Soul of the Community: A First Attempt to Assess Attachment to a Community

Anna Quach* Jürgen Symanzik† Nicole Forsgren Velasquez‡

Abstract
In this article, we work with data from the Soul of the Community survey project that was conducted by the Knight Foundation from 2008 to 2010. Overall, 26 communities across the United States with a total of more than 47,800 participants took part in this study. Each year, around 200 different questions had to be answered by each participant. One key variable is attachment to one's community. In our article, we provide an initial assessment via various Machine Learning Algorithms which factors may have an effect on attachment.

Key Words: Machine Learning; Data Mining; Predictive Analytics; Random Forests; CART; Archetypes; Knight Foundation

1. Introduction

The goal of our analyses is to find factors that foster attachment to one's community. Prior to our analyses, data cleaning in the three data sets[1] provided by the Knight Foundation[2] was conducted, resulting in excluded variables and cases (see below). Following this cleaning, analyses reveal the important factors that impact attachment to particular communities as well as attachment to communities as a whole. We also examine if there are differences in attachment between communities as well as demographics. Using archetypal analysis, we identify similarities among people that are attached or not attached to their community. Finally, we compare our results to previous research. Figure 1 shows a map of the 26 communities involved in the "Soul of the Community" (SOC) survey project[3]. The map reveals that most communities are located in the eastern United States. Participating communities range from cities with more than one million inhabitants (such as Philadelphia, PA) to rural communities with less than 20,000 inhabitants (such as Milledgeville, GA). The number of people surveyed in each community and the sample size following data cleaning can be seen in the dot chart in Figure 2.

Our article is arranged as follows: In Section 2, we describe our data cleaning steps. An overview of machine learning algorithms follows in Section 3. Results for Random Forests, Recursive Partitioning and Regression Trees (RPART), Least Absolute Shrinkage and Selection Operator (LASSO), and an Archetypal Analysis are provided in Sections 4 through 7, respectively. A graphical summary of some of the most interesting variables via dot charts is provided in Section 8. We finish

*Utah State University, Department of Mathematics and Statistics, 3900 Old Main Hill, Logan, UT 84322–3900, USA. Phone: 208 316 2798, E–mail: aquach4@hotmail.com

†Utah State University, Department of Mathematics and Statistics, 3900 Old Main Hill, Logan, UT 84322–3900, USA. Phone: 435 797 0696, Fax: 435 797 1822, E-mail: symanzik@math.usu.edu

‡Utah State University, Department of Management Information Systems, 3900 Old Main Hill, Logan, UT 84322–3900, USA. Phone: 435 797 3479, Fax: 435 797 2351, E-mail: nicolefv@gmail.com

[1] http://streaming.stat.iastate.edu/dataexpo/2013/

[2] http://www.knightfoundation.org/

[3] http://www.soulofthecommunity.org/

with a documentation of all variables of interest and our conclusion in Section 9. The software we used is summarized in Section 10. Appendix A contains two tables with predictor variables and additional variables that summarize all variables and their abbreviations used in this article.

2. Data Cleaning

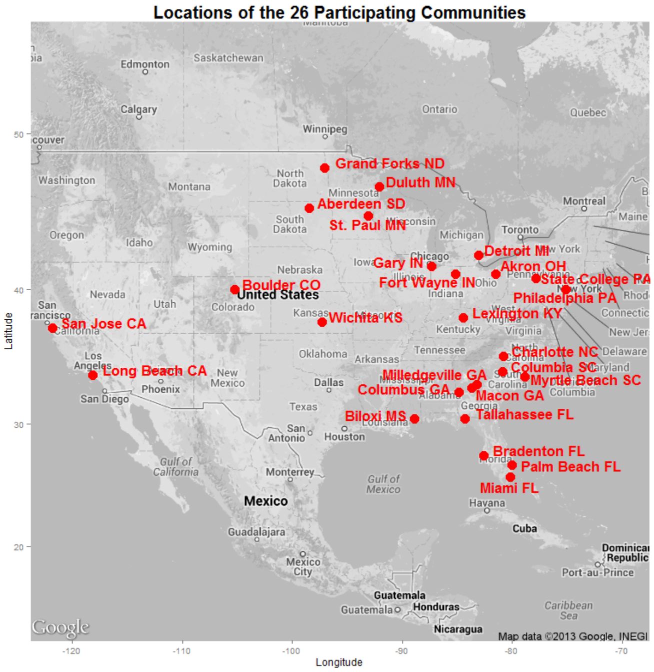


Figure 1: Map of the communities involved in the SOC project.

A number of variables and cases were removed prior to our analyses. Variables were removed for the following reasons: (i) Variables with a large number of missing responses (more than 45%) among the cases were excluded. (ii) When variables were provided as 5–level variables and aggregated 3–level variables (variables names ending with an r), the aggregated 3–level variables were removed as they provided less nuanced information. (iii) Variables not observed in all three years were excluded in the other year(s) for comparison purposes. (iv) All index variables (see Table 1) were removed, assuming that the variables that were aggregated into an index variable would show up together if the index variable is an important predictor variable. (v) Finally, all variables were removed that form the basis for "Community Attachment" (which is one of our main response variables). For (iv) and (v), Principal Component Analysis (PCA) was conducted. Ultimately, 55 variables were retained for analysis from the original 179 (2008), 195 (2009), and 229 (2010) variables, respectively.

After the removal of variables, cases were removed for the following reasons: (i) Cases with at least one missing value in the remaining variables were removed. (ii) Answers such as "don't know", "refuse to answer", or "did not answer the question" in the survey were replaced as missing and then were handled according to (i). Figure 2 shows the effect of data cleaning for the sample sizes in each community in each year. Although steps (i) and (ii) sound rigorous, in most communities/years, only a few cases had to be deleted. Notice that communities with considerable decreases in sample size after data cleaning were mostly urban communities (such as Philadelphia, PA, in 2008 and Charlotte, NC, in 2009). An explanation of why we see these dramatic changes in urban communities following data cleaning would be interesting, but has not been investigated here.

