STAT 685 Project

James Cadena, Thomas Nadobny, Mike Pfahler, and Huiting Sheng

May 2021

1 Abstract

The study explored origin and destination data in the Greater Houston area to find differences in travel patterns between weekdays and weekends. To measure differences, the study identified origin and destination nodes that have a strong interaction using ratio analysis. Using the origin and destination nodes identified in the ratio analysis, the study used Stochastic Block Models (SBM) to reveal the location, total clusters, and total nodes in each cluster.

2 Statement of Research

The study's stated hypothesis is that there is a difference in transportation behaviors between weekdays and weekends. Using origin and destination location data collected from mobile phones, the research is interested in revealing hidden subgroups within the location data, otherwise known as community detection. The research will focus on differences between weekday and weekend travel origins and destinations before the COVID-19 pandemic.

To find differences in weekday and weekend transportation behaviors, the study will assess the strength of interactions between origin and destination nodes. Interaction strength will be measured using ratio analysis. Ratios will be calculated from total mobile devices that moved between origin and destination nodes which indicate the strength of the interaction. Any interactions exceeding a defined threshold will be kept for the Stochasic Block Model. The research will leverage Stochastic Block Models for bipartite networks developed by Wyse et al. (2015). In the research completed by Wyse et al. (2015), a bipartite network consists of two types of nodes, Type A and Type B, which may be linked to each other but not within. A matrix of the interactions between node Type A and node Type B will be created where the rows and columns are simultaneously evaluated. In the study for origin and destination data, origins will be node Type A and destinations will be node Type B. The goal is to identify and estimate subgroups K and G that are in node Type A and B, respectively.

3 Data Overview

This study used origin destination data which represents travel patterns by humans from SafeGraph Inc. The location of the population is the Houston Metropolitan area or Houston Metro consisting of the eight counties: Harris, Fort Bend, Montgomery, Barzoria, Galveston, Liberty, Waller, and Chambers counties. The timeframe was from January 1, 2020 through January 31, 2020, this is prior to both the US declaring a Public Health Emergency and the World Health Organization declared COVID-19 a Pandemic.

4 Data Preprocessing

The hypothesis investigates if the behavioral transportation patterns of humans differ between weekdays and weekends in Houston. It is expected that holiday transportation patterns would differ from nominal transportation patterns. Therefore, the holidays: New Year's Day and Martin Luther King Day, were omitted. This allows the transportation patterns to be more homogeneous and systematic without the confounding affects such as special holidays. Additionally, transportation patterns where the person did not travel far, so the origin is the remained as the destination, were omitted. Furthermore, we exempted the Greater Houston counties, the remaining counties were Harris County and Fort Bend County.

Instead of total counts of humans traveling between an origin and destination, the fraction or ratio of humans leaving their residence by origin and destination is the key parameter of interest. These ratios were aggregated over the month of January per day and separated between weekends and weekdays. This ratio is similar to an odds ratio, therefore it will be transformed with the logistic distribution. Finally, the transformed ratios are continuous data, we can assume the Gaussian distribution is appropriate and then use the Gaussian model-based clustering to infer the number of latent clusters and cluster memberships among the origin and destination nodes.

Count = number of people traveled from origin(n) to destination(m)

Device Count = total number of devices at origin(n)

Completely Home Device Count = total number of devices which remained at origin (Did not travel)

$$Ratio = \frac{Count_{nm}}{DeviceCount_n - CompletelyHomeDeviceCount_n}$$

Thresholds for the ratio parameters are utilized in the analysis. Ratios which exceed the value of a threshold value of 0.15 will be used for the model-based clustering algorithm. The ratio threshold forced the model to use the most frequent origin to destination transportation patterns. A secondary side effect enabled the algorithms to quickly iterate because the majority of the ratios were less than the threshold, hence omitted. The ratios exceeding the threshold were used to identify the latent clusters among the origin and destination nodes.

