Data Science for Social Good - Summer 2017

CAN TRAFFIC SENSORS DETECT VEHICLE CRUISING?

Brett Bejcek — DSSG Fellow
Anamol Pundle — DSSG Fellow
Orysya Stus — DSSG Fellow
Mike Vlah — DSSG Fellow
Valentina Staneva — Data Scientist
Vaughn Iverson — Data Scientist
Steve Barham, City of Seattle — Project lead

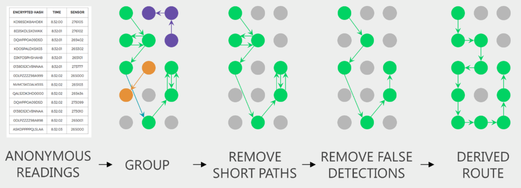
PROJECT GOALS

1. Identify Aggregated Cruising
2. Differentiate Parking and For-Hire Vehicles
3. Protect Privacy

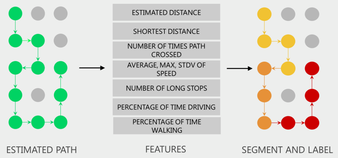
Vehicle cruising by cars that are searching for parking or waiting to be hailed contributes to significant congestion on city streets but is not easily measurable.

DATA PROCESSING AND CLASSIFICATION

1 ESTIMATE PATHS The first step in our analysis is to derive probable routes from the scattered sensor readings. Identifying precise trips or paths would be very difficult to accomplish due to the data consistency and reliability issues. We estimate path information through a series of algorithms that group readings, remove paths that are too short to process, and clean false detections, resulting in a derived route that is mapped the actual street grid in Seattle.



2 EXTRACT METADATA Metadata features are extracted from the estimated paths. These represent travel characteristics that can be used for classification and machine learning.



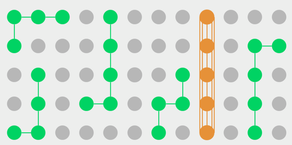
3 LABEL VEHICLES-FOR-HIRE We differentiate vehicles-for-hire from other traffic using a simple algorithm, prior to machine learning. If a vehicle leaves the sensor grid (is not detected) for more than 15 minutes, and if it does so more than 5 times throughout a day, it is likely either a for-hire vehicle or a bus. Bus drivers can then be filtered out by their "dispersion ratio" (total number of times a vehicle is detected divided by the number of unique sensors that detected it). If this ratio is around 1, the traveler did not cover the same ground many times, and so cannot be a bus driver.

FOR-HIRE VEHICLE EXAMPLE

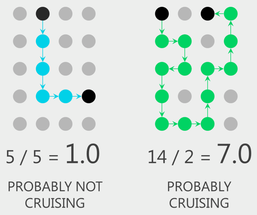
4 LARGE GAPS IN READ TIMES (5 TRIPS)
UNIQUE SENSORS / TOTAL READS =
22 / 22 = 1.0 [HIGH DISPERSION]

BUS EXAMPLE

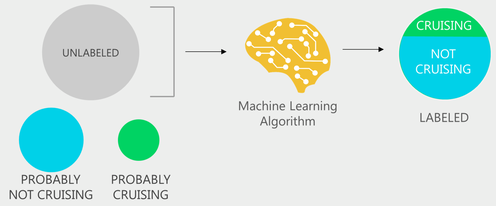
4 LARGE GAPS IN READ TIMES (5 TRIPS)
UNIQUE SENSORS / TOTAL READS =
5 / 25 = 0.2 [LOW DISPERSION]



4 SEMI-SUPERVISED MACHINE LEARNING Cruising classification was completed in multiple steps. First, the distance ratio (shortest distance between start and end sensor hits / routed distance) was used to classify the cases as either probably cruising or not probably cruising. As illustrated below, a trip that follows the most direct path is probably not cruising, and a path that meanders significantly is probably cruising. A distance ratio of 7.0 suggests that the path traveled 6 times more than necessary, contributing to traffic and congestion along the way.



Approximately 40% of the segments were labeled probably not cruising and approximately 13% were labeled probably cruising. This left 47% of the data unlabeled. We trained the labeled data using three models- decision trees, logistic regression, and gradient boosting classifier- to identify which other features aside from distance ratio are meaningful in classifying a trip as cruising or not cruising. We found that the gradient boosting classifier has the best model consistency, accuracy, AUC ROC, precision, recall, and f1 score. Using this model, the remaining 47% of data was labeled.



5 AGGREGATE FOR HEATMAP We developed aggregation scripts to summarize the results. To protect privacy, only the aggregate data stream and heatmap would be available for analysis. We developed a web app to demonstrate potential uses. The algorithms and data would be a potential candidate for use in a linked data repository with strong governance, such as the University of Washington Transportation Data Collaborative.

CASCADIA URBAN ANALYTICS COOPERATIVE

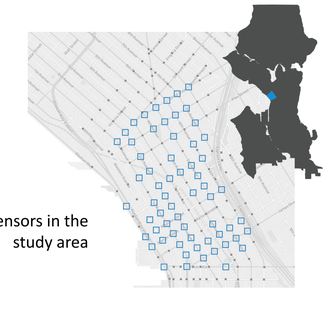
Seattle Department of Transportation

Microsoft

SAMPLE TRAFFIC ANALYSIS SENSOR DATA

The project leverages and repurposes existing traffic analysis sensor data used to calculate travel times and help the city optimize the traffic signals along important corridors. The traffic analysis sensors detect unique identifiers of mobile devices, which are hashed (anonymized) and salted (anonymized differently daily), allowing devices to be paired among locations for the day. There are over 200 sensors in Seattle, however for the purposes of this study, we only used sample data from 64 sensors in the downtown central business district for one-week period.

INCOMPLETE GRID	BIG DATA	SENSOR DETECTION
SENSORS ONLY COVER **37%** OF THE GRID	SENSORS PRODUCE **200K** OBSERVATIONS / HR	SENSORS DETECT **38%** OF DEVICES W/ WIFI ON



Sensors in the study area

RESULTS

Approximately 35% of the sample data was labeled as cruising and visualized. Of that amount, activity attributed to vehicles-for-hire is in the range of 10% or fewer. We found that the number of time crossed, average speed, and percentage of time driving were the most important features considered by the gradient boosting classifier.