New York City Bus Data: Methodology

The Open Bus
2016

1 Introduction

Many public bus systems around the world have begun to equip buses with GPS locators, providing real time locational data for some or all vehicles in the system's fleet. Such data streams are normally used in real time, to allow users to know when the next bus will arrive. The Open Bus (TOB) repurposes this data in order to provide retrospective statistics on bus performance. Collected GPS data enables the public to access extremely rich information related to the regularity with which buses arrive. Data can be compared against publicly released schedules so that reliability can be calculated. This document provides a summary of the methodology used to collect and process bus data for New York City.

2 Data Collection

GPS data is observed at the bus stop level. A looping code automatically checks every bus stop across New York City approximately every 3 minutes. At each check, the system returns a binary observation: 1 if there is a bus at, or approaching the stop; 0 otherwise. This system provides panel data for every bus stop in New York City, with timestamps indicating the time the stop was observed.

Data collection began on March 1st, 2015. Data collection has been executed continuously since that date. Collection relies on the proper operation of GPS equipment, public servers, and internet connectivity, meaning interruptions in any of this infrastructure will cause periods of incomplete data. These events are rare and are easily identifiable in the data as periods void of any observations.

3 Processing

The 3 minute gap between observations allows for the possibility that some buses will be "missed" as they pass stops. Consider the following example, three consecutive stops (stops 1, 2 and 3) are observed, at 1:00 pm stop 1 is observed to have a bus present but stops 2 and 3 do not. The next observation is 1:03 pm, at which time stop 3 has a bus present, while stops 1 and 2 do not. This suggests the bus must have passed by stop 2 at some point between 12:00 and 12:03 pm. Such instances are "back-filled" such that a bus is attributed to stop 2 at 12:00 pm. In theory, this method allows data to be nearly perfectly extracted, with only slight error in imputing the exact time a bus arrived at stops for which the precise time was not observed directly. This method also allows data collection to be robust to momentary lapses in GPS infrastructure as such instances can be back-filled.

4 Web Interface

The web interface provides averaged statistics for one month intervals, with the month indicated on each page. Data displayed on web pages is drawn exclusively from Monday to Friday, 6 am to 10 pm. Average headway is calculated by taking the average time elapsed between buses for each stop, and then averaging across stops.

5 Raw Data

Raw data will be posted in the first week of each month, providing data collected in the previous month. Raw data is provided in .csv format. Raw data includes observations covering 24 hours per day, 7 days per week within the month indicated. There is an observation for every minute in which a bus was recorded somewhere along the route. Observations are rounded to the closest minute. Times are represented by Unix time, which represents the number of elapsed seconds since 12:00 am, January 1, 1970. Each variable represents a unique stop along the particular route. For any stop at a particular time, a 1 represents a bus at, or approaching the stop at that time, a 0 represents no bus present at that time.

Stop IDs are assigned by the Metropolitan Transportation Authority. Stops are ordered consecutively, as they would appear when riding a bus. Stop IDs can be spatially referenced using the MTA's Bus Time service.

6 Raw Data Codebook

time : The number of elapsed seconds since 12:00 am, January 1, 1970
stop_* :

1 indicates a bus was either approaching or at the stop

0 otherwise

* = Bus stop id number, as assigned by the MTA