Applications of Trajectory Data in Transportation: Literature Review and Maryland Case Study

Nikola Marković, Przemysław Sekuła, Zachary Vander Laan, Gennady Andrienko, Natalia Andrienko

Abstract

This paper considers applications of trajectory data in transportation, and makes two primary contributions. First, it provides a comprehensive literature review detailing ways in which trajectory data has been used for transportation systems analysis, distilling existing research into the following six areas: demand estimation, modeling human behavior, designing public transit, measuring and predicting traffic performance, quantifying environmental impact, and safety analysis. Additionally, it presents innovative applications of trajectory data for the state of Maryland, employing visualization and machine learning techniques to extract value from 20 million GPS traces. These visual analytics will be implemented in the Regional Integrated Transportation Information System (RITIS), which provides free data sharing and visual analytics tools to help transportation agencies attain situational awareness, evaluate performance, and share insights with the public.

Keywords: transportation, trajectory data, literature review, visual analytics, machine learning, big data

1. Introduction

Numerous detailed trajectory datasets have recently become available, including Global Positioning System (GPS) traces from cell phones and vehicles, anonymized Call Detail Records (CDR) from cell phone providers, and data from arrays of Bluetooth and Wi-Fi detectors that re-identify devices over time. As vast amounts of spatiotemporal data becomes more ubiquitous, transportation agencies have an opportunity to leverage these resources to improve analysis techniques and answer important questions more efficiently. However, since this data often needs to be purchased, agencies should be well-informed about potential benefits in order to assess its value to their organization. While the literature concerned with trajectory data is rich, it is naturally fragmented and often focuses on technical contributions in niche areas, which makes it hard for government agencies to assess its specific application to transportation domains. To overcome this issue, the current paper explores trajectory data from the perspective of a transportation agency, seeking to synthesize existing approaches and also present new applications for transportation systems analysis. It is worth noting that a review
paper by Andrienko et al. (2017) also seeks to bring trajectory data closer to practice; however, it focuses in particular on visual analytics approaches that may be useful for transportation agencies. The authors conclude that it is necessary to establish collaboration between the visual analytics and transportation research communities, and seek to do so in the current paper.

The trajectories analyzed in this paper were obtained from a major GPS company in North America. It provides Internet services and mobile applications informing users about traffic conditions, which are estimated based on terabytes of GPS data collected daily from millions of mobile phones, cars, trucks, vans and other fleet vehicles. In 2016 the Maryland State Highway Administration (SHA) purchased GPS traces of all trips recorded in Maryland during four months of the previous year. The SHA subsequently contracted the Center for Advanced Transportation Technology (CATT) to determine the value of the trajectory data for transportation system analysis and evaluate the cost/benefit trade-off. With the goal of enabling SHA and other government agencies to accurately assess the value of these datasets, this paper provides a thorough overview of potential use-cases in various domains of transportation engineering. In particular, we make two contributions:

- We provide a comprehensive literature review detailing innovative uses of trajectory data in transportation systems analysis. The review covers studies that exploit different trajectory datasets (GPS traces, CDR, Bluetooth and Wi-Fi detectors) in six areas of transportation engineering: demand estimation, modeling human behavior, designing public transit, traffic performance measurement and prediction, environmental impact, and safety analysis. This review can serve as a single reference point for government agencies trying to decide whether purchasing trajectory data would be beneficial.

- We visually explore a set of 20 million GPS traces in Maryland and present novel transportation applications using the dataset and an array of visualization and machine learning techniques. Furthermore, we summarize and discuss best-practices for investigating trajectory data. The visual analytics developed in this paper will be made available to transportation agencies via the Regional Integrated Transportation Information System, an automated data sharing, dissemination, and archiving system developed and maintained by the CATT Lab at the University of Maryland. (RITIS 2016).

The next section provides a comprehensive literature review of various uses of trajectory data in transportation, while the following two sections describe the dataset and methods used to analyze it. After presenting and discussing results for the state of Maryland, we conclude by summarizing the findings and discussing future steps.

2. Literature Review

Applications of trajectory data in transportation are synthesized into six areas, each of which is discussed in a separate subsection. It is worth noting that the following review does not seek to
provide an exhaustive overview of the literature, but highlight some of the relevant work in order to illustrate applications of trajectory data in different areas of transportation engineering.

2.1. Demand estimation

At the core of demand modeling and transportation planning is the problem of estimating the number of trips that take place between specific locations (Iqbal et al., 2014). Traditional four-step models begin by estimating trip productions and attractions based on census and household travel survey data, and use simple models (e.g., gravity) to link the two together to generate Origin-Destination (O-D) trip tables (McNally, 2007). More advanced activity-based models attempt to better quantify travel demand with respect to individuals and activities, which consequently requires better data, typically in the form of detailed surveys (Zhang et al., 2011). Regardless of the approach, accurately estimating demand is a necessary first step in developing trustworthy models that can answer important policy questions.

The traditional data sources used to estimate demand are census and travel survey data, which are sometimes combined with traffic counts from roadside sensors (Iqbal et al., 2014; Siла-Nowicka et al., 2016). While these datasets contain valuable information, their use is sometimes hindered by non-representative sampling and misreported responses on surveys (Kuwahara and Sullivan, 1987; Groves, 2006), as well as difficulties with reconstructing the trips between O-D pairs based on sparse vehicle count data (Van Zuylen and Willumsen, 1980; Lo et al., 1996). Given the importance of determining O-D pairs and the shortcomings of traditional methods, mobility data offers an appealing, more direct approach to inferring demand.

2.1.1. O-D matrix based on trajectories: an example

The methodology employed in Toole et al. (2015) is an innovative example of how trajectory data can be utilized to estimate demand. The proposed approach begins with a preprocessing procedure, where CDR data points representing timestamps of phone calls and text messages are mapped to locations, either through triangulation methods or simply by locating the nearest cell tower. Individual anonymous users’ locations are then tracked over time to form trajectories through space, making the data functionally similar to GPS trajectory data, but with less spatial resolution. From this point, the goal is to mine the data to extract the number of trips that take place between locations, a process that involves making assumptions about how to define important locations and assign meaning to the set of movements over time. The authors use an algorithm from Zheng and Xie (2011) to transform the detailed trajectory data into more manageable trajectories of stay locations, where a stay location represents a place in which a cell-phone user spends significant amounts of time. The region is then divided into a set of zones, and stay points are assigned to the zone that encompasses them, meaning that the stay-point trajectories represent trips between zones. After carefully discarding users who do not use their phones frequently enough to accurately characterize their travel behavior and scaling the results to reflect the total population, an estimate of daily O-D trip tables can be produced.
2.1.2. Additional considerations

It should be noted that more realistic demand models consider time-of-day effects (e.g., AM/PM Peak, Off Peak) and the types of trips taken (e.g., Home-Work, Home-Other), which can also be extracted from mobility data. These considerations are tackled with varying levels of complexity in Wang et al. (2012), Iqbal et al. (2014), Çolak et al. (2015), and Alexander et al. (2015), but further discussion of extracting meaning from the data will be postponed until the next subsection.

Other demand estimation research includes application of GPS traces to (a) infer O-D patterns on freeways and turning ratios for roundabouts as well as other locations where surveys are not practical or effective (van Vuren et al., 2011), (b) develop statewide O-D matrices for trucks in Florida (Zanjani et al., 2015), (c) calibrate an agent-based demand model (Sharman and Roorda, 2010), (d) assess taxi demand (Castro et al., 2012), and (e) perform dynamic demand estimation (Yang et al., 2014, 2016; Moreira-Matias et al., 2016). The interested reader is referred to Lee et al. (2014) for additional discussion of emerging data sources for estimating travel demand, including Bluetooth detectors and transit fare cards.

2.2. Modeling human behavior

Quantifying human behavior is a key component of demand modeling and transportation planning, since understanding why people travel and the specific choices they make in the process (e.g., mode and route choice) can be useful for shaping policies that positively impact the overall transportation system. This subsection focuses on two specific aspects of human behavior: assigning context to travel movements and choice analysis.

