Boston University MSSP Portfolio

Jack Carbaugh

December 2022

Contents

1 Speak for the Trees - Demographics Analysis

1.1 Introduction

The client for this partner project is Speak for the Trees, a Boston organization focused on providing trees to local communities to improve the health of local citizens and the environment of the area. In particular, they are focused on providing trees to local communities that may be under-served or under-canopied. In order to perfect this pursuit, the client wants the team to analyze tree canopy, demographic, environmental, and health data in an integrated setting.

The progression of the project first involved mapping the spatial tree data, along with potential demographic, environmental, and health factors. Then, predictions will be made of tree canopy coverage using demographic data, as well as environmental data. Finally, the effect of tree data on health will be analyzed. The group I was a part of mainly focused on the relationship between tree canopy coverage and demographic data.

1.2 Data and Methods

The main source of data provided by Speak for the Trees was a raster file of Boston, where each pixel represented one square meter. Each pixel had 3 possible values: 0 would represent that the square meter mostly consisted of a non-green surface, 1 would represent trees covering most of the area, and 2 would represent a majority non-tree green surface. Using the QGIS software, it was possible to create a shape file over varying region types of Boston that described the percentage of tree canopy coverage. For demographics data, the group wanted to focus on variables such as minority percentage, income, population density, and age. The demographics data was obtained from public sources such as the 2020 Census and the ACS 2019 5-year survey. Once again using QGIS, the group was able to define these variables over specific regions. For this project's analysis, the census tract was used as the shape size.

After collecting all necessary data and properly formatting it within shape files, the group moved to R with the goal of building a model predicting tree canopy coverage from the demographic variables. Beforehand, though, it was necessary to account for potential spatial correlation within the data. This was done using kriging, which would allow for the smoothing of the different variables across the Boston area. A spherical model was assumed, as we would expect neighboring regions to have similar demographic variables.

Variograms were also utilized to determine which variables fit the spherical model most clearly. The variograms with the clearest sills were minority percentage, income, and population density. Because of this, a linear regression model was fit utilizing these variables to see how tree canopy coverage could be predicted.

1.3 Results and Discussion

The linear regression model's coefficients were then analyzed for significance. The main result was significance in population density and median income, with population density having a positive effect and median income with a negative effect. Thus, the model result showed that more densely populated, lower income census tracts on average have a higher percentage of tree canopy.

This result was not close to our initial hypothesis, as one would think lower population density areas would have room for more trees. It was also a prevailing theory of our client that trees were a form of luxury more likely to appear in affluent regions. The highest tree canopy regions in Boston after kriging were in Southern Boston, so our team's rationale of this result was that the underdeveloped regions of Southern Boston were beginning to be filled in through gentrification.

1.4 Conclusion

The post-kriging data and model results were then sent to separate groups that combined our findings with environmental and health data. Overall, while the result was not expected, the project was still successful in its approach to collect these different areas of tree-related study into one analysis.

For our group, the experience was informative in working with spatial statistics, as well as the data acquisition process and the ability to learn new software (QGIS).

2 Glycocalyx Comparisons

2.1 Introduction

The endothelial glycocalyx is a layer of proteoglycans and glycoproteins lining portions of the aqueous outflow pathway of the eye. Our client was interested on this topic because very little is known about the glycocalyx within the aqueous outflow pathway. Our goal was to determine the typical height and coverage distribution in normal eyes, including whether the glycocalyx differs based on the amount of flow of certain radial segments of the outflow pathway, as well as changes in the height and coverage distribution of glycocalyx in a model of glaucoma (laser-treated eyes in this experiment).

This left us with three tasks: Comparing heights and coverage of specific locations between different flow regions, comparing heights and coverages of different locations within individual flow regions, and finding further statistical evidence to prove the difference.

2.2 Data and Methods

The data provided was split into height data (measuring the height of the glycocalyx), and coverage data (measuring the percent coverage of the glycocalyx on a given cell). Each glycocalyx observed was from one of 4 regions, and one of 7 outflow locations. The outflow locations are different spots along the aqueous outflow pathway of the eye, while the region described control eyes of different outflow amounts, as well as the lasered eyes. There also were variables for which of the monkey subjects the data came from, as well as which eye (left or right).

As the client wanted to compare the heights and coverage of the glycocalyx between different regions and outflow locations, our first attempt was to run several t-tests comparing the difference in means across the groupings. While this provided a quick and easy result to compare to our summary statistic visualizations, this process ignored the potential effects of the monkey subject and side of eye the data came from.

Adjusting the approach, the group decided to move to a multilevel model so that all the variables could be analyzed. The type of eye was used as the random effect, as well as the interaction between region and outflow location in order to make comparison easier. The model produced allowed for ease of comparison to the baselines, but overall comparisons were still difficult. The model was still presented to the client regardless, as it allowed them to predict future measurements of the glycocalyx.

