Using Data Mining to Analyze Donation Data for a Local Food Bank

S. Jiang, L. Davis, H. Tavares De Mleo, and J. Terry
Department of Industrial and Systems Engineering, North Carolina A&T State University, Greensboro, NC, USA

Abstract - *Food insecurity is one of the difficult situations a lot of American communities face today. Hunger, particularly experienced by children has serious impacts on the society. Fighting hunger cannot solely depend on the government assistance programs. Non-profit organizations such as Feeding America play a very important role in this effort. These organizations heavily rely on food donations. However, it is not easy to understand donation and hence presents challenges for those organizations to plan and manage their resources. In this research, data mining techniques were applied to analyze donation data from a local food bank and useful information was generated to help the food bank manage their resources.*

Keywords: Data Mining, Donation, Food Bank

1 Introduction

IT'S hard to imagine that one in six Americans struggle with food and many of them are children [1]. Everyday in America, millions of people are unable to provide proper meals for themselves, making food security or food insecurity - the availability of food and the accessibility to it [2] - a big concern. According to a recent research, in year 2011, about 17.9 million households (14.9% of US population) were food insecure, an increase of 4% from the year before [3]. Furthermore, 16.7 million children lived in food insecure household in 2011. Food insecurity problems vary among different states. According to Feeding America, a non-profit organization that fights hunger in America, seven states had statistically significant higher household food insecurity rates than the U.S. national average in 2009- 2011. The state of North Carolina with a rate of 17.1%, is one of them. In 2010, 27.6% of the North Carolina children lived in food insecurity households. Child hunger presents several problems including health, education, and workforce and job readiness problems [4]. US government has food assistance programs to help fight the hunger. The Supplemental Assistance Nutrition Program (SNAP), Women, Infants, and Children (WIC) and the Emergency Food Assistance Program (TEFAP), are the top three programs has the most participation. One in four Americans participates in at least one of the food assistance programs yearly [4]. Unfortunately, food insecurity problems are way more serious than government assistance programs alone can handle. A lot of non-profit organizations are also working hard to fight the hunger. Among them, Feeding America, formerly known as America's Second Harvest, is the nation's largest hunger-relief charity engaged in the fight to end hunger. Its mission is to feed hungry Americans through a network of associated food banks. The Feeding America organization assists local food banks in securing and distributing food, raising funds and acquiring more donors, sharing best practices amongst food banks and other agencies, as well as advocating and inspiring individuals and the government to take action in ending hunger. Over 200 food banks under the Feeding America network are serving counties across the country and are supplying food to over 37 million Americans.

North Carolina has several food banks that are a part of the Feeding America network. The North Carolina Association of Feeding America Food Banks consists of six food banks and one food shuttle organization: Food Bank of Albemarle, Food Bank of Central and Eastern North Carolina, Manna Food Bank, Second Harvest Food Bank of Metrolina, Second Harvest Food Bank of Northwest North Carolina, Second Harvest Food Bank of Southeast North Carolina, and the Inter-Faith Food Shuttle. The food banks of North Carolina communicate public awareness about hunger issues, initiate fundraising events to collect donations, as well as distribute such food donations statewide.

In 2011, the North Carolina Food Banks distributed over 121 million pounds of food to 10 million North Carolinians in need. The North Carolina Association of Feeding America Food Banks works in all 100 counties in the state and have nearly 2,700 partner agencies. These agencies include church pantries, soup kitchens, shelters for the homeless and abused, childcare facilities and programs, and senior meal programs. Practically 170,000 individuals receive assistance from one of those partner agencies every week. The utilization of food banks has been steadily increasing since the early 1980s [6]. The food bank that is the focus of this study is the Food Bank of Central and Eastern North Carolina (FBCENC) that serves the largest population in the state.

The Food Bank of Central and Eastern North Carolina (FBCENC) serves 34 of the 100 counties in North Carolina and is the largest food bank in the area. The FBCENC is

comprised of six branches located in the Wilmington, Durham, Raleigh, Sandhills, Greenville, and New Bern areas. The New Bern branch was recently established within the past two years. The headquarters of the FBCENC is operated under the Raleigh branch and is located in Wake County. The FBCENC distributes over 150,000 pounds of food to 800 partner agencies. Partner agencies consist of emergency food programs such as soup kitchens, food pantries, homeless shelters, elderly nutrition programs and recognized churches. These partner agencies serve more than 500,000 individuals at risk of hunger across the 34 counties.

