Seattle Department of Transportation

CAN TRAFFIC SENSORS DETECT VEHICLE CRUISING?

University of Washington eScience Institute
Data Science for Social Good Project
Summer 2017

PRELIMINARY
FINDINGS
DRAFT 8/23/2017

Seattle
Department of
Transportation

PROJECT SUMMARY

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. This project demonstrates the ability to identify relative prevalence of vehicle cruising within a sample of anonymous traffic analysis sensor data. The project uses data science techniques to process, analyze, model, and aggregate large, noisy data into reproducible cruising metrics.

We created a processing and classification pipeline which labels approximately 35% of total discernable data as cruising. Of that amount, activity attributed to vehicles-for-hire is in the range of 10% or fewer. These preliminary results are aggregated to block segments and hourly time periods to generate cruising heatmaps.

This project was conducted through the summer Data Science for Social Good (DSSG) program and led by Steve Barham from the Seattle Department of Transportation. The University of Washington eScience Institute hosts the program, bringing together students and researchers to tackle real-world data science projects that have a social good impact. The program is supported by generous funding by the Microsoft Corporation in the context of the Cascadia Urban Analytics Cooperative and the Alfred P. Sloan Foundation. The City, four full-time data science fellows, and two data science faculty at the eScience Institute worked together over a 10-week period.

Project Lead	Steve Barham, City of Seattle
DSSG Fellows	Brett Bejcek, Anamol Pundle, Orysya Stus, and Mike Vlah
Data Scientists	Valentina Staneva and Vaughn Iverson

BACKGROUND

The City of Seattle seeks to improve travel reliability, optimize the use of the right-of-way, improve the parking experience, reduce emissions, and lower transportation costs. We know that cruising can have a significant impact on congestion, but very few data exist for the City to analyze. This project seeks to create new traffic analysis data by identifying two types of cruising, anecdotally known to be significant.

Cruising-for-Parking

Vehicles searching for parking spots after they have already arrived at their destination can contribute to a significant portion of traffic in congested areas. A recent study by INRIX suggests that Seattle drivers spend 9 minutes per trip, and 58 hours per year, searching for parking.