The final approach that was taken involved using ANOVA and Tukey HSD tests. The Tukey test in particular allowed for locating significant differences amongst the wide range of group pairings the client required. While the assumptions of this test were strict (equal variance and normality amongst subgroups), most subgroups met them. A sensitivity analysis was run to check if those that didn't meet the assumptions significantly changed the results; the results barely changed upon this exclusion, so all subgroups were included. The Tukey tests

that were performed allowed for a clear answer to the client's questions, and provided easy to comprehend confidence intervals.

2.3 Results and Discussion

Utilizing the above methods, we mainly found significant differences in the height measurements of the glycocalyx between outflow locations when isolating for region. Even when isolating across different regions, the heights of the glycocalyx differed in mean across multiple pairs of outflow locations. More isolated significant differences were also found across the coverage's Non-Lasered region or SC outflow location, and the height's CC or ESV outflow location. These differences, similarly seen in the random effects of the mixed effect model, would seem to be due to lower than expected height measurement in the ESV location of the Control eyes, height measurement in the CC location of the Non-Lasered eyes, and coverage measurement in the SC location of the Non-Lasered eyes, as well as higher than expected coverage measurements in the ESV location of the Non-Lasered eyes.

2.4 Conclusion

The results were slightly unexpected to our client. The amount of differences of the glycocalyx's heights within the outflow locations was quite clear from simple summary statistics and visualizations. However, her hypothesis was that the different flow regions would make a bigger difference than they did overall. Despite this, the goal of the project was simply to understand the glycocalyx better, so it was successful in that regard.

Through this project, the team was able to try out multiple difference in mean testing methods. The entire project was also very thorough, as many different techniques had to be utilized, and new aspects analyzed. This project also required a vast amount of visualization, as all of the different possible comparisons needed to be made interpretable.

3 Air Cleaner

3.1 Introduction

Another partner project was brought by a company that produces an air cleaning machine. The main purpose of the device is to purify the air of Covid-19 virus, as well as other allergens and particulates. The cleaner would mainly be utilized in a meeting room or classroom setting to make mask-less conversation safe and worry-free. The air is cleaned by sucking in air through the side, putting it through a high-grade HEPA filter, and shooting air up towards the ceiling to be dispersed.

The client had a well tuned simulation of their product, where the spread of Covid-19 air flow could be measured throughout the room. Simulations where run with and without the air cleaner on in order to compare. Our team's initial goal was to utilize this simulation data in order to estimate the probability of infection with and without the air cleaner on. Thus, and overall goal was to statistically back the air cleaner's effectiveness in air purification.

3.2 Data and Methods

The main data initially provided was the simulation data. Two kinds of simulations were run: steady state (where the conditions of the room are kept constant), and transient (where the conditions are allowed to vary over each time point). The transient data was too massive to easily work with, and the conditions were likely stable enough to make the steady state sufficient. The main variable of interest from the simulation was the density of particles within a space.

Unfortunately, there was a steep learning curve dealing with the simulation data, as both the Ansys and Paraview software needed to be moderately understood. Even when retrieval of the data was possible, the team had continuous issues with understanding the exact content of the simulation. There were also problems verifying the virus infection rate, as the simulation only measured data along the surface of the mouth of the model persons. Through research, we found that it would be more accurate to measure the amount of particles in a bubble around the persons' heads, which was not easy to accomplish.

Luckily, a second source of data was introduced to the team in a lab test of the air cleaner's performance against smoke, dust, and pollen particles. The test was done three times for the different particles, and the machine was given 20 minutes to clean the air. The result was CADR, a measurement of the amount of space in the room able to be cleaned by the air cleaner. An analysis of the lab results showed a concern in the effectiveness of the client's device, as it seemed to give a lower than expected rating, which would be insufficient for the room used in the simulation. To check this result, the team studied the lab's testing process further, as well as attempted to check the raw data. The raw data was checked both manually and using a poisson model.

3.3 Results and Discussion

The resulting manual checks were not able to reproduce the lab test results. This was partially due to a lack of clarity on the testing company's part, as their calculation methods were not publicly available. The poisson model, however, was able to properly predict the results utilizing the raw data. This model allowed for better visualization of the data as well, along with better understanding of the testing process.

Despite the concerns of the simulation, the results were still utilized to help understand the infection rate. This was done utilizing the well known Welles Riley Model. While the model also has its own concerns, and some variables had to be estimated as they were not easily obtainable from the simulation, the model was still able to mostly confirm the results the simulation attempted to provide

3.4 Conclusions

As can be seen, the goals of the project changed significantly as time progressed, as the simulation became too esoteric to work with, and the previous goal became too lofty. Even without solid tangible results, the process helped the client better understand their air cleaner device, as well as the statistical methods and tests used to evaluate its effectiveness. Our group was able to assist them in locating potential weaknesses in their simulation, as well as improvements that can be made in the design of the device.

