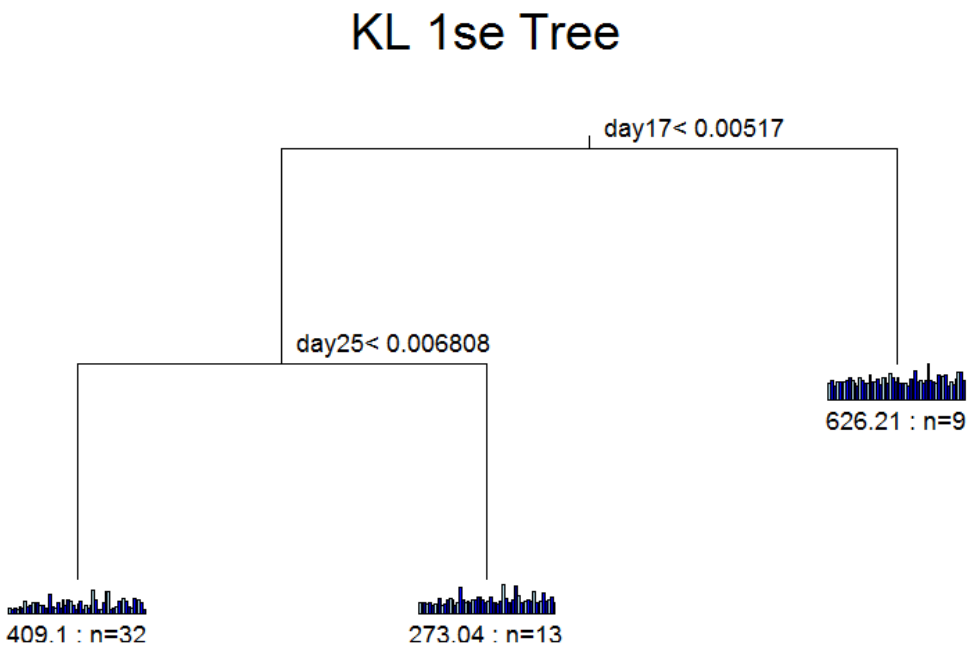


Figure 7: Kullback-Leibler Tree Following Minimum CVRE Rule

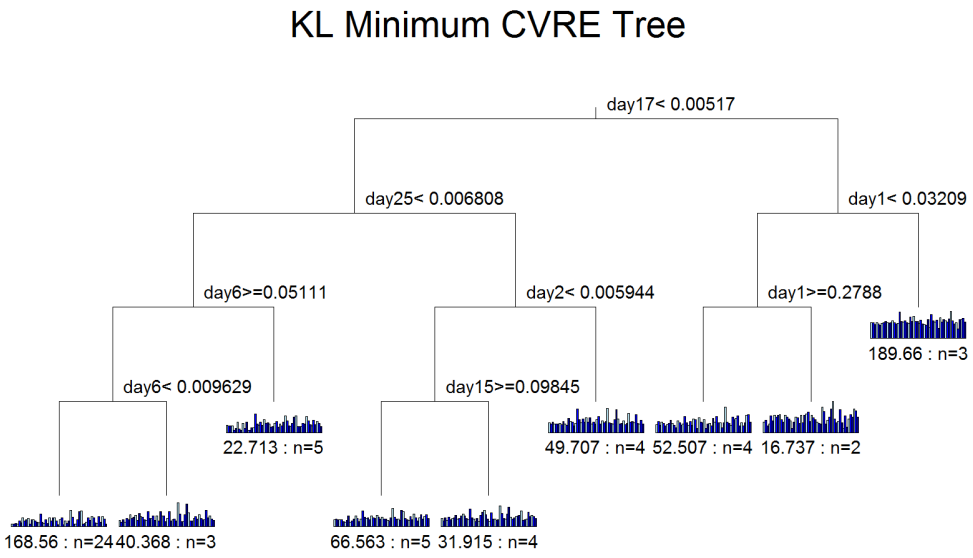
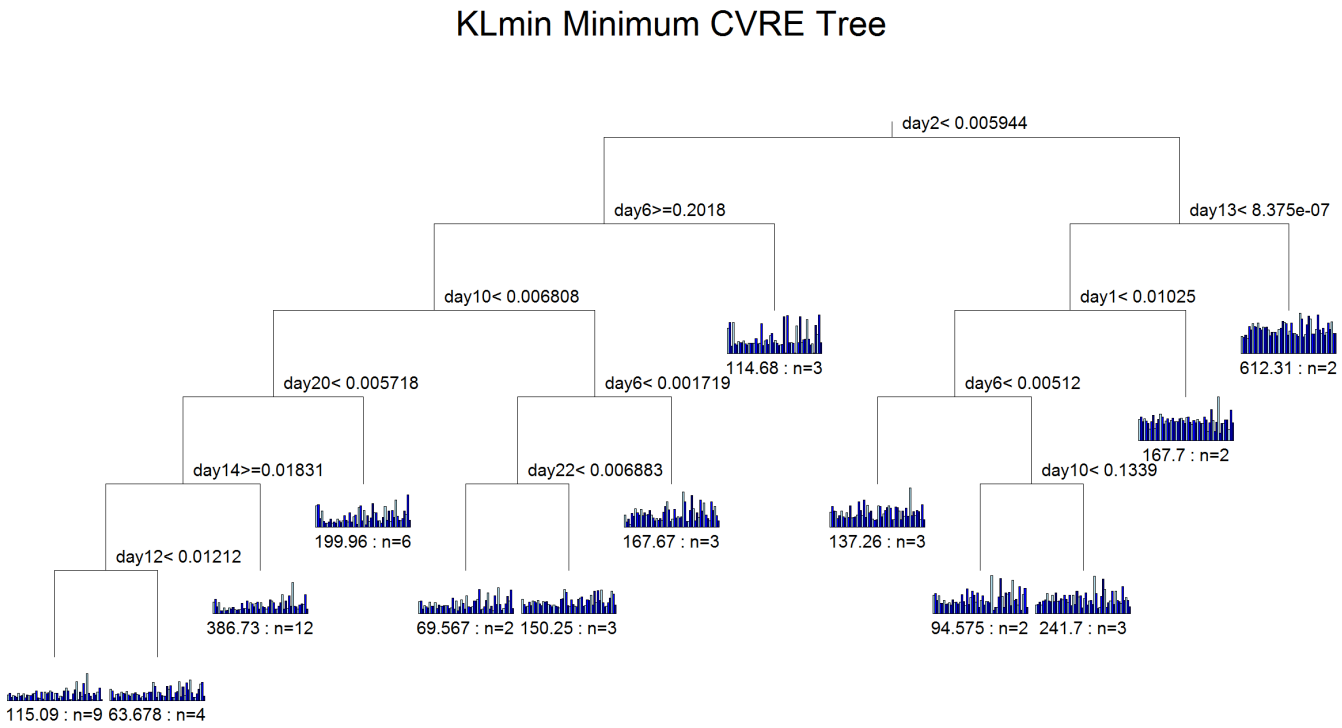


Figure 8: KL_{min} Tree Following 1se Rule

Figure 9: KL_{min} Tree Following Minimum CVRE Rule



5.3 SIMILARITY OF TREES

The six trees are all exploratory attempts to cluster the underlying campaigns, yet their sizes vary dramatically. How similar are the clusters? The Jaccard coefficient (see *Jaccard Coefficient* above) provides an objective measure, scaled from 0 (completely dissimilar) to 1 (identical clustering). Table 2 summarizes comparisons between the trees.

Table 2: Jaccard Coefficients for the Six Multivariate Regression Trees

	Euclid 1se	KL 1se	KLmin 1se	Euclid Min CVRE	KL Min CVRE	KLmin Min CVRE
Euclid 1se	1.00					
KL 1se	0.24	1.00				
KLmin 1se	0.30	0.36	1.00			
Euclid Min CVRE	0.48	0.16	0.15	1.00		
KL Min CVRE	0.13	0.53	0.25	0.11	1.00	
KLmin Min CVRE	0.18	0.15	0.15	0.23	0.18	1.00

Euclidean clusters according to both rules (1se and minimum CVRE) are generally similar at 0.48, as are both Kullback-Leibler distance clusters (0.53). The KL_{min} procedure is highly volatile, and is not even particularly analogous according to the separate rules. The other similarity metrics are mostly quite low, suggesting that the clusters differ significantly according to the different rules and distance metrics.