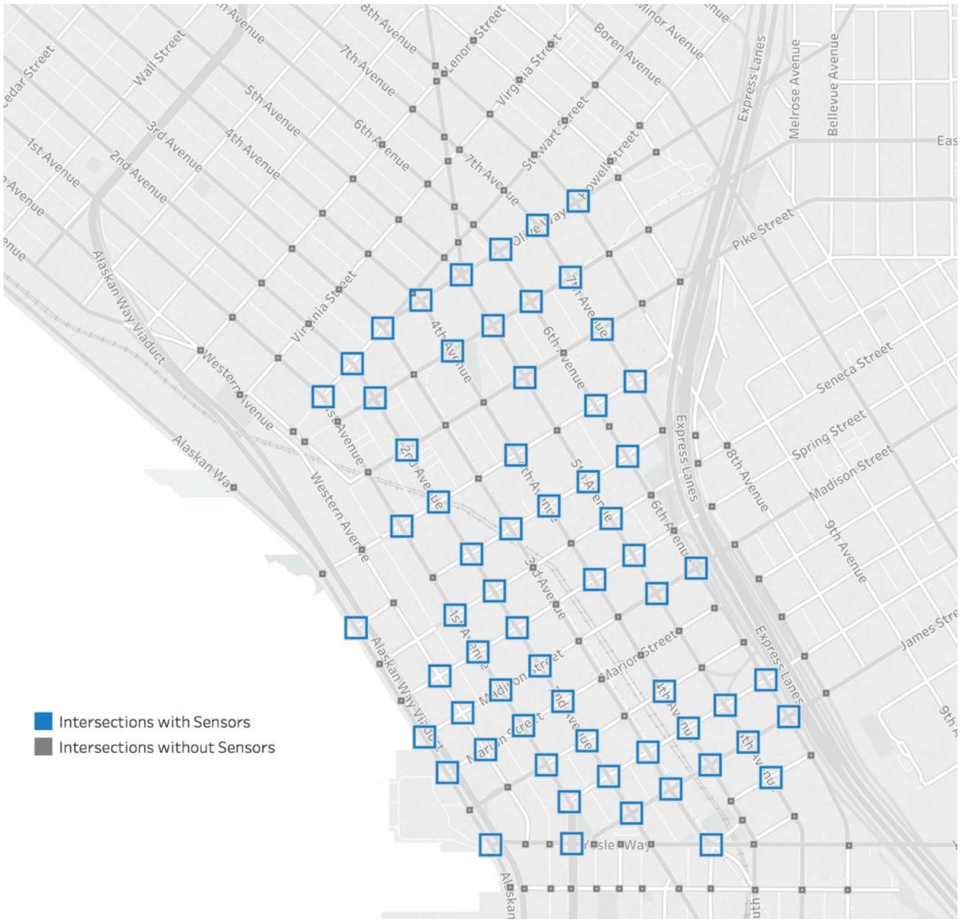
Cruising-for-Hire

Vehicles for-hire (taxis, for-hire vehicles, and app-based Transportation Network Companies) queue in motion, waiting to be hailed, or on the way to pick up another passenger, also known as dead-heading. For every for-hire trip, there is inevitably a measurable amount of travel without a passenger, or deadheading. A study by Bruce Schaller estimated that Taxi dead-heading in New York adds 7-8 miles for every 10 miles of fare trips, and TNC dead-heading adds 12-13 miles. TNCs do not share detailed trip data, but it is conceivable that deadheading may be in the interest of the companies because when there are more cars queued on the streets, the pickup wait times for passengers will be lower.

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. The sensors cover 37 percent of the grid within our study area.

The data from traffic analysis sensors are noisy and scattered. Due to the variability and unreliability of the data, recreating precise trip information from the anonymous sensor readings is not feasible. Sensors cannot pinpoint exact locations, only that the detected device is within the sensor range (up to 2 blocks), thus the data set includes many readings that are "false positive". Our analysis also suggested that the sensors only detect signals approximately 38% of the time, thus the data set excludes may readings that are "false negative".

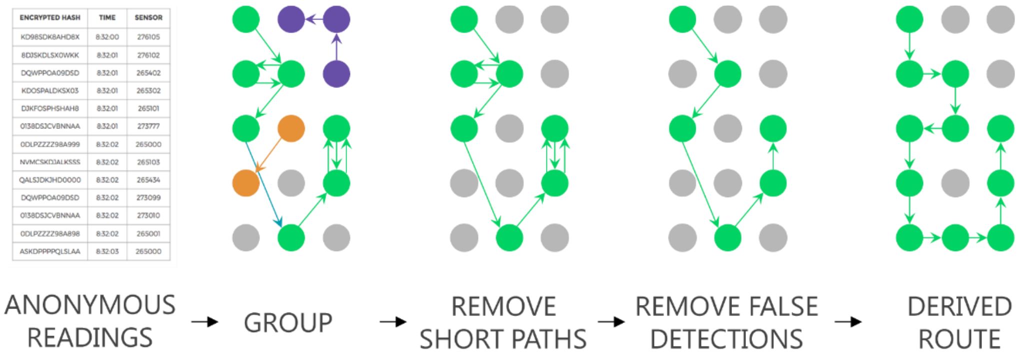


Study Area Sensor Grid

DATA PROCESSING AND CLASSIFICATION

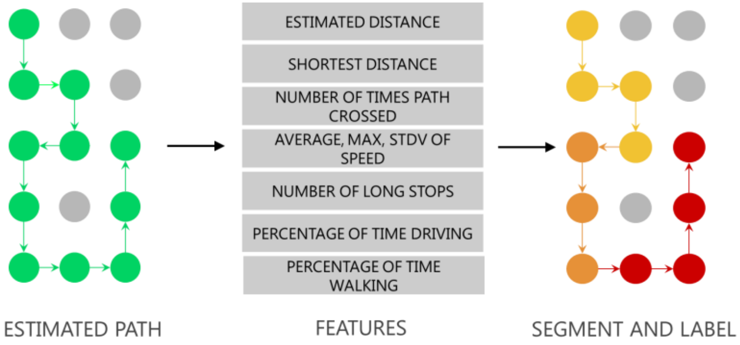
Despite the technical challenges with the sample data set, we have found that traffic analysis insights can be gained by performing data cleaning, analytical operations, and data science operations.