$$Threshold \geq 0.15$$

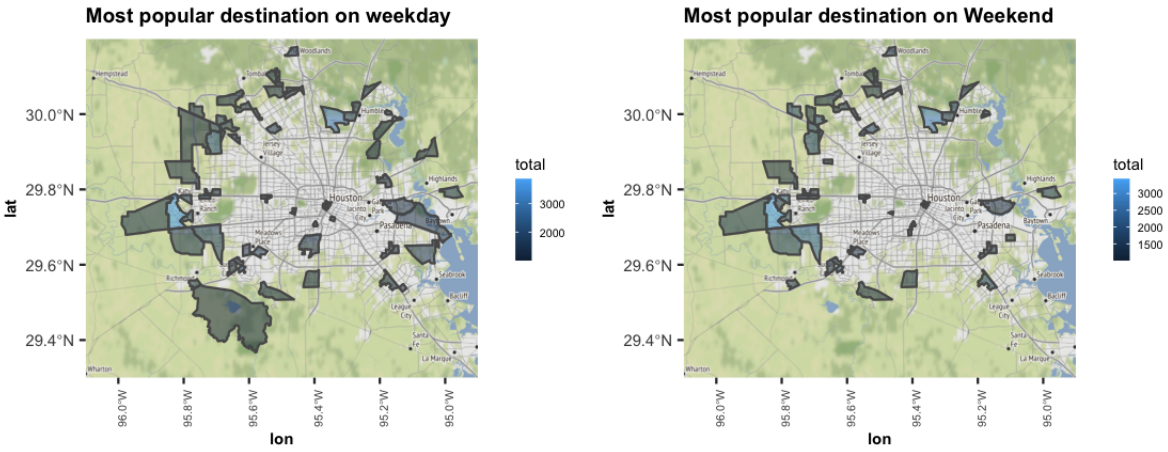
There are other important features or predictor variables that were in this study besides the ratio variable. The required distance traveled in miles is an important feature in transportation behavior of humans. We anticipate larger distances are inversely associated with the ratio or frequency of travel between a given origin and destination. So, farther transportation distances should be less common.

5 Visualization

First, let's take a look at the most popular destinations (places where over 1000 people visit daily) that people travel to on weekdays and weekends. We have removed the records where origin and destination are same then grouped by destination and date, and calculated the number of daily visits.

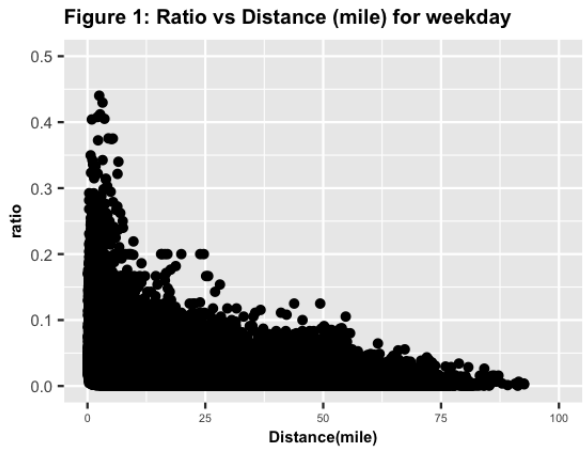
Below are the maps showing the most popular places. We can see that there are lots of places that are both popular on weekdays and weekends, however during weekdays there are more locations that have high popularity. Upon further examination, the popular places are actually industrial areas (e.g. the Southeast

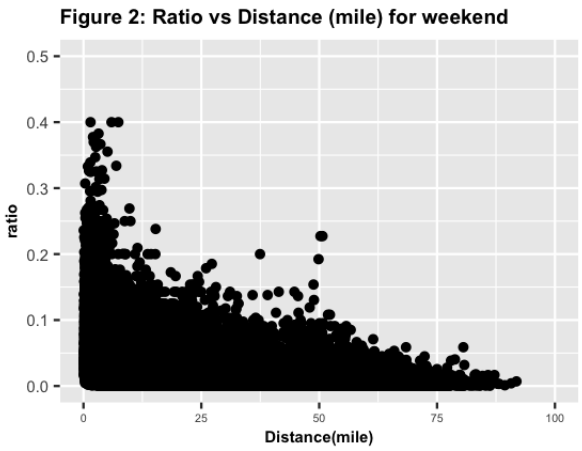
Harris area), School/University Area(e.g University of Houston) and offices area(e.g. downtown office area). This logic checks out as most people go to work or school Monday through Friday.

