Predicting Park Visitation in the Twin Cities Metropolitan Area in Minnesota Using Geolocated Social Media Data

 **Abstract**

 In this paper, building on “Geolocated social media as a rapid indicator of park visitation and equitable park access” by Hamstead et al. (2018)[3], we examine potential predictors for municipal park visitation in the Twin Cities Metropolitan area (TCMA) in Minnesota as approximated by geolocated social media (GSM) data from the Twitter and Flickr posts. We discover significant correlations between the GSM data and park area, length of foot trails, and demographic variables. However, our models for the Twitter and Flickr GSM data also exhibit significant differences, pointing the way for potential research into the effect of platform demographics on GSM.

1. **Background**

Policymakers have an interest in promoting equality of access to public parks. To aid in facilitating equitable access, we study park visitation rates in the Twin Cities Metropolitan area (TCMA) in Minnesota and determine factors that could contribute to accessibility. First discussed in [6], recent research has found that geolocated social media (GSM) data can serve as a useful measure of park visitation [4,5]. For example, GSM data can be used in modeling tourism to nature-based parks [6] and predicting national park and lake visitation rates [4,5]. Our primary purpose is not to validate the usefulness of GSM data as a predictor of park visitation, but rather to work from the assumption that the GSM data serve as a useful measure of park visitation and determine potential predictors of the GSM data in municipal parks in the TCMA.

Given the nature of the questions to be examined, a multiple regression model would be appropriate. In fact, there are many variables that could potentially influence the GSM count data in a particular park. We build on the conclusions of [3], which took the GSM methods developed in [4,5,6] and applied them to municipal parks in New York City (NYC). Because our research question is similar to that of [3], we hypothesized that the variables found to be significant in predicting visitation of the NYC parks in [3] would also be significant in predicting visitation of the TCMA parks. We also hypothesized that the presence of a dog park and length of foot trails within 0.25 miles of a park would be significantly positively correlated with visitation, while distance from the TCMA’s downtown areas would be significantly negatively correlated with visitation. All the response and explanatory variables considered can be found in Table A1 in the appendix.

1. **Methods**

**2.1. Data Collection**

Data were collected on 1581 parks in the TCMA, including characteristics of each park and the average number of posts on Twitter and Flickr per day geolocated at each park. Unique combinations of users and post dates on each platform, or Twitter-User-Days (TUD) and Photo-User-Days (PUD), respectively, are used as proxies for park visitation rates.

All Flickr data were obtained through Flickr’s Application Programming Interface (API). Flickr’s API provides geotagged data on photographs. Similar to the Flickr data, all Twitter data was obtained through Twitter’s Streaming API. Unlike the Flickr data, the Twitter data are only representative of a random 1% of Twitter’s publicly accessible tweets [3].

 Locations and dates of the Flickr/Twitter photos/tweets were recorded to represent when and where each photo/tweet was posted; it is important to note that a photo could have been taken and then posted on a later date, i.e., the date posted may not actually be representative of the true date the photo was taken. Also, while the Flickr data were available from 2005 to 2014, the Twitter data were only available from 2012 to 2014.

**2.2. Model Selection**

As discussed above, we hypothesized that the findings in [3] would be applicable to the TCMA parks. Therefore, in constructing our full model, we used a pre-selective process and combined the variables included in the final models of [3] with several other variables that we believed may have an effect on park visitation that were not included in their analysis. Then, we applied the Lasso, using the one standard error (“1se”) criterion implemented in the glmnet package in R [1] for the penalty parameter ($λ$) selection. We created two models, one for our TUD-based response variable (Model T) and the other for our PUD-based response variable (Model P) (see Table A2 in the appendix for some key statistics for the Lasso models).

**2.3. Variable Transformations**

 *Response variables*: The original TUD and PUD were significantly skewed, so we applied normalizing transformations to these variables. Specifically, we considered the log mean TUD per capita by taking the log of the mean TUD per year per surrounding neighborhood population, and treated PUD in a similar way (see Figures A1-A2). To ensure that the log transformation works, we removed zero observations prior to the transformation.

