RESEARCH REPORT

Institute of Transportation Studies | UNIVERSITY OF CALIFORNIA

Mobile Device Data Analytics for Next-Generation Traffic Management

Jane Macfarlane, Ph.D. Director Smart Cities and Sustainable Mobility, Institute of Transportation Studies, University of California, Berkeley
Anthony Patire, Ph.D. Research & Development Engineer, California PATH, University of California, Berkeley
Kanaad Deodhar, Graduate Student Researcher, Institute of Transportation Studies, University of California, Berkeley
Colin Laurence, Associate Data Scientist, Lawrence Berkeley National Laboratory

November 2021

Berkeley Institute of Transportation Studies

Report No.: UC-ITS-2020-24 | DOI: 10.7922/G2SX6BGF

Technical Report Documentation Page

1. Report No. UC-ITS-2020-24	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Mobile Device Data Analytics for Next-Generation Traffic Management		**5. Report Date** November 2021
		6. Performing Organization Code ITS Berkeley
7. Author(s) Jane Macfarlane, Ph.D; Anthony Patire, Ph.D; Kanaad Deodhar; Colin Laurence		**8. Performing Organization Report No.** N/A
9. Performing Organization Name and Address Institute of Transportation Studies, Berkeley 109 McLaughlin Hall, MC1720 Berkeley, CA 94720-1720		**10. Work Unit No.** N/A
		11. Contract or Grant No. UC-ITS-2020-24
12. Sponsoring Agency Name and Address The University of California Institute of Transportation Studies www.ucits.org		**13. Type of Report and Period Covered** Final Report (July 2019–December 2020)
		14. Sponsoring Agency Code UC ITS

15. Supplementary Notes
DOI:10.7922/G2SX6BGF

16. Abstract
Quality data is critically important for research and policy-making. The availability of device location data carrying rich, detailed information on travel patterns has increased significantly in recent years with the proliferation of personal GPS-enabled mobile devices and fleet transponders. However, in its raw form, location data can be inaccurate and contain embedded biases that can skew analyses. This report describes the development of a method to process, clean, and enrich location data. Researchers developed a computational framework for processing large scale location datasets. Using this framework several hundred days of location data from the San Francisco Bay Area was (a) cleaned, to identify and discard inaccurate or problematic data, (b) enriched, by filtering and annotating the data, and (c) matched to links on the road network. This framework provides researchers with the capability to build link-level metrics across large scale geographic regions. Various applications for this enriched data are also discussed in this report (including applications related to corridor planning, freight planning, and disaster and emergency management) along with suggestions for further work.

17. Key Words Transportation planning, mobility applications, GPS data, smartphones, data quality, data fusion, data cleaning, pipeline processing, cloud computing,		**18. Distribution Statement** No restrictions.	
19. Security Classification (of this report) Unclassified	**20. Security Classification (of this page)** Unclassified	**21. No. of Pages** 37	**21. Price** N/A

Form Dot F 1700.7 (8-72) Reproduction of completed page authorized

About the UC Institute of Transportation Studies

The University of California Institute of Transportation Studies (UC ITS) is a network of faculty, research and administrative staff, and students dedicated to advancing the state of the art in transportation engineering, planning, and policy for the people of California. Established by the Legislature in 1947, ITS has branches at UC Berkeley, UC Davis, UC Irvine, and UCLA.

Acknowledgments

This study was made possible with funding received by the University of California Institute of Transportation Studies from the State of California through the Public Transportation Account and the Road Repair and Accountability Act of 2017 (Senate Bill 1). The authors would like to thank the State of California for its support of university-based research, and especially for the funding received for this project.

Disclaimer

Mobile Device Data Analytics for Next-Generation Traffic Management

Jane Macfarlane, Ph.D. Director Smart Cities and Sustainable Mobility, Institute of Transportation Studies, University of California, Berkeley
Anthony Patire, Ph.D. Research & Development Engineer, California PATH, University of California, Berkeley
Kanaad Deodhar, Graduate Student Researcher, Institute of Transportation Studies, University of California, Berkeley
Colin Laurence, Associate Data Scientist, Lawrence Berkeley National Laboratory

November 2021

Berkeley Institute of Transportation Studies

Report No.: UC-ITS-2020-24 | DOI: 10.7922/G2SX6BGF

Table of Contents

Table of Contents

List of Tables

List of Figures

Executive Summary

Executive Summary

The availability of detailed location data from personal mobile devices and fleet transponders has increased significantly in recent years. Carrying rich, detailed information on travel patterns, this data has already seen broad adoption in the transportation planning industry, for example, informing corridor studies and travel demand models. This report outlines an architecture and computational framework for the ingestion, processing, and analysis of raw location data, that can be used to transform GPS points from mobile devices into actionable insights about the transportation network. This method was used to generate traffic data for the San Francisco Bay Area and applied to a recent bridge closure to examine changes in local traffic patterns.

The core focus of this work is to provide tools for research in transportation planning by enabling organizations and researchers across California to effectively take advantage of location data. For example, researchers will be able to better understand regional travel flows and understand the congestion impacts of emergencies and disasters, two applications which are explored in this report. Though this framework has been developed with proprietary location data and roadway network information provided by HERE Technologies, it can be generalized to operate on any type of geospatially referenced data and open-source maps, e.g., OpenStreetMap.

Future studies can build upon the foundation presented in this report, most notably this work is being used to validate an urban-scale parallel discrete event simulator, Mobiliti, which is currently being developed at the Lawrence Berkeley National Laboratory and the Smart Cities and Sustainable Mobility Research Center at the Institute for Transportation Studies at UC Berkeley. The framework developed through this effort lays the groundwork for public agencies to build more efficient, safer, and more effective transportation systems.

Contents

Introduction

Urban transportation planners and researchers often rely on models of travel demand to support their policy and operational planning. A travel demand model is a synthetic population travel profile that defines the number of trips, start time of the trips, and origin/destination of each trip taken by persons in the urban region. They often include information about the purpose of the trip. Building these models involves conducting large surveys of the population and analyzing information from focused traffic counts. They are very complex to build and require a significant amount of funding to develop. Consequently, they are often only conducted every five or more years. These travel demand models are then used to estimate traffic conditions on the road network using either simulation tools that emulate the actions of vehicles on the road network or traffic assignment tools that optimize the routes that the vehicles should take under specific constraints, such as travel time for the individual driver. The transportation planner can then use the results to predict when and where congestion may arise in the region during a typical day and take appropriate steps to mitigate it.

As GPS-enabled smartphones, telematics units, and other mobile devices have proliferated, an unprecedented amount of mobile device data is being generated. The location data is collected by a variety of organizations, including Google, Apple, and fleet management companies. Navigation applications along with a wide variety of smartphone applications regularly collect location data. Organizations, like Google and Apple, use the data to optimize routing which considers current traffic conditions and predicted congestion patterns. Fleet management companies track the movement and delivery of their goods and often specify driver routing. This type of travel path data is rarely made available for public use and a market has emerged for purchasing and post-processing the raw GPS location data.

The value of the GPS location data lies in its currency. Unlike a five-year-old travel demand model, this data provides up-to-date indicators of travel behavior which can improve our understanding of where travelers are going and when congestion is occurring. As such, it has significant value to transportation planners and researchers provided that the data is reduced, processed properly and integrated into road network models that generate relevant transportation-related metrics. Example metrics include regional and city level vehicle miles travelled, time series of the link congestion, and congestion hot spots.

There are many challenges with accessing and using this data for transportation research and planning purposes. The size of the data poses a significant issue for many organizations that are not used to working with large datasets. A single day of GPS data for a region like the Los Angeles basin can be several hundred megabytes or more. The data only represents a small portion of the vehicles present on the roadway (called its penetration rate) and to generate meaningful statistics many days of GPS data must be aggregated to attain relevant geospatial coverage. To process this amount of data in a reasonable period requires cloud services, such as Amazon Web Services (AWS). In addition, the raw data itself can have significant veracity issues. The precision of GPS data can be affected by trees and buildings by as much as tens of meters. As such, processing the data involves a lot of physics-based data validation, such as ensuring that speeds and locations of vehicles

adhere to reasonable expectations. Consequently, building a processing pipeline (a set of data processing elements where the output of each element is the input of the next) to transform the raw data into a usable form can require a diverse team of software engineers with experience that may not traditionally be resident in organizations that can benefit from this type of data.