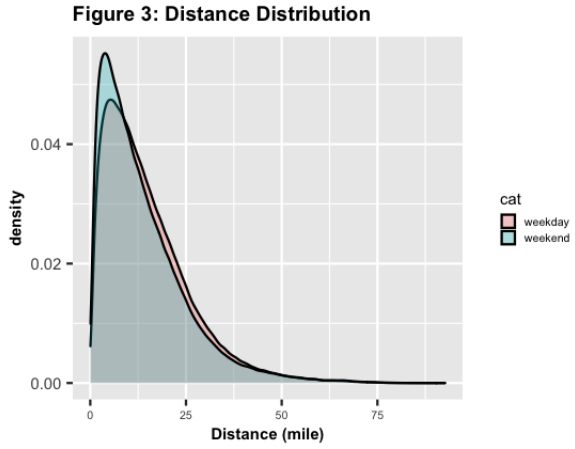


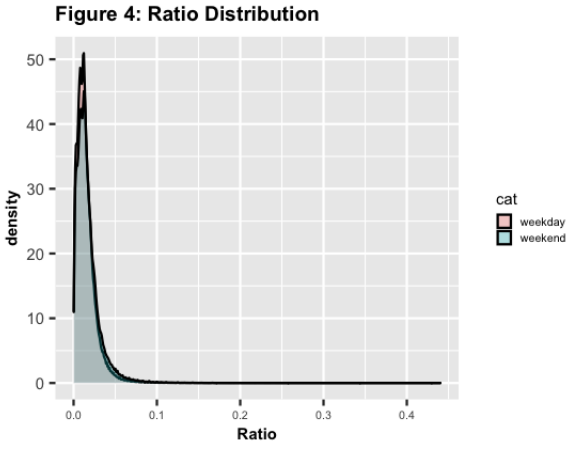
Next, let's see if there are any relationships between the ratio and the distance between the origin and destination, and if there are any differences between weekdays and weekends. To do this we added another predictor variable, distance in miles, and created Scatter plots to illustrate the relationship.

Below are the ratio vs distance scatter plots for weekdays and weekends. Both of them have very similar patterns. As expected, the data shows people tend to stay near their home when they leave the house. As distance increases, the number of people traveling longer distances decreases. The distribution plot also shows that most of people stay closer to their house on Weekend, which indicate that people may like relax at home and only do some grocery shopping during the weekend.The ratio distribution plots shows ratio is high-right skewed, within a census block, people spread out and do not visit the same places as the others in their block.



