Final Report

Group 1: Jiawen Chen, Brooke Felsheim, Elena Kharitonova, Xinjie Qian, and Jiarui Tang

4/30/2022

Introduction

As human populations are rising across the world, so is the proportion of people that live in urban areas. Estimates from the *UN World Urbanization Prospects* indicate that over 4.2 billion people (55% of the global population) currently live in urban areas, and by 2050, an additional 2.5 billion people (68% of the global population) could be living in urban areas[1]. More people living in urban areas calls for more space-, cost-, and energy-efficient systems of transportation as an alternative to cars. One such promising transportation alternative is the implementation of bicycle sharing programs.

Bicycle sharing programs are transportation schemes that allow individuals to rent bicycles on a short-term basis for either a set rate or for free. Most bicycle sharing programs have many computer-controlled bicycle rack "hubs" dispersed across a city that keep bikes locked and release them for use when a user enters the appropriate information/payment from a station or an app (Figure 1). A user can then ride the bike and return it to any other bicycle hub that is part of the same program. Many cities across the world have begun implementing bicycle sharing programs, including Chapel Hill, which has a Tar Heel Bikes sharing system[2]. Systems like these provide convenient, inexpensive, and eco-friendly transportation options for individuals residing in a city.

Successful implementations of bike sharing programs depend on proper management of these systems. It is important for a bike sharing program to provide a stable supply of rental bikes to its population so its users feel that they can rely on the system for their transportation needs. The analysis of bike sharing data allows for a better understanding of the demand of rental bikes in a city, which, in turn, can help inform a city about how to provide appropriate supplies of rental bikes for its population. For our project, we were interested in predicting the number of bikes rented within a given bike sharing system given information about weather, time of day, and date. We were also interested in assessing the most important variables for predicting bike rental counts. To answer these questions, we fit and evaluated a negative binomial generalized mixed model, a conditional inference tree, and a random forest model, using data from three publicly available bike sharing demand datasets.

The first dataset we use is a London bike sharing demand dataset downloaded from Kaggle[3] and provided by Transport for London[4]. This dataset contains hourly bike rental count observations over two years, from Jan 04 2015 - Jan 03 2017. The first full consecutive year of data was used as the training set in the analysis, and the second full consecutive year of data was held out as a test set.

The second dataset we use is a Seoul bike sharing demand dataset downloaded from the UCI Machine Learning Repository[5] and provided by the Seoul Metropolitan Government[6]. This dataset contains hourly bike rental counts over one year, from Dec 1 2017 - Nov 30 2018. This was used as an independent test set.

The third dataset we use is a Washington, D.C. bike sharing demand dataset downloaded from Kaggle[7] and

[1] United Nations, Department of Economic and Social Affairs, Population Division (2018). World Urbanization Prospects: The 2018 Revision, Online Edition.

[2] https://move.unc.edu/bike/bikeshare/

[3] https://www.kaggle.com/datasets/hmavrodiev/london-bike-sharing-dataset

[4] https://cycling.data.tfl.gov.uk

[5] https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

[6] https://data.seoul.go.kr

[7] https://www.kaggle.com/datasets/marklvl/bike-sharing-dataset

provided by Capital Bikeshare[8]. This dataset contains hourly bike rental counts over two years, from Jan 01 2011 - Dec 31 2012. This was used as an independent test set.
Each dataset contained hourly observations of bike rental count data. To simplify our analysis, we chunked the hourly data into three time blocks: [0:00 - 8:00), [8:00 - 16:00), and [16:00 - 24:00). Additionally, because temperature and humidity can be correlated with time of day, we chose to use the maximum and minimum daily temperature and humidity measurements for each 8-hour data point in order to avoid any issues of colinearity.
We created an R package named `bikeSharing` that includes methods for training and evaluating the negative binomial glmm and random forest models, as well as the processed source data from all three sets.
The below code can be used to install the package and load the package library. The zipped package source data, `bikeSharing_1.0.0.tar.gz` can be found in the Github repository for our project[9].

```
if(!require("bikeSharing", quietly = TRUE))
  install.packages("package/bikeSharing_1.0.0.tar.gz", repos = NULL)
library(bikeSharing)
```

Once the package is loaded, the bike sharing data from all three sets becomes easily accessible through the variable names `london`, `seoul`, and `dc`. To directly load them into the R environment, one can simply run:

```
data("london", "seoul", "dc")
```

All three datasets contain bike rental count data as well as 11 additional weather-, time-, and date-related variables that were used as predictors in our models:

- Hour chunk (00:00 - 8:00, 8:00-16:00, 16:00-24:00)
- Weekend status (Yes/No)
- Holiday status (Yes/No)
- Season (Winter, Spring, Summer, Autumn)
- Minimum daily temperature (C)
- Maximum daily temperature (C)
- Minimum daily humidity (%)
- Maximum daily humidity (%)
- Wind speed (m/s)
- Presence of any rain or snow (Yes/No)
- Date (mm-dd)

The way that these variables are used within our models will be further described in the Methods section. For all of the analyses performed, the `london` dataset was divided into training and testing sets, where the training set contained all "Year 1" data (Jan 04 2015 - Jan 03 2016), and the testing set contained all "Year 2" data (Jan 04 2016 - Jan 03 2017).

```
london_train <- london[london$Year == "Year 1",]
london_test <- london[london$Year == "Year 2",]
```

[8]https://ride.capitalbikeshare.com/system-data
[9]https://github.com/brookefelsheim/bios735-group1

Methods

Negative Binomial Generalized Linear Mixed Model

Let y_{ij} be the number of bikes rented at hour chunk j of day i. Thus i ranges from 1 to 365, and j ranges from 1 to 3, corresponding to hour chunks [00:00 - 8:00) [8:00-16:00), [16:00-24:00). We assume that the number of bikes rented for a given hour chunk within a specific day follows a negative binomial distribution, so $Y_{ij} \sim NB(\mu_{ij}, \theta)$ using the Hilbe parameterization, so:

$$P(Y_{ij} = y_{ij}|\mu_{ij}, \theta) = \frac{\Gamma(y_{ij} + \theta)}{\Gamma(y_{ij} + 1)\Gamma(\theta)} \left(\frac{\theta}{\theta + \mu_{ij}}\right)^{\theta} \left(\frac{\mu_{ij}}{\theta + \mu_{ij}}\right)^{y_{ij}} \tag{1}$$

The mean of y_{ij} for each day i at hour chunk j is μ_{ij}, which we assume follows a negative binomial generalized linear mixed model with a random intercept. Thus it is determined from the following model:

$$\log(\mu_{ij}) = x_{ij}^T \beta + b_i \tag{2}$$

Let x_{ij} = (1, I(HourChunk$_{ij}$ = [8:00,16:00)), I(HourChunk$_{ij}$ = [16:00,24:00)), Weekend$_i$, Holiday$_i$, I(Season$_i$ = Spring), I(Season$_i$ = Summer), I(Season$_i$ = Winter), Min_Temperature$_i$, Max_Temperature$_i$, Min_Humidity$_i$, Max_Humidity$_i$, Wind_Speed$_{ij}$, Rain_or_Snow$_{ij}$)^T.

HourChunk$_{ij}$ is a categorical variable corresponding to the hour chunk j of day i, with the reference hour chunk being [0:00,8:00), Weekend$_i$ is a binary variable with 1 if day i is a weekend, 0 if it is not, Holiday$_i$ is a binary variable with 1 if day i is a holiday, 0 if it is not. Season$_i$ is a categorical variable corresponding to which season day i is in, with the reference season being Autumn. The Min_Temperature$_i$ and Max_Temperature$_i$ are the minimum and maximum temperature of day i, respectively. Similarly, The Min_Humidity$_i$ and Max_Humidity$_i$ are the minimum and maximum humidity of day i, respectively. Wind_Speed$_{ij}$ is the average wind speed of hour chunk j of day i. Rain_or_Snow$_i$ is a binary variable with 1 if during day i hour chunk j there is any rain or snow, 0 if there is not. Thus $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{13})'$.

In Equation (2), b_i is the unobserved random effect that the day i has on the number of rented bikes at any given hour chunk. It is assumed that $b_i \sim N(0, \sigma_\gamma^2)$. It is assumed that b_k and b_j are independent for any $k \neq j$.

Thus, of interest is the estimate θ, β, and σ_γ^2. Equations 1 and 2 can be combined to obtain the following likelihood equation for a given day (i).

$$L(\theta, \beta, \sigma_\gamma^2|\mathbf{y}, \mathbf{b}) = \prod_{i=1}^{365} \prod_{j=1}^{3} p(y_{ij}) = \prod_{i=1}^{365} \left(\prod_{j=1}^{3} f(y_{ij}|\mu_{ij}) \right) f(b_i|\sigma_\gamma^2) \tag{3}$$