The donations received by the FBCENC are generated from local food drives, deliveries from partner food banks, and individual and business donations. The FBCENC also receives food and monetary donations from the government through the TEFAP and SNAP programs. In addition, the FBCENC will also purchase food to fulfill the unmet demand.

One of the challenges the FBCENC faces is to manage their resources to effectively fight the hunger. Although the FBCENC does receive food from various sources, majority of them are from donations as seen in Fig. 1.

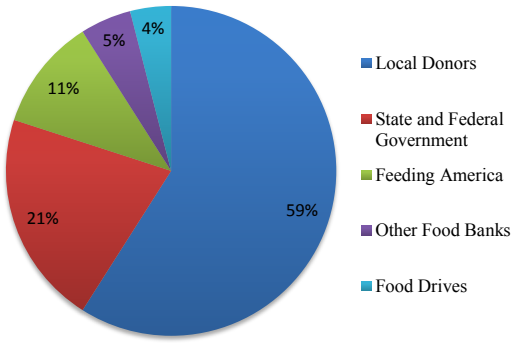

Fig. 1 FBCENC Food Sources

It is clearly from Fig. 1 that donations are very critical to the FBCENC mission. Therefore, understanding donations and detecting and discovering the trend of donation become a very important task for the FBCENC. Unfortunately, donation data are huge in volumes, and complex in nature, making them difficult to analyze. It is a typical data rich but information poor situation. Therefore, appropriate tools are needed to examine and analyze the donation data, and to discover important patterns.

Data Mining, an increasing important tool in extracting information from large amounts of data [7,8], can be applied in this situation.

This study aims to apply data mining techniques to explore the donation data and to use visualization tool present information to the food bank.

2 Method

2.1 Data collection

Historical donation data from FBCENC were retrieved. It contains 88,133 records of the food received by the Food Bank for the fiscal years of July 2006-07 to June 2010-11. To ensure only the donated records used, the data was filtered to remove the purchased records from the set. This decreased the records of food donations to 87,604. The key fields in the dataset include: posting date, donor name, gross weight, Category code, storage code as seen in Table 1.

TABLE 1
KEY FIELDS IN THE DATASET

Key Fields	Description
Posting Date	Date item received
Donor Name	Name or title of source
Gross Weight	Total mass of item
UNC_Product_Category_Code	Donor classification
UNC_Storage_Requirements_Code	Storage classification
FBC_Product_Type_Code	Food classification
Branch_Code	Branch
FBC_Product_Category_Code	Classification of receipt

A closer look at the data revealed there are many issues that needed to be addressed before the analysis. For instance, on certain days, one donor might have a positive gross weight and on the same day the same number but negative gross weight would also appear. This would have an impact on the frequency of donation although the total gross weight remains the same. Other indications of potential data error entry can be seen as a negative weight of donation. Clearly, data needed to be preprocessed before any analysis could be done.

2.2 Data preprocessing

Screening process was conducted to examine the sample for any unusual observations, missing values, and outliers.

2.3 Exploratory data analysis

Although there have been many high powered and yet expensive data mining tools commercially available, a simple and cheap alternative us needed for this study. JMP® is such a tool [9]. Since the goal of this study was to explore the donation data and provide insights to the management of the food bank, an exploratory analysis was conducted using JMP®.

First, a frequency analysis was conducted since the food bank is interested in donation frequency. Specifically, they are interested in getting information on frequent donors, occasional donors, and one-time donors. Second, a trend

analysis was done on the longitudinal data to detect any patterns that may exist.

2.4 Stochastic modeling

From the food bank's viewpoint, it would be useful if the number of new donors each month can be predicted. In this study, we applied Markov chain analysis in JMP® to solve this problem.