The alternative for these organizations is to purchase data that has been post processed to generate selected types of information, or analytics, of particular interest. Data analytics is generally an exploratory process in which looks for patterns in real-world data. The problem with this solution is that the post-processing information is generally considered to be the intellectual property of the vendor and the purchaser cannot determine the original source of the data. Making decisions based on data without understanding how, where and when the data was collected can introduce embedded biases and equity issues. Beyond these processing issues, the cost of these purchases can often be beyond the means of city and researchers' budgets for data.

Location data generated by mobile devices present additional problems. For example, the data is often a generated by a variety of mobile device types, including consumer devices such as smartphones and personal navigation devices, as well as telematics devices on trucks and delivery vehicles. While the reported average vehicle speeds might appear to be within reasonable ranges, the source of the data must be considered. Because trucks often drive slower than private vehicles, if the share of data collected from trucks is higher than their actual share on the road then the analytics produced may reflect slower highway speeds than reality. This will bias the predicted travel times for private vehicles and indicate the presence of congestion patterns that may not exist.

This research project focused on driver responses to different traffic scenarios using large-scale GPS data from mobile devices. It established a consistent methodology for processing GPS-based travel data that can then be used to understand transportation system performance and optimization. Key features of the travel paths, or trajectories, were identified to enable their classification and identify congested or free-flow traffic conditions. Large-scale patterns in the data, such as baseline activity, anomalies, locations of blockages, and potential network inefficiencies were identified. The computational framework described in this report will provide a core foundation for future transportation research.

Large-Scale Mobility Data Exploration

Data can be used to look for patterns in mobility data that reflect driver behavior in context of a particular event. such as: What routes would drivers take if a bridge in the Bay Area were closed unexpectedly? How would congestion patterns change? What congestion mitigation plans should cities near the bridge employ if the bridge is unexpectedly closed?

Mobility data can present unique challenges for data scientists, as the size of the data files can be quite large compared to other transportation datasets. The quality of the raw data must be evaluated, and the data must be cleaned based on knowledge of the collection process and the devices that were used for data collection. Because of the scale of the data, it may require cloud-based services which can be expensive and require specialized expertise.

While the goal of this research was to use data science techniques to generate actionable information for transportation planning purposes, a key focus was to provide a computational foundation for future research within the University of California Institute of Transportation Studies. The aim was to build a robust core processing pipeline for mobility data that can be leveraged and expanded over time. As such, the data analytics framework — called *Advance* — was designed to build data transformation pipelines [1] that can be used economically and collaboratively.

Advance: A Flexible Data Processing Framework

Advance is a framework for concise scripting of a data transformation process that can be incrementally built and easily debugged. The design of Advance partitions the processing infrastructure from the data transformation code itself. It encourages simple small data transformations that can be reused in the same pipeline or integrated into other pipelines. Fundamental to Advance is the notion of multithreading that allows the computation algorithm to make the best use of the host machine and process the data as quickly as possible to support creative data exploration. Furthermore, the large scale for which the processing architecture is designed — millions of vehicle paths over hundreds of days and over large geographic regions — allows for analyzing travel patterns beyond the scales generally found in the typical corridor or local municipal studies.

Advance has several key features:

- It maintains **provenance** of the data so that anyone *receiving* data from the pipeline can easily determine the processing steps that were applied to the raw data.
- It is **flexible**. New steps can easily be introduced, and steps can be reordered quickly.
- It is **multilingual**. New analytics can be written in any computer language.
- It uses **multithreading** on the host machine where possible to reduce processing time.

- It **does not require continued reprocessing** of data.
- It is **simple and easy to use** so that students could quickly and easily contribute to the processing pipeline.
- It can be **easily ported to a cloud environment** once the core pipeline has been established.

Details regarding the construction and operation of the Advance framework can be found in Appendix A.

The architecture of Advance allows for a hybrid environment in which the analytics can be implemented on a local server or on cloud-services. This is particularly important in the learning environment where students, who may not be computer science focused, can experiment, and pursue investigative paths without concern for the cost of computing time.

Enriching Mobility Data

Mobility data enrichment is implemented using an ordered set of scripts that are managed by the Advance framework. The scripts create metrics for identifying problematic data, add actual road links that the vehicle may have taken, and determine the locations of intermediary stops. The geospatial extent of the datasets used for this study is illustrated in Figure 1 and Figure 2 and are described further in Appendix A, along with the basic data cleaning scripts used to remove bad data.

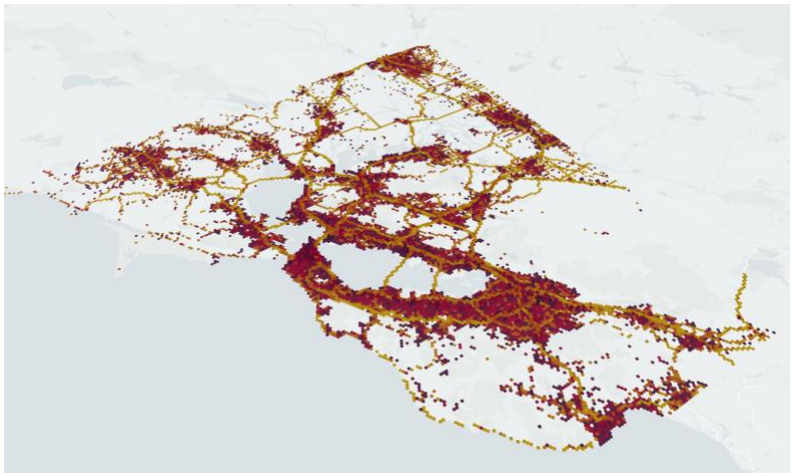

Figure 1. Heatmap of individual GPS points showing the geospatial extent of mobility data for the San Francisco Bay Area

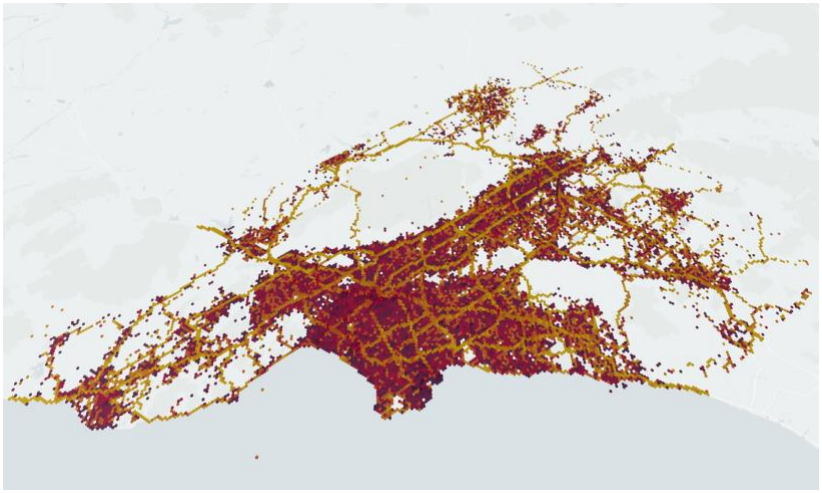

Figure 2 Geospatial extent of mobility data for the Los Angeles Basin

Roadway Link Matching

A necessary capability for extracting information from mobility data is to match the GPS locations to an actual road link so that the path of the mobile device on the road network is revealed. The computational algorithms that attempt to define which road link is being used by the mobile device is called map-matching. Map-matching is a complex problem for several reasons. A GPS point has inherent limits in accuracy and its location can also be affected by its environment. As such, its actual position on the road network can be ambiguous. Adding to the complexity of this problem is the variety of sampling rates — time between each GPS point — that are often found in location datasets. Because of the complexity involved in map matching, software solutions are usually proprietary to companies that work with geospatial data, such as map makers. An open-source map matcher, OSRM [2], that also does routing is available that uses the OSM network. Due to the requirements of our project that focus on driver behavior on the Mobiliti network, we chose to develop a link matcher that uses additional details of road characteristics that are not available in the OSM network.