Since the random effects b_i are unobservable, this means they must be integrated out of the above expression to obtain the likelihood, so:

$$L(\theta, \beta, \sigma_\gamma^2|\mathbf{y}, \mathbf{b}) = \prod_{i=1}^{365} \left[\int \left(\prod_{j=1}^{3} f(y_{ij}|\mu_{ij}) \right) f(b_i|\sigma_\gamma^2)\, db_i \right]$$

$$= \prod_{i=1}^{365} \left[\int \prod_{j=1}^{3} \frac{\Gamma(y_{ij} + \theta)}{\Gamma(y_{ij} + 1)\Gamma(\theta)} \left(\frac{\theta}{\theta + e^{x_{ij}^T\beta + b_i}}\right)^{\theta} \left(\frac{e^{x_{ij}^T\beta + b_i}}{\theta + e^{x_{ij}^T\beta + b_i}}\right)^{y_{ij}} \frac{1}{\sqrt{2\pi\sigma_\gamma^2}} \exp\left(-\frac{b_i^2}{2\sigma^2}\right) db_i \right]$$

So thus, the log likelihood is found to be

$$l(\theta,\beta,\sigma_\gamma^2)=\sum_{i=1}^{365} log\Big[\int\prod_{j=1}^{3}\frac{\Gamma(y_{ij}+\theta)}{\Gamma(y_{ij}+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+e^{x_{ij}^T\beta+b_i}}\right)^{\theta}\cdot\left(\frac{e^{x_{ij}\beta+b_i}}{\theta+e^{x_{ij}\beta+b_i}}\right)^{y_{ij}}\frac{1}{\sqrt{2\pi\sigma_\gamma^2}}\exp\left(-\frac{b_i^2}{2\sigma_\gamma^2}\right)db_i\Big]$$

This log-likelihood will be maximized to obtain estimates for $\theta,\beta,\sigma_\gamma^2$ through an MCEM approach. Assuming that the b_i's were known, we first define the complete data log likelihood as:

$$\log L_C(\theta,\beta,\sigma_\gamma^2|\mathbf{y},\mathbf{b})=\log\left[\prod_{i=1}^{365}\left(\prod_{j=1}^{3}f(y_{ij}|\mu_{ij})f(b_i|\sigma_\gamma^2)\right)\right]=\sum_{i=1}^{365}\log\left(\prod_{j=1}^{3}f(y_{ij}|\mu_{ij})f(b_i|\sigma_\gamma^2)\right)$$

$$\text{So, } l_C(\theta,\beta,\sigma_\gamma^2|\mathbf{y},\mathbf{b})=\sum_{i=1}^{365}\left(\sum_{j=1}^{3}\log f(y_{ij}|\mu_{ij})+\log f(b_i|\sigma_\gamma^2)\right)$$

Thus, for the MCEM algorithm, we will be maximizing the expectation of the complete data log likelihood, otherwise known as the Q-function at step t. So the Q-function is defined as the following:

$$\begin{aligned}Q(\theta,\beta,\sigma_\gamma^2|y,\theta^{(t)},\beta^{(t)},\sigma_\gamma^{2\,(t)})&=E[l_C(\theta,\beta,\sigma_\gamma^2|\mathbf{y},\mathbf{b})|y,\theta^{(t)},\beta^{(t)},\sigma_\gamma^{2\,(t)}]\\&=E\left[\sum_{i=1}^{365}\left(\sum_{j=1}^{3}\log f(y_{ij}|\mu_{ij}^{(t)})+\log f(b_i|\sigma_\gamma^{2\,(t)})\right)\right]\\&=\sum_{i=1}^{365}\left[\int\left(\sum_{j=1}^{3}\log f(y_{ij}|\mu_{ij}^{(t)})+\log f(b_i|\sigma_\gamma^{2\,(t)})f(b_i|y,\theta^{(t)},\beta^{(t)},\sigma_\gamma^{2\,(t)}\right)db_i\right]\end{aligned}$$

Where $\mu_{ij}^{(t)}=e^{x_{ij}^T\beta^{(t)}+b_i}$ and $f(b_i|y,\theta^{(t)},\beta^{(t)},\sigma_\gamma^{2\,(t)})$ is the posterior distribution of b_i given the observed data and the current parameter estimates. Let $Y_i=(y_{i1},y_{i2},y_{i3})^T$. The density function for b_i given $Y_i,\beta^{(t)},\theta^{(t)},\sigma_\gamma^{2\,(t)}$ is:

$$\begin{aligned}f(b_i|Y_i,\theta^{(t)},\beta^{(t)},\sigma_\gamma^{2\,(t)})&\propto f(Y_i|b_i,\theta^{(t)},\beta^{(t)},\sigma_\gamma^{2\,(t)})\cdot f(b_i|\sigma_\gamma^{2\,(t)})\\&\propto\prod_{j=1}^{3}\left(\frac{\theta^{(t)}}{\mu_{ij}^{(t)}+\theta^{(t)}}\right)^{\theta^{(t)}}\cdot\left(\frac{\mu_{ij}^{(t)}}{\mu_{ij}^{(t)}+\theta^{(t)}}\right)^{y_{ij}}\cdot\exp\left(-\frac{b_i^2}{2\sigma_\gamma^{2\,(t)}}\right)\\&\propto\sum_{j=1}^{3}\left[(y_{ij}\log\mu_{ij}^{(t)}-(y_{ij}+\theta^{(t)})\log(\mu_{ij}^{(t)}+\theta^{(t)}))-\frac{b_i^2}{2\sigma_\gamma^{2\,(t)}}\right]\end{aligned}\tag{4}$$

If we could sample from this posterior distribution of b_i, we could approximate this integral using a montecarlo approach. So,

$$Q(\theta,\beta,\sigma_\gamma^2|y,\theta^{(t)},\beta^{(t)},\sigma_\gamma^{2\,(t)})=\frac{1}{M}\sum_{i=1}^{365}\sum_{k=1}^{M}\left(\sum_{j=1}^{3}\log f(y_{ij}|\mu_{ijk}^{(t)})+\log f(b_{ik}|\sigma_\gamma^{2\,(t)})\right)$$

Where b_{ik} is one of the M samples of the posterior distribution b_i and $\mu_{ijk}^{(t)}=e^{x_{ij}^T\beta^{(t)}+b_{ik}}$.

To sample from this distribution posterior distribution of b_i, we employed the metropolis hastings algorithm with a random walk. So at set t, for each new b_i, we considered $b_i^* = b_i^{(t)} + \epsilon$ where $\epsilon \sim U\left(-\frac{1}{4}, \frac{1}{4}\right)$. From this, the Metropolis Hastings ratio $R(b_i^{(t)}, b_i^*)$ was determined whether or not to accept this new b_i^*. For each new b_I^*

$$R(b_i^{(t)}, b_i^*) = \frac{f(y_{ij}|\theta, \beta, \sigma_\gamma^2, b_i^*) f(b_i^*|\sigma_\gamma^2)}{f(y_{ij}|\theta, \beta, \sigma_\gamma^2, b_i^{(t)}) f(b_i^{(t)}|\sigma_\gamma^2)}$$

Thus, the new value for the $b_i^{(t+1)}$ is:

$$\min\left\{1, \frac{P(b_i|Y, \theta, \beta, \sigma_\gamma^2)}{P(b_i^{(r)}|Y, 1/\theta, \beta, \sigma_\gamma^2)}\right\} \tag{5}$$

.

$$b_i^{(t+1)} = \begin{cases} b_i^* & \text{with probability } \min(R(b_i^{(t)}, b_i^*), 1) \\ b_i^{(t)} & \text{otherwise} \end{cases}$$

```
glmm_fit <- MCEM_algorithm( beta_initial = c(8.3, 1.5, 1.5, -0.25, -0.50, 0,
                                             0, -0.25, 0, 0, 0, 0, 0, -0.25),
                            theta_initial = 10,
                            s2gamma_initial = 0.2,
                            M = 1000,
                            burn.in = 200,
                            tol = 10^-4,
                            maxit = 100,
                            data = london_train
                            )
```

```
str(glmm_fit)
```

Machine learning models

Next, we applied two machine learning methods to predict the bike count. The first model we used is a conditional inference tree. In traditional partitioning, all possible splits are investigated to find the best split, which results in overfitting and selection. The conditional inference tree embeds the partition step with a permutation test, thereby enabling this method to be robust to covariates of different scales. Furthermore, it is capable of stopping when no significant correlation exists between the covariates and the response (Hothorn, Hornik, and Zeileis 2006). We emplyed the ctree funciton in the partykit to fit the conditional inference tree.

We visualized the three layer conditional inference tree in order to verify the relationship between bike count and variables. It is apparent that people tend to rent fewer bikes between midnight and 8 am. Maximum temperatures and humidity also affect bike rental rates. The splits that are employed in the conditional reference tree provide well reasoned explanations of the data structure, which means high accuracy predictions are made.