1) Estimating 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. WeWe 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 (one-way and two-way) in Seattle.



ENCRYPTED HASH	TIME	SENSOR
KD98SDK8AHD8X	8:32:00	276105
8DJSKDLSX0WKK	8:32:01	276102
DQWPPOA09DSD	8:32:01	265402
KDOSPALDKSX03	8:32:01	265302
DJKFOSPHSHAH8	8:32:01	265101
0138DSJCVBNNAA	8:32:01	273777
ODLPZZZZ98A999	8:32:02	265000
NVMCSKDJALKSSS	8:32:02	265103
QALSJDKJHD0000	8:32:02	265434
DQWPPOA09DSD	8:32:02	273099
0138DSJCVBNNAA	8:32:02	273010
ODLPZZZZ98A898	8:32:02	265001
ASKDPPPPQLSLAA	8:32:03	265000

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



ESTIMATED DISTANCE
SHORTEST DISTANCE
NUMBER OF TIMES PATH CROSSED
AVERAGE, MAX, STDV OF SPEED
NUMBER OF LONG STOPS
PERCENTAGE OF TIME DRIVING
PERCENTAGE OF TIME WALKING
ESTIMATED PATH
FEATURES
SEGMENT AND LABEL

3) Vehicle-for-Hire Labeling. 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 =

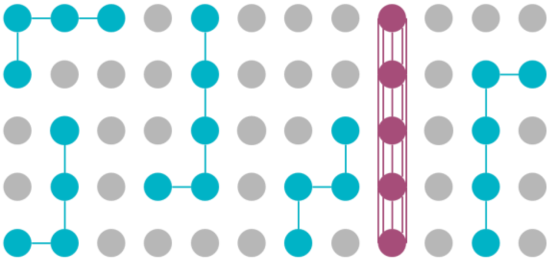
22 / 22 = 1.0 [HIGH DISPERSION]

BUS EXAMPLE

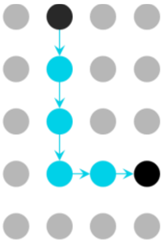
4 LARGE GAPS IN READ TIMES (5 TRIPS)

UNIQUE SENSORS / TOTAL READS =

5 / 25 = 0.2 [LOW DISPERSION]

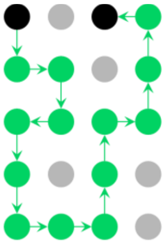


4) Multi-Step Classification, Semi-Supervised Machine Learning. Cruising classification was completed in multiple steps. First, the distance ratio (shortest distance between start and end sensor hits / routed distance) was used to classify the cases as either probably cruising or not probably cruising. 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.