For the purposes of this project, we developed new link-matching software — the Smart Cities Link Matcher (SCLM). SCLM provides a mechanism to determine link level metrics, including average speed estimates. The cleaned vehicle paths are the inputs for this software that is also implemented as a script and managed by the Advance framework [Appendix].

The algorithm for transforming the cleaned GPS data into link-matched roadway paths is illustrated in the following series of steps and associated Figures.

Step 1: The process begins with a set of GPS locations, as shown in Figure 3. These are expressed as linear segments as shown in Figure 4. The data cleaning scripts generate a start point and end point for each segment.

Figure 3. Original GPS points

Figure 4. Example set of GPS segments in San Francisco. In this trajectory, the data was collected every 90 seconds.

Step 2: For each GPS point, identify all roadway links within 10 meters as candidate links (highlighted in blue in Figure 5).

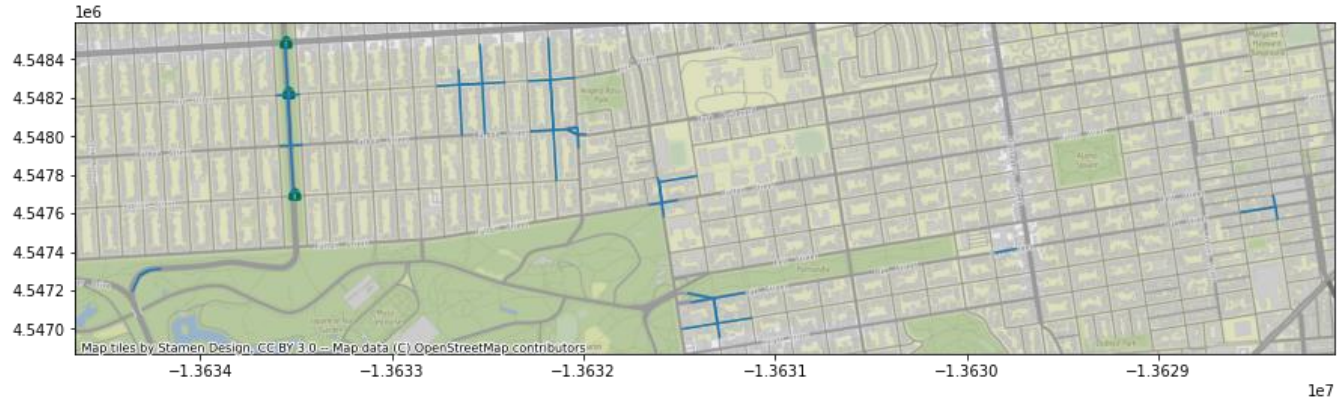

Figure 5. All links within 10 meters of each GPS point

Step 3: Calculate Dijkstra's shortest free-flow travel-time path between all nodes of subsequent sets of candidate links. GPS data from a device that is sampled at high frequency, such as once a second, provides good information for selecting the next link in the path. GPS data from a device that is sampled at low frequency, like once every 90 seconds, creates gaps in our perception of where on the road network the device went. This makes it more difficult to select the next link in the path. The Dijkstra algorithm generates a possible path.

Step 4: Extract the road links from the full network based on the paths identified in the previous step, as shown in Figure 7.

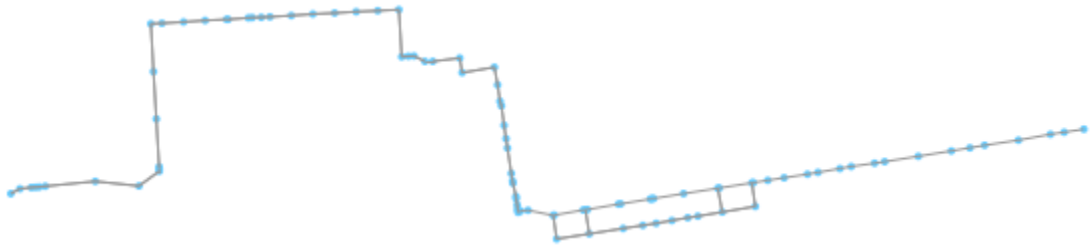

Figure 6. Subsetted road network

Step 5: Select the most likely path in the small subsetted road network, based on the starting point of the trip and ending point of the trip and a similarity metric. The similarity metric chosen was the difference between the geometric length of the distances between the GPS points and the geometric length of all the candidate links. The route with the smallest difference is chosen as the candidate path. Figure 7 shows the selected path.

Figure 7. Final selected path

Step 6: Associate the original GPS information with links in the selected path. For each link in the selected path, the GPS points closest to that link are identified. Links are tagged with the GPS reported speeds for calculating link-level metrics.

The output of the SCLM is the conversion of all raw GPS locations into sets of road links and their associated metrics. Figure 8 shows the active links — meaning that mobile devices were on them — between the hours of 8am and 9am on February 6, 2019.

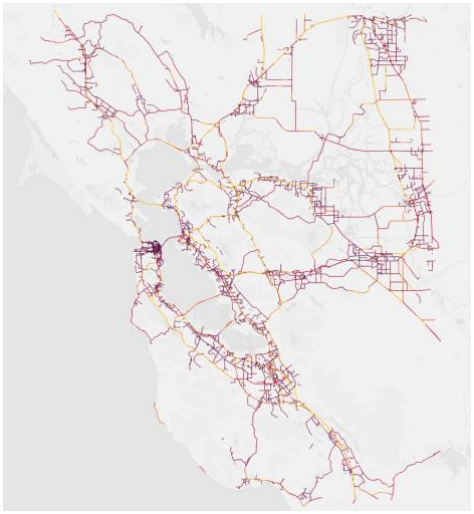

Figure 8. Active Links in the Bay Area 8am - 9am 2/6/2019

Though effective in the vast majority of cases, there are opportunities to improve and refine this algorithm.[1] Several edge cases, including vehicle paths with U-turns, should be addressed in future work in order to improve the accuracy of the link-matching and subsequent analysis. In addition, alternate routing mechanisms might be considered in contrast to the use of Dijkstra's shortest-path solely based on free-flow travel time as drivers do not always take the shortest path.

Extracting Driver Behavior Indicators

Once link-matching has been completed for a collection of GPS locations, various other metrics can be calculated. For example, using properties of the roads that the vehicle most likely took, it is simple to determine the number of unique named streets traversed, the road types along the route, and the heading changes. These metrics can be indicators that drivers took circuitous routes to reduce their travel time. This could also indicate that the driver might be taking a route suggested by a navigation app that is using residential roads to bypass traffic on a major arterial.

Furthermore, approximate entry and exit times for each link along the map-matched path are calculated by interpolating between GPS points and assuming a constant ratio between reported speed and free-flow for links without associated GPS points. This interpolation-based approach enables the estimation of entry times and traversal speeds for every link in the path even when the GPS sampling rate is low.

The generated metrics can also be used to check the accuracy of the link-matching. For example, by comparing the length of the initially assumed vehicle trajectory to the length of the final path, a significant difference can indicate a mismatched path. Interpolation can also help detect mismatched paths or erroneous paths, by identifying discrepancies between the spatial and temporal ordering of GPS points and associated roadway links. Such situations can arise due to GPS errors, or edge cases — such as when a vehicle exits a highway, then doubles back along a frontage road. If many GPS points must be removed for the interpolation approach to work, it suggests that the map matched path does not match the GPS trajectory very well. In such cases the entire trajectory is flagged as problematic. The entry and traversal times for every link of the non-problematic trajectories can then be aggregated.

[1] The algorithm required the use of several geospatial tools, including a number of Python packages and the use of a PostgreSQL/PostGIS database. Trajectories are ingested into SCLM as GeoJSONs, and the final selected path is returned as a JSON of links and associated GPS information. The routing algorithms are handled within the Python script, while many of the spatial operations to select candidate links for each GPS point, and to associate GPS points to links in the selected path, are handled by the database.

Using Large-Scale Mobility Data Analytics for Transportation Planning

There are many potential opportunities for integrating this large-scale, mobility data analytics into transportation planning. The link matching capability creates the opportunity to develop speed tables for links across the region and because it is temporal as well as geospatial, estimates of speed changes over time can be generated. This type of data, along with historical data, is typically used by traffic congestion models to provide a view of congestion and propose alternative routes to reduce travel times. It has significant implications for travel demand modelling, corridor management, transportation simulation, and for disaster and emergency planning, which are explored in this section.