2.5 Cluster analysis

Cluster analysis is a widely used tool that can explore data and group observations into clusters based on certain criteria. Two types of cluster analysis are often used. The first is the hierarchical cluster analysis and the second is the k-means cluster analysis. Typically, hierarchical cluster analysis is used in the early stage of the research where exploration of data is the main concern. K-means cluster analysis, on the other hand, is an individual directed technique [7,8].

Given the large number of donors, it is useful to find a reasonable way to group them together. In this study, a cluster analysis was used. Both gross weight and frequency were selected as the criteria to get the clusters. Since we are exploring the donation data before further analysis can be done, hierarchical cluster analysis was used instead of k-means cluster analysis. Ward's minimum variance method was used for the hierarchical cluster analysis. Ward's method measures the distance between two clusters using the ANOVA sum of squares between the two clusters added up over all the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation. The sums of squares are easier to interpret when they are divided by the total sum of squares to give the proportions of variance (squared semipartial correlations) [8]. JMP® was used to conduct the hierarchical cluster analysis.

3 Results

In this study, an exploratory analysis was conducted to examine the donation data for the local food bank. A stochastic model was also built to predict new donors for the food bank. Cluster analysis was done to study donors. The following provide detailed results of these analyses.

3.1 Frequency analysis

Donation data were examined to analyze the frequency. It can be seen from Fig.2 that majority of donors are occasional donors and only a few of them are frequent donors. To further investigate this, four groups were formulated as: (1) Group 1: the first 10% of the most frequent donors (145 times); (2) Group 2: between the 3rd quantile (36) and the first 10% of the most frequent donors (145); (3) Group 3: between 1 and the 3rd quantile (36) ; and (4) Group 4: One time donors (1).

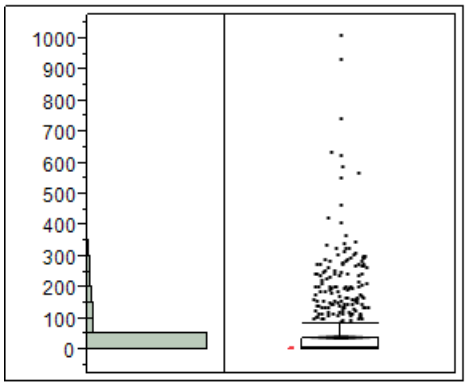

Fig. 2 Boxplot of donation frequency

Table 2 provides the descriptive statistics of the gross weight for each of the four groups. A visualization of the information as represented by a tree map can be seen in Fig. 3.

TABLE 2
DESCRIPTIVE STATISTICS OF GROSS WEIGHT BY EACH GROUP

Donor Group	Number of donors	Average	Standard Deviation	Percentage
Group 1	101	390793.41	944,09.76	20.31%
Group 2	153	677940.91	2509026.69	53.37%
Group 3	490	94471.78	154961.65	23.82%
Group 4	283	17252.60	65,832.58	2.51%

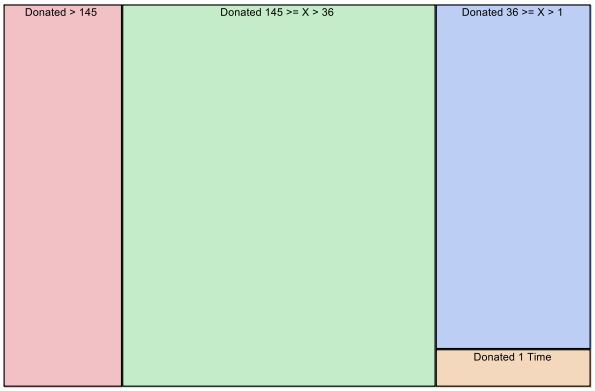

Fig. 3 Tree map of Gross_weight by each donor group

It is clear that more than half of the total gross weight came from Group 2 (between the 3rd quantile (36) and the first 10% of the most frequent donors). One-time donors (Group 4) only contributed about 2.5% of the gross weight even though there were 283 of them, ranking the second among four groups. Hence, caution needs to be taken when using frequency of donations as a measure.

3.2 Trend analysis

Given the data set spans five fiscal years, we also looked at the overall trend of amount (in terms of gross weight) of donation over the years as seen in Fig 4.