2.2.1. Context of travel movement

Given detailed mobility datasets, intelligent data mining strategies can be utilized to derive meaning and context from the locations visited. Zheng (2015) provides perhaps the most thorough overview of the field, distilling trajectory data mining into the following phases: (a) preprocessing, which includes trajectory compression, stay-point detection, trajectory segmentation and map matching, (b) data management, which involves indexing and storing the data so it can be retrieved quickly and (c) pattern mining, which involves clustering (by time, shape, and segment), classifying, and detecting outliers. The last phase is particularly interesting for transportation, because its application involves grouping similar trip origins, destinations, times of day, trip durations, and sections of road, in order to extract prevailing patterns and answer transportation-related questions. Much of the literature in this area is concerned with the performance of different clustering algorithms, of which popular ones include hierarchical (Bastani et al., 2011), centroid (Ashbrook and Starner, 2003; Moreira-Matias et al., 2016), and density (Castro et al., 2012) models. Although the theory behind the different algorithms can be complex, analysis is becoming increasingly more user-friendly, with fully-documented open-source software implementations available for public use, such as the data mining query tool M-Atlas (Giannotti et al., 2011).
The aspect of trajectory mining most relevant to demand estimation is the stay-point detection process, which helps identify locations at which an individual spends significant amounts of time (Zheng and Xie 2011; Zheng 2015). The implication is that these points represent locations that are important to the individual, and repeated observations during different times of day can indicate home or work locations, which are particularly useful for demand modeling. For example, Andrienko, Andrienko, Fuchs and Jankowski (2016) use trajectory data to detect and classify significant locations (e.g., home, work, or social) in a way that respects user privacy, employing a visual analytics approach (e.g., Andrienko et al. 2007, 2012, 2013) and demonstrating its capabilities on a benchmark dataset and location data from Twitter. Other examples include Schneider et al. (2013) and Jiang et al. (2015), which utilize the concept of network motifs to investigate and describe the types of locations where cell phone users spend extended periods of time in Paris and Boston, respectively. Advanced applications of these techniques are discussed in Widhalm et al. (2015), where the authors organize stay locations according to activity, and provide a methodology to order activity sequences based on the locations and timestamps of CDR data. Note that keeping track of activity sequences has important implications for advanced activity-based demand models, which require more data than typical demand models to ascertain trip-chains and activities. This approach is employed by Jiang et al. (2015), where the authors build upon previous research to specify a detailed framework for activity-based modeling, while relying on their ability to extract activity-based patterns from Singapore CDR data.

2.2.2. Choice analysis

The other aspect of modeling human behavior that can be improved through mobility data is describing choice behavior (e.g., mode and route choice). In transportation, people’s behavior is usually addressed with discrete choice models, where users consider a set of mutually exclusive alternatives and choose the one that maximizes their utility. Given a set of observations about travel behavior from some segment of the population, a transportation modeler seeks to find model parameters that best describe the observed behavior (Ben-Akiva and Lerman 1985). Consequently, detailed travel survey data (e.g., the National Household Travel Survey in the United States) is vital to discrete choice modeling, but is laborious to acquire and may become outdated after a few years. Accordingly, mobility data provides an opportunity to observe how people behave, from which discrete choice models can be calibrated, verified, or shown to be flawed. This can be illustrated through the following two studies.

Xu and González (2016) combine CDR, Waze GPS data and a handful of other sources to investigate the impact of special events on a city’s travel patterns, focusing on the 2016 Olympics in Rio de Janeiro, Brazil. They estimate the O-D demand prior to the Olympics, and creatively utilize the Olympic event schedule, stadium capacities, Airbnb and hotel information to account for additional destinations and demand from tourists. After building the demand model to account for the Olympics, they explore choice behavior in the form of mode shift and traffic routing strategies, noting the overall system implications associated with the different choice behaviors. Another example is
Lima et al. (2016), where the authors use GPS traces of 526 vehicles to investigate routing behavior and check whether people take the lowest-cost paths, which is commonly assumed in traffic assignment. They cluster start and end points to find important locations, cluster trajectories to determine possible routes, and discover that most users take the same path in the majority of situations, which often is not the minimum cost path. Studies like these help determine whether choice models that are based on utility maximization actually match real-world behavior.

2.3. Designing public transit

Public transit systems provide an effective way to help relieve congestion, reduce emissions, and transport people efficiently in areas where significant travel demand exists between common origins or destinations (Pinelli et al., 2016). Transit planning consists of selecting system characteristics (e.g., station locations, routes, fleet size, service frequencies, fares) in order to provide satisfactory service at minimal cost (Guihaire and Hao, 2008). The problem of optimal transit system design has been widely addressed in the literature, examples of which include Chien and Schonfeld (1997), Chien and Spasović (2002), Fan and Machemehl (2004), and Cevallos and Zhao (2006).

2.3.1. Trajectories as input to optimization models

The traditional transportation network optimization techniques rely on aggregate O-D matrices (Pinelli et al., 2016), which we have already discussed in the demand estimation subsection of this paper. We reemphasize that, in addition to traditional survey/land use/traffic count methods, these O-D matrices can be estimated by mining trajectory data. Note that this approach may be particularly useful in developing countries where survey data may not be available, and in cities where travel survey data quickly becomes outdated due to rapid population growth. In such cases, trajectory data may help provide reasonable aggregate demand estimates to feed existing transit network optimization models. An example of this approach is found in Berlingerio et al. (2013), where the authors use CDR to propose route changes to a transit system in Abidjan, Ivory Coast, resulting in estimated average travel time reductions of up to 10% across the city.

2.3.2. Data-driven approach

There is another, more data-driven approach to transit planning. Rather than reducing trajectory data to a set of important O-D locations and using these O-D matrices to feed an optimization model, the data-driven approach seeks to use the trajectory data directly to infer optimal transit routes. One of the most complete examples of this approach is found in Pinelli et al. (2016), where the authors propose a methodology to design a new transit network in Abidjan, Ivory Coast, using the aforementioned cell phone data. The premise is based on the idea that a transit network’s service should reflect the spatial and temporal patterns of people’s movement, such that supply appropriately matches demand. Based on patterns that emerge from the massive amount of cell phone data points (referred to as m-trails), a set of potential routes are selected and then refined by employing other utility-maximization strategies. Upon selecting the routes, the authors use linear programming to find optimal service frequencies.
2.3.3. Additional studies

Additional work that considers transit design from this perspective include Bastani et al. (2011), who optimize routes for a proposed flexible shuttle service based on taxi trajectories. They identify important trips through agglomerative clustering and use them to build a directed graph for route-finding, demonstrating the ability to significantly reduce trip mileage. Similarly, Chen et al. (2014) plan bidirectional night routes for buses by clustering taxi GPS traces, building and pruning bus route graphs, and proposing a bidirectional probability-based spreading algorithm to determine the best route. They illustrate the methodology on a real-world dataset consisting of 1.57 million taxi trips from Hangzhou, China, concluding that their approach outperforms the existing night bus service. Another method includes Wang et al. (2016), who form trajectories from bus smart card data and propose new bus lines based on travel patterns mined from the trajectories, using data from the Beijing Public Transport Group to demonstrate the approach. Additional applications include planning airport shuttle bus stop locations from GPS data based on $k$-means clustering and optimization (Liu et al., 2014), and designing demand-oriented bus systems through clustering, integer programming, and a proposed routing algorithm using taxi GPS traces (Lyu et al., 2016).

2.4. Traffic performance measurement and prediction

Trajectory data can be used both to analyze historic performance of a traffic system and to help predict future traffic states. Upon discussing the related work, we point out existing challenges in this area.

2.4.1. Quantifying past performance

Transportation agencies require traffic data in order to quantify system performance, inform policy decisions, and identify areas of improvement (Brennan Jr et al., 2013). Important performance indicators include congestion-related measures, such as travel times over different time periods, travel time reliability, vehicle/person throughput, occupancy, and total vehicle delay, all of which depend on the ability to accurately capture data. Traditional traffic sensors such as induction loop detectors and radar/microwave detectors are useful for obtaining vehicle counts, but have more difficulty estimating travel time distributions because these fixed sensors measure only spot-mean speed (Kesting and Treiber, 2013). There are many intelligent techniques that can be used to overcome this drawback of traditional detector data (Lee and Coifman, 2011; Lu et al., 2014; Coifman, 2015), but trajectory datasets offer an alternative, direct approach for measuring travel times. Rather than inferring travel times based on point measurements and constant-speed assumptions, these datasets can be used directly to calculate travel time distributions, quantify congestion measures, and serve as a ground truth for other sensor data (Mudge et al., 2013; Kim and Coifman, 2014; Young, 2014; Brennan Jr et al., 2015). State and local agencies can leverage these probe vehicle data and existing methodologies to develop mobility reports, an example of which is the Maryland Mobility Report (Mahapatra et al., 2015).
2.4.2. Real-time predictions

In addition to quantifying past performance of a transportation network, traffic data can be used for real-time traffic state predictions, provided that data feeds are available in real time. With some exceptions (e.g., [Wedin 2015]), literature in this area tends to focus on data assimilation techniques, which seek to optimally blend predictions from traffic models and field measurement observations, each of which contain some unknown levels of uncertainty ([Evensen 2009]). While significant data assimilation research has been performed using stationary sensors ([Dailey 1999]; [Coifman 2001]; [Guo et al. 2009]; [Allström et al. 2014]), trajectory-based measurements provide new opportunities for traffic state estimation. For example, [Wei et al. 2010] investigate the performance of a Kalman filtering approach to travel time estimation using data from a fleet of GPS-enabled probe vehicles, with traditional traffic sensors serving as ground truth measurements. Recognizing that GPS and loop detector data sources contain complementary information, others consider assimilation techniques that merge data collected from both fixed and moving measurements ([Chu et al. 2005]; [Yang 2005]; [Herrera and Bayen 2008]; [Xia et al. 2016]), while focusing on different aspects of the assimilation problem and application areas (e.g., [Yang 2005] considers arterial traffic in the context of disruptive events).