Travel Demand Modeling & Corridor Planning

The methodology developed by this research project can be used to enhance activity-based travel demand models. Travel demand models help agencies understand regional travel flows, and plan for infrastructure investments. They describe the origins, destinations, purposes and times of day of travel demand within a region. For example, aggregated data from sources such as Google can be used to generate origin-destination flows along specific corridors, which can provide detailed metrics on average speeds, estimates of vehicle volumes, and regional trip flows on any given link. These metrics can then be used to evaluate different transportation demand management (TDM) strategies [3, 4], and give insights to the current road network's performance and potential future network changes. Figure 9 and Figure 10 are trip maps that use the processed GPS data to show the trips that access two different road corridors in the Bay Area region.

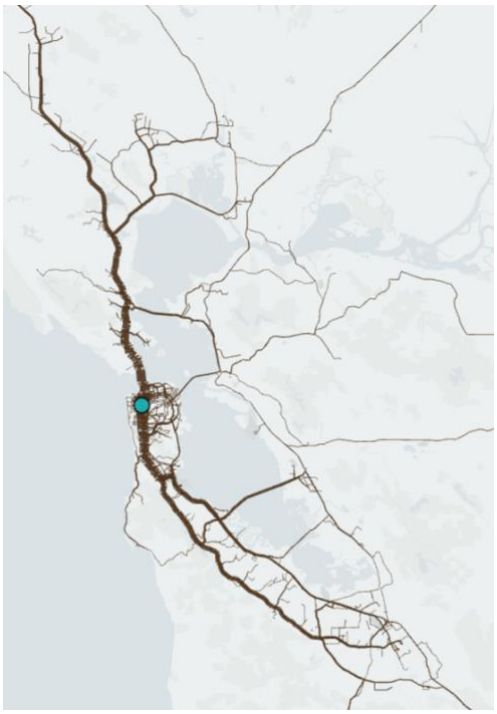

Figure 9. Trajectories for Vehicles Using Southbound 19th Street, denoted by the blue dot, in San Francisco. Line thickness is logarithmically proportional to number of vehicles using a road.

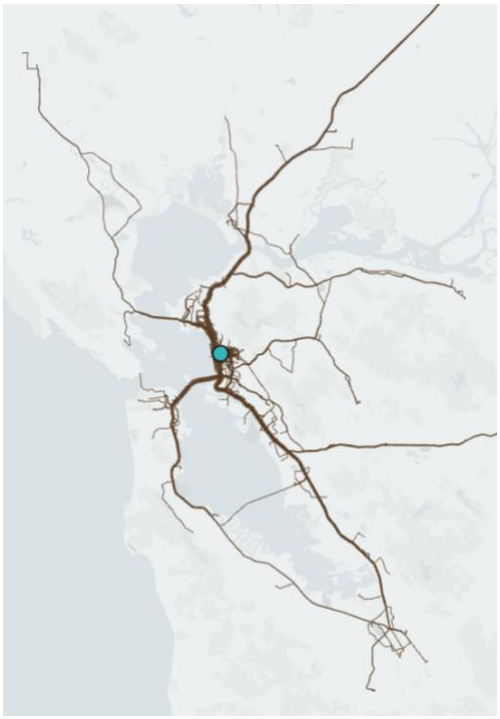

Figure 10. Trajectories for Vehicles Using Eastbound University Avenue in Berkeley

Freight Related Dynamics

The particular mobile location data used for this project tagged information from consumer devices, such as personal cell-phones and built-in GPS units in privately-owned vehicles, and information from commercial fleet devices. Commercial vehicles will often travel at lower speeds than privately-owned vehicles and can result in biasing the data analytics. This difference may also highlight roadway features that fleet drivers are experiencing.

Figure 11 shows network links with distinct private/commercial differences across the entire Bay Area. Line width of a link indicates the total vehicle flow on the link and the color indicates the speed difference between privately-owned and commercial vehicles. The data is from the PM peak, 3pm until 7pm, averaged across multiple days. We note that State Route 37 just west of Vallejo, which has only one lane in either direction, is one of the few major routes where both vehicle classes travel at roughly the same speed. In general, we find that privately-owned vehicles travel faster than commercial vehicles. One notable exception to this is around toll plazas, where fleet vehicles travel 20kph faster or more, which can be seen in Figure 12. The speed comparisons can help to identify and highlight locations across the region where substantial differences exist between car and truck speeds that could present opportunities for improving traffic flows. For example, commercial vehicles travel much slower on roads with steep inclines, such I-580 over Altamont Pass to the east of Livermore, where commercial vehicles travel roughly 20kph slower than private vehicles. Our particular

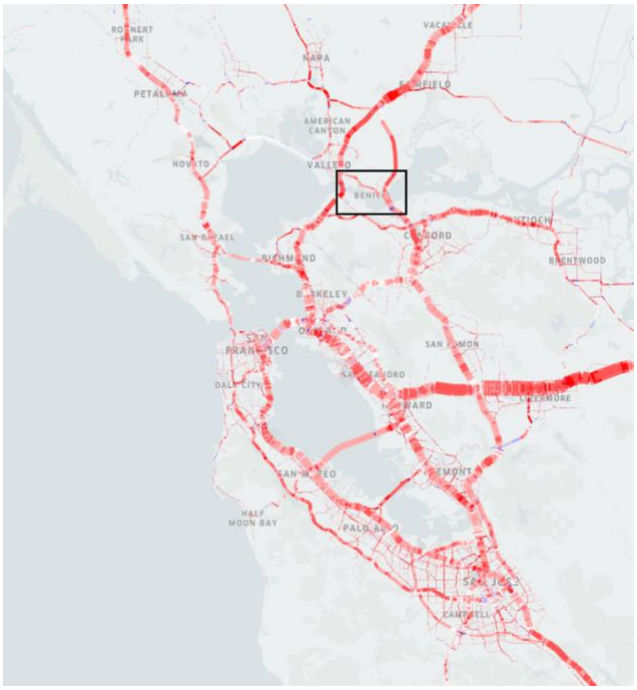

Figure 11. Speed difference between commercial and private vehicles for all links in the San Francisco Bay Area. Thicker lines indicate higher total vehicle flows. Red indicates average privately-owned vehicle speeds are faster while blue indicates commercial vehicles speeds are faster. The black box outlines the area shown in Figure 12.

dataset only distinguishes between private and commercial vehicles; future processed datasets could include more detailed breakdowns, such as specific vehicle make and model, that could help identify additional road network characteristics that impact driver behavior.

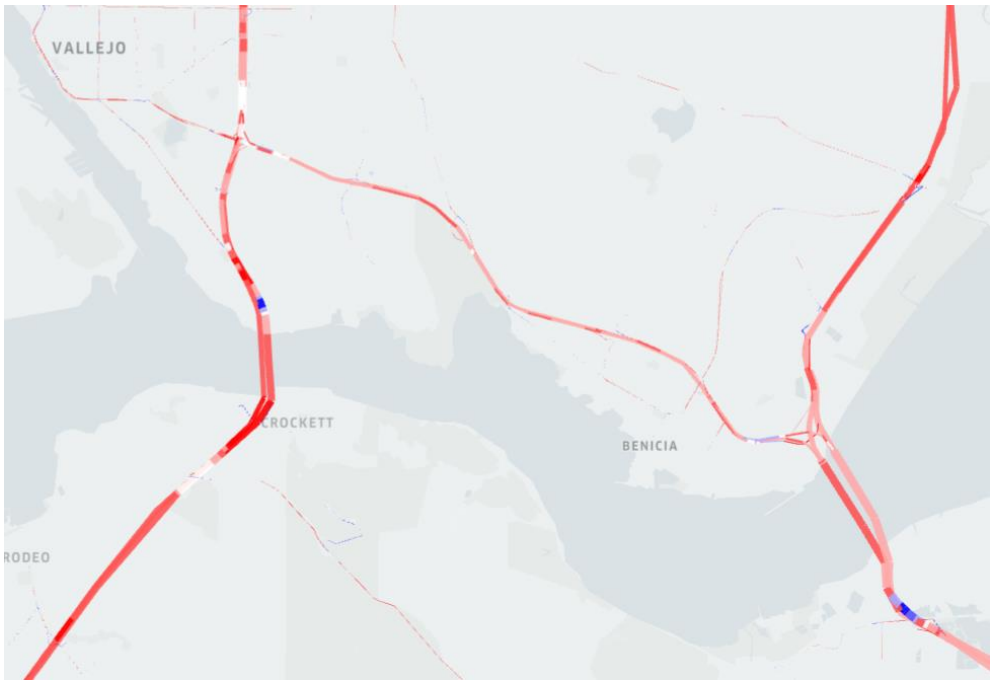

Figure 12. Zoomed in view of differences between privately-owned and commercial vehicles in the region around the Carquinez and Benicia-Martinez bridges.