Figure 3 provides a graphical representation of the variables and cases that were

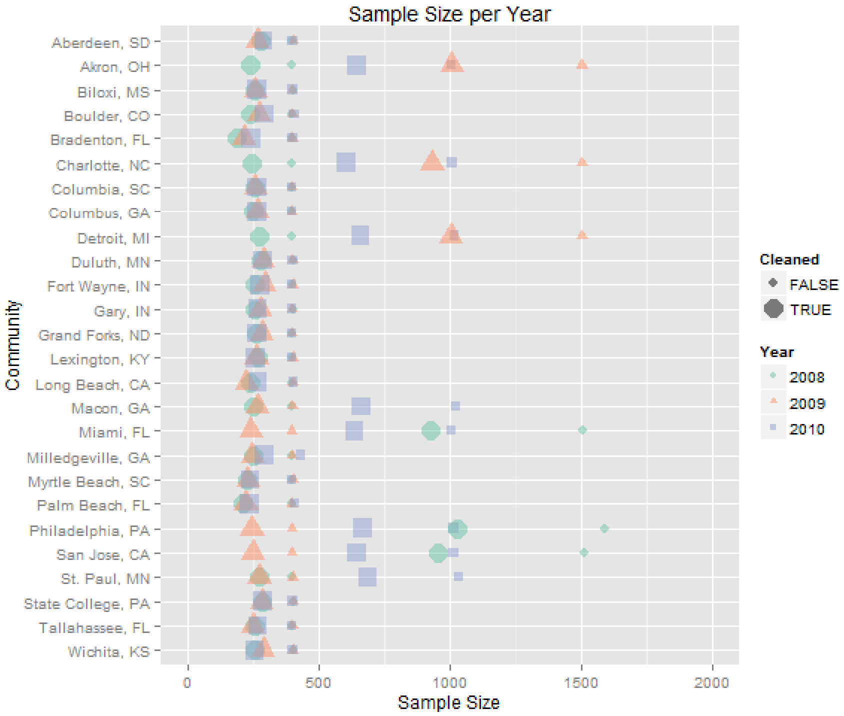


Figure 2: Dot Chart of the sample size in each year before and after data cleaning.

Table 1: Table of the formulation of the 15 index variables found using Principal Component Analysis.

Index Variable	Derivations from Means of Variables
Community Attachment	(Loyalty + Passion)/2
Community Loyalty	(qce1 + qce2 + q6a)/3
Community Passion	(q3a + q3b)/2
Basic Services	(q7cr + q7dr + q7kr)/3
Leadership	(q7lr + q15abr)/2
Education	(q7fr + q7gr)/2
Safety	(q18r + q19r)/2
Aesthetics	(q7ar + q7br)/2
Economy	(q7er + q9r + q10r + q14r + q15r + q15aar)/6
Social Offerings	(q7hr + q7ir + q7mr)/3
Community Offerings	(Basic Services + Aesthetics + Safety + Economy + Social Offerings + Education + Leadership)/7
Civic Involvement	(q22ar + q22br + q22cr + q22dr)/4
Openness	(q8ar + q8br + q8cr + q8dr + q8er + q8fr)/6
Social Capital	(q23r + q24r + q25r + q26r)/4
Community Domains	(Community Offerings + Involvement + Openness + Social Capital)/4

removed from further analyses. Overall, the largest number of cases were removed from the 2010 data set, but the original sample size that year was also approximately 50% larger than the sample sizes in 2008 and 2009.

3. Machine Learning Algorithms

Algorithms that we used to predict the attachment level in each of the years were: LASSO (least absolute shrinkage and selection operator), Random Forests, RPART (Recursive Partitioning And Regression Trees), and Multiple LDA (linear discriminant analysis).

- LASSO is a variable selection method for regression that shrinks some coefficients and sets others to 0. LASSO minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Therefore, LASSO tends to produce some coefficients that are exactly 0 and give interpretable models because of the constraint. LASSO is also numerically stable.

- Random Forests is a powerful statistical classifier that uses bootstrap samples (repeated sampling with replacement from the learning set) and randomness in the tree-building procedure. The Random Forests algorithm begins with growing a forest of many trees to a data set. A tree is grown on each independent bootstrap sample from the training data. At each node, a small number of randomly selected variables is used for binary partitioning and the best split on the selected variables is found. The trees are grown to maximum depth and each tree is used to predict the observations that were not in the bootstrap sample). The predicted class of an observation is calculated by the majority vote of the out-of-bag predictions for that observation.