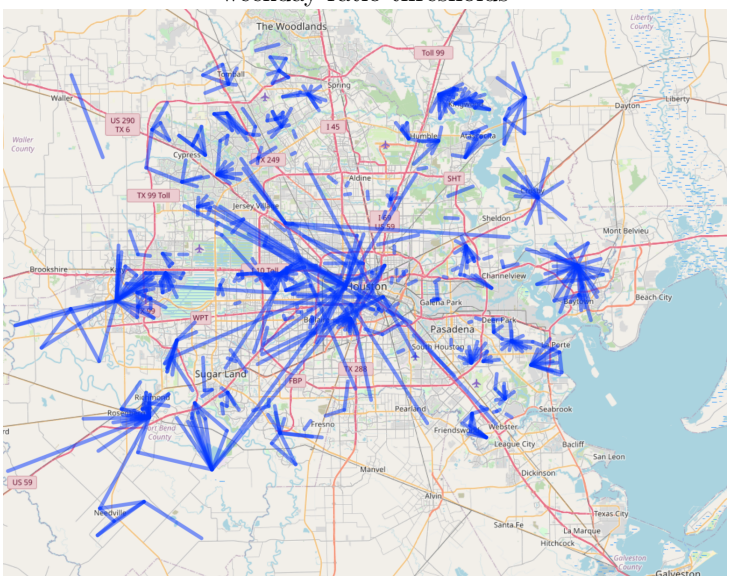






In the above distance-ratio plots, there are some points that show a large ratio which indicate more people from one census block group go to the same place than other places. We take threshold 0.15 and plot them on the maps and see where are they. On weekdays, most of those people go to their nearby schools, which make sense as people living in the same census block group likely go to the same school. On Weekends, most of those people go to places of entertainment, however lots of people go to their nearby schools as well, which may indicate people go to school in order to participate in their children's activities.

weekday ratio thresholds



weekend ratio thresholds

6 Methods

For the purposes of community detection, the number of clusters in the origin and destination nodes are estimated based on the strength of the interactions between the nodes, which in this research is the calculated Ratio variable. The greater the value of the Ratio, the higher the interaction amongst the origin and destination nodes. A minimum threshold of 0.15 was used for the Ratio values to ensure that only the strongest interactions were included. The primary method of analysis used for this research was to bi-cluster the origin destination matrix using the Stochastic Block Model (SBM) with a greedy search with exact Integrated Complete Likelihood (ICL).

The SBM provides a model-based approach to bi-clustering of two-mode networks, and in particular, focuses on modeling interactions between the nodes rather than the nodes themselves. As referenced in Wyse et al. (2015), bi-clustering is beneficial for two-mode networks as it is widely applicable and is able to identify "allusive insights" which more classical network measures are unable to determine. Through this

method, the number of clusters in both the origin and destination nodes can be estimated while simultaneously partitioned based on the clustering criterion, the ICL. The ICL is derived from the allocation vector which provides clustering of data points to the component densities, i.e. their likelihoods. These are then taken into account when determining the number of clusters for the node sets, where the values of K and G which are most supported by the data produce the largest ICL. A greedy search is used to optimise the exact ICL iteratively and is preferred for computational efficiency.

To determine the ICL and estimate the SBM for the data, we must first apply a Bayesian formulation. In this research, the primary prior distribution chosen was Gaussian, however, a Bernoulli was also considered but ultimately determined inadequate. This overall method was applied to the origin destination matrix through an original R formula titled bibloco.fit, created by and utilized in Wyse et al. (2015). To apply bibloco.fit to the data, the set of Gaussian parameters first had to be determined. For this research, the following parameters were chosen:

$$\text{alpha} = 2$$

$$\text{kappa} = 1/\text{Variance of Ratio}$$

$$\text{delta} = 2$$

$$\text{gamma} = 4$$

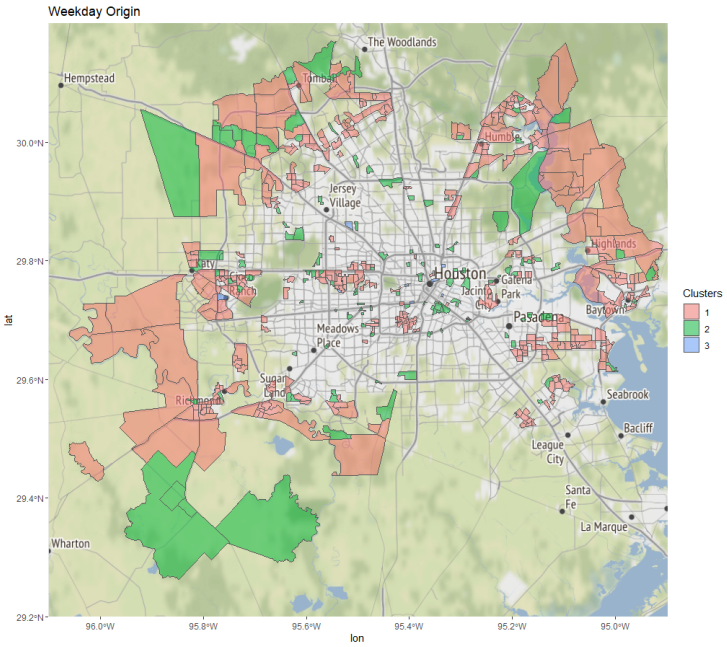
$$\text{xi} = \text{Mean of Ratio}$$

These parameters were created as a vector object in R and then plugged into the bibloco.fit formula, along with other variables which identified the prior distribution as Gaussian, the greedy search, non-sparsity, etc. Below is the input of the bibloco.fit formula used this research for the weekday origin destination data:

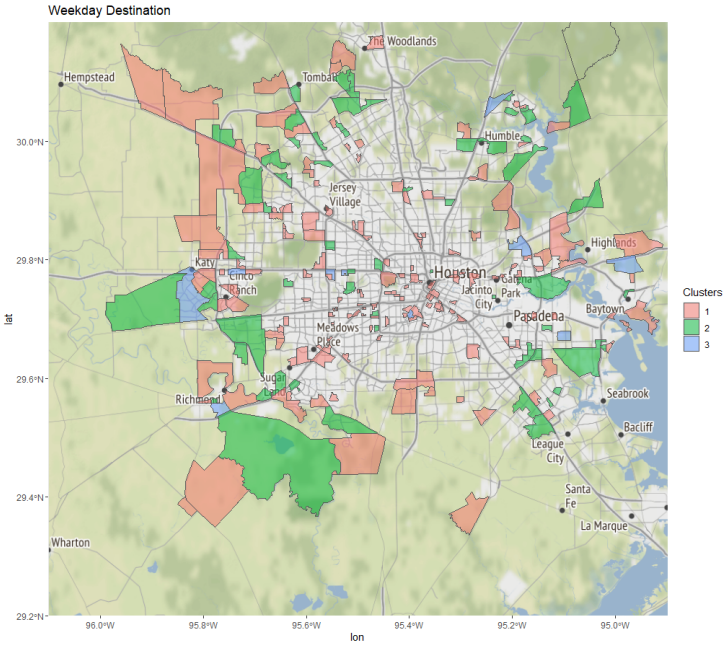
```
fit_non_sparse_weekday = bibloco.fit(gaussian_weekday_adjusted,
                                     model.type = "Gaussian",
                                     model.params = params,
                                     alg.type="greedy",  init.type = 0,
                                     n.runs = 10,
                                     init.groups =
                                     dim(gaussian_weekday_adjusted),
                                     n.restarts = 2,  greedy.fast = FALSE,
                                     greedy.merge =TRUE, keepICL = FALSE,
                                     sparse=FALSE, delta=0.05,
                                     merge.thresh =  0., n.margin = 2)
```