*Explanatory variables*: Mainly, the power transformations suggested by the Box-Cox method were applied to achieve symmetry of skewed variables. Also, we standardized our demographic variables (White, Black, and Hispanic population of the neighborhood surrounding a park) in such a way as to be robust to outliers using median and interquartile range (IQR) for fair comparisons. Table A1 in the appendix contains a full list of the variables and transformations we used.

1. **Results**

**3.1. Summary of Models T and P**

After model estimation and selection using Lasso, we have the following models whose coefficient estimates of the chosen variables are summarized below. Note that the p-values or confidence intervals are not reported for these estimates as the Lasso is known to produce biased estimates and thus the standard errors are not very meaningful [2].

|  |  |
| --- | --- |
| **Model T** | **Model P** |
| *Common Explanatory Variables* |
| log(Shape\_Area\_100m2) | 0.86355 | log(Shape\_Area\_100m2) | 0.16842 |
| $$\sqrt[3]{Dist2WtrMC}$$ | -0.06438 | $\sqrt[3]{Dist2WtrMC}$  | -0.22721 |
| Playgrnd\_dummy | 0.1827 | Playgrnd\_dummy | 0.23315 |
| $$\frac{PerWhite-median(PerWhite)}{IQR(PerWhite)}$$ | 0.10438 | $$\frac{PerWhite-median(PerWhite)}{IQR(PerWhite)}$$ | 0.16813 |
| pervacant | 0.32588 | pervacant | 0.29746 |
| Large\_Road | 0.03342 | Large\_Road | 0.01889 |
| $$\sqrt[3]{Dist2MSP}$$ | -0.19195 | $$\sqrt[3]{Dist2MSP}$$ | -0.38569 |
| Trl100mMC | 0.00426 | Trl100mMC | 0.0113 |
| DogPark\_dummy | 0.35785 | Dogpark\_dummy | 0.14126 |
| *Unique Explanatory Variables* |
| log(greenar) | -0.37448 | WaterCtMC | 0.003834 |
| HardCrt | 0.13624 |
| $$\frac{PerBlack-median(PerBlack)}{IQR(PerBlack)}$$ | 0.12365 |
| $$\frac{PerHisp-median(PerHisp)}{IQR(PerHisp)}$$ | 0.00862 |
| log(propvalue1000) | 0.00332 |
| $$\sqrt[3]{Dist2Bike}$$ | -0.00111 |
| $$\sqrt[3]{Dist2Bus}$$ | -0.16635 |
| log(PopDensity\_per100m2)  | 0.16308 |

As we hypothesized, almost every variable present in the TUD-based model of [3] is also present in Model T. Only two were eliminated, including the number of water features within 0.1 miles of each park (WaterCtMC) and length of bike routes within 0.25 miles of each park (Bike100mMC). Note that both had coefficients that were very close to zero in the model presented in [3].

Interestingly, while most of Model T’s coefficients have the same sign as the coefficients in the TUD-based model of [3], our model finds positive correlations between TUD and the Black and Hispanic populations of the surrounding neighborhood (PerBlack and PerHisp, respectively), whereas in [3], a negative correlation and no significant correlation, respectively, were reported (although it bears noting that the coefficient for our Hispanic population variable is small). This could indicate that disadvantaged communities have better access to parks in the TCMA than they do in NYC.

 Our hypothesis regarding the consistency of variable selection fared somewhat worse for Model P than it did for Model T. Five variables present in the PUD-based model of [3] are not present in Model P, including the Black and Hispanic population in the surrounding neighborhood and distance to the nearest public transit (Dist2Bus). The lack of demographic variables could again indicate that disadvantaged communities have better access to parks in the TCMA than they do in NYC.

 All of Model P’s coefficients have the same sign as their equivalents in Model T. However, there are some interesting differences in magnitude. For instance, the correlation coefficient between park area and PUD is roughly four times the correlation coefficient between park area and TUD. Moreover, the correlation coefficient between the presence of a dog park (DogPark\_dummy) and PUD is less than half the correlation coefficient between the presence of a dog park at TUD. Overall, PUD appears less responsive to its predictors than TUD; this likely reflects the limitations of PUD, as we will discuss in Section 4.