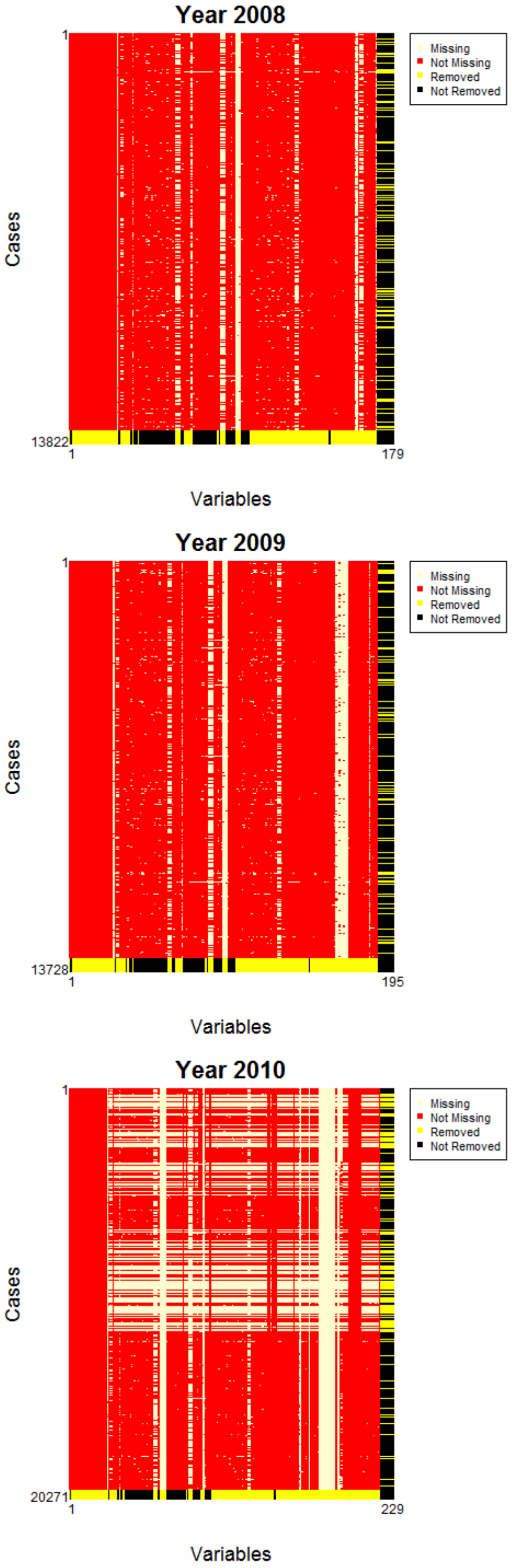


Figure 3: Heatmaps showing missing data. Also shown are cases and variables removed from further analyses (for reasons explained in the text).

- RPART is a decision tree that builds classification or regression models and is widely used for problems where the variables do not meet the usual assumptions. The decision tree is constructed by either splitting or not splitting each node on the tree into two daughter nodes. To do the splitting on continuous variables, the cutoff value is the value that is best in discriminating between the groups in question and partitioning is similar for discrete variables. This procedure is done recursively by applying the same criteria to the subgroups.
- LDA is a statistical technique that finds a linear combination of features to separate or classify two or more classes. LDA is one of the commonly used techniques for data classification and dimensionality reduction. LDA uses information from the independent variables to achieve the clearest possible separation or discrimination between or among groups. LDA maximizes the ratio of the between-class variance to the within-class variance in order to maximize separability.

Each algorithm's misclassification error rate and references are given in Table 2. 10-fold cross validation is used to tune the parameter in order to minimize the misclassification error. The misclassification error rate for each statistical method was found to be approximately equal. We chose to use (i) Random Forests to predict attachment because it has the lowest error rate and (ii) RPART because its error rate is not much worse than Random Forest, its simplicity, and the added interpretation it provides over Random Forests. (iii) LASSO was also used in further analyses because it is less computationally expensive than Random Forests and it is a popular analysis method. In Section 6, the heatmaps allow conclusions similar to those found with these two other algorithms.

Table 2: Table of the machine algorithms used in predicting a categorical response variable.

	Error Rate			
Algorithm	2008	2009	2010	References
LASSO	Doesn't do classification; however, the classes can be treated as a continuous variable where LASSO does well.			(Tibshirani, 1996)
Random Forests	0.31	0.30	0.30	(Breiman, 2001)
RPART	0.35	0.35	0.36	(Breiman et al., 1984)
LDA	0.32	0.33	0.31	(Venables and Ripley, 2002)

4. Random Forests Results

Tuning the parameter, the number of splits at each node, the resulting values of 6, 7, 4 for year 2008, 2009, and 2010, respectively, were chosen by having the largest accuracy value. With this information, the number of splits at each node values were used along with 500 trees, where the overall misclassification error rate is approximately constant and unchanging, in each year in the final model. The variable importance plot for each year can be found in Figure 4. The variable importance for classification problems is measured by the mean decrease in accuracy. Variables with larger mean decrease in accuracy are ranked higher in importance in predicting attachment. To find a cutoff point, we look for drops or large gaps between the points and keep the variables before the drop or gap.

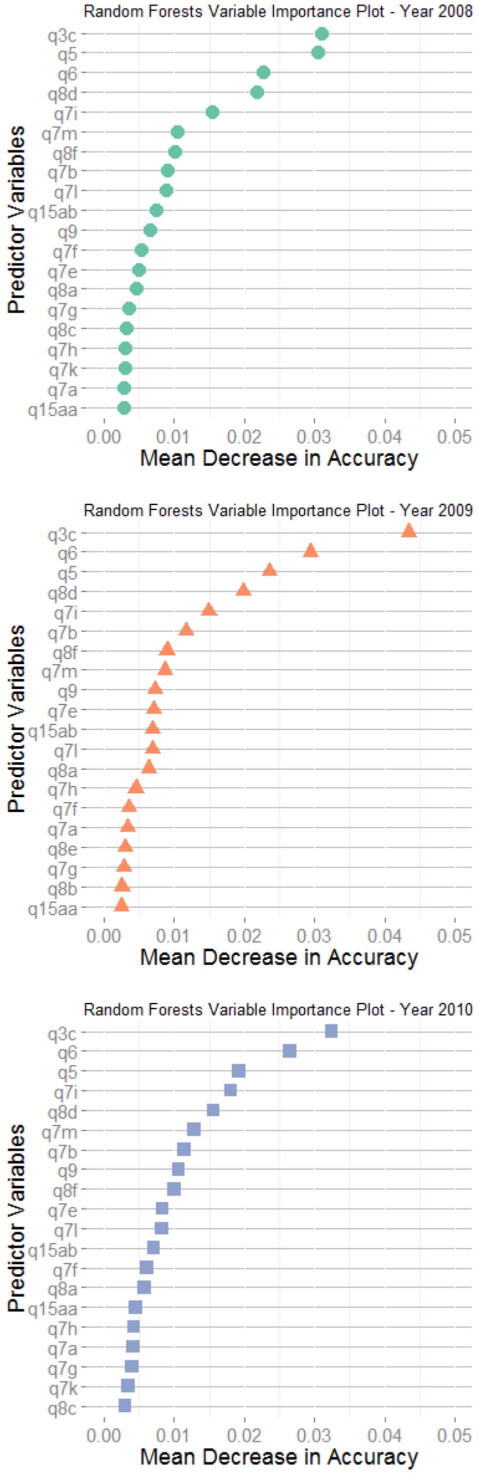
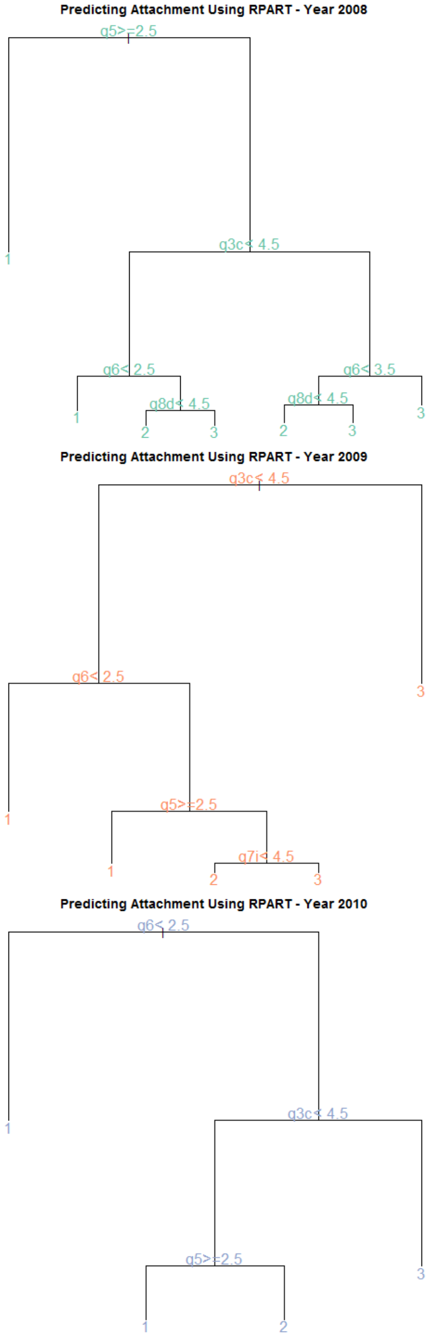