Listing 1: Weekday bibloco.fit Formula Input

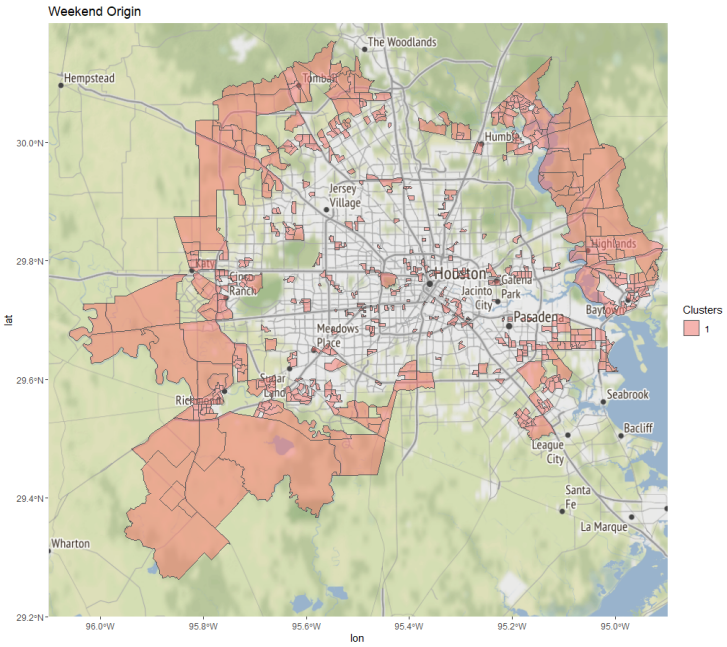
The output from the bibloco.fit model was the maximum ICL generated along with the estimated values of K and G, i.e. the number of clusters for origin and destination, and the cluster sizes. Additionally, plotting values are produced for the clusters which allow for visualizations to be created based on the origin destination relationship. The outputs from the weekday origin destination data were compared to that of the weekend data to determine differences in transportation behaviors.



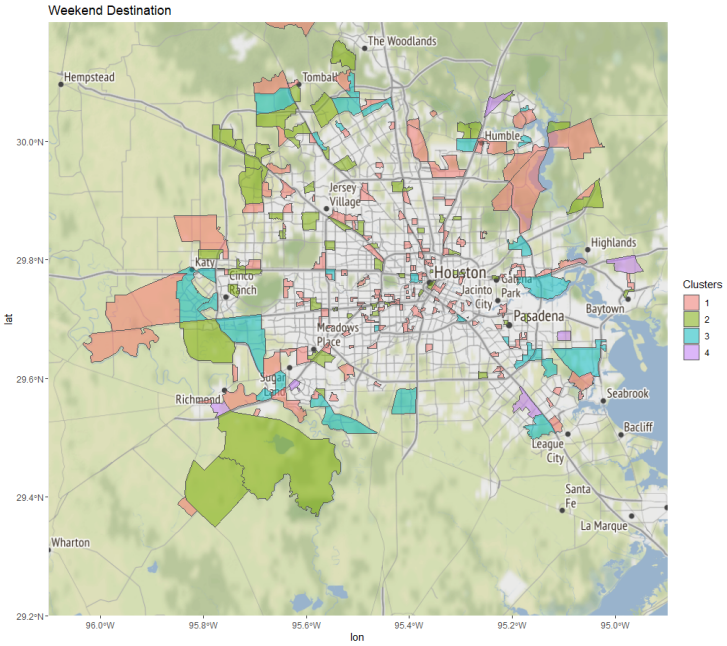
Weekday origin data output: 3 clusters with sizes (352, 133, and 6).



Weekday destination data output: 3 clusters with sizes (169, 54, and 10).



Weekend origin data output: 1 clusters with size (536).



Weekday origin data output: 4 clusters with sizes (163, 63, 31, and 6).

7 Results and Conclusion

Based on the results, the study found that in January 2020, there were different clusters in the origin and destination data for weekdays and weekends. The weekday origin and destination output generated three separate clusters while the weekend origin and destination output generated one and four clusters, respectively. The study did find that the further away an origin node is to a destination node, the weaker the relationship. Future studies could utilize the methods developed here for other analyses including mobility

in different socioeconomic locations and impacts of mobility due to the COVID-19 stay-at-home orders.

References

Jason Wyse, Nial Friel, Pierre Latouche (2015), Inferring structure in bipartite networks using the latent blockmodel and exact ICL

Vincent Brault, Mahendra Mariadassou (2015), Co-clustering through Latent Bloc Model: a Review. Journal de la Societe Franxaise de Statistique, Vol 156 No. 3

Carsten F. Dormann (2021), Using bipartite to describe and plot two-mode networks in R