2.4.3. Challenges

Although GPS and cell-phone data add substantial value to traditional stationary traffic sensors, there are aspects that make the analysis of these datasets more cumbersome. One such difficulty is that recorded locations have to be associated with known road segments before the data can be utilized fully. Thus, trajectory data vendors often take care of the complex map matching process (at least when aggregating traffic characteristics for road segments), but there is also an array of techniques and algorithms that have been developed to facilitate this process. For example, [Jensen and Larsen 2014] describe an assortment of map-matching algorithms, and use them as the basis for traffic state estimation. The interested reader is directed to [Zheng 2015] for a discussion of map-matching approaches, which are separated into algorithms characterized by additional information utilized (geometric, topological, probabilistic, advanced) and the range of sample points considered (local/incremental, global).

There are some additional considerations to point out with respect to using trajectories to quantify and predict system performance. First, the corresponding travel times only represent individual probe vehicles, so it is important to ensure that a given mobility dataset contains enough samples to be representative of traffic on the road. Additionally, the types of vehicles that produce location data may be biased (e.g., towards large trucks or fleet vehicles), and their behavior may not realistically represent actual traffic patterns. These types of issues are addressed in [Hellinga and Fu 1999] and [Yamamoto et al. 2006], and it is important to be aware of them. Finally, the specific type of technology used to capture location data is more important in this particular application of trajectory data than in other applications described in this paper. Namely, GPS data is more spatially accurate than CDR or Wi-Fi detection, and this is especially relevant when trying to map locations to road
segments. However, temporal accuracy is important too, and GPS data that is collected irregularly or over long time periods may be problematic, as described in a study that investigated vehicle tracking in Chicago via GPS and Wi-Fi technology (Eriksson, 2013). Less granular data may be appropriate for estimating demand across different parts of a city, but can cause unexpected results for quantifying system performance, particularly in urban settings where there are numerous cross streets and intersections or freeway overpasses (Yim and Cayford, 2001).

2.5. Environment

The transportation sector was responsible for 26% of all 2014 greenhouse gas emissions in the United States, mostly from burning fossil fuel for vehicles, trains, planes, and ships (EPA, 2016). Thus, transportation agencies are often interested in (a) quantifying their environmental impact, and (b) developing strategies to make operations more efficient and shift reliance away from fossil fuels. Both can be aided by the use of trajectory data.

2.5.1. Quantifying emissions

An important way to quantify the environmental impact of traffic is through transportation emissions models, which can be approached from macroscopic or microscopic vantage points. Macro-level models (e.g., EMEP, EEA) base the emissions calculations on aggregate flows and average vehicle speeds along transportation networks (Bandeira et al., 2014). In contrast, micro-level models focus on individual vehicles’ accelerations and decelerations (Zegeye et al., 2009), which produce more accurate emissions estimates than macroscopic models (Ahn and Rakha, 2008), examples of which include CMEM (An et al., 1997) and VT-Micro (Rakha et al., 2004). Since trajectory data can be used to improve demand estimation techniques, its application to macroscopic emissions modeling yields more accurate estimates. Similarly, since micro-level emissions models rely on knowledge of vehicle accelerations, trajectory data can be used to calculate these inputs directly rather than relying on estimates from microsimulation experiments (e.g., Treiber et al. 2008, Feng et al. 2011, Yao et al. 2013). From either perspective, trajectory data provides an opportunity to better quantify existing emissions resulting from transportation operations.

2.5.2. Mitigating emissions

One attempt to reduce greenhouse gas emissions is by developing vehicles which use alternative energy sources instead of fossil fuels (Buekers et al., 2014). Electric vehicles are one such alternative, but are hindered by a lack of necessary infrastructure for conveniently recharging (Rauh et al., 2015). Thus, in an attempt to promote adoption of electric vehicle and related technologies, cities and planning agencies may be interested in determining how to best locate recharging/refueling infrastructure. A handful of recent studies suggest that trajectory data may be beneficial for achieving these goals, including Sellmair and Hamacher (2014), Li et al. (2015), and Yang et al. (2017). These examples use taxi GPS traces from Germany and two cities in China as inputs to facility location problems, seeking to determine optimal locations for charging infrastructure under various constraints.
and objective functions (e.g., maximize charging stations per taxi stand, minimize cost, minimize infrastructure investment). For example, Yang et al. (2017) use a large GPS trajectory dataset from Changsha, China to determine optimal charging station locations for electric taxis, formulating the optimization problem as an integer program that seeks to minimize infrastructure investment.

Another approach to mitigating transportation’s environmental impact is by encouraging alternative mode choices, such as transit, cycling, and walking. As cities seek to embrace this challenge, trajectory data can be used to help determine average pedestrian and bicycle flows on different network links (which are not captured in typical demand models) in order to track progress and also help identify potential safety concerns. For example, Zheng, Li, Chen, Xie and Ma (2008), Zheng, Liu, Wang and Xie (2008), and Stenneth et al. (2011) use supervised learning to ascertain travel mode from GPS data, with Stenneth et al. (2011) claiming 93.5% classification accuracy when using the transportation network, bus and train stop information, and other relevant features. Other related examples in this area include using GPS or Bluetooth data to identify patterns in pedestrian movements (Van der Spek et al., 2009; Ellersiek et al., 2013; McArdle et al., 2014), cyclist behavior (Broach et al., 2012; Strauss et al., 2015), and even develop bicycle route choice models (Hood et al., 2011; Charlton et al., 2011).

2.6. Safety

Trajectory data has recently been used in a number of innovative applications focusing on emergency response and cyclist/pedestrian safety. In addition to these existing applications, the detailed spatio-temporal data presents opportunities for other safety-oriented analyses that have not yet been considered.

2.6.1. Emergency response

Hara and Kuwahara (2015) demonstrate how trajectory data may be useful during emergencies by using probe vehicle and smartphone GPS data to assess network conditions after a 2011 earthquake in Japan and make recommendations for disaster management. In response to the devastating earthquake, Song et al. (2014) develop a methodology for probabilistically modeling human movement using GPS traces to help better respond to future disasters. Similarly, Kusano and Inoue (2013) and Ikeda and Inoue (2016) determine optimal evacuation routes after natural disasters, with Ikeda and Inoue (2016) employing a multi-objective genetic algorithm to jointly optimize evacuation distance, time, and safety.

2.6.2. Cyclist and pedestrian safety

A separate branch of safety research leverages GPS, Wi-Fi and Bluetooth trajectory data to provide insight into cyclist and pedestrian safety. For example, Langford et al. (2015) use GPS traces from a bike-sharing system in Knoxville, Tennessee over a two-year period to quantify the behavior of cyclists on both traditional and electric bikes. Their results indicate that cyclists on both types of bikes exhibit similar alarming behavior, including riding on the wrong side of the road on
45% of road segments, and violating traffic signals 70% of the time. Similarly, Dozza and Werneke (2014) investigate bicycle risk by analyzing GPS traces, calculating incident rates through simple odds ratios, and concluding that crash risk is greatest at intersections and on roads that are in poor condition. Other research which applies trajectory data in this general area includes analyzing the accuracy with which a bike’s position can be located (Lindsey et al., 2013), processing bicycle GPS data using Kalman filtering and locally-weighted regression (Luo and Ma, 2014), categorizing bicycle environments from GPS data using support vector machines (Joo et al., 2015), and combining GPS traces with bicycle count data to infer high-risk areas for cycling injuries (Strauss et al., 2015). These analyses provide methodological frameworks and recommendations that may be useful for transportation agencies looking to design bike lanes or improve bikeshare safety.

