Abstract

Arlington County’s Multi-use trails (MUT’s) are widely considered some of the best in the nation in terms of sheer breadth and usability. However, no real study has been conducted on exactly how the MUT’s are used by Arlington’s population. The county currently funds and manages the MUT’s as recreational facilities, which means that their maintenance and general management is at a lower priority than Arlington’s roads. However, the MUT’s also serve as key elements of the nonmotorized transportation network for many bicycle commuters. As a result, the County’s treatment of the resource as recreational may lead to misplaced design and management priorities. For example, after heavy snowfall, the County can take up to a week to clear the MUT’s, with a notable impact on commuting cyclists. Obtaining better understanding of trail use patterns can help Arlington County better allocate funds and resources to particular trail segments, while better understanding of location specific patterns can help the County better prioritize maintenance activities during different times of the year.

Hypothesis/Goal

The 17 different trail segments monitored by the County are spread throughout Arlington, ranging from the less dense, more residential areas in the western area to the dense, urban core in Rosslyn that is closely linked with Georgetown in the District of Columbia. The goal of this project is to determine which Census-based features of our dataset most strongly correlate with overall usage patterns. That data could then be used by Arlington and other jurisdictions to determine what other areas are prime candidates for new MUT construction. While weather is expected to be the strongest factor influencing overall traffic patterns, our goal is to determine which secondary factors most strongly influence bike traffic: area demographics, per capita income, time of day, holidays and other work interruptions.

Dataset

The data used to test the hypotheses consists of three categories: sensor data, weather data, and census data.

Sensor Data:

The raw sensor data was from 17 different sensors spread across the county. While most of the sensors could differentiate cyclists and pedestrians, one sensor (Four Mile Run) was only able to provide undifferentiated data, so for our analysis of total bike traffic, this sensor was excluded from the sample population.

The raw data was separated into datasets for each individual sensor, and required extensive pre-processing and cleaning to normalize and combine. Please see the R-file in the project folder to see the exact work that went into that.

The outputted data from that process had these characteristics:

Date, Time (15 minute intervals), Total (Bike and Pedestrian total for either direction), PedIN (Pedestrians IN, as determined by sensor’s direction settings), PedOUT, BikeIN, BikeOUT, Location (sensor location name), CounterID (number corresponding to sensor), PedTOTAL (total count of pedestrians), BikeTOTAL(total count of bicycles, this is our main output that we are predicting for, based on guidance from my Virginia Tech partner), TotalOUT, and TotalIN.

Weather Data:

A ground-level assumption for this study is that weather accounts for the most variation in total bike traffic, particularly in a temperate, widely varying climate such as Arlington, VA. While weather is not a factor/feature that we are studying, it is a huge factor that must be taken into account to determine other contributing factors. The weather data was gathered from Weather Underground and was combined into a single file for 2009 to 2013 and merged with the bike data via R’s rbind command (see processing file). The outputted weather data had the following characteristics:

EST(Date), Max, Mean, and Min Temperatures, Max, Mean, and Min Dew Points, Max, Mean, and Min Humidity, Max, Mean, and Min Air Pressure, Max, Mean, and Min Visibility, Max, Mean, and Min Windspeed, Max, Mean, and Min Wind Gusts, Precipitation, Cloud Cover, Events (rain, sleet, snow, etc), and Wind Direction.

Census Data:

The Census’s annual household surveys and ten-year Census data provide the foundation of data we are comparing against based on sensor location. By comparing total bike traffic to location attributes including household income, total population, education levels, rentals vs owner occupied levels, etc we hope we can find the characteristics of an area that would make it most likely to heavily utilize a bike path. This could help Arlington County ensure that any new MUTs it constructs will provide a strong return on investment through citizen engagement and satisfaction. The Census data provided has the following characteristics:

Counter ID, Model(type of sensor), Population (of area near counter), CycleComm (number of cycle commuters according to Census Household Survey), CycleCommM (number of male cycle commuters), CycleCommF, InLaborForce (number of people in labor force), NotInLaborForce, TotalInc (total income of all respondents summed together), HousingUnits (total number of housing units), UnitsOcc (number of occupied units), UnitsVac (number of vacant units), OwnerOcc (number of units occupied by owner), RenterOcc (number of units occupied by renter), MovedIn2010-12 (lived in area less than three years), MovedIn2000-09 (moved in within last decade), Vehicles0 (households with no vehicles), Vehicles1 (households with 1 vehicle), Households, NumStudents3up (number of children age 3 and up), Hsgrad (number of High School graduates), Bachelor (number of highest education Bachelor’s), Master (number of people highest education obtained Graduate), Foreign (number of foreign born), Noncitizen (number of non-US citizens), IncomePerCap (income per capita), MeanHousInc (mean household income).

Approach:

After initial combining, the data is further processed into a normalized matrix measuring 4380 rows by 74 columns based on cumulative time slices for each day: Midnight to 7AM, 7AM to 10AM, 10AM to 4PM, 4PM to 7PM, and 7PM to Midnight. This was designed to capture “rush hour” for commuting days in two distinct slices - 7 to 10AM, and 4 to 7PM.

We then explore the data using linear regression and summaries. Eventually we will test the prediction using the “bagging technique”. But for now, I’m still in the exploration phase to determine which Census factors most strongly predict for overall bike traffic…