Figure 10 provides more information. The scatterplots suggest weak to medium correlation between the Euclidean distance and the two versions of Kullback-Leibler. Table 3 confirms this intuition. Euclidean distance is only 0.3 correlated with Kullback-Leibler, and only 0.18 correlated with KL_{min}. The issue is one of compactness. The longest campaign is fifty-nine days, but many of the responses are concentrated over only a handful of days, resulting in a

sizeable number of days with zero responses. In fact, the median campaign has forty-two zero days. This is not an issue for Euclidean distance, but the minor tweaks required for Kullback-Leibler distance (see Proximity Measure above) begin to add up and shift the structure of the underlying campaign. Since the clusters are effectively different, the following section evaluates their individual performances as a predictive measure.

Figure 10: Scatterplots of the Distances between Campaigns via the Three Metrics

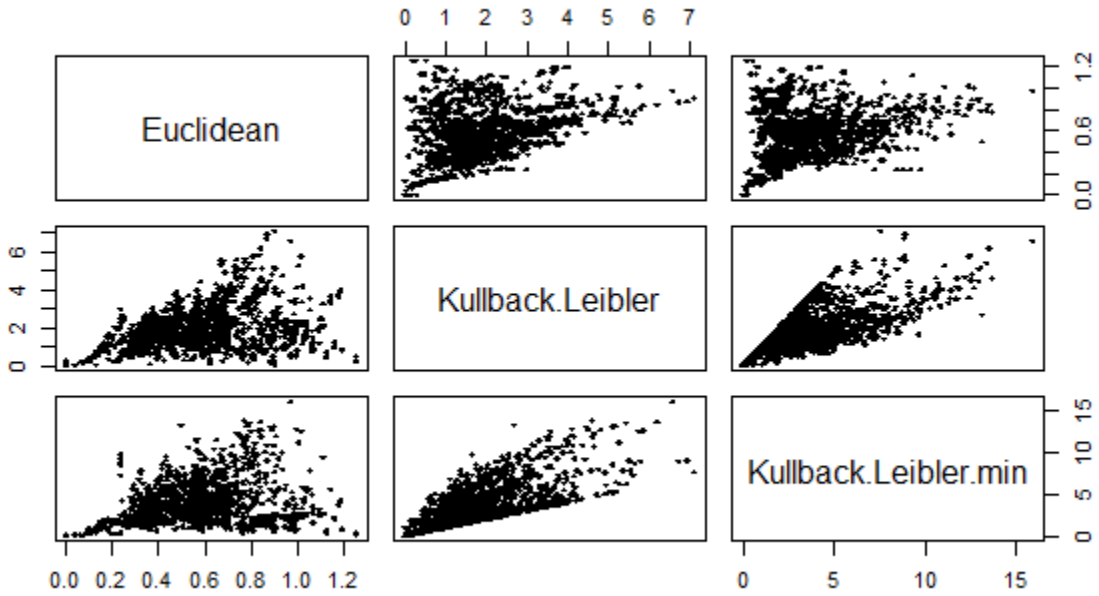


Table 3: Correlations between Distances According to the Three Metrics

	Euclidean	KL	KLmin
Euclidean	1.00		
KL	0.30	1.00	
KLmin	0.18	0.69	1.00

5.4 EVALUATION OF TREES

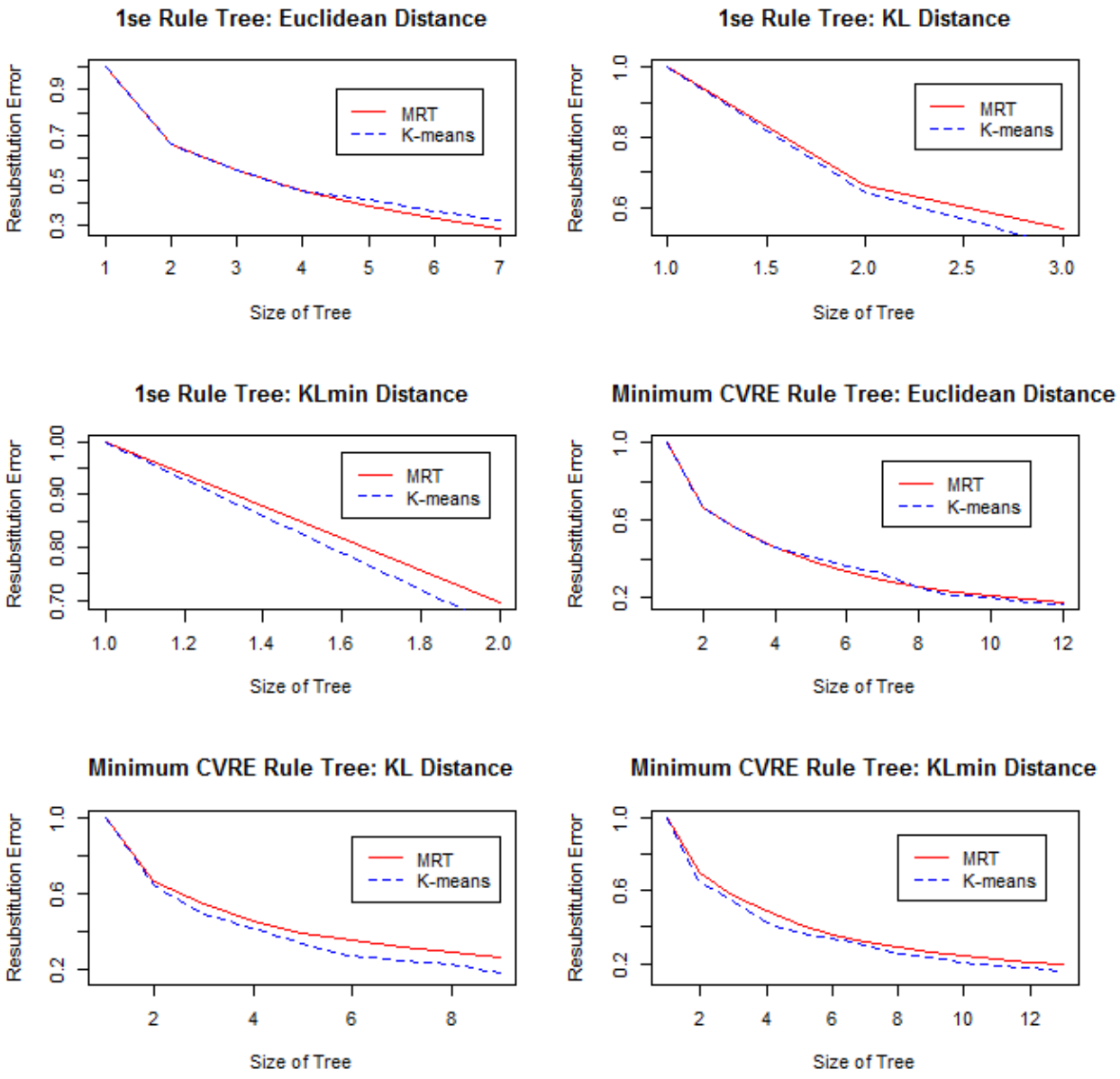
Table 4 summarizes the CVRE and its standard error for the 500 iterations that were used to create each tree. By design, the minimum CVRE trees have lower relative error than the 1se trees. The Euclidean trees clearly outperform the Kullback-Leibler metrics by a significant margin (more than 0.2 less CVRE under both rules). In addition, the CVRE has considerably less fluctuation using Euclidean distance: the standard error of the CVRE is about half the size of the Kullback-Leibler measures. The Euclidean distance is thus less error-prone, and simultaneously the accuracy of that measure is less in doubt. The 1se tree is almost as accurate as the minimum tree (0.522 compared to 0.482), and has the further advantage of being less complex.

Table 4: Average CVRE and SE for the Six Trees after 500 iterations

	1se Rule		Minimum CVRE Rule	
	CVRE	SE	CVRE	SE
Euclidean	0.522	0.068	0.482	0.067
KL	0.781	0.124	0.689	0.107
KL_{min}	0.843	0.138	0.760	0.142

A useful appraisal of the effectiveness of MRT should involve comparison to another technique. Figure 11 shows the re-substitution error for k-means clustering and MRT across all six trees. MRT results in lower error than k-means only in one instance, the 1se rule Euclidean tree.