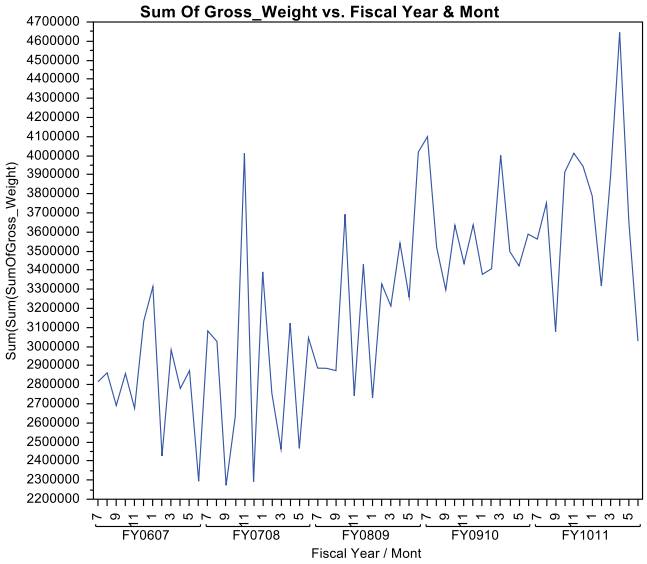

Fig. 4 Donation over the five years

Overall, there is an increasing trend in terms of amount of donation. However, large oscillation was found for data in certain years indicating more efforts need to be spent to investigate the story behind it. Another observation was made on the number of donations over the years as seen in Fig. 5.

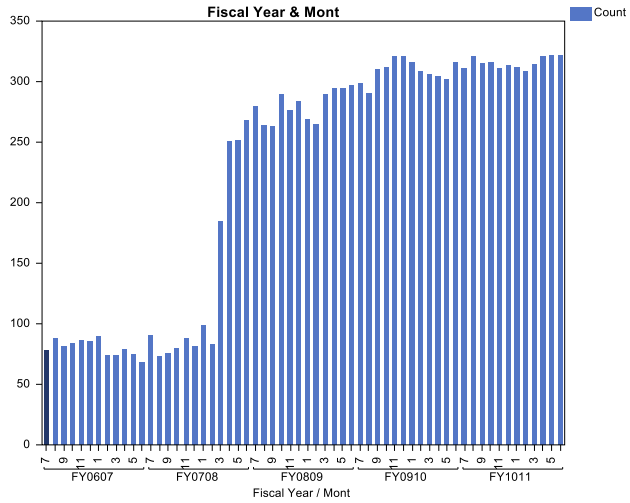

Fig. 5 Number of donations over the five years

There is an obvious jump from the first two years to the next three years in terms of the number of donations. Within the first two years, the number of donations each month is relatively the same and the same was noticed for the last three years. Further work is needed to discover the reason behind this. We also examined the total donations as measured by the gross weight. Figure 6 provides a snapshot of donation for each fiscal year. It clearly shows an upward trend which is consistent with the observation on the donation frequencies. Fig. 7 provides the monthly donation for each fiscal year.

From the graph, it seems that a similar pattern can be detected for each fiscal year since the lines seem to be parallel. This will be investigated further in the future research.