Transportation Simulation

Transportation simulations are often used for planning and evaluating the operations of transportation infrastructure. In such simulations, a virtual model is created of a region's transportation network, and travel within the region throughout the day is simulated. The Smart Cities and Sustainable Mobility Research Center in the Institute of Transportation Studies at UC Berkeley and Lawrence Berkeley National Laboratory are currently developing an urban-scale transportation simulation known as *Mobiliti*. Mobiliti uses high-performance computing resources at the National Energy Research Scientific Computing Center (NERSC) to simulate the movement of vehicles throughout the San Francisco Bay Area road network and provides estimates of the associated congestion, energy usage, and productivity loss [5]. The framework developed by this research project was used to produce link-level congestion and usage metrics across hundreds of days of data to improve the foundational models that drive the simulation — such as the travel demand model that estimates the typical trips that occur during an average day in the region — and validate the simulated network dynamics. The regional simulation results can also be aggregated to generate a wide variety of metrics that can provide insights for city planners.

Table 1 shows the percentage difference between the percentage shares of simulation-estimated vehicle miles traveled (VMT) across each roadway class by county and those calculated based on GPS-derived VMT. The distribution of VMT across various roadway classes and counties throughout the Bay Area provides insight into how well the GPS data matches the simulation data. In this case, we see a good alignment between simulation results and raw data for most of the road classes and counties, with all discrepancies under 8 percent. This is an important validation for the accuracy of the network loads produced by the Mobiliti simulation.

Table 1. Normalized Percentage Difference Between Simulation and Mobile Device Data: Single Day Share of Total Regional VMT by Roadway Type and County in the Bay Area

Roadway Class	VMT by County									
	Alameda	Contra Costa	Marin	Napa	San Francisco	San Mateo	Santa Clara	Solano	Sonoma	Total
Freeway	-7.6	-1.2	0	0	0.7	0.5	-0.5	-6.8	-0.9	-15.9
Arterial	1.4	0.5	0.4	-0.1	1.6	1.2	5.4	0.2	0.2	10.7
Local	1	0.7	0.2	0.1	0.6	0.5	1.5	0.3	0.3	5.2
Total	-5.3	-0.1	0.5	0	3	2	6.4	-6.3	-0.4	0

Applying the Framework to Disaster and Emergency Planning

On February 7, 2019, the Richmond-San Rafael Bridge, a major East-West roadway link across the San Francisco Bay, typically carrying 80,000 vehicles a day, was unexpectedly closed between 10:30 am and 8:15 pm due to a structural failure on the upper bridge deck. As a result of this incident, typical traffic flows across the northern San Francisco Bay region were interrupted and shifted, as travelers were forced to use alternate routes [6].

The GPS data collected from this incident by HERE Technologies offered a unique opportunity to understand how drivers' behaviors changed due to a significant disruption in the network. Given that this was an isolated and relatively minor incident (minor in terms of the number of hours the bridge was closed), which only affected a single major roadway link, the impacts were primarily from the route changes and late arrivals caused by this closure, rather than more widespread impacts that might occur from a major natural disaster. However, it does provide some insights into the potential effects of a larger scale event.

By comparing the data-based travel patterns of February 7 to a baseline — the average of four other Thursdays before and after the event (January 24, January 31, February 14, and February 21) — changes in travel patterns and the roadway network can be identified.

Visualizing the changes in average link speeds for the approximate duration of the closure (11am-8pm) shows the impacts of the bridge closure. Figure 13 shows the speeds on links across the region on the day of the closure and the day prior to the closure versus the speeds on those links during the baseline. Each point, representing a single link, is colored based on proximity to the Richmond-San Rafael Bridge, with the darkest reds representing the links closest to the bridge. Points lying along the diagonal line represent road segments that experienced no change in speed; those above the line saw an increase in average speed while those below the line saw slower speeds. Immediately, it is clear that several dozen links close to the bridge (dark red dots) experienced dramatic reductions in speed on the day of the closure. However, it is equally important to note that many links experienced increases in speed, with general variability far greater on the day of the closure than the day before.

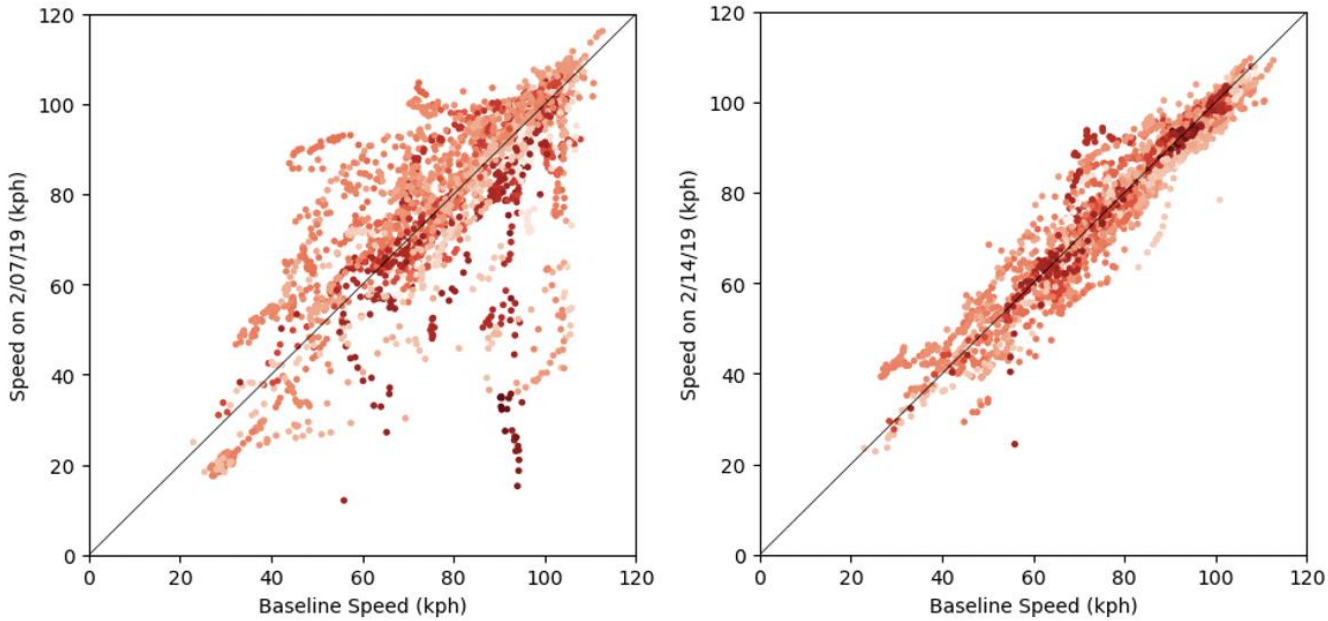

Figure 13. Speeds on links around the Richmond-San Rafael Bridge from 11am-8pm between February 7th & Baseline (Left) and February 14th and Baseline (Right). Darkest reds indicate links closer to the Bridge