From a pedestrian and urban planning perspective, Koshak and Fouda (2008) use GPS traces to characterize human movement in order to address the issue of excessive pedestrian density during special religious events in Saudi Arabia. Likewise, Johansson and Helbing (2010) analyze trajectories during a crowd disaster to characterize how pedestrian dynamics change from low to unsafe crowd densities. Although their empirical data is extracted from video, it is nonetheless trajectory data that can be treated similarly to datasets collected from other technologies.

### 2.6.3. Additional applications

Trajectory data may have additional safety applications that are not yet well-researched, particularly with regard to traffic safety. One example involves using trajectory data to predict the frequency (or probability) that an accident will occur, either along general road networks or specific locations such as work zones. A large body of existing literature proposes models that explain crash frequency in terms of explanatory variables such as traffic demand, time of day, weather, lane closures, etc. using either traditional statistical estimation (e.g., Venugopal and Tarko, 2000; Khat-tak et al., 2002; Chen and Tarko, 2012; Yang et al., 2013) or machine learning approaches (e.g., Xie et al., 2007; Ma and Kockelman, 2006; Li et al., 2008). Based on the idea that trajectory data can be used to estimate traffic demand and the fact that demand is a key explanatory variable used in most crash frequency models (e.g., Abdel-Aty and Radwan, 2000; Li et al., 2008; Yang et al., 2015), we note that trajectory datasets may be useful for transportation agencies seeking to understand how to improve traffic safety, e.g., by changing road characteristics or determining where to position response teams (Sekula et al., 2017). Another example involves using trajectory data to glean information about drivers’ behavioral tendencies. Hatipkarasulu et al. (2000) use GPS trajectory data to explore car-following behavior between vehicles, characterizing how they interact in terms of positions, speeds, and accelerations. Similarly, Chen et al. (2010) use NGSIM trajectory datasets to calibrate a MITSIM mirosimulation model, while Colombaroni and Fusco (2014) use artificial neural networks to calibrate a microsimulation model using GPS data. Based on these car-following models, it may be possible for transportation agencies to characterize driver populations in terms of average time or space headways, identify unsafe or overly-aggressive tendencies, and recommend appropriate countermeasures.
Figure 1: A sample trip with relatively few waypoints and descriptive statistics for 6.4 million trips recorded in October. Trip lengths are computed based on great-circle distances between waypoints.

3. Data

The dataset used in this paper consists of GPS trajectories from 20 million trips recorded during February, June, July and October of 2015. Each trip consists of an origin and destination, as well as a number of intermediate waypoints, each of which has a corresponding time stamp (see Figure 1 for a sample trip). Insight into the dataset is provided by summarizing characteristics of trips recorded during the month of October. Namely, the median trip duration and length are about 18 min and 7 miles (Figure 1), while the median time lapse and spacing between consecutive waypoints are approximately 1 second and 28 meters respectively (Figure 2). About 77% of the trips are internal to Maryland, while the remaining 23% have at least one waypoint outside Maryland (Figure 3). The same visual indicates that the vast majority of trips correspond to vehicles (which are subdivided into three weight classes) while about 1% of all the trips are pedestrian movements. In addition, Figure 3 shows that most trips pertain to fleet vehicles, and that fewer trips are observed over the weekend. In total, the raw GPS traces include 1.4 billion waypoints which requires 112 GB of storage space.

3.1. Preprocessing

Since GPS data includes measurement errors, the recorded waypoints are not necessarily located along the physical road network. In addition, the granularity of data is not always high enough to include a waypoint along every single road link (or a traffic message channel) that a vehicle traverses.
(a) Waypoint locations (October 2015)

Figure 2: Visual representation of October waypoints and descriptive statistics based on a sample of over 360 million waypoints from October. Spacing is expressed in great-circle distances, while statistics is computed after removing outliers (e.g., unrealistic displacements due to device-related errors).

(b) Waypoint statistics

Figure 3: Summary of trip attributes for 6.4 million trips that took place in October 2015: geospatial characteristics, mode, provider type, vehicle weight classes, and split of trips for days of the week.
Therefore some preprocessing is needed in order to map match waypoints to the road network and reconstruct road-based routes. This was done using the OpenStreetMap (Haklay and Weber, 2008) routing tool (Figure 4), which applies a hidden Markov model to find the most likely road-based route from a time-stamped sequence of latitude/longitude pairs (Newson and Krumm, 2009). The computationally-intensive map matching was carried out in parallel on a 10-core computer, and took about 3 days to process all 20 million trips. Since map matching results in trajectories that include a significant amount of additional information (i.e., data about every road link that a vehicle traverses), the corresponding dataset increased in size from the initial 112 GB to over 5 TB. However, after removing redundant information (i.e., keeping only one node per road link), the remaining dataset was reduced from 5 TB to 700 GB.

3.2. Database

In order to efficiently store and query the large dataset we utilized PostgreSQL 9.6, an open-source database that has several useful features for analyzing spatio-temporal data. First, it comes with PostGIS spatial database extender, which adds support for geographic objects and allows location queries to be run in the Structured Query Language (SQL). It is also highly integrated with QGIS, which is an open source Geographic Information System (GIS) that was extensively used in this study. Additionally, it includes a number of built-in solutions to facilitate data manipulations, such as table inheritance mechanism, spatial indexing, and advanced spatial queries. Finally, it is widely-used for processing spatial data, which results in a sizable online community and support. The primary
disadvantage of using PostgreSQL is the limitation regarding parallel queries (i.e., the planner will not conduct a parallel query if it involves any data writing). However, this limitation will likely be removed in future releases of PostgreSQL.

3.3. Penetration rate

Because the trajectory data represents only a subset of vehicles on the road, it is important to roughly quantify the penetration rate (PR) of the analyzed trips. Doing so may help indicate the extent to which the sample is representative of overall traffic, and also provide insight into the total number of vehicles traveling on road segments between fixed traffic sensors. To perform rough PR estimates, we compared GPS traces and data from 47 automatic traffic recorder (ATR) stations in Maryland, which typically provide hourly vehicle counts without differentiating between vehicle types. The average hourly PRs at 47 locations are provided in Figure 5, which indicates that average PRs at these 47 locations vary from 0.85% to 5.52%, with a median of 1.86%. This implies that observed trips capture one in every 54 vehicles. Additionally, the same figure shows that the average hourly GPS counts at the 47 locations vary from 4.46 to 98.5 vehicles, with a median of 19.5 vehicles. Note that the previously-discussed map matching procedure is essential to accurately estimate GPS trajectory penetration rates, and that alternative approaches may significantly underestimate their representativeness.

4. Methods

We employ an array of machine learning algorithms and data visualization techniques to extract value from 20 million GPS traces and effectively communicate our results with transportation agencies. Here we provide an overview clustering algorithms used in the analysis, as well as software solutions that the authors found particularly useful in analyzing and visualizing trajectory data. This
4.1. Density-based clustering

DBSCAN (density-based spatial clustering of applications with noise) is a widely-applied clustering algorithm (Ester et al., 1996), which identifies each data point as a core point, border point, or outlier based on two input parameters: $\varepsilon$ and $\text{MinPts}$. $\varepsilon$ is a radius parameter that defines the $\varepsilon$-neighborhood $N(\varepsilon)$ around each point, and $\text{MinPts}$ represents the minimum number of data points in $N(\varepsilon)$ required to form a core point. Clusters are built around core points (which represent high-density areas) by iteratively adding density-connected points. DBSCAN does not require the number of clusters as an input parameter, can easily find arbitrarily-shaped clusters, is robust with respect to outliers (which are treated as noise and do not affect existing clusters), and is also implemented in many libraries which facilitates its application. However, one of the disadvantages is that it is very sensitive to input parameters (Karypis et al., 1999), where small changes to the radius and distance parameters can yield different clustering results. In addition, the definition of distance should be carefully considered because it naturally affects the results (e.g., see Pelekis et al., 2012 for a related discussion of similarity measures for trajectories). This paper utilizes DBSCAN for constructing isochrones based on trajectory data. Finally, a related density-based clustering algorithm OPTICS (ordering points to identify the clustering structure) (Ankerst et al., 1999) is used in other applications described in the following section.