5 / 5 = 1.0

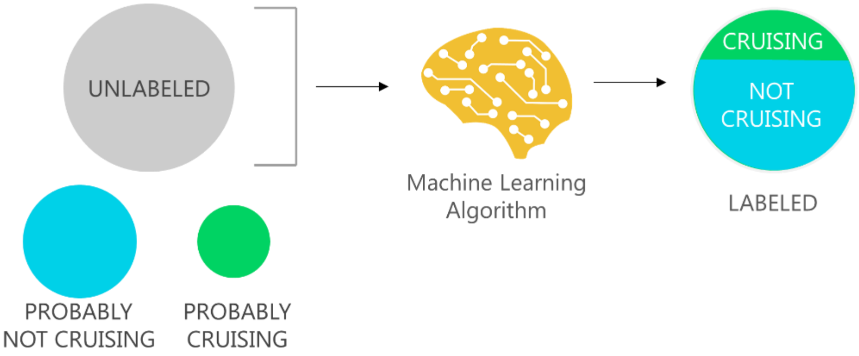
PROBABLY NOT CRUISING



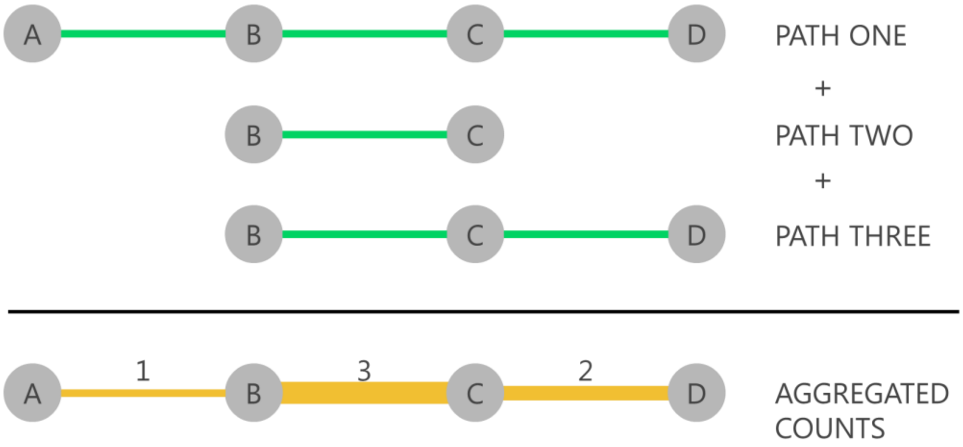
14 / 2 = 7.0

PROBABLY CRUISING

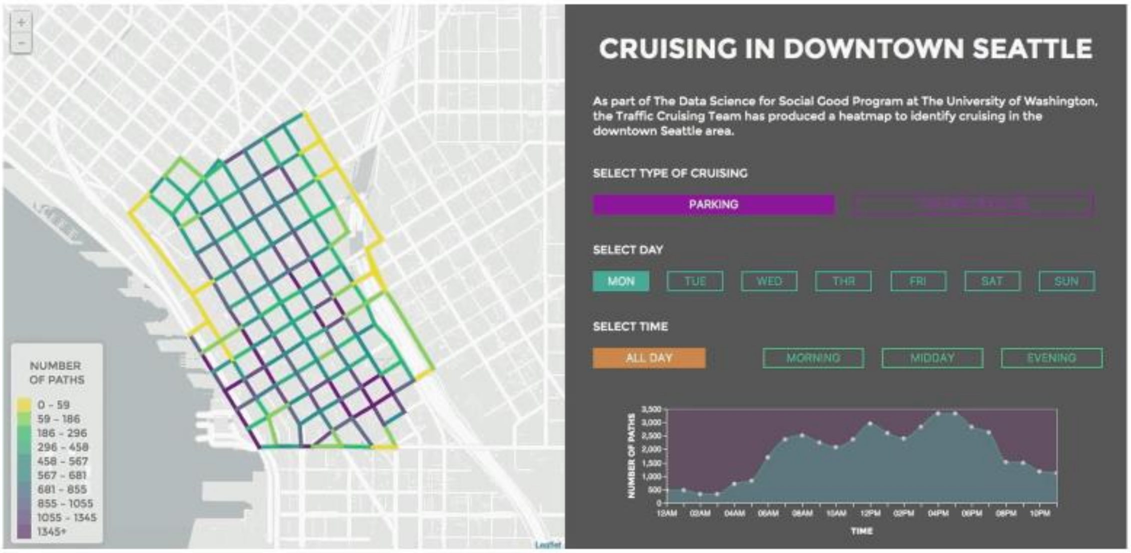
We trained the labeled data using three models- decision trees, logistic regression, and gradient boosting classifier- to identify which other features aside from distance ratio are meaningful in classifying a trip as cruising or not cruising. We found that the gradient boosting classifier has the best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



5) Aggregation and Heat Map. We developed aggregation scripts to summarize the results. To protect privacy, only the aggregate data stream and heatmap would be available for analysis and potential public consumption.