A random forest model was then applied to predict the number of bikes that would be rented. Random forest is composed of many decision trees as opposed to a single tree, resulting in a more accurate result. Incorporating the randomness allows random forest to protect against overfitting and can be applied more effectively to other data sets. The disadvantage of the random forest is its computational complexity. To fit the random forest, we created a method `train_random_forest()` within our `bikeSharing` R package that leverages the the `train()` function within the `caret` R package. The optimal tuning parameter mtry was determined using a 5-fold cross validation. We visualized a random selected tree.

```
rf_fit <- train_random_forest(data = london_train)
```

Random forest sample trees share similar splits with conditional inference trees. The estimated bike rental count is much higher in the hour chunks that are not 0-8am. Additionally, the estimated count is high when the temperature is more than 22.25 degrees.

Variable Importance in machine learning models

To investigate which variables affect the prediction most, we calculate the importance of the variables in the conditional inference tree and the random forest. In conditional inference tree, we calculated the mean decrease in accuracy when deleting a variable.

In random forest, we calculate the increase in MSE when deleting a variable. The code to do this was implemented as a function `plot_rf_importance` in the `bikeSharing` R package.

```
plot_rf_importance(london_train)
```

The top important variables of the conditional inference tree and random forest are similar. The hour chunk is the most important variable in both models. Additionally, the maximum temperature has a significant impact on the model prediction. Additional details of the variables are discussed in the discussion section. We further compare the performance of random forest and conditional inference tree using the second year bike renting data in London. Conditional inference tree results in a $R^2 = 0.81$ and random forest has $R^2 = 0.91$. Due the similarity in these two models, we selected random forest tree in the further comparison.

Results

```
glmm_model_fit(glmm_fit, london_train, scale_to_reference_mean = "no",
               reference = london)
glmm_model_fit(glmm_fit, london_test, scale_to_reference_mean = "no",
               reference = london)
glmm_model_fit(glmm_fit, dc, scale_to_reference_mean = "yes",
               reference = london)
glmm_model_fit(glmm_fit, seoul, scale_to_reference_mean = "yes",
               reference = london)
```

```
rf_model_fit(rf_fit, london_train, scale_to_reference_mean = "no",
             reference = london)
rf_model_fit(rf_fit, london_test, scale_to_reference_mean = "no",
             reference = london)
rf_model_fit(rf_fit, dc, scale_to_reference_mean = "yes",
             reference = london)
rf_model_fit(rf_fit, seoul, scale_to_reference_mean = "yes",
             reference = london)
```

Discussion and Conclusion

In conclusion, factors including hours, holidays, min temperature, max temperature, min humidity, max humidity, average wind speed and presence of rain/snow all have inference on the number of bike rented in the three cities. Hours from 4pm to midnight have the largest effect on the number of bike rented. Hours from 8am to 4pm, winter season, rain/snow also have large impact on the bike rent count. Random Forest have better prediction result on all the training and testing data, compared to the results of GLMM.

In addition to the conclusion above, we also have some findings that need to discuss more. The RMSE for Seoul is much larger than that in the other two cities. The results on the test dataset are much worse than in

the training dataset. This may imply that our model may have some limitations across cities. The possible reason is that in our model, we only considered the effects of days, hours and weather on the bike count but did not take information about cities into consideration (e.g. population, GDP), which may cause bias in prediction.

To get a better understanding on how our model predict on the sequences of bike count in different dataset, we plotted out the predicted bike count and the ground truth for each of the four datasets. Below is the longitudinal plots of the predicted bike count and the true bike count in the three cities.

To understand the bias and variance of our prediction results, we also drew the boxplots of predicted bike count and the true bike count of the four datasets. The plots are shown below:

Overall, our result is consistent with the published research by Sylwia et al. (2021) in "Impact of environment on bicycle travel demand-Assessment using bikeshare system data". In their study, they used the data from the bike sharing system in Cracow, Polan and conducted an ordinal least square regression model to analyze the effect of daily air temperature, daily rainfall, public holidays, and school holidays on the daily number of bike rented from bike sharing system. Their study result indicated that weather conditions, especially air temperature and daily rainfall, have large impact on the number of bike sharing system, which is consistent with ours result. Compared to their research, our study further found that the maximum temperature and the minimum humidity have more impact on the bike count. We did not restrict our study on daily level but cut one day into different hour chunks and found that different hour chunks in one day can also have large effect on the number of bike rented from bike sharing system.

Reference

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." Journal of Computational and Graphical Statistics, 15(3), 651–674. doi:10.1198/106186006X133933.

Sylwia P., Mariusz K., Carmelo D (2021). "Impact of environment on bicycle travel demand—Assessment using bikeshare system data." Sustainable Cities and Society, 67, [102724]. https://doi.org/10.1016/j.scs.2021.102724