4.2. Software

- V-Analytics (formerly Descartes and CommonGIS) is a free visual data exploration and visual analytics software that facilitate exploration, analysis and modeling of different kinds of spatio-temporal data: events, time series, trajectories and situations. The system includes a variety of interactive visualization techniques (Andrienko and Andrienko, 2006), supports necessary transformations of spatio-temporal data (Andrienko et al., 2013) and integrates a number of computational methods, adapted for analysis in space and time. Particularly, tools for clustering trajectory data with a library of suitable similarity measures are integrated (Andrienko et al., 2009). We used V-Analytics to do quick exploratory analysis, compare performance of different clustering algorithms, and obtain high-quality visuals.

- QGIS is an open-source GIS tool developed through the Open Source Geospatial Foundation (QGIS Development Team, 2015). QGIS was used to prepare the majority of maps and map-based animations in this work. We found QGIS particularly useful due to its interface with PostgreSQL for easy preparation and management of large datasets, its interface with Python for programmatic manipulation of maps and their appearance, and the large online community that provides support and numerous plug-ins written in Python and C++.
5. Illustrative Case Study

In this section we showcase numerous applications of trajectory data in transportation. While some case studies pertain to applications that have been previously studied in the literature (e.g., demand estimation, public transit), they are still included for completeness and to illustrate new ways of effectively visualizing results. More importantly, this section presents several innovative applications of trajectory data: measuring accessibility via isochrones, identifying candidate locations for speed cameras, and enforcing vehicle weight limits.

5.1. O-D matrices

As argued in the literature review, demand modeling and transportation planning relies on estimating the number of trips that take place between specific locations (Iqbal et al., 2014). To illustrate this task, we map the origins and destinations of the 20 million trips to geographic areas of different sizes (i.e., traffic analysis zones, zip codes, counties and states), and derive the corresponding O-D matrices. While dense O-D matrices are somewhat difficult to visualize, those with fewer entries can be visually explored using open-source software Circos (Krzywinski et al., 2009), shown in Figure 6. For example, Figure 6(a) depicts trips between Maryland and other states, where the green ribbons denote trips originating in Maryland and ending in other states. This visual indicates that most trips originate and end in few neighboring states (i.e., Virginia, Pennsylvania), which are ordered clock-wise based on the total number of trips. It also shows that the number of trips going in and out of Maryland is balanced, which can be observed by comparing the two outermost concentric circles that are of approximately same length and color pattern. Figure 6(b) visualizes the subset of these trips that traverse Maryland, and indicates that a notable number of trips that originate and end in a neighboring state (e.g., District of Columbia, Delaware, Virginia) still use the Maryland infrastructure. Figure 6(c) shows a county-based O-D matrix for trips internal to MD and suggests that most trips originate and end within the same county, which is an expected result because the median trip length is about 7 miles (Figure 1). As previously mentioned, we can also derive more granular O-D matrices (i.e., for zip codes and traffic analysis zones), which can easily be explored via interactive applications (e.g., GIS, web).

Recall that the analyzed GPS traces represent only a sample of all vehicles on the road, and thus need to be scaled by appropriate expansion factors (e.g., through knowledge of the PR as discussed in Section 3) to estimate actual traffic volumes. Accordingly, the O-D matrices obtained from the GPS traces can also be scaled; we can use the same expansion factor for the entire O-D matrix to obtain a rough estimate of the overall volumes, or apply machine learning techniques to obtain custom expansion factors for different O-D pairs, vehicle types, times of the day (peak vs. off-peak), and days of the week. In addition, it is possible to correct for heavy vehicle biases, which may be observed in GPS data (Figure 3). This can be done by determining PR of trajectory data for different types of vehicles (i.e., passenger cars vs. trucks), which is possible with traffic sensors that can differentiate vehicle types.
(a) Trips between MD and other states

(b) Trips traversing MD

(c) Trips between counties in MD

Figure 6: O-D matrices visualized with Circos [Krzywinski et al., 2009].
(a) Trips from Washington to Baltimore (left) and from Baltimore to Washington (right)

(b) Trip statistics

Figure 7: Trajectories of trips between Washington and Baltimore beltways that took place during October. Boxplots show travel times for trips departing within the specified hour of the day or day of the week. The median travel times are connected to emphasize the trend, while the bottom and top edges of a box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, while the outliers are plotted individually using “+” symbol.
5.2. An O-D pair

Rather than considering an entire O-D matrix at various levels of granularity, it is sometimes useful to focus on a specific O-D pair. To illustrate this, we consider trips between Washington and Baltimore (Figure 7), and use GPS traces between this O-D pair to visually explore flow patterns, travel time variability and split rates amongst three major routes. Figure 7(a) shows the raw trajectories as well as aggregated trips for days and links between neighboring polygons. Interestingly, both beltways and I-95 (the middle road) show clear weekly patterns, whereas I-295 (East-most road) has stable load in both directions with no weekly patterns. Moreover, Figure 7(b) visualizes travel times between the Washington and Baltimore beltways broken down by hour of day for weekday/weekend and day of week. On weekdays, the morning peak occurs for trips departing at 7-8 AM, while the afternoon peak is observed for trips departing at 4-5 and 5-6 PM. A very different travel pattern is observed on weekends, during which travel times are much steadier and also shorter than on weekdays. Additionally, the GPS trajectories between this O-D pair allow us to estimate turning movements at every intersection, which is often needed as an input to traffic microsimulation models. Since turning movements along the corridor connecting Washington and Baltimore is of special interest to local agencies, we developed an application which facilitates a detailed analysis of turning movements (Figure 8). The application allows a user to select a road link along the I-95 corridor and obtain statistics related to the trajectories that traverse it in a specified direction and time period. Some of the statistics reflect information about the turning movements, composition of vehicles, trip durations, and major origins/destinations/O-D pairs given different levels of spatial aggregation. This application can be used to compare turning movements of different vehicle types during various time periods, and consequently calibrate traffic simulation models for the region.

5.3. Trip generators and isochrones

In addition to analyzing GPS traces between O-D pairs, it is instructive to consider origins and destinations separately. For example, Figure 9 shows trip origins, which are spread over the entire state of Maryland. While the sheer number of data points obscures any patterns, creating and overlaying a simple heat map representing origin frequency shows that many of the trips originate at only a handful of locations. The main trip generators (i.e., areas with the highest density of origins) are downtown Baltimore, Baltimore-Washington International Airport, and the stretch between Bethesda and German Town. Upon identifying the main trip generators, we can query trips that originate in these areas and visualize mobility at different times of day through isochrones, which we construct based on GPS trajectories. It is worth noting that isochrones based on GPS traces provide a different accessibility measure than the isochrones based on theoretical travel times. The latter tell us where people could travel within a certain period of time, while the former indicate where (many) people have actually traveled within certain period of time. In the reminder of this subsection we explain application of density-based clustering in the design of trajectory-based isochrones.

Trajectory datasets often contain anomalous waypoints, which may skew mobility statistics and visualizations. Here we describe a density-based clustering approach that helps identify these out-
Figure 8: Interface of an application that allows a user to compute traffic splits at every intersection along the I-95 corridor between Washington and Baltimore. The snapshot shows distribution of traffic moving northwest from the selected (purple) link. Brighter colors indicate more traffic coming from the selected road link.

Figure 9: Heat map indicates several major trip generators, including Baltimore downtown, Baltimore Washington International airport – Fort Mead, Bethesda – German Town.
liers based on the spatial distribution of the dataset using the previously-described DBSCAN algorithm. For some applications it is important to characterize the properties of different clusters or differentiate between core and border points, but here we are primarily interested in whether each latitude/longitude location is an anomaly or not.

To showcase this approach, we consider a set of trips originating from a single location and use the DBSCAN algorithm to identify outliers for 10, 20, 30, and 40 minute trips. As an example, we focus on a set of approximately 3,000 trips beginning from the Port of Baltimore, which consists of 218,302 total points (95,155 within 10 min, 141,586 within 20 min, 164,053 within 30 min, and 178,257 within 40 min). Using the scikit-learn Python implementation of DBSCAN (Pedregosa et al., 2011), we cluster the points for different combinations of input parameters (Figure 10(a)), remove the points that algorithm identifies as outliers, and visualize the results in the form of isochrones. Figure 10(b) shows the results of running DBSCAN with $\varepsilon = 1.1$ km, $\text{MinPts} = 60$ points for all points within 10 min travel time of the origin, with outlier points colored red, non-outlier points colored brown, and a concave hull connecting the boundary points of the relevant region. Note that, if the algorithm had not removed the marked outliers, the concave hull would have included these points too, suggesting inflated levels of mobility. This procedure is repeated for 20, 30, and 40 minute trips, and the concave hulls bounding the non-outlier points are plotted in Figure 10(c). The shape of the different concave hulls reflects the fact that mobility is greatest along the main highways, which matches our intuition. A similar visual is provided while considering only trips with 5-5:30 PM departures from the Port of Baltimore. Figure 10(d) indicates reduced mobility during the observed peak period. Finally, as a validation of the outlined approach, we note that isochrones for heavy vehicles designed based on a traditional method (Figure 11) show very similar patterns to those observed in Figure 10(c).