We developed an interactive web map to show relative amounts of cruising on a street-by-street level during different times of day and different days of the week. Users can visualize different days of the week, times of day, and differentiate between cruising for parking and cruising for hire.



CRUISING IN DOWNTOWN SEATTLE

As part of The Data Science for Social Good Program at The University of Washington, the Traffic Cruising Team has produced a heatmap to identify cruising in the downtown Seattle area.

Cruising Heat Map Web Application

PRELIMINARY RESULTS

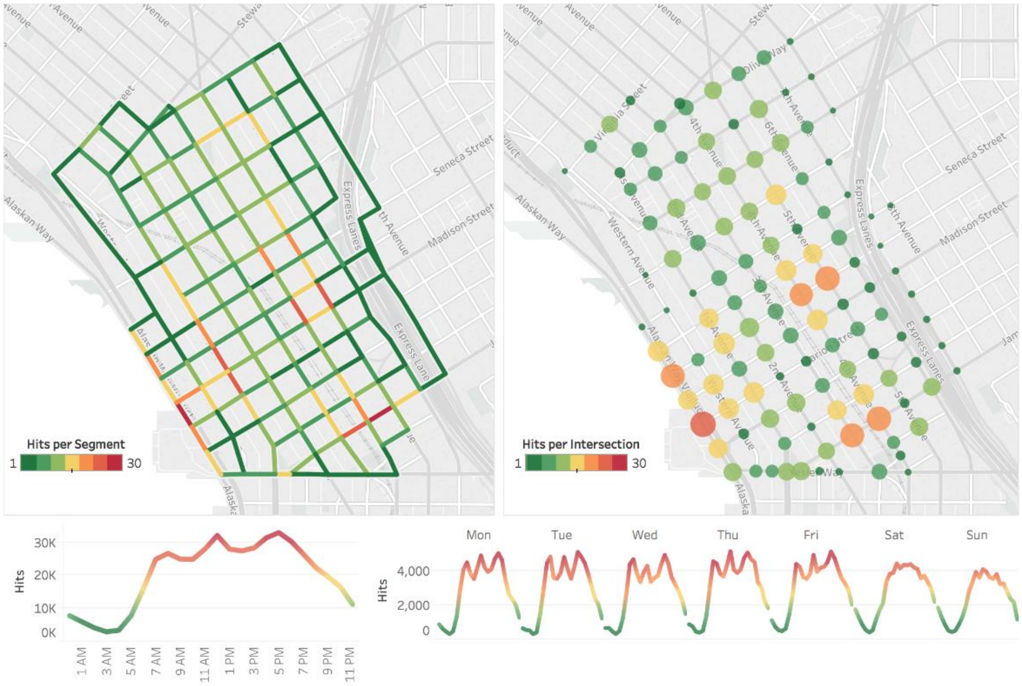
Approximately 35% of the sample data was labeled as cruising and visualized. Of that amount, activity attributed to vehicles-for-hire is in the range of 10% or fewer. We found that the number of time crossed, average speed, and percentage of time driving were the most important features considered by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

The intensity of cruising for parking as a proportion of total traffic fluctuates with time of day. It follows a similar pattern to proportional parking occupancy data collected by SDOT manual surveys. Cruising in the Central Business District exhibits expected triple peak weekday patterns (spikes during mornings, lunch, and evenings commutes).

Preliminary results for the sample week are visualized below. These charts illustrate the spatial and temporal variance. An expanded study period will be needed to baseline and index the cruising measurements. It is noted that the outer edges of the study area will not register the full volume of cruising because of vehicles traveling outside of the grid. We will attempt to expand the boundaries of the study area in future analysis.