Figure 4: Random Forests Variable Importance Plot for years 2008, 2009, and 2010.

In the year 2008 results, we could keep either two variables — q3c (the community has a good reputation to outsiders or visitors who do not live here) and q5 (if you had the choice of where to live would you rather ...) — or four variables — adding q6 (how would you compare how the community is as a place to live today compared to 5 years ago?) and q8d (families with young children). In 2009, we should keep at least one variable, qc3, and in 2010 we should keep two variables, q3c and q6. q3c seems to be the most important variable in all years, suggesting some consistency in data from all three years.

5. RPART Results

As mentioned previously, RPART is of interest because the error rate was not much worse than Random Forests' misclassification error rate, it is more simple and adds interoperability since it looks only at one tree. The tuning parameter for RPART is the threshold complexity parameter, which is similar to an advisory parameter or in other words, it prunes the tree. Like Random Forests, RPART looks for the best split and the variable with the best split in separating people into three groups (attached, neutral, not attached) is also the most important variable. Once the data is separated, the process of finding the best split is repeated to each sub-group and so on until no improvement can be made. Results are shown in Figure 5 for each year. Notice that each year first splits on a different variable, however, there are similar variables found in the models for the three years. In addition, those five different variables found in all three years are also ranked important in Random Forests.



Predicting Attachment Using RPART - Year 2008

Predicting Attachment Using RPART - Year 2009

Predicting Attachment Using RPART - Year 2010

Figure 5: RPART results in predicting attachment. Values in the terminal stand for the following: 1- Not Attached, 2-Neutral, 3-Attached

The year 2008 tree first splits at q5 (if you had the choice of where to live would you rather ...). People with q5 values greater than or equal to 2.5 (levels 3 and 4) go to the left and people with q5 values less than 2.5, i.e., levels 1 and 2, go to the right. Overall, people that are either attached, neutral, or unattached to their community have similar thoughts and attitudes towards their community.

6. LASSO Results

LASSO has a max steps tuning parameter and the default value is used on all cases in the clean data set and then on the 26 participating communities for each year. The top four variables with the greatest absolute coefficient value in predicting attachment were investigated further. The heatmaps in Figure 6 show the results of the most important predictors overall and in each community. The variables are ordered by the frequency with which they are ranked as the four most important variables among all communities and in each community in the three years. Observe that the most important predictor variable for a community is colored in dark red. Our analyses show that q3c is the most important variable in predicting attachment with q6, q8d, and q7i following. The heatmaps also show that there are differences among the communities in what is important in influencing whether individuals are attached, neutral, or not attached to their community.