Lastly, both Models T and P supported our hypothesis regarding additional variables not discussed in [3]. Specifically, the presence of a dog park (DogPark\_dummy) and the length of foot trails within 0.25 miles of a park (Trl100mMC) are both positively correlated with TUD, while the distance from the TCMA’s downtown areas (Dist2MSP) is negatively correlated with TUD.

**3.2. Model Diagnostics**

It is crucial to check the model assumptions on the residuals, in particular, the constant variance and normality assumption, to ensure that the results obtained in the previous section are reliable. Looking at the residual plots for Models T and P (see Figures A3 and A5 in the appendix), the assumption of constant variance is not severely violated. On the other hand, the normality assumption may be violated as the histograms and Q-Q plots (see Figures A4 and A6 in the appendix) suggest that the distribution may be slightly skewed. In addition, the Shapiro-Wilk test returns a p-value of 4.479x10-4 and 6.201x10-9 for Model T and Model P, respectively, which amounts to significant evidence that the residuals may not be normally distributed. Nevertheless, the violation does not appear to be very severe from the histograms, so we conclude that our models may be useful in practice.

**4. Discussion**

Overall, our findings generally support our hypotheses. That is, our models are more or less consistent with the ones in [3]. However, we have also identified presence of a dog park, the length of trails within 0.25 miles of a park, and distance from the downtown areas as new key variables potentially affecting the park visitation. The magnitudes of the coefficients of our explanatory variables were sometimes smaller or larger than expected compared to [3], but only a few variables—most notably Black and Hispanic population in Model T—exhibited coefficients of a different sign than expected. Model P gave us the most surprises overall, containing fewer predictor variables than we expected, for reasons that it is somewhat difficult to glean. However, the fact that the assumption of normally distributed residuals is slightly violated for both Models T and P indicates that our findings should be treated with some skepticism.

 Moreover, the differences between Model T and Model P point to potential limitations of the GSM data as a proxy measure of park visitation. When we removed zeroes for our TUD and PUD response variables, we ended up with a dataset for PUD around 60% of the size of our dataset for TUD; this is likely a result of Flickr’s being a much less popular platform than Twitter, which could account for the generally lower-magnitude coefficients in Model P than in Model T. Moreover, there are potential demographic differences to consider between the users of Flickr and Twitter. Twitter is a significantly more “casual” social network than Flickr, and thus likely has a younger, poorer userbase, which could affect the kinds of parks that are posted about on each respective platform. Further examination of the effects of demographic differences between platforms’ users on GSM could be a promising avenue for future research.

**References**

[1] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1), 1—22.

[2] Goeman, J., Meijer, R., and Chaturvedi, N. (2018). *L1 and L2 Penalized Regression Models*. Package Version 0.9-51. Available at:

https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf

[3] Hamstead, Z.A., Fisher, D., Ilieva, R.T., Wood, S.A., McPhearson, T., and Kremer, P. (2018). Geolocated social media as a rapid indicator of park visitation and equitable park access, *Computers, Environment and Urban Systems* 72, 38—50.

[4] Keeler, B.L., Wood, S.A., Polasky, S., Kling, C., Filstrup, C.T., and Downing, J.A. (2015). Recreational demand for clean water: evidence from geotagged photographs by visitors to lakes, *Frontiers in Ecology and the Environment* 13(2), 76—81.

[5] Sessions, C., Wood, S.A., Rabotyagov, S., and Fisher, D.M. (2016). Measuring recreational visitation at U.S. National Parks with crowd-sourced photographs. *Journal of Environmental Management* 183, 703—711.

[6] Wood, S.A., Guerry, A.D., Silver, J.M., and Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation*.* *Scientific Reports* 3:2976, 1—7.