As a team, we were able to learn new software, use different types of infection models, and better understand simulation data. In terms of the process, we were able to learn how to properly switch focus if the initial goal was not obtainable. The team was also able to spend significant time evaluating testing processes, and critique method usage.

4 College Football Attendance

4.1 Introduction

College football within the NCAA's Division I Football Bowl Subdivision (FBS) is one of the most popular sports in the United States, and consists of teams constructed by some of the country's largest and most competitive universities. Due to the popularity of the sport, these programs can be a major source of revenue for the universities. In particular, ticket sales and other sponsorships are major part of this revenue, all being linked to attendance to the games. Thus, this project is mainly meant to better understand when fans attend college football games, in order to assist universities in increasing their attendance numbers, and thus increasing revenue.

4.2 Data and Methods

This analysis utilizes the College Football Games (2000 to 2018) posted by Jeff Gallini on Kaggle. This dataset has a majority of the factors expected to have an effect on attendance, such as weather data, game data, and team performance data (as well as attendance numbers themselves). The dataset logs over 6000 games for 63 home teams. Even among NCAA Division I FBS, the highest tier of college football, there is a great difference in the size of pedigree of school football programs. Because of this, attendance numbers are expected to vastly differ based on the school in question; this will require the use of a multilevel linear model with grouping based on school.

Due to the sizable variance in size, location, and recruiting prowess of Division I FBS schools, a multilevel linear model grouping by school will be used to predict the stadium fill rate of college football games. Fixed effects will include the amount of precipitation during the game, whether or not the game is Non-conference or not, and whether or not it is a "Big Game," such as a rivalry or championship game. The amount of games into the season will interact with the minimum temperature during the game, as all locations will generally get colder as the season progresses from late summer to winter. Record will interact with tailgating, as there was likely correlation seen in the EDA. Record will also be a random effect, as fan bases can range from expecting success every year, expecting failure every year, or being surprised based on the team's usual quality.

The variable used to represent attendance is fill rate, which is calculated by dividing the attendance number by the stadium capacity. Using fill rate instead of raw attendance numbers will provide a better standardization for smaller schools. Due to some stadiums allowing for overflow capacity, the fill rate will occasionally be greater than 1. To fix this, the fill rates were scaled to fit within the range of 0 to 1, and then a logit transformation was used to get closer to normality.

4.3 Results and Discussion

While many of the fixed effects seem to be significant, not as many are impactful in determining the fill rate of stadiums. As expected, precipitation deters people from attending games, but temperature strangely has a negative effect, meaning people are more likely to attend colder games. The number of games played was negative, telling us that more people attend games early in the year, which is expected as this is when schools usually increase their marketing push, and fans are hungry for football after the near 7 month break. Fans also are slightly more likely to attend games featuring nonconference opponents, potentially due to these games providing rarely seen marquee matchups

The random effects allow us to understand which teams' fan bases are most affected by their record. Interestingly, the schools with the greatest random coefficients for record are mainly mid-sized schools, while the groups with weak effects are mainly the stronger programs. These mid-sized schools occasionally get "flash in the pan" years where they're incredibly strong, but will have much weaker years more often. This volatility is likely the cause of the high random effect. The strong schools, meanwhile, are all usually competitive, meaning that their high record does not matter as much, as they mainly get consistent legacy attendance.

To improve this analysis, having a greater sample size, particularly among the schools with highly rated tailgating, could be quite helpful. It would also be interesting to add more school-related variables to this dataset, such as funding numbers, student population data, etc. Many of the variables studied in this dataset are either out of the schools control, such as weather, or are correlated with simply fielding a stronger team.

4.4 Conclusion

From this study, we can see that increasing tailgating quality, or perhaps simply a stronger pre-game experience, may lead to higher game attendance. Along with this, it seems that a winning culture attracts fans, along with stronger rivalries.

In general, this analysis allowed for practice with mixed effect models, new visualization tools, and the process needed for a self-directed study.

5 Heart Rates and Familiarity

5.1 Introduction

In previous studies, it has been shown that the presence of an unfamiliar object can be linked to a particular response in heart rate. This response involves the suppression of the heart rate and its variability upon interacting with the unfamiliar object, and then rising back to normal levels after a short time frame. While this response has been analyzed with a physical unfamiliar experience, the client is hypothesizing that this response should occur upon a sonic experience as well, in particular one involving musical lessons.