While the proposed density-based clustering approach represents an innovative application of trajectory data to quantify mobility, DBSCAN’s results are very sensitive to the input parameters, where the best input parameters depend heavily on the size and specific distribution of the dataset (e.g., note different parameter values reported in Figure 10(a) and different number of points mentioned in the previous paragraph). Consequently, the proposed approach suffers from excessive parameter tuning and the need for visual sanity checks. As an extension of the proposed approach, one could try to develop a method to automatically adjust parameter setting for different case studies. Development of such a method would certainly represent a challenging task.

5.4. Designing public transit

Public transit operates most efficiently when it provides services that appropriately match customers’ spatial and temporal demand. Since GPS traces capture spatio-temporal patterns, they can be used to improve public transit by comparing existing transit routes with actual trips in a metropolitan region. To illustrate this application, we focus on trips in the Annapolis, MD region and cluster their O-D pairs using the OPTICS algorithm (Ankerst et al., 1999). The raw trajectories of Annapolis trips are visualized in Figure 12(a), while clustered O-D pairs are color-coded and shown in Figure 12(b). The map-matched trajectories are then overlaid onto the existing Annapolis transit
| Period | Isochrone (min) | ε (km) | MINPts (pts) |
|--------|----------------|--------|--------------|
| all    | 10             | 1.1    | 60           |
|        | 20             | 1.3    | 20           |
|        | 30             | 1.4    | 10           |
|        | 40             | 1.6    | 5            |
| peak   | 10             | 0.5    | 20           |
|        | 20             | 0.9    | 15           |
|        | 30             | 1.1    | 5            |
|        | 40             | 1.5    | 5            |

(a) Parameter setting for DBSCAN

(b) A 10-min isochrone and outliers (all)

(c) Port of Baltimore isochrones (all)

(d) Port of Baltimore isochrones (peak)

Figure 10: DBSCAN is used to construct isochrones from trip waypoints using the outlined parameter settings. After filtering waypoints based on density, the isochrone is obtained by constructing a concave hull which connects the boundary points. Reduced mobility is observed when isochrones are constructed only based on trips with 5-5:30 departures, which represents peak period in the observed area.

Figure 11: Traditional isochrones for heavy-vehicles from OpenRouteService [Neis and Zipf 2008] are used to validate the proposed clustering-based approach.
network in Figure 12(c), applying a linear heat map in order to emphasize the most-traveled routes. This visual comparison of important trajectories and the transit network reveals that some highly-traveled routes are currently not covered with the transit system. This simple visual comparison may be useful for facilitating discussion with the City of Annapolis about modifying bus routes to best accommodate additional trips. Furthermore, given sufficient interest in a full transit system evaluation or re-design, the GPS traces could be used in conjunction with an array of data mining, operations research and microsimulation techniques to explore spatio-temporal characteristics of trips, optimize routes and service frequencies, and evaluate potential savings.

5.5. Traffic performance measurement and congestion alleviation

State transportation agencies are required to calculate and report annual performance metrics to the Federal government ([MAP-2I, 2016]), which depend on traffic speeds and volumes along the state transportation network. Speeds can be estimated reasonably well based on data collected from probe vehicles, but accurate traffic volume information is available only for a handful of roads links where ATR stations are located. Consequently, transportation agencies resort to estimating volumes based on annual daily average volumes and historic speed data, which may not result in accurate estimates. Since trajectory data represents a sample of traffic volumes, it can be used to estimate volumes throughout the observed road network, either through expansion factors or more advanced machine learning techniques. A simplistic approach would consist of counting the number of GPS trips along road links, and scaling it with the expansion factor of about 54 (Figure 5). This would result in rough estimates of traffic volumes for the observed time period, which could be improved by looking at additional factors, such as speeds, weather, temporal and road characteristics, incident reports, etc.

From a different perspective, we can use trajectory data to help mitigate traffic congestion, which occurs when many people travel to common locations at the same time. Using the OPTICS algorithm, we can detect groups of trips with similar destinations and arrival times (Figure 13). Thus, by identifying the main trip attractors (e.g., National Institutes of Health and Walter Reed Medical Center, which are located across the street from each other), transportation agencies or interested organizations may be able to introduce control measures (e.g., managed lanes, dynamic tolling, carpooling, flexible working hours) that help shift demand and reduce congestion.

5.6. Safety

Detailed trajectory data can reveal speed profiles of millions of anonymized drivers, which has important safety implications. We compute average speeds between all consecutive waypoints in our data set (which includes 1.4 billion GPS points), and focus on ones with higher than average speeds. Figure 14 shows a heat map that indicates locations where higher speeds are recorded with greater frequency. The result could be readily used by agencies in charge of deploying speed cameras and radar patrols, which would likely help improve safety and reduce property damage. However, we stress here that trajectory data is anonymized and speed violations cannot be traced back to
Figure 12: Clusters of trips in Annapolis can be used to modify bus transit network in order to accommodate additional movements. The last visual shows commonly traveled routes that currently are not covered with the transit system.
individuals; the goal is to identify segments of the road network that may be good candidates for safety improvements.

5.7. Weight control

Some truckers may overload their vehicles in order to increase their productivity, which results in excessive pavement and environmental damages. For example, pavement damage attributed to overweight trucks in California was roughly estimated at $20–$30 million per year (Santero et al., 2005). An effective way of reducing this damage is to implement weigh-in-motion systems (Marković et al., 2015, 2017), which are designed to detect and fine overweight trucks. However, an issue with these systems is that they are inroad facilities, which once deployed in a transportation network remain in their locations for several years. Thus, truckers quickly learn the locations of these systems and can start taking detours in order to avoid them (Cottrell Jr, 1992), which can lead to increased pavement and environmental damage due to more vehicle miles traveled (Bešinović et al., 2013).

Trajectory data can reveal route choices of millions of anonymized drivers, which can be used to investigate the extent to which truckers are avoiding weigh-in-motion systems. As an illustrative case study, we consider two weigh-in-motion systems in Maryland and compute the percentage of vehicles that take immediate detours (Figure 15). The results indicate that large trucks are not bypassing the systems, whereas 0.6 – 1.8% of other vehicles are deviating from the main road in the immediate vicinity of the weigh-in-motion systems and then returning to the main road afterwards. This may suggest an evasion problem, but would need to be further verified by looking into the travel times along the two routes when the potential evasions occurred. It is noteworthy that considering
Figure 14: Heat map of locations with higher than average speed recordings indicates candidate locations for implementation of speed cameras. After an initial analysis at the regional level (Figure 14(a)), an analyst can focus on a particular road road segment and explore directional speed profiles (Figure 14(b)). Clearly, color thresholds can be changed to narrow down candidate locations.

additional alternative routes would likely provide a better picture of potential evasive strategies. Again, we stress that trajectory data is anonymized and potential evasions cannot be traced back to individuals; the objective is to identify areas that may be good candidates for additional weight control, which would reduce excessive damages and also improve safety for all the road users.

6. Discussion

The previous section showcased a variety of trajectory data applications using 20 million GPS trips recorded in the state of Maryland. These applications were chosen to show how trajectory datasets may be useful for transportation agencies, and thus include both existing and novel approaches. The three innovative applications presented in this paper are summarized as follows:

- We propose an approach for constructing isochrones via density-based clustering/filtering of trajectory data, which yields a different measure of accessibility than isochrones calculated from travel times. The proposed isochrones indicate locations where many people have traveled within a specified time period, whereas the latter show locations where people could theoretically travel in the same period of time. Accordingly, some less-visited (perhaps unsafe or unpopular) neighborhoods may be excluded from the trajectory-based isochrones, thus providing a different picture of accessibility to various facilities (e.g., supermarkets, gas stations). Another advantage of designing isochrones based on trajectory data is that it can be carried out without information about the transportation network and historic travel times along various road links.
Figure 15: Examining potential evasion of weigh-in-motion systems at MD 32 East (left) and US-301 North (right) along immediate detours. Locations of weigh-in-motion systems along the main routes are indicated with pentagrams.
• We propose using trajectory data to identify locations with frequent high-speed observations, which may be ideal candidates for deploying speed cameras or radar patrols. This approach involves applying a heat map to identify road segments with the highest density of consecutive waypoints with greater than average speeds. This simple approach can be easily implemented even without reconstructing road-based trajectories, provided the data is highly granular (e.g., 1 second time lapse between consecutive waypoints). The result could be readily used by agencies in charge of enforcing speed limits, which would likely help improve safety and reduce property damage.

