

# Sage and One Bus Away

Tim Willis, Kevin Chu, Alli Adamonis, Dan Meyering

*Abstract: To use King County Metro's data in order to provide a statistical analysis of certain routes and stops on-time percentage using the statistics package inherent in Sage as well as various imported libraries. Our goal is to clearly display applicable graphs, interacts and statistics of the historical Metro data for use by Metro itself and One Bus Away users and developers.*

## Introduction

As college students, more often than not we find ourselves taking the bus. The University of Washington is such a densely populated area that it is difficult to find parking and with the U-Pass it is simply more convenient to take the bus to campus than pay more for parking. One tool that has become indispensable for most bus riders is a UW project called "One Bus Away" (OBA). OBA uses King County Metro's (KCM) real-time Google Transit Feed Specification (GTFS) feed to monitor where any given



bus is at a certain time. Though possibilities are endless with OBA, it immediately makes wondering where your bus is and when it will finally come less of a daunting question. OBA has been implemented in various interfaces (Web, Phone, SMS, native iPhone and Android apps) making it extremely easy to use and it also contains a set of tools for developers wanting to experiment with the program.

While OBA will tell you where a certain bus is, it does not track what percentage of buses are early, on time or late to certain stops. This

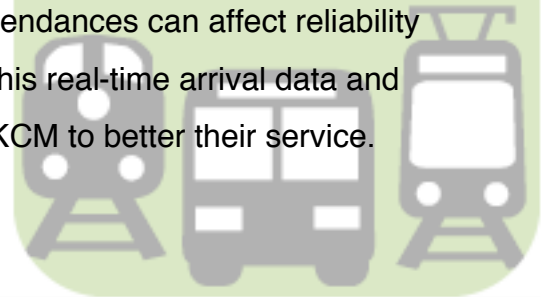
functionality could be particularly useful for transit agencies in order to see how outside factors affect service reliability. Rush-hour trips and traffic are obvious hindrances to reliability, but concerts and other events that have high attendances can affect reliability as well. Using OBA, Python and Sage, we hope to index this real-time arrival data and draw statistical conclusions from it that could be used by KCM to better their service.

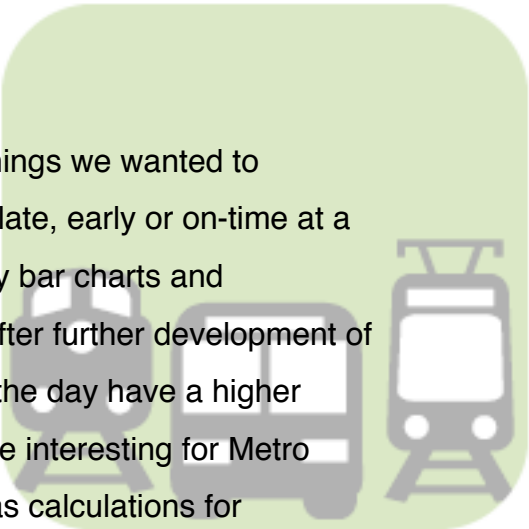
### ***Methods and Mathematics***

As mentioned prior, OBA provides various services for developers hoping to experiment with the code. In addition to implementing a Java interface, it also maintains a REST server that has various methods including “arrivals-and-departures-for-stop” and “current-time” which return the current arrivals and departures for a given and the official KCM time, respectively. Other methods are included as well, but those two are what we use in order to implement our code. Through querying the server at intervals of 3 mins, we’re able to construct a comprehensive database of arrival times that can be analyzed for later use. This database is stored in a text file and compiled over some long period of time, be it a day or a week.

Once the database is completed, in order to read the data into Sage or statistical software, we need to convert it from its current format of concatenated XML tags. This is done through simple string modification, line by line, to create a couple of meaningful lists. These lists can then be read into Sage’s comprehensive statistical functionality.

The main mathematical ideas we used were from statistics in junction with the Sage Notebook to assess the KCM data. By implementing different statistical analysis tools we were able to numerically and graphically represent the data collected in an attempt to correctly describe how many and when busses were likely to be on time, early or late. Statistics can be easily altered and biased so we wanted to test our conclusions with varying statistical methods including scatterplots, bar charts, and confidence intervals in order to maximize the strength of the conclusions we found.





When initially deciding on this project one of the first things we wanted to accomplish was being able to calculate the number of buses late, early or on-time at a particular stop. Using the historical data we created frequency bar charts and histograms in order to visually represent the data. However after further development of the project, our group thought that finding out which times of the day have a higher probability of buses being late, early or on-time would be more interesting for Metro travelers as well as Metro itself in many different ways such as calculations for anticipated arrival times for future bus schedules. Therefore we created a scatterplot that when the user chooses a time interval, it depicts what busses were late and which ones were early, giving a good representation of the probability a bus will be on time. Another useful statistical method we implemented was a confidence interval for the mean scheduled minus actual arrival time for a particular stop. The confidence interval calculates an interval for the approximate expected arrival of the bus with a certain amount of confidence, giving a good estimate for how late or early the bus will arrive.

### ***Future Projects and Potential Research***

For our project we utilized the Sage Notebook and although the Sage Notebook has an immense amount of features that can immediately be utilized or easily accessed or imported into it, we did find some issues and came up with some additional functionality we would love to see implemented into the notebook. In particular, since our project revolved around statistics in the notebook, we would absolutely like to see easier interaction among the statistical packages that can be imported into Sage such as SciPy (Pylab), MatLab, and RPy and the Sage Notebook interface and its implicit statistical functions. Consequently, after importing libraries into a Sage Notebook Worksheet, the tendency for the worksheet to not always recognize that imported package caused unnecessary errors and the need to continually import the same packages which is frustrating especially for basic programmers. Also, the statistical functions implicit in Sage tend to be extremely basic with restrictions on even labeling the axis or titling a graph. Further development of the statistical capabilities in the Sage

Notebook would be extremely helpful for existing users and would attract new users who natively use other math software.

In hindsight, the way we implemented this project could have been more integrated with OBA. Instead of having to wait however long to compile a database of arrival times, simply using already stored data from KCM might have been a bit easier. OBA has the database functionality built-in, it is simply a matter of having them get back to you after you contact them to get it set up. Some of the processes that seem a bit daunting to us might be simpler that way. With such comprehensive database functionality, it would be nice to analyze data from any given stop or route based on a certain parameters, like time of day, day of the year, etc. The ultimate goal would be to build something that KCM would want to use to research the on-time percentage of their busses, somewhat of a “missing link” between all the data that OBA provides and what KCM could do to better their service.