The client is concerned with the change in heart rate of children during musical lessons, with interest in the differences of the familiarity of the music. The experiment involved holding musical lessons for a group of children aged 3 - 4 over the course of 8 weeks. Each child was equipped with a heart rate monitor that took a record of BPM every 3 seconds. While at the start of the experiment the musical tracks were unfamiliar, by the end many would become familiar to the children. The experiment also involved transition periods between some musical episodes, which are intended as moments of rest and reflection, as well as a time to exchange props. Videos of the experiment were taken, and can be referred to to see how the children were reacting in certain moments.

5.2 Data and Methods

The client's data is built in a longitudinal or panel format, with measurements at specified time points for 7 subjects. The time measurements aligned with the three second intervals of the heart rate monitor, and were thus evenly spaced. The experiment was conducted during approximately 1 hour classes over each week for 8 total weeks, so gaps in the time series occur. This also meant that some children were not present during all of the meetings, leading to some missing data and an unbalanced data set. The dependent variable is the heart rate of the child measured in BPM. The state variable is categorical and describes whether the experiment was in its familiar, unfamiliar, or transitional stage, and is the predictor variable of most interest to the client. Other categorical variables that were sparsely utilized were for the episode (a small section) of the experiment, what kind of props were being used, and which musical material was being played.

The project began with extensive visualization of the data, as this time series data can be viewed in multiple ways, and is essential to understanding the phenomenon during the experiment. The main goal of the visualizations was to show how the heart rates changed across different episodes of familiarity. One of the most interesting visual takeaways was a noticeable alignment in trend of the heart rates across children. Despite the children having differing mean heart rates, they seemed to increase and decrease in tandem, and with similar variance.

The first analysis used in this study was a change-point analysis. One of

the main areas of interest for the client is being able to locate regions of change in heart rate in the children, and seeing if these changes align with changes in familiarity. Change-point analysis is a well known method that can help with this, and can locate the time in which a change in mean heart rate occurs. Once a set of change-points are found, it can be seen if they align with a change in familiarity, episode, or other event. Non-parametric methods of change-point analysis were used when possible, in order to work around the non-stationarity of the data. The PELT algorithm utilizing CROPS penalty was used, and the heart rate data was pooled over all the children. Segmentations were then selected based on the elbow plot, with decision preference given to segmentations with more change-points.

A second analysis was later done with the purpose of testing for mean difference in heart rate across the state groups. An ANOVA analysis on pooled data over episodes was done as an initial check. A more robust analysis was performed by building a linear mixed effect model. The state variable, as the main predictor of interest, was the only fixed predictor, while the child was set as a random effect. To account for the serial correlation and the small subject sample size, a variety of adjustments were utilized, such as adding an AR1 correlation structure, post-hoc robust variance matrix estimation, and bias-reduced linearization. While model fit was weak, this was due to only including categorical variables, and was not the goal of the model. Instead, it allowed for a group mean comparison to be completed.

5.3 Results and Discussion

Using change-point analysis, time points in the data that relate to a change in the structure of the heart rate signal were found. These change-points could be observed further using the client's video of the experiment to understand what was happening at these points, or could highlight particular moments in familiarity or episode change. Upon returning to the videos, the client was able to notice isolated examples of oddities involving the props they utilized. The change-points for the mean are distinct, and frequently highlight changes in episode, which is where the states of the experiment would change as well. The change-points for the variance are less distinct, but still occasionally key in on moments of exceptionally high or low variance.

Using the coefficient estimates and their standard error adjustments from the random effects linear mixed model, a significant difference at the 5 percent level is not found between any of the means of the heart rates in the varying familiarity/transitional states. Overall, an effect on heart rate due to familiarity of the episodic musical track is not distinctively apparent, and is not distinguishable from random sampling error. While the initial ANOVA attempt showed a significant difference between the mean heart rates of the unfamiliar and familiar states, the linear mixed effect model with the help of the the serial correlation adjustments gave a more robust result that is better trusted.

As this particular study was seen as exploratory, the client was able to take suggestions on the experiment. The client became interested in the synchro-

nization of the heart rates during the experiment, and will look into potential dependence among the students and instructor during a proceeding trial. Considerations also needed to be made on setting a clearer baseline heart rate to compare to, as well as locating the exact time period of the heart rate change phenomenon that the client is interested in.

5.4 Conclusion

The change-point analysis allowed for greater understanding of the heart rate data, and was able to highlight potential points of interest within the extensive time series data. The linear mixed effect model estimated the mean differences between heart rates of the experiment states, but did not find any significant differences.

This project required the use of many different visualization techniques that focused on particular aspects of the data; even some minor animations of the data were built. The project required research on a plethora of new modeling techniques and practices in order to meet the specifications of the data and client requirements.