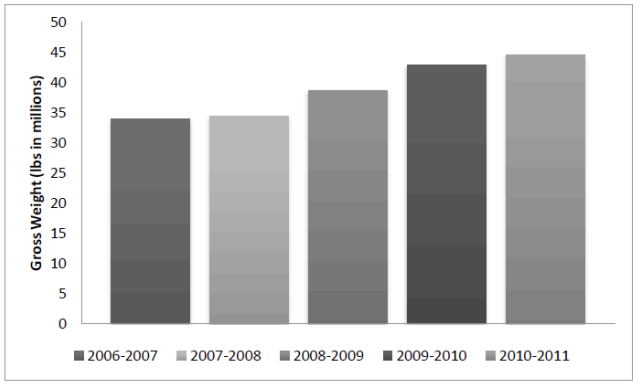

Fig. 6 Donations (Gross Weight) for each fiscal year

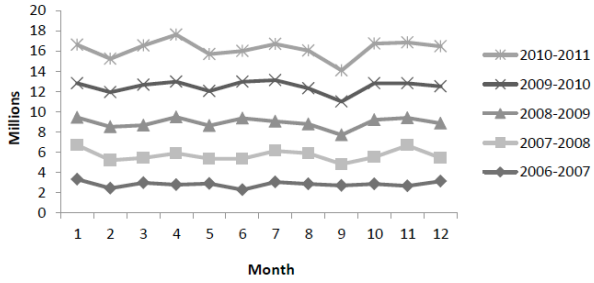

Fig. 7 Monthly Donations (Gross weight) over the five years

3.3 Stochastic modeling

Markov chain analysis was conducted to predict the number of new donors each month to help them plan and manage their resources more effectively. First, a transition matrix was developed. Since the goal is to predict the number of new donors, only one-time donor data were used. Four states were created based on the number of one-time donors in a month: State 1: no more than 2 donors; State 2: more than 2 but no more than 5 donors; State 3: more than 5 but no more than 8 donors; and State 4: more than 8 donors. Based on the data, a frequency matrix was generated as shown in Table 3. Using the frequency matrix, a transition matrix was then developed as seen in Table 4.

TABLE 3
FREQUENCY MATRIX

State	1	2	3	4
1	1	7	2	0
2	6	10	8	5
3	0	9	1	1
4	3	3	1	2

TABLE 4
TRANSITION MATRIX

State	1	2	3	4
1	0.1000	0.7000	0.2000	0.0000
2	0.2069	0.3448	0.2759	0.1724
3	0.0000	0.8182	0.0909	0.0909
4	0.3333	0.3333	0.1111	0.2222

Stationary vectors were acquired using MATLAB and results indicated that after some time (25), the system would be in a stationary state with the following probabilities:

Prob(State 1) = 0.1641;
Prob(State 2) = 0.4978;
Prob(State 3) = 0.2036;
Prob(State 4) = 0.1341.

This forecasting model will provide some insights to the food bank as they plan their resources in the future.

3.4 Cluster analysis

Hierarchical cluster analysis was conducted on the donation data to understand donors. Figure 8 shows the dendrogram based on the results. From the dendrogram, it can be seen 20 clusters can be obtained using the scree plot. Even though this is still a very large number, it is a good starting point given the complex nature of the donation data.

4 Discussion and Conclusion

Food insecurity and hunger are critical issues among the communities of the United States. Food banks such as the Food Bank of Central and Eastern North Carolina, aid the communities by providing food and other necessities to partner agencies whom then distribute to those in need. The survival of most food banks, including the FBCENC, rely heavily on receiving donations from the community. Being so, the amount of donations received fluctuates over time and can be difficult to predict. This instability increases the difficulty for a food bank to properly plan, distribute, and ration donations to the partner agencies. The purpose of this study was to apply data mining techniques to explore donation data and use visualization tools to provide meaningful information to the FBCENC.

Donation data were first preprocessed since the screening process revealed various problems with the data entry. Feedback has been given to the food bank to help improve data quality in the future. Although it is almost inevitable to have data errors given the large amount of data and human involvement in the entering process, it is still important to take some precautions to reduce those errors. With the recorded data, preprocessing them has proven to be important once again.

Exploratory data analysis was conducted to uncover data patterns. Frequency analysis revealed there exists relatively large number of one-time donors (the second largest group) and yet the total amount of donation from them only provides a small percentage of the total gross weight. We recommend the FBECNC to take this into consideration in their strategic planning.