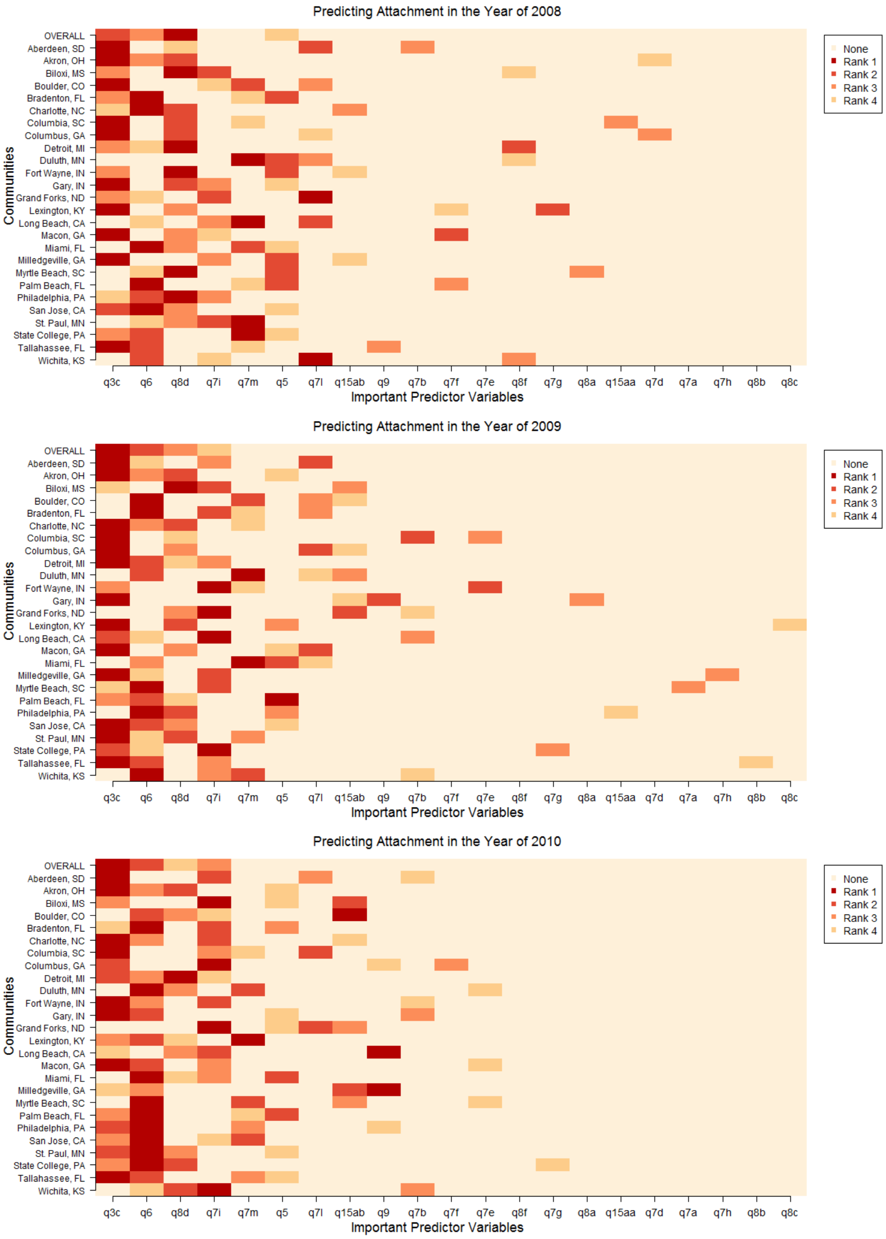


Figure 6: Heatmap of the LASSO results revealing the ranking of the predictor variables among all communities and in each community in each year. Variables are sorted from most frequent to least frequent occurrence in all communities and all three years combined.

While most of the important predictor variables consistently occur in a community in all three years, there are changes over time. Some of these changes are further examined in Section 8.

7. Archetypal Analysis

We used Archetypal Analysis to identify clusters within the data in the year 2008 and investigate if these clusters describe attachment to their community. Archetypal analysis represents each individual in the data as a mixture of individuals of pure type or archetypes (Cutler and Breiman, 1994). The variables used in archetypal analysis are: q3c, q6, q8d, q7i, q7m, q5, q7l, as determined by the results from our LASSO analysis. Cases missing in any of the variables were removed giving a sample size of 12,685 cases. The number of archetypes to keep is done by referencing a scree plot (a plot of the residual sum of squares for each archetype). In our case, the scree plot suggested a 7 pseudo-cluster; however, cluster membership in the 7-cluster solution performed poorly.

Therefore, we kept a 3 pseudo-cluster solution, which performed well. A visualization of our archetypal analysis in the base year, 2008, and an interpretation of the pseudo-clusters are shown in Figure 7. In the triangular figure (left), we see that those in archetype 1 are dominated by those who are unattached to their community. Those in archetype 2 or 3 display a mixture of attachment status. To describe these pseudo-clusters, we reference the dot chart (right). The points in the dot chart represent the percentile of each variable in the archetype as compared to the overall data set. For example, for archetype 1 variable q3c, the percentage is 7%, indicating that the q3c value in archetype 1 is the 7th percentile of the 12,685 cases in the 2008 data.The Kruskal-Wallis rank sum test (not shown here) indicated that there are significant differences between the means of the pseudo-clusters for each variable used in the archetypal analysis. Table 3 shows column percentages of a cross tabulation between the pseudo-clusters and attachment status. The cross tabulation of the groups are significantly different from each other according to the chi-square test.

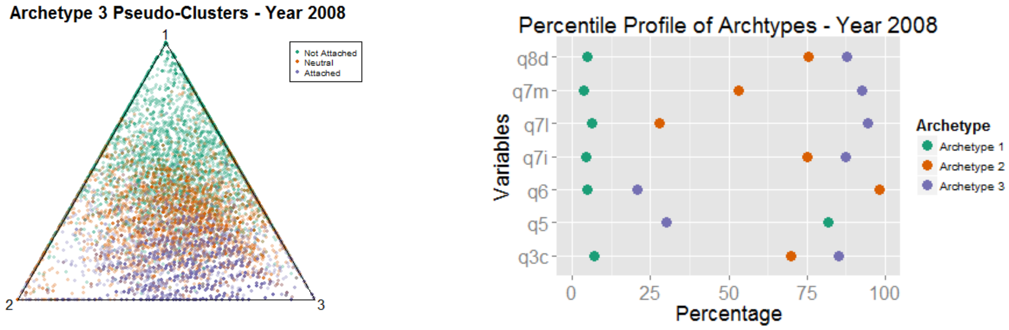


Figure 7: Graphical representation of the archetype 3 pseudo-cluster solution (left) and graphical representation to aid with the interpretation of the archetype 3 pseudo-cluster solution (right).

Table 3: Two-way table of attachment group and cluster assignment. Percentages in each column sum up to 100%. The p-value is the result of a chi-square test.

Levels	Pseudo-Cluster 1	Pseudo-Cluster 2	Pseudo-Cluster 3	P-value
Attached	2%	38%	40%	<0.0001
Neutral	20%	43%	45%	
Not Attached	78%	19%	15%	

The first archetype primarily has mostly people that are not attached to their community and archetypes 2 and 3 have about an approximately equal percentage of attached, neutral, and not attached people.