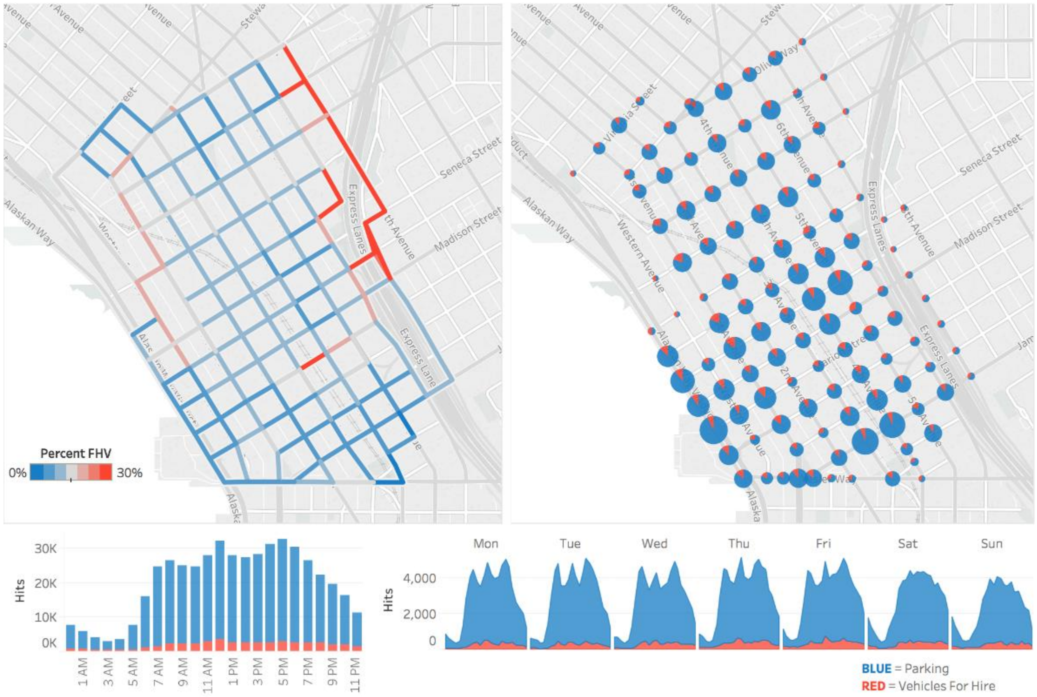
TRAFFIC CRUISING RELATIVE HEATMAP

Aggregated cruising observations, draft analysis by the Data Science for Social Good Traffic Cruising Project



TRAFFIC CRUISING PARKING AND FOR-HIRE HEATMAP

Aggregated cruising observations, draft analysis by the Data Science for Social Good Traffic Cruising Project



PRIVACY AND DATA GOVERNANCE

A key component of this project was to contemplate steps to protect user privacy even though identifiers have been anonymized and the raw data no longer exist. We attempted to measure the error in the data, which provides a degree of anonymity. We aggregated results to a reasonable resolution (hour and street segment). We also recognize that there is a way for users to opt-out of data collection and their devices will be ignored.

The algorithms used to generate estimated paths and aggregations would be a potential candidate for use in a linked data repository with strong governance, such as the University of Washington Transportation Data Collaborative. Thus, subscribers would be able to consume the aggregated heat map information but the algorithms and anonymized data would not be accessible.

BENEFITS AND NEXT STEPS

The City of Seattle has very little data regarding vehicle traffic cruising and its impact on Seattle streets. This project can provide data, where none has previously existed, to inform transportation policies, infrastructure investments, and management decisions with the end goal of reducing cruising and making it easier for travelers to reach their destination.

In addition to supporting the City's mobility and parking programs, real-time cruising information could be used by third-party mobile applications to help predict the availability of parking and efficiently direct vehicles to their destinations.

The City will spend some time analyzing the results and reviewing cruising activity and patterns over a longer study period, for example, over the course of the year. Part of the initial analysis would be to develop a baseline that can be used to compare with seasonal and other variable changes in transportation activity.

Next steps could include expanding the geographic areas of the analysis, obtaining and incorporating additional ground truth data, identifying biases in relation to overall traffic volume, and correlating with parking transaction data.