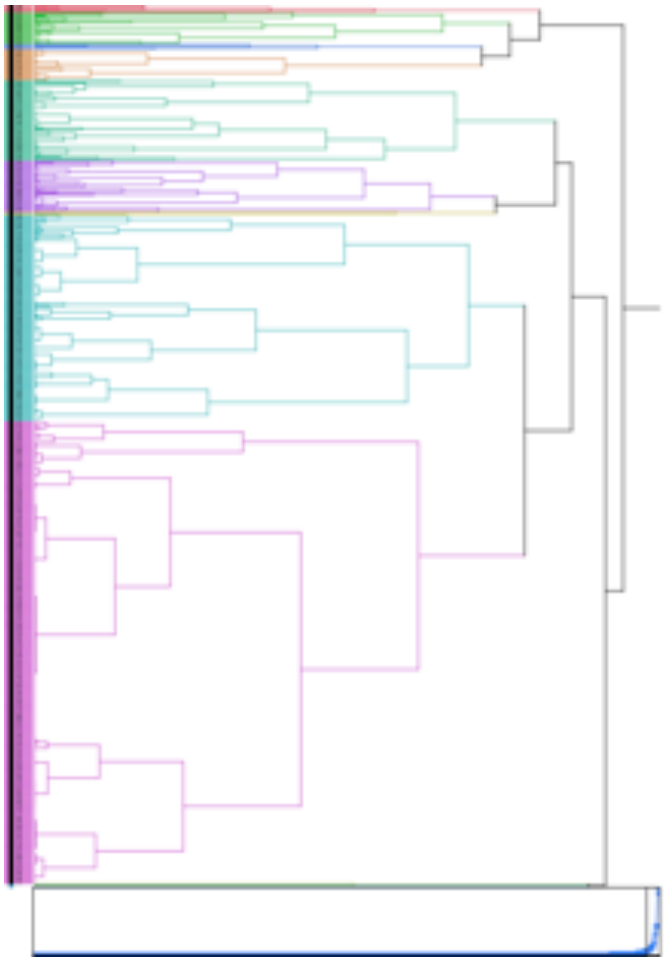

Fig. 8 Dendrogram of the cluster analysis

Trend analysis revealed an overall increasing trend in donations and yet large oscillation was observed for certain years. It is important to conduct further investigation on those data and adjust strategies accordingly. It is also interesting to notice that similar patterns were discovered from the monthly data for all five fiscal years. More research needs to be done to study those patterns since it will provide guidance for the FBCENC plan their resources in the monthly basis.

Markov chain analysis was conducted to predict number of one-time donors. Results indicated that a stationary system could be achieved over time. This will provide some insights to the FNCENC and help their strategic planning.

Finally, a hierarchical cluster analysis was conducted to understand donors. About 20 clusters were formulated based on the analysis. This number is still too large and yet did reveal useful information about donors. Further research need

to consider revising selection criteria and with better understanding of donation data, a k-means cluster analysis may be needed. The goal of this study is to explore the donation data and provide initial thoughts on how these data can be used to help the FBECNC. Future research will also include predictive modeling where multiple regression, logistic regression, decision tree, and neural network can be applied to build predictive models for donation data and provide the FBECNC more information in their effort to fight the hunger in America.

5 Acknowledgment

The authors would like to thank the food bank of Central and Eastern North Carolina (FBCENC) for supplying the data for this research. This research is partially funded by the National Science Foundation (CMMI 1000018). Points of view or opinions stated in this paper are ours and do not necessarily reflect the official position or policy of the National Science Foundation.

6 References

[1] M., Nord, A. Coleman-Jensen, "Household *Food Insecurity in the United States*." 2009.

[2] Barrett, C. B. (2010). Measuring food insecurity. *Science, 327(5967),* 825-828.

[3] Feeding America. “Hunger and poverty statistics,” retrieved at http://feedingamerica.org/hunger-in-america/hunger-facts/hunger-and- poverty-statistics.aspx

[4] Feeding America. “Children Hunger Facts,” retrieved at http://feedingamerica.org/hunger-in-america/hunger-facts/child- hunger-facts.aspx

[5] Feeding America. “Hunger in America 2010 National Report”, retrieved at http://feedingamerica.issuelab.org/resource/hunger_in_america_2010_ national_report

[6] Tarasuk, V. S., and Beaton, G. H., 1999, "Household Food Insecurity and Hunger Among Families Using Food Banks," *Canadian journal of public health*. *Revue canadienne de sante publique,* **90,** 109-113.

[7] M. Kantardzic, “Data Mining: Concepts, Models, Methods, and Algorithms,” Wiley, 2011.

[8] J. Han, M., Kamber, J. Pei, “Data mining: concepts and techiques,” 3rd edition, Morgan Kaufman, 2011

[9] SAS Institute, “JMP 10 Modeling and Multivariate Methods,” 2012.