8. Dot Chart

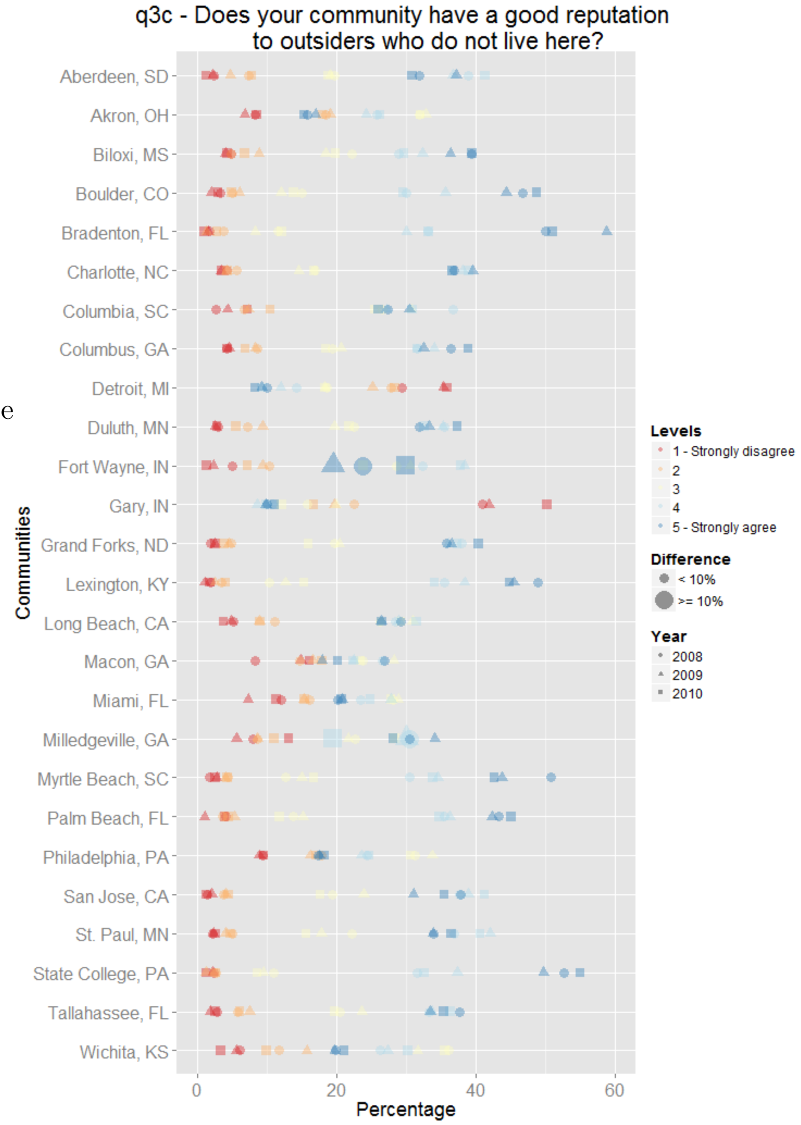


Figure 8: Dot chart of the most important predictor variable.

Our analyses revealed that there are mainly three variables that are important in determining attachment status. These variables are: q3c, q6, and q8d. We further investigated these variables in all three years by dot charts (see Figures 8, 9, and 10). The values are column percentages of the cross tabulation between the class levels and the communities of each variable. There are differences greater than 10% in the class levels between the three years for some communities in some of

