**Project Report 3: Estimating Yearly Visitors to Germantown Metropark**

Client Name: Five Rivers Metroparks

Meeting Time: N/A

**Background**

The client wants to estimate the total number of yearly visitors to Germantown metropark while utilizing as few traffic counters as possible. Currently each entrance to the park is equipped with a traffic counter. The client states that it is expensive to maintain all the counters. As a cost-cutting measure, the client would like to remove some of the counters while still estimating the total visitors to the park with accuracy of the actual value of the yearly total.

**Data**

The data consist of total number of monthly visitors to Germantown metropark from 2012 to 2017. There are seven different gates to the park and the number of visitors are broken down by gate. Nature Center gate was permanently closed in October of 2015 and is excluded from the analysis. But the number of visitors using that gate in previous years is included in the total visitor count for the final analysis.

**Method**

Total attendance is sorted by quarter (Jan-March, April-June, July-Aug, Sep-Dec) per gate per year. Total visitors to the park for each quarter is separately calculated. *Leave One Out Cross Validation (LOOCV)* process with simple linear regression in statistical software package R is utilized to carry out the analysis with quarterly total attendance for the park for each year as the dependent variable and quarterly attendance for each gate for each year as the independent variable. Separate models are implemented in the *training* process for each quarter and each gate totaling 24 different models (4 quarters per year and 6 different gates).

In *LOOCV*, data is divided into subsets and a single subset is designated as validation data while the remaining subsets are used as training data. This process is repeated such that each subset is used once as the validation data. The goal of this cross-validation process is to test the model’s ability to predict new data that was not used in estimating it.

In this project, each iteration of *LOOCV* for each gate consists of designating that gate’s quarterly attendance for a specific year (independent variable) and that gate’s total quarterly attendance for all years (dependent variable) as validation data and using the data for same two variables from all other years as training data. *Mean Absolute Error (MAE)* from each training iteration is divided by average quarterly park attendance from the respective quarter to calculate *average* *MAE*. This value is utilized to measure the accuracy of the prediction.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gate | Q1 | Q2 | Q3 | Q4 | Total |
| Main Park | 0.08 | 0.06 | 0.10 | 0.06 | 0.07 |
| Trail Center | 0.15 | 0.19 | 0.14 | 0.09 | 0.15 |
| Dam Parking | 0.11 | 0.11 | 0.13 | 0.16 | 0.13 |
| Sled Hill | 0.23 | 0.11 | 0.13 | 0.13 | 0.14 |
| NC Entrance | 0.10 | 0.08 | 0.21 | 0.06 | 0.12 |
| Old Mill | 0.05 | 0.14 | 0.14 | 0.15 | 0.13 |

**Table 1: Mean Absolute Error per Average Attendance**

**Results**

As shown in figure 1, Only the Main Park gate has average MAE within accuracy. Thus, a simple linear regression model with Germantown Park’s total quarterly attendance for each year as the dependent variable and Main Park gate’s quarterly attendance for each year as the independent variable is utilized to make predictions on the total yearly attendance.



**Figure 1: MAE per Average Attendance**

Simple linear regression models for each quarter are as follows:

Quarter 1:

Quarter 2:

Quarter 3:

Quarter 4:

To calculate the total park attendance for a specific quarter, substitute the number of visitors using the Main Park gate in that quarter to and solve for . This value is the total park attendance for that quarter with error. Adjusted R2 in this case represents the proportion of variance for a dependent variable that’s explained by the independent variable.