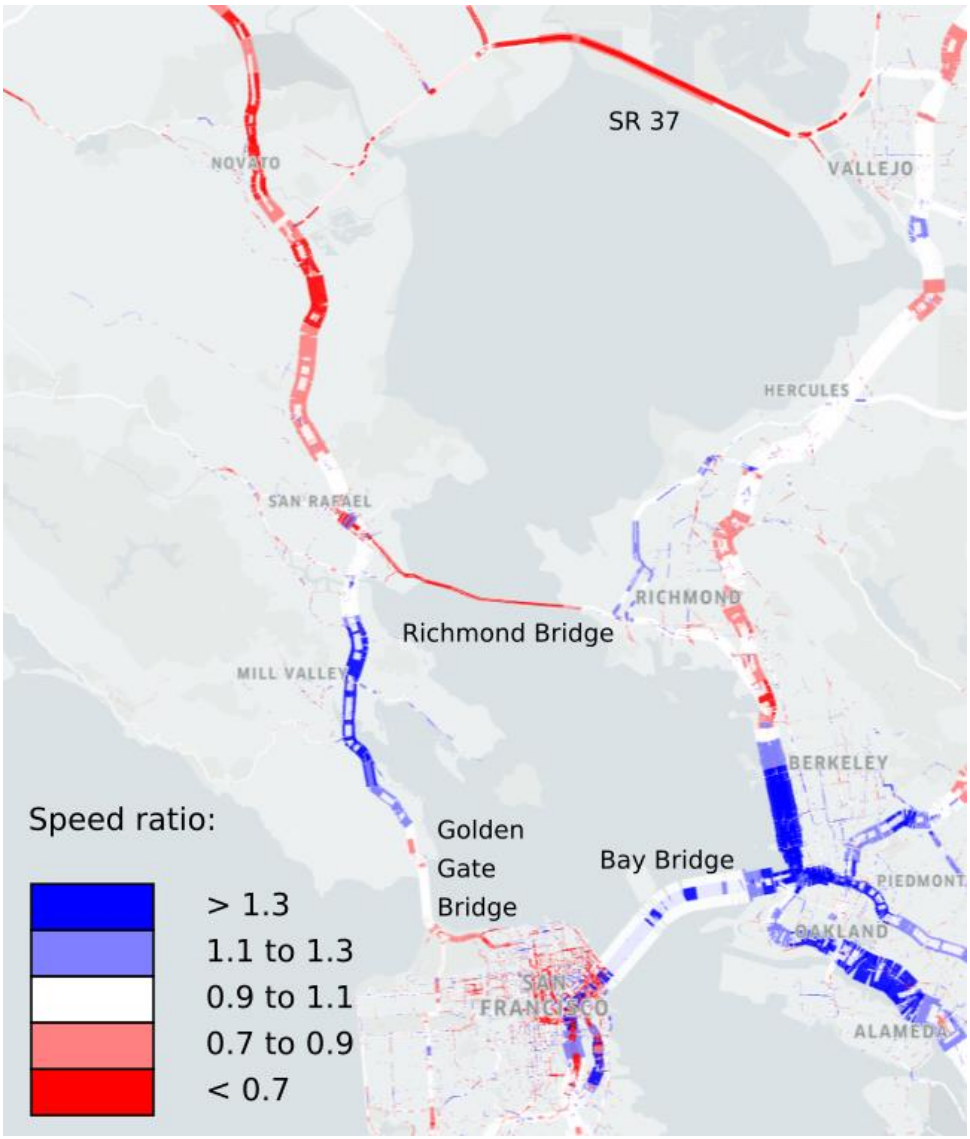

Figure 14. Speed changes between February 7 & Baseline during the PM peak (3PM-7PM). Red indicates slower speeds on February 7, blue indicates higher speeds, line thickness indicates vehicle volume.

Looking at the larger region, Figure 14 maps the speed changes across the Bay Area during the PM peak (3 pm-7 pm). There are two East-West routes across the Bay that could provide substitute capacity for the Richmond-San Rafael Bridge: State Route (SR) 37, across the top of San Pablo Bay, and the San Francisco-Oakland Bay Bridge. State Route 37 experienced significant reductions in average speed during the closure, while the Bay Bridge saw a minor increase in average speed. This likely indicates that SR 37 experienced a significant increase in traffic volumes due to the bridge closure, while the Bay Bridge was relatively unaffected suggesting that the structure had sufficient capacity to absorb the added traffic. We also expect that a selection of drivers changed

their normal behavior, e.g., waiting out the congestion and leaving much later to go to their destination. This is being modeled in a subsequent analysis.

Many roads to the east of the Bay Bridge through Oakland experienced significant increases in speed during the bridge closure, despite many of them being part of potential reroutes for traffic that had previously planned to take the bridge — suggesting that the bridge closure may have caused reductions in travel demand.

Examining vehicle movements on SR 37 on the day of the bridge closure, we identified a 10 percent shift for both mean cumulative heading changes per kilometer (a measure of how often a vehicle turns in a given distance) and total functional class changes (such as switching from a freeway to a local road). Future research could focus on developing these indicators, and additional metrics, to quantify detailed changes in drivers' routing choices and shifts in travel patterns.

The capacity to analyze detailed changes in traffic flows in response to disasters will help evacuation planning efforts by identifying roadways experiencing the greatest increases in traffic, and which destinations individual vehicles move toward. As climate change continues to exacerbate wildfires in California, and the state's aging infrastructure makes network disruptions more likely, this information could prove critically important to evacuation and network design strategies.

Conclusion

As the availability of location data generated by mobile devices increases, the capacity for transportation agencies to effectively use this data to better understand transportation networks becomes critically important. The methodology established in this report enables the large-scale, rapid transformation of widely available GPS data into a useful format that contains information about traffic dynamics, roadway network usage, and the behavior of drivers across a region. This methodology has significant potential to better inform traffic simulations, travel demand models, and emergency and disaster management.

This work represents the initial step towards building a robust but flexible approach to the processing of large mobility datasets. Future work will focus on computational efficiency improvements to scripts involved in the processing architecture, which could enable use of even larger datasets and simplifications in the toolset, to reduce the challenges associated with implementing new datasets. Additionally, the link-matching script could be improved to deal with a greater number of edge cases found in the travel paths, which would improve the integrity of the link-path data and increase the information content of the raw GPS data.

Appendix A

Advance - A Data Transformation Processing Framework

Advance is a framework It allows for concise scripting of a data transformation process that can be incrementally built and easily debugged [1]. The design of Advance partitions the processing infrastructure from the data transformation code itself. It encourages simple small data transformations that can be reused in the same pipeline or integrated into other pipelines. Fundamental to Advance is the notion of multithreading in order to make the best use of the host machine and process the data as quickly as possible to support creative data exploration.

The framework forms a data transformation pipeline as a series of steps. The core mechanism is to process a step and make the results of each step available as input to the next step. The artifacts of each step are preserved in directories that are numbered according to their position in the pipeline. When the output of a step generates incorrect results or the code for the step does not complete due to an error, the artifacts can simply be deleted, the code for the particular step can be modified and the pipeline can be restarted. Previously successful steps are skipped, and the computation quickly picks up at the step that was last successfully completed. When a step fails the results are preserved in directories prefixed with "tmp_". This isolates incomplete step data and ensures that the step is re-processed when the problem is resolved. As such, reprocessing of data that has already proven robust is avoided.

A data pipeline can use the Advance framework by establishing a primary Ruby script that imports Advance and includes the data transformation steps. While Advance is written in Ruby and is implemented as a Ruby script, the data transformations are not required to be written in Ruby. Each step describes a command to be run on a folder of files or a single file. These commands can be one of a collection of prepackaged Advance scripts, unix commands (like split, cut, etc.), or scripts/commands written in other languages. Advance invokes these scripts one by one as would a person typing at the command line. Advance logs the exact command that is invoked so that it can be run separately to check the output manually and to debug failures.

Advance steps are composed of three components:

- a step processing type,
- a descriptive slug describing the step (as a Ruby symbol), and
- the command that transforms the data.

For example:

single :unzip_7z_raw_data_file, "7z x {previous_file}"

single :split_files, "split -l 10000 -a 3 {previous_file} gps_data_"

multi :add_local_time, "cat {file_path} | add_local_time.rb timestamp local_time US/Pacific > {file}"

The current step processing functions in Advance are single and multi. Single applies the command to the last output, which should be a single file. Multi speeds processing of multiple files by doing work in parallel using another Ruby gem called TeamEffort [7].