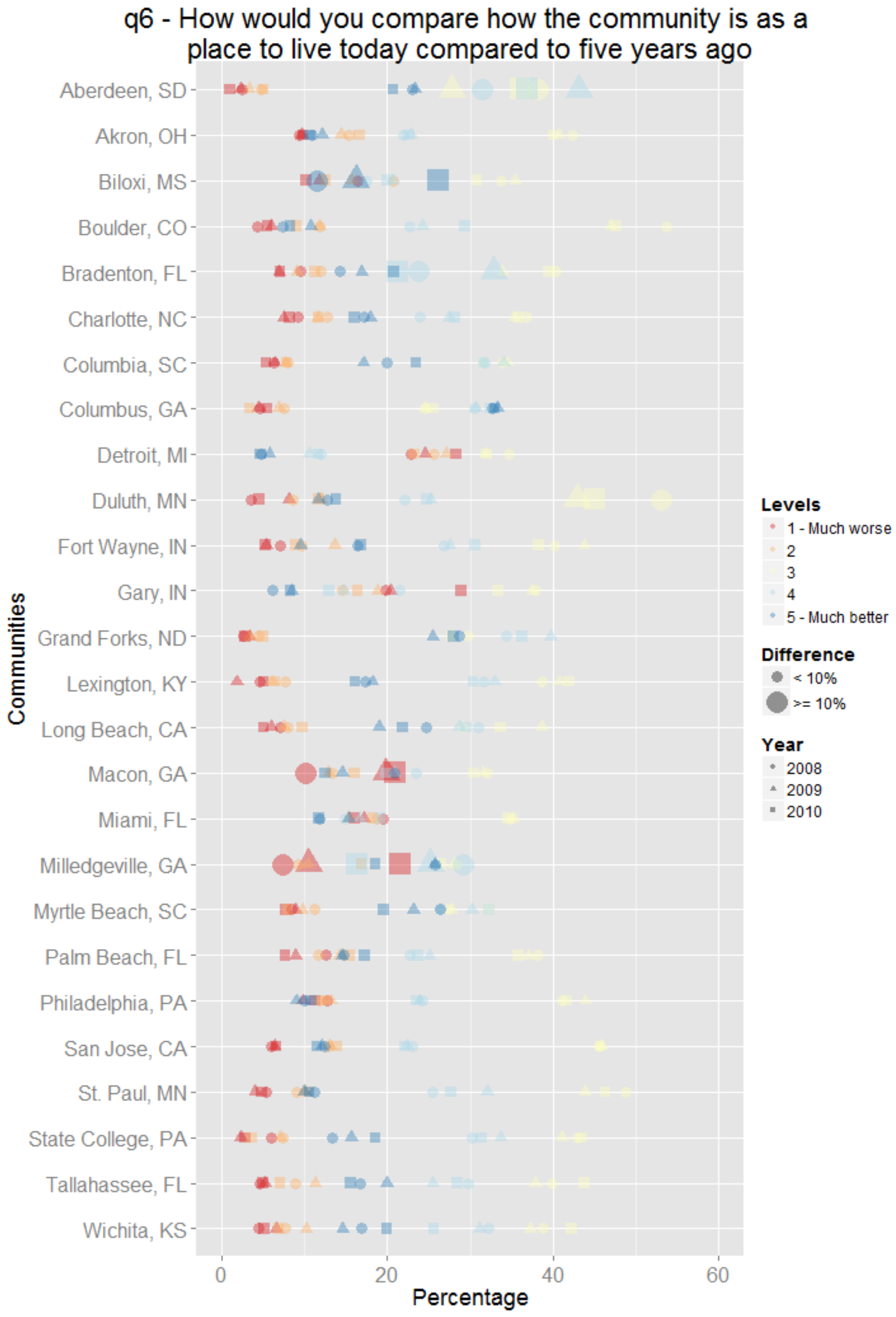


Figure 9: Dot chart of the second most important predictor variable.

the predictor variables. It would be necessary to verify whether these are true changes over time or just a side effect of relatively small sample sizes. Moreover, there are some communities that are rather unusual (when compared to the other communities) with respect to some of the predictor variables. For example, in Gary, IN, more than 40% of the participants in each year strongly disagree with q3c (the community has a good reputation to outsiders or visitors who do not live here).
In Detroit, MI, approximately 50% of the participants chose 1 (much worse) or 2 (second to worse) as their answer to q6 each year (how would you compare how the community is as a place to live today compared to five years ago). Apparently, this

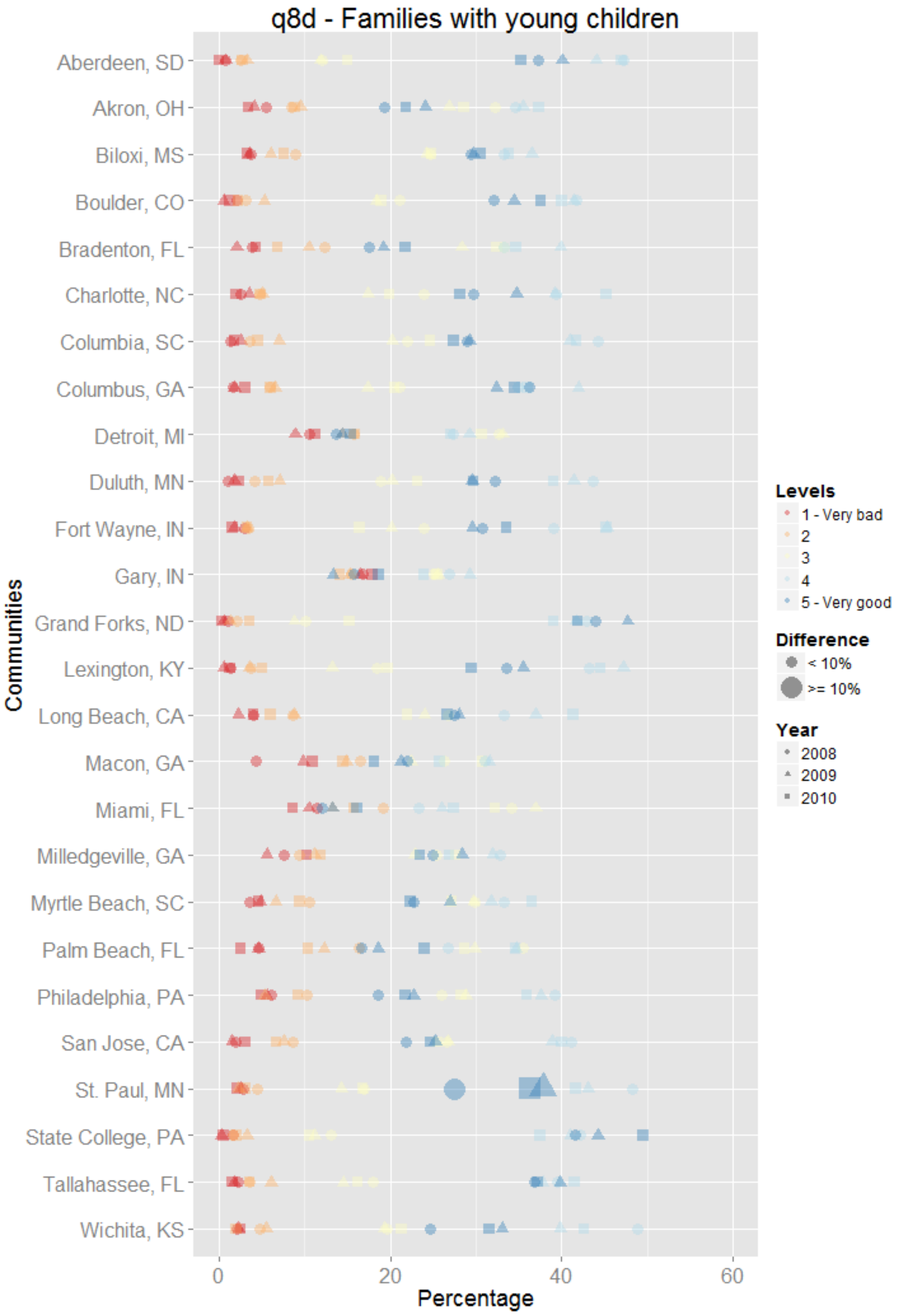


Figure 10: Dot chart of the third most important predictor variables.

can be interpreted as the long–term decline of this city that eventually resulted in its bankruptcy in July 2013[4].