**Appendix**

|  |  |
| --- | --- |
| **Variable and Description**  | **Transformation Applied** |
| **TUD\_mean/TotPop**(Response variable for the mean TUD per capita from 2012 to 2014)  | **Natural Log**Applied to symmetrize the distribution; zeroes were removed beforehand to make the transformation work. |
| **PUD\_mean/TotPop**(Response variable for the mean PUD per capita from 2005 to 2014) | **Natural Log**Applied to symmetrize the distribution; zeroes were removed beforehand to make the transformation work. |
| **Shape\_Area\_100m2** (Park area in hundreds of meters squared) | **Natural Log**Applied to symmetrize the distribution. |
| **greenar** (Total greenspace area in meters squared) | **Natural Log**Applied to symmetrize the distribution. |
| **WaterCtMC** (Number of water features within 0.1 miles of park) | **No Transformation Required** |
| **Dist2WtrMC** (Distance to water bodies in hundreds of meters) | **Cube-Root Transformation**Applied to symmetrize the distribution. |
| **HardCrt** (Number of hard court features [i.e. tennis, basketball courts]) | **No Transformation Required** |
| **Playgrnd\_dummy** (Presence or absence of a playground) | **No Transformation Required** |
| **PerBlack** (Percentage of the neighborhood population that is Black) | **Robust Standardization**Applied for fair comparisons of demographic variables.(PerBlack-median(PerBlack))/IQR(PerBlack)  |
| **PerWhite** (Percentage of the neighborhood population that is White) | **Robust Standardization**Applied for fair comparisons of demographic variables.(PerWhite-median(PerWhite))/IQR(PerWhite)  |
| **PerHisp** (Percentage of the neighborhood population that is Hispanic) | **Robust Standardization**Applied for fair comparisons of demographic variables.(PerHisp-median(PerHisp))/IQR(PerHisp)  |
| **pervacant** (Percentage of vacant property in the neighborhood) | **No Transformation Required** |
| **propvalue1000** (Property value in thousands of dollars) | **Natural Log**Applied to symmetrize the distribution. |
| **Bike100mMC** (Length of bike routes within ¼ miles of park in hundreds of meters) | **No Transformation Required** |
| **Dist2Bike** (Distance to nearest bike route in hundreds of meters) | **Cube-Root Transformation**Applied to symmetrize the distribution. |
| **Dist2Bus** (Distance to nearest bike route in hundreds of meters) | **Cube-Root Transformation**Applied to symmetrize the distribution. |
| **Large\_Road** (number of interstate and state recognized roads with ¼ mile) | **No Transformation Required** |
| **Dist2MSP** (Distance to nearest downtown area in hundreds of meters) | **Cube-Root Transformation**Applied to symmetrize the distribution. |
| **PopDensity\_per100m2** (Neighborhood’s population density per hundred meters) | **Natural Log**Applied to symmetrize the distribution. |
| **Trl100mMC** (Length of trail within ¼ miles of park in hundreds of meters) | **No Transformation Required** |
| **DogPark\_dummy** (Presence or absence of a dog park) | **No Transformation Required** |

**Table A1**: Complete list of response and explanatory variables in our full model and transformation applied.

|  |  |  |
| --- | --- | --- |
| **Key Statistic for Lasso** | **Model T** | **Model P** |
| Penalty Parameter ($λ$) | 0.0138332 | 0.06018926 |
| Mean Squared Error (MSE) | 1.250257  | 1.295003 |

**Table A2**: Penalty parameter and MSE values for Model T and Model P using Lasso.



**Figure A1**: Histograms of mean TUD per capita and its log transformation.



**Figure A2**: Histograms of mean PUD per capita and its log transformation.

**Model T Residual Diagnostics**



**Figure A3**: Plot of Model T’s residuals vs. fitted values and order of observations (time order).



**Figure A4**: Histogram and Q-Q plot of Model T’s residuals.

**Model P Residual Diagnostics**



**Figure A5**: Plot of Model P’s residuals vs. fitted values and order of observations (time order).



**Figure A6**: Histogram and Q-Q plot of Model P’s residuals.