Creating an Advance pipeline script is expected to be an incremental process. This approach encourages a methodical, single step at-a-time approach to data processing. The user starts with a single step, runs the script and checks the results. When the output is validated, the user adds the next step. Once the next step is added, the user simply restarts the script. Previously successful steps are skipped and the script moves on to the first incomplete step.

Specifying script input and output is accomplished by identifying the files that are to be transformed -> the input files, and the output of the data transformation -> the output files. Advance provides a few tokens that can be inserted in the command string for this purpose:

- {input_file} indicates the output file from the previous step. It is also used to indicate the first file to be used and it finds that file in the current working directory.
- {file_name} indicates an output file name, which is the basename from {input_file}. Commands often process multiple files from previous steps, generating multiple output files. Those output files are placed in the step directory.
- {file_name_without_extension\} is simply {file_name} with the extension removed. This is useful when one is transforming a file from one type — with an extension — to another type, with a new extension.
- {input_dir} indicates the directory of the previous step.

Example Script

#!/usr/bin/env ruby

require "advance"

include Advance

ensure_bin_on_path # ensures the directory for this script is on the path so that related scripts can be

referenced without paths

single :unzip_7z_raw_data_file, "7z x {previous_file}" # uses 7z to inflate a file in the current directory

single :split_files, "split -l 10000 -a 3 {previous_file} gps_data_" # split the file into smaller files that are 10K lines

multi :add_local_time, "cat {file_path} | add_local_time.rb timestamp local_time US/Pacific > {file}"

adds a local_time column to a csv

Running a pipeline then becomes as simple as creating a directory that contains the single initial file, and invoking the script from that directory.

Advance makes a transformation on the data and stores the result in a new directory. It zips the directory before it advances to the next step to limit storage requirements resulting from the transformations. The key word "advance" is associated with the gem and using the defined script creates the new directory for the transformation. If the step does not have "advance" as a key word it just runs the identified script.

Also included in Advance is a static statement. Static means there is no transformation on the data — it provides a capability to generate a new set of files that are modifications of the existing files in that step. This was included primarily for making visualization files, namely GeoJSON files.

Mobility Data

The data used for this study consisted of location data from the San Francisco Bay Area and Los Angeles Basin, and was provided by HERE Technologies under proprietary license. In the Bay Area, the data spans 279 days from November 25, 2018 to August 28, 2019. In the Los Angeles Basin, the data spans 455 days from May 31, 2018 to August 29, 2019. Each day is represented in a CSV file with the format presented in Table 2. The sources of the location data are a mix of private and commercial navigation devices, including mobile phones and fleet telematics devices, with each trip tagged either as private or commercial. The data in these CSVs are not limited to vehicles but could include data from pedestrians and bicyclists using navigation applications. As such, we must define what kind of analytics is necessary to filter the data appropriately for this project. In a typical data file from the Los Angeles Basin, from June 5, 2018, two dozen different providers reported data from over 100,000 unique devices. The geospatial extent of the data for both the Bay Area and Los Angeles Basin is shown in Figure 1 and Figure 2, respectively.

Table 2. Raw Data Format

Variable	Description
PROBE_ID	Device ID
SAMPLE_DATE	Date and time in Coordinated Universal Time (UTC) of the data point
LAT	Latitude
LON	Longitude
HEADING	Heading in degrees
SPEED	Speed in kilometers per hour
PROBE_DATA_PROVIDER	Source of the data, distinguishing between commercial/fleet and consumer devices

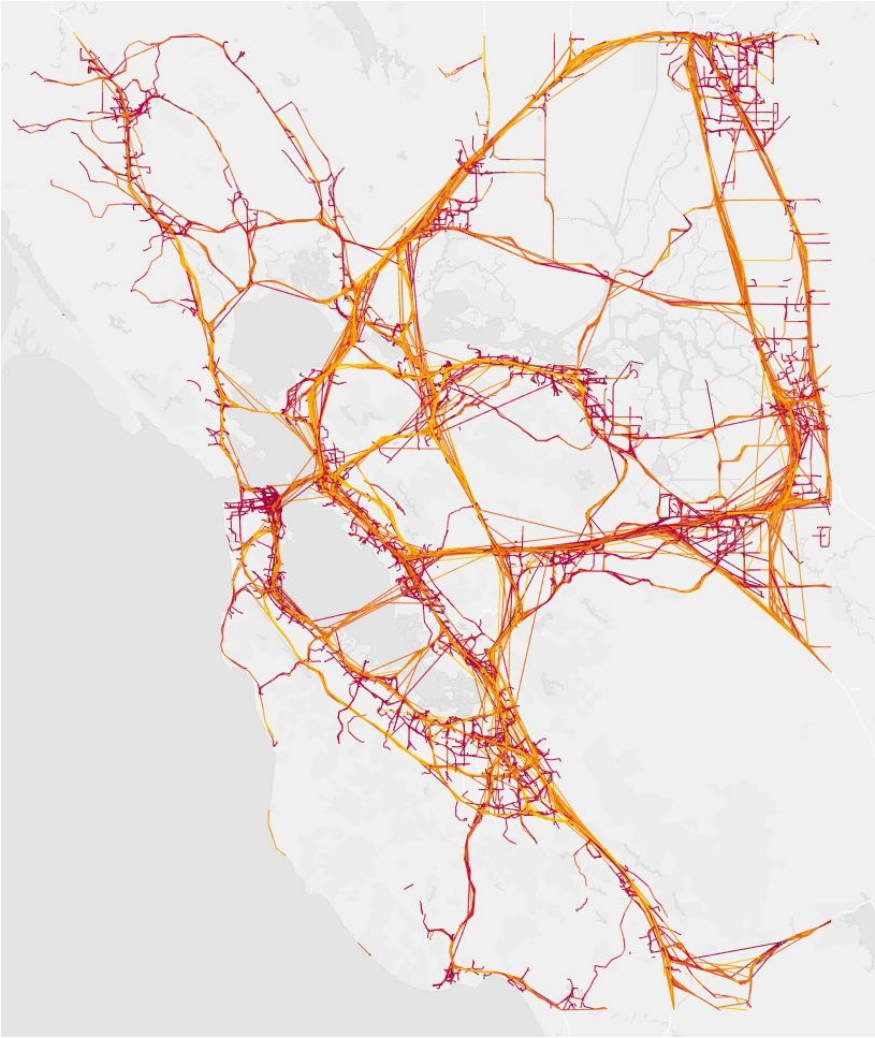

Figure 15. Cleaned vehicle trajectory outputs from pipeline, 12am-1am on February 6, 2019

Reconstructability

Reconstructability addresses the problem of recovering traveled geospatial paths on a transportation network from time sampled location traces. As described in Macfarlane and Xu [8] the geospatial road density sets a fundamental constraint on the sampling frequency. If the constraint is violated, then the path cannot be reconstructed without an inference about the path of the device. For example, a GPS device moving along a limited access freeway has a sampling interval of 30 seconds; it will generally have a highly reconstructable path, since there will usually be at least one GPS sample between each exit, making the path clear from GPS data alone. However, if the same device travels through closely spaced city streets, the device may often travel several blocks within the 30 second sampling interval, resulting in multiple paths being possible to connect the GPS points, which means the path would have poor reconstructability. The analyst must determine what kinds of inference should be made based on the context of the analytics being pursued.

Reconstructability carries significant implications for the usefulness of point-location data: if all links along a pathway are associated with a GPS point (and reported speed), road conditions can be estimated for every single link. Conversely, if very few links along a pathway have associated GPS points, the estimated congestion of the transportation network involves far more guesswork. The dependency of reconstruction on the density of the road network means that there will be multiple paths between the time ordered GPS points and as a consequence the inference that chooses the path becomes significantly less robust. SCLM uses Dijkstra's travel-time shortest path with quality metrics for its inferencing. With this approach, trajectories of all sampling rates can be ingested into the link-matcher, simplifying the problem and the processing workflow.