[4] http://www.detroitnews.com/article/20130719/METRO01/307190045

9. Documentation & Conclusion

Predictor variables found in one of the machine learning algorithms in predicting attachment are documented in Table 4. The table also lists some of the variables that make up the index variables in Table 1. Table 5 lists the remaining variables and descriptions used to make up the index variables in Table 1. Variables found to be more important than others appear in bold.

Overall, we found that q3c was the most important variable in predicting attachment for most communities, but not all. Those that answered strongly agree to q3c (the community has a good reputation to outsiders or visitors who do not live here) were attached to their community, and those that answered strongly disagree were not attached to their community. The next four most important variables (in numerical order) were q5, q6, q7i, and q8d. Consistent results were obtained from the three machine learning algorithms. However, slight variations in the exact ranking could be found.

In contrast, a previous study[5] concluded: "The Knight Foundation's Soul of the Community project found that there is a strong, positive correlation between residents' attachment to their community and economic growth in that community. It also found that the qualities that most attach people to the place they live are aesthetics (the natural and manmade beauty of a place), social offerings (exciting opportunities to socialize with old friends and make new ones), and openness (how welcoming a place is to diverse groups of people)." However, our findings differed considerably. From the two variables (q7a and q7b) that make up aesthetics, neither one was among our top–five variables; from the three variables (q7h, q7i, and q7m) that make up social offerings, only q7i was among our top–five variables; and from the six variables (q8a, q8b, q8c, q8d, q8e, and q8f) that make up openness, only q8d was among our top–five variables. Neither q7i nor q8d was our single most important predictor of attachment.

Overall, people who have positive things to say about their community were also attached to their community and wanted to stay within their neighborhood, and those who were negative about their community were not attached to their community and would move to another city or state altogether if they could.

10. Software and R Packages Used

All data analyses and visualizations were done in R. R is a free software environment for statistical computing and graphics that can be downloaded at[6]. The following R packages were used: archetype (Adele Cutler's version that is not available in R), caret (random forests, rpart, lda)[7], ggmap, ggplot2[8], lars, maps, plyr, randomForest[9], RColorBrewer, rpart[10], scales (archetype), stats.

[5] http://www.soulofthecommunity.org/content/loving-where-you-live-key-successful-community

[6] http://www.r-project.org/

[7] http://cran.r-project.org/web/packages/caret/caret.pdf

[8] http://docs.ggplot2.org/current/

[9] http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[10] http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

A. Appendix

Appendix A summarizes the predictor variables in Table 4 and it lists additional variables related to the index variables in Table 5.

Table 4: Table of predictor variables. Unless specified differently, all variables in this table used levels 1 to 5 with 1 - Very bad ... 5 - Very good (and 1 - Low, 2 - Medium, 3 - High for the aggregated 3–level "r" variables).

Variable	Description
q3c	The community has a good reputation to outsiders or visitors who do not live here 1 - Strongly disagree ... 5 - Strongly agree
q5	If you had the choice of where to live would you rather 1 - stay in your neighborhood 2 - move to another neighborhood 3 - Move outside of your community 4 - Move to another city and state
q6	How would you compare how the community is as a place to live today compared to five years ago 1 - Much worse ... 5 - Much better
q7a(r)	The availability of outdoor parks playgrounds and trials
q7b(r)	The beauty or physical setting
q7c(r)	The highway and freeway system
q7d(r)	The availability of affordable housing
q7e(r)	The availability of job opportunities
q7f(r)	The overall quality of public schools in your community
q7g(r)	The overall quality of the colleges and universities
q7h(r)	Having a vibrant nightlife with restaurants clubs bars etc
q7i(r)	Being a good place to meet people and make friends
q7k(r)	The availability and accessibility of quality health care
q7l(r)	The leadership of the elected officials in your city
q7m(r)	How much people in your community care about each other
q8a(r)	Young talented college graduates looking to enter the job market
q8b(r)	Immigrants from other countries
q8c(r)	Racial and ethnic minorities
q8d(r)	Families with young children
q8e(r)	Gay and lesbian people
q8f(r)	Senior citizens
q9(r)	How would you rate economic conditions in your community today
q15aa(r)	Now is a good time to find a job in my area 1 - Strongly disagree ... 5 - Strongly agree (1 - Low 2 - Medium 3 - High)
q15ab(r)	The leaders in my community represent my interests 1 - Strongly disagree ... 5 - Strongly agree (1 - Low 2 - Medium 3 - High)

Table 5: Table of additional variables that make up the index variables in Table 1.

Variable	Description
qce1	Taking everything into account how satisfied are you with the community as a place to live
qce2	How likely are you to recommend the community to a friend or associate as a place to live
q6a	And thinking about five years from now how do you think the community will be as a place to live compared to today
q3a	I am proud to say I live in this community
q3b	The community is the perfect place for people like me
q10r	Right now do you think that economic conditions in your community as a whole are getting better or getting worse
q14r	Based on what you know or have seen, would you say that, in general, your company or employer is (answer from Sd)
q15r	How likely are you to agree that your job provides you with the income needed to support your family?
q18r	How would you rate how safe you feel walking alone at night within a mile of your home
q19r	How would you rate the level of crime in your community
q22ar	Performed local volunteer work for any organization or group
q22br	Attended a local public meeting in which local issues were discussed
q22cr	Voted in the local election
q22dr	Worked with other residents to make change in the local community
q23r	How many formal or informal groups or clubs do you belong to in your area that meet at least monthly
q24r	How many of your close friends live in your community
q25r	How much of your family lives in this area
q26r	How often do you talk to or visit with your immediate neighbors

References

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Cutler, A., Breiman, L., 1994. Archetypal Analysis. Technometrics 36 (4), 338–347.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the LASSO. Journal of the Royal Statistical Society 58 (1), 267–288.

Venables, W. N., Ripley, B. D., 2002. Modern Applied Statistics with S, 4th Edition. Springer, New York, NY.