• We propose using trajectory data to examine whether trucks evade weigh-in-motion systems by changing their travel paths, which would enable agencies to decide if they should take additional measures to enforce weight control. The approach consists of examining the percentage of vehicles taking detours around the weigh-in-motion systems, which highlights locations where a substantial number of potential violators are observed and would allow agencies to deploy mobile patrols to deter truckers from overloading their vehicles and taking alternative routes. This would help reduce the excessive pavement and environmental damage associated with overweight vehicles, which can cost millions of dollars annually.

Since characteristics of trajectory data can significantly influence its applicability, we provide a discussion about some possible challenges that transportation agencies should be aware of when purchasing trajectory data. The following is a list of potential data-related issues that agencies may want to discuss with data vendors in order to obtain a more complete picture about applicability of a specific dataset. Some general recommendations to transportation agencies interested in acquiring trajectory data are included as well.

• **Sampling rate.** The average time lapse between consecutive waypoints significantly affects applicability of trajectory data, and agencies should try to acquire data with the highest granularity possible (e.g., with the median or average time lapse of 1 second). For example, a large time lapse between waypoints may not influence estimation of O-D matrices, but it could make reconstruction of road-based trajectories a significant challenge, especially in dense urban areas where it may be impossible to determine which route a vehicle took. Thus, it is important to request information about the granularity of data and assess how it would influence the anticipated analysis before actually acquiring data. Also, requesting road-based trajectories in addition to raw data, may save agencies quite a bit of time and resources needed for map matching, which was discussed in Section 3.1.

• **Spatial precision.** The number of decimal numbers used to report waypoint latitudes/longitudes is another factor that can influence applicability of trajectory data. For example, rounding a waypoint location to four decimal numbers introduces an error of about 11 m. While this error would not necessarily prevent us from reconstructing road-based trajectories or studying
demand, it would significantly affect any speed estimates and its use in microsimulation models. Assuming that the median spacing between two consecutive waypoints is 28 m (Figure 2), location errors of 11 m would make speed estimates meaningless. The same applies to computing vehicle acceleration/deceleration rates that are needed for microsimulation models used to estimate emissions, such as CMEM \cite{An1997} and VT-Micro \cite{Rakha2004}. Thus, agencies should request latitudes/longitudes expressed with six decimal numbers, and still account for the errors that are inherent to GPS technology.

- **Division of trajectories into trips.** Transportation agencies should be aware that GPS companies may reset a trip whenever the vehicle is idle for a specified period of time (e.g., 10 minutes). When this occurs within the boundaries of a state for which data was purchased, an analyst can still chain consecutive trips by looking at the unique device identifications. However, when a trip gets reset once it leaves the state, than the information about subsequent lags of the trip is lost. This is probably the reason that Figure 2 does not include any trips going to the West Coast, as such a long trip would necessitate stops long enough to reset the trip. To overcome this problem and gain better insight into long-distance trips, agencies from multiple states could jointly purchase data for an entire region (e.g., East Coast or all of USA), which also may be more cost efficient due to economies of scale.

- **Population bias.** Transportation agencies should be aware of the bias in data towards certain types of vehicles. For example, the dataset discussed in this paper is biased towards delivery trucks (Figure 3). This may not represent a major issue if the observed region includes a network of ATR stations that can differentiate between different vehicle types. In this case, an analyst can determine the penetration rates of different types of vehicles (passenger cars vs. trucks) and account for any bias in further analysis. However, when such a network of sensors is unavailable, correcting for the bias becomes a challenge and may limit applications of trajectory data (e.g., estimation of an O-D matrix becomes a challenge). Therefore the government agencies interested in purchasing trajectory data should also account for the availability of other data sources that would enable them to correct for the aforementioned bias in data.

- **Unique device identifications.** Each trip in a trajectory dataset includes an identification (ID) of the device it was recorded from. Device IDs enable an analyst to chain consecutive trips of the same vehicle and thereby reconstruct its movement over a longer period of time, which provides a better insight into mobility patterns. However, data vendors may decide to periodically change device IDs (e.g., at midnight) for privacy or some other reasons, which clearly limits the analysis. Thus, transportation agencies interested in purchasing trajectory data should inquire about vendor’s policies with respect to resetting device IDs and account for its implications on their analyses. Additional issues that analysts should be aware of are occasionally duplicated or swapped device IDs, which may arise when resetting device IDs. These and other issues related to trajectory data are discussed in Andrienko, Andrienko and Fuchs \cite{Andrienko2016}.
7. Conclusions

This paper synthesizes innovative applications of trajectory data in transportation, which is relevant to government agencies looking to introduce this type of data into their analyses and decision making processes. We provide a comprehensive literature review discussing applications of trajectory data in six areas of transportation systems analysis: demand estimation, modeling human behavior, designing public transit, traffic performance measurement and prediction, environment and safety. Additionally, we perform an extensive analysis of 20 million GPS trajectories in Maryland, demonstrating both existing and new applications of trajectory data in transportation. We employ an array of techniques encompassing data processing and management, machine learning, and visualization, and describe best-practices for using them to extract value from trajectory data, thus allowing transportation agencies to estimate the time and effort needed to introduce this type of data into their modeling efforts. As trajectory data becomes more prevalent and acquisition costs decrease, we believe that this type of data will become an invaluable resource to transportation agencies across the world. To this end, we are working to incorporate visual analytics into the Regional Integrated Transportation Information System (RITIS, 2016), which is widely used by government agencies across the United States to share data, obtain situational awareness, evaluate performance, and disseminate insights with the public.

Acknowledgments

The authors would like to thank Subrat Mahapatra and the Maryland State Highway Administration for their support throughout this project. Help from the I-95 Corridor Coalition and the City of Annapolis are also appreciated. This support is gratefully acknowledged, but it implies no endorsement of the findings.

References

Abdel-Aty, M. A. and Radwan, A. E. (2000), ‘Modeling traffic accident occurrence and involvement’, Accident Analysis & Prevention 32(5), 633–642.

Ahn, K. and Rakha, H. (2008), ‘The effects of route choice decisions on vehicle energy consumption and emissions’, Transportation Research Part D: Transport and Environment 13(3), 151–167.

Alexander, L., Jiang, S., Murga, M. and González, M. C. (2015), ‘Origin–destination trips by purpose and time of day inferred from mobile phone data’, Transportation Research Part C: Emerging Technologies 58, 240–250.

Allström, A., Bayen, A. M., Fransson, M., Gundlegård, D., Patire, A. D., Rydergren, C. and Sandin, M. (2014), ‘Calibration framework based on bluetooth sensors for traffic state estimation using a velocity based cell transmission model’, Transportation Research Procedia 3, 972–981.
An, F., Barth, M., Norbeck, J. and Ross, M. (1997), ‘Development of comprehensive modal emissions model: Operating under hot-stabilized conditions’, Transportation Research Record: Journal of the Transportation Research Board (1587), 52–62.

Andrienko, G., Andrienko, N., Bak, P., Keim, D. and Wrobel, S. (2013), Visual analytics of movement, Springer Science & Business Media.

Andrienko, G., Andrienko, N., Chen, W., Maciejewski, R. and Zhao, Y. (2017), ‘Visual analytics of mobility and transportation: State of the art and further research directions’, IEEE Transactions on Intelligent Transportation Systems.

Andrienko, G., Andrienko, N. and Fuchs, G. (2016), ‘Understanding movement data quality’, Journal of location Based services 10(1), 31–46.

Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D. and Giannotti, F. (2009), Interactive visual clustering of large collections of trajectories, in ‘Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on’, IEEE, pp. 3–10.

Andrienko, G., Andrienko, N. and Wrobel, S. (2007), ‘Visual analytics tools for analysis of movement data’, ACM SIGKDD Explorations Newsletter 9(2), 38–46.

Andrienko, N. and Andrienko, G. (2006), Exploratory analysis of spatial and temporal data: a systematic approach, Springer Science & Business Media.