While this approach is reasonable, it will likely result in paths that did not happen in the real world. The metrics that compare the path with the original GPS trajectory are intended to provide some level of confidence in the link matched path. For example, a delivery truck may take a very circuitous route that is not necessarily represented in the reported location data. For this situation, a temporal metric can be used to infer alternate paths. As such, reconstructability can reflect a measure of "trustworthiness" of the path. A reconstructability ratio is defined that is equal to the ratio of links in the link-matched path that are tagged as reconstructable. In this context, a link is reconstructable if at least one of two conditions are met: first, that the link has at least one associated point, which means that there is at least one GPS point which is closer to the given link than to any other link, and second, that there is no other link that the vehicle could have used. For example, if a section of a freeway between two exits is made up of three links, and the first and third links have associated points, the second link is automatically reconstructable even if it does not have any associated points. A trajectory with a higher reconstructability has fewer places where the link matcher had to infer the correct path, and as such is more likely to be accurate.

Reconstructability ratios were calculated for a full day of Bay Area trajectories and grouped by mean GPS sampling interval. Predictably, trajectories that had faster sampling rates had far higher reconstructability ratios, which are shown in Figure 16. Trajectories with a 1 second sampling interval had a reconstructability ratio of 0.95, which dropped to 0.77 for a 10 second sampling interval, 0.47 for a 30 second sampling interval, and 0.34 for a 90 second sampling interval. As was defined in [8], these results are dependent on the network graph density. The dataset used for this study had varying degrees of graph density. Future work can consider the impact of the designated region on the reconstructability, for example less dense cities could have much higher reconstructability with the same variability of sampling rates. These types of considerations can support raw data provider selections and data purchase decisions.

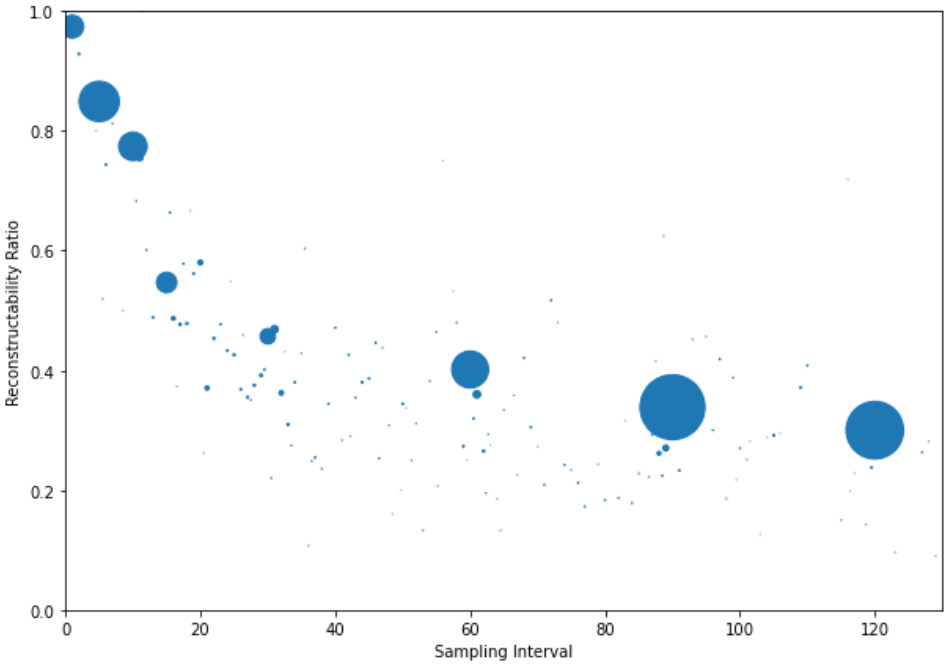

Figure 16. Reconstructability ratio versus sampling interval of trajectories

Scaled Implementation in the Cloud

The volume of data involved with this project necessitated the use of tools that enabled rapid data entry and processing. Though several high-performance computing frameworks were tested, in its final form, the processing architecture of this project involves the pipeline written in Ruby using the Advance Gem, cleaning and analysis scripts, including link-matching, written in Python, and several Amazon Web Services tools, including the Relational Database Service (RDS), the Simple Storage Service (S3), and the Elastic Container Service (ECS) as shown in Figure 17. The Advance Ruby Gem enables a high degree of flexibility for testing and exploratory work, by simplifying the addition and removal of specific analysis scripts. In this way, researchers, students, and professionals can easily add their own analysis or transformation steps to the existing architecture.

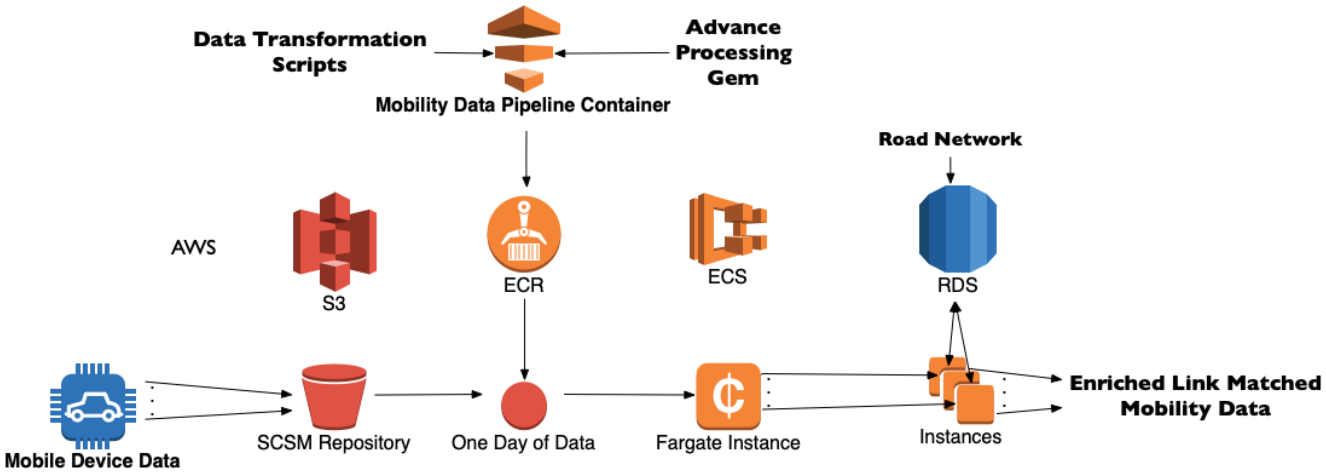

Figure 17. Data transformation and enrichment pipeline implemented with AWS

References

[1] Macfarlane J, Felkins K. Advance Ruby Gem. 2020. Available at https://rubygems.org/gems/advance/versions/0.1.1

[2] OSRM API Documentation [Internet]. Project-osrm.org. 2020 [cited 20 January 2021]. Available from: http://project-osrm.org/docs/v5.23.0/api/#

[3] Lemp, J. and Komanduri, A., 2019. Plan For the Future of Transportation Using Location-Based Services Data. [online] Medium. Available at: https://medium.com/csdataanalytics/plan-for-the-future-of-transportation-using-location-based-services-data-19db912f8a19 [Accessed 5 January 2020].

[4] Sana, B., Castiglione, J., Cooper, D. and Tischler, D., 2017. Using Google's Aggregated and Anonymized Trip Data to Support Freeway Corridor Management Planning in San Francisco, California. Transportation Research Record: Journal of the Transportation Research Board, 2643(1), pp.65-73.

[5] Chan, C., Wang, B., Bachan, J. and Macfarlane, J., 2018. Mobiliti: Scalable Transportation Simulation Using High-Performance Parallel Computing. 2018 21st International Conference on Intelligent Transportation Systems (ITSC).

[6] Sanfrancisco.cbslocal.com. 2019. All Lanes Reopened On Richmond-San Rafael Bridge After Concrete Chunks Fell. [online] Available at: https://sanfrancisco.cbslocal.com/2019/02/07/falling-concrete-closes-richmond-san-rafael-bridge-traffic [Accessed 10 January 2020].

[7] Felkins K. Team Effort Ruby Gem. 2020. Available at https://rubygems.org/gems/team_effort/versions/1.1.1

[8] Macfarlane, J. and Xu, B., 2017. Temporal Sampling Constraints for GeoSpatial Path Reconstruction in a Transportation Network. Proceedings of the 10th ACM SIGSPATIAL Workshop on Computational Transportation Science.