Figure 11: Comparison of MRT to K-means Clusters



6 DISCUSSION

Lennert-Cody suggests that Kullback-Leibler divergence is an effective clustering technique for multivariate response data. However, the analysis above suggests that the method is not particularly robust, and relies on very precise binning of the data. When natural

divisions exist (e.g. days) that result in zero or near-zero proportions in some bins, the metric loses accuracy. As a result, using the classic Euclidean distance is preferable in those situations.

Lane's assertion of over-pruning when using the 1se rule does not appear to be supported by the data applied here. CVRE between the two Euclidean trees is comparable, and the 1se tree is generally to be preferred for its increased simplicity. Rather than asserting a general rule, it appears that the optimum tree is best chosen on a case-by-case basis, validated with later study. In the present context, the Euclidean 1se tree is more than sufficient to summarize the exploratory clusters.

The Euclidean 1se tree provides some important information about the underlying campaigns. The far right cluster represents intuitively what we might hope for in terms of participation, with a more even spread of responses across most of the days. The students appear to be participating regularly, and it is also the largest cluster! Adjacent nodes to the far right represent a deadline pressure, with many responses concentrated at the end of the campaign. Alternately, the two leftmost nodes represent a sharp decline in interest or participation in the project, with most responses happening in the first couple days. The clusters could be used to determine which teachers to invite back into the MZ program for the following year (based on which teachers were most successful at promoting participation). Conversely, a further analysis of the procedures implemented by teachers in the far right cluster could provide a variety of professional development best practices that new teachers to the program would benefit from.

A caveat does need to be included, however. The Euclidean 1se tree creates two clusters with only two campaigns in each. Such tiny clusters are often indicative of over-fitting. These tiny clusters exist even after pruning back the tree from its minimum CVRE. Though it is

the best of the options explored, care should be taken in testing and applying the Euclidean 1se tree to ensure that it is not unduly influenced by the dataset explored in this paper.

Given more time, this study could be improved via application of the 632+ bootstrap estimator to CVRE (Merler & Furlanello, 1997). Using both real and simulated data, Merler et al show that a weighted estimate of CVRE between the re-substitution error examined above and a bootstrap estimate of the same (drawing sample observations for the training set **with replacement**) has the dual advantages of lower standard error estimates for CVRE and mitigation of the possibility of over-fitting that is not uncommon with regression tree methods. Application of the method would require a significant extension to the code of the existing *mvpart* library in R, though it would be a useful further investigation.

Ultimately, the most important next step is to compare the exploratory clusters created here to academic performance, preferably at the individual level. Do particular patterns of technology use correspond to increased engagement or retention of science and math curriculum? If so, they should be endorsed and publicized to teachers and independently verified in classroom trials. Even small gains in educational achievement could have a major impact on overall student success in some quite challenging subjects.

REFERENCES

Borcard, Daniel, François Gillet, and Pierre Legendre. *Numerical Ecology with R*. New York: Springer, 2011. Print.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984. Print.

De'ath, Glenn. "Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships." *Ecology* 83.4 (2002): 1105-117. Print.

Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. "On Clustering Validation Techniques." *Journal of Intelligent Information Systems* 17.2/3 (2001): 107-45. Print.

Lane, Stephen E. "Topics in Functional Data Analysis." Diss. The U of Melbourne, Australia, 2012. Web. 14 May 2014.

Lee, Taiyeong, Yongqiao Xiao, Xiangxiang Meng, David Duling. "Clustering Time Series Based on Forecast Distributions Using Kullback-Leibler Divergence." Web. 16 May 2014. http://forecasters.org/wp/wp-content/uploads/gravity_forms/7-2a51b93047891f1ec3608bdbd77ca58d/2013/06/ISF2013_LEE_TSClustering.pdf

Lennert-Cody, Cleridy E., Mihoko Minami, Patrick K. Tomlinson, and Mark N. Maunder. "Exploratory Analysis of Spatial–temporal Patterns in Length–frequency Data: An Example of Distributional Regression Trees." *Fisheries Research* 102.3 (2010): 323-26. Web. 01 Nov 2013.

Merler, Stefano, and Cesare Furlanello. "Selection of Tree-Based Classifiers with the Bootstrap 632+ Rule." *Biometrical Journal* 39.3 (1997): 369-82. Print.