Andrienko, N., Andrienko, G., Fuchs, G. and Jankowski, P. (2016), ‘Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces’, Information Visualization 15(2), 117–153.

Andrienko, N., Andrienko, G., Stange, H., Liebig, T. and Hecker, D. (2012), ‘Visual analytics for understanding spatial situations from episodic movement data’, KI-Künstliche Intelligenz 26(3), 241–251.

Ankerst, M., Breunig, M. M., Kriegel, H.-P. and Sander, J. (1999), OPTICS: Ordering points to identify the clustering structure, in ‘ACM Sigmod Record’, Vol. 28, ACM, pp. 49–60.

Ashbrook, D. and Starner, T. (2003), ‘Using GPS to learn significant locations and predict movement across multiple users’, Personal and Ubiquitous Computing 7(5), 275–286.

Bandeira, J. M., Fontes, T., Pereira, S. R., Fernandes, P., Khattak, A. and Coelho, M. C. (2014), ‘Assessing the importance of vehicle type for the implementation of eco-routing systems’, Transportation Research Procedia 3, 800–809.

Bastani, F., Huang, Y., Xie, X. and Powell, J. W. (2011), A greener transportation mode: Flexible routes discovery from GPS trajectory data, in ‘Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems’, ACM, pp. 405–408.
Ben-Akiva, M. E. and Lerman, S. R. (1985), *Discrete choice analysis: Theory and application to travel demand*, Vol. 9, MIT press.

Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F. and Sbodio, M. L. (2013), AllAboard: A system for exploring urban mobility and optimizing public transport using cellphone data, *in* ‘Joint European Conference on Machine Learning and Knowledge Discovery in Databases’, Springer, pp. 663–666.

Bešinović, N., Marković, N. and Schonfeld, P. (2013), Optimal allocation of truck inspection stations based on k-shortest paths, *in* ‘Transportation Research Board 92nd Annual Meeting’, number 13-3123.

Brennan Jr, T. M., Remias, S. M., Grimmer, G., Horton, D., Cox, E. and Bullock, D. (2013), ‘Probe vehicle-based statewide mobility performance measures for decision makers’, *Transportation Research Record: Journal of the Transportation Research Board* (2338), 78–90.

Brennan Jr, T. M., Remias, S. M. and Manili, L. (2015), ‘Performance measures to characterize corridor travel time delay based on probe vehicle data’, *Transportation Research Record: Journal of the Transportation Research Board* (2526), 39–50.

Broach, J., Dill, J. and Gliebe, J. (2012), ‘Where do cyclists ride? A route choice model developed with revealed preference GPS data’, *Transportation Research Part A: Policy and Practice* 46(10), 1730–1740.

Buekers, J., Van Holderbeke, M., Bierkens, J. and Panis, L. I. (2014), ‘Health and environmental benefits related to electric vehicle introduction in EU countries’, *Transportation Research Part D: Transport and Environment* 33, 26–38.

Castro, P. S., Zhang, D. and Li, S. (2012), Urban traffic modelling and prediction using large scale taxi GPS traces, *in* ‘International Conference on Pervasive Computing’, Springer, pp. 57–72.

Cevallos, F. and Zhao, F. (2006), ‘Minimizing transfer times in public transit network with genetic algorithm’, *Transportation Research Record: Journal of the Transportation Research Board* (1971), 74–79.

Charlton, B., Hood, J., Sall, E. and Schwartz, M. A. (2011), Bicycle route choice data collection using GPS-enabled smartphones, *in* ‘Transportation Research Board 90th Annual Meeting’, number 11-2652.

Chen, C., Li, L., Hu, J. and Geng, C. (2010), Calibration of MITSIM and IDM car-following model based on NGSIM trajectory datasets, *in* ‘Vehicular Electronics and Safety (ICVES), 2010 IEEE International Conference on’, IEEE, pp. 48–53.
Chen, C., Zhang, D., Li, N. and Zhou, Z.-H. (2014), ‘B-Planner: Planning bidirectional night bus routes using large-scale taxi GPS traces’, *IEEE Transactions on Intelligent Transportation Systems* **15**(4), 1451–1465.

Chen, E. and Tarko, A. (2012), ‘Analysis of crash frequency in work zones with focus on police enforcement’, *Transportation Research Record: Journal of the Transportation Research Board* (2280), 127–134.

Chien, S. I.-J. and Spasović, L. N. (2002), ‘Optimization of grid bus transit systems with elastic demand’, *Journal of advanced transportation* **36**(1), 63–91.

Chien, S. and Schönfeld, P. (1997), ‘Optimization of grid transit system in heterogeneous urban environment’, *Journal of Transportation Engineering* **123**(1), 28–35.

Chu, L., Oh, S. and Recker, W. (2005), Adaptive Kalman filter based freeway travel time estimation, *in ‘84th TRB Annual Meeting, Washington DC’*, Citeseer.

Coifman, B. (2001), ‘Improved velocity estimation using single loop detectors’, *Transportation Research Part A: Policy and Practice* **35**(10), 863–880.

Coifman, B. (2015), ‘Empirical flow-density and speed-spacing relationships: Evidence of vehicle length dependency’, *Transportation Research Part B: Methodological* **78**, 54–65.

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiratta, S. R. and González, M. C. (2015), ‘Analyzing cell phone location data for urban travel: Current methods, limitations, and opportunities’, *Transportation Research Record: Journal of the Transportation Research Board* (2526), 126–135.

Colombaroni, C. and Fusco, G. (2014), ‘Artificial neural network models for car following: Experimental analysis and calibration issues’, *Journal of Intelligent Transportation Systems* **18**(1), 5–16.

Cottrell Jr, B. H. (1992), The avoidance of weigh stations in Virginia by overweight trucks, Technical report, Virginia Transportation Research Council.

Dailey, D. J. (1999), ‘A statistical algorithm for estimating speed from single loop volume and occupancy measurements’, *Transportation Research Part B: Methodological* **33**(5), 313–322.

Dozza, M. and Werneke, J. (2014), ‘Introducing naturalistic cycling data: What factors influence bicyclists safety in the real world?’, *Transportation Research Part F: Traffic Psychology and Behaviour* **24**, 83–91.

Ellersiek, T., Andrienko, G., Andrienko, N., Hecker, D., Stange, H. and Mueller, M. (2013), Using Bluetooth to track mobility patterns: Depicting its potential based on various case studies, *in ‘Proceedings of the Fifth ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness’*, ACM, pp. 1–7.
EPA (2016), Inventory of U.S. greenhouse gas emissions and sinks: 1990-2014, Technical report.

Eriksson, J. (2013), Evaluation of traffic-flow monitoring technologies: Cicero-midway smart corridor case study, Technical report, FHWA-ICT-13-007.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, in ‘Kdd’, Vol. 96, pp. 226–231.

Evensen, G. (2009), Data assimilation: The ensemble Kalman filter, Springer Science & Business Media.

Fan, W. and Machemehl, R. B. (2004), Optimal transit route network design problem: Algorithms, implementations, and numerical results, Technical report, SWUTC/04/167244-1.

Feng, T., Arentze, T. and Timmermans, H. (2011), Instantaneous emission modeling with GPS-based vehicle activity data: Results of diesel trucks for one-day trips, in ‘Proceedings of the Eastern Asia Society for Transportation Studies’, Vol. 2011, Eastern Asia Society for Transportation Studies, pp. 147–147.

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S. and Trasarti, R. (2011), ‘Unveiling the complexity of human mobility by querying and mining massive trajectory data’, The VLDB Journal The International Journal on Very Large Data Bases 20(5), 695–719.

Groves, R. M. (2006), ‘Nonresponse rates and nonresponse bias in household surveys’, Public opinion quarterly 70(5), 646–675.

Guihaire, V. and Hao, J.-K. (2008), ‘Transit network design and scheduling: A global review’, Transportation Research Part A: Policy and Practice 42(10), 1251–1273.

Guo, J., Xia, J. and Smith, B. L. (2009), ‘Kalman filter approach to speed estimation using single loop detector measurements under congested conditions’, Journal of Transportation Engineering 135(12), 927–934.

Haklay, M. and Weber, P. (2008), ‘Openstreetmap: User-generated street maps’, IEEE Pervasive Computing 7(4), 12–18.

Hara, Y. and Kuwahara, M. (2015), ‘Traffic monitoring immediately after a major natural disaster as revealed by probe data – A case in Ishinomaki after the Great East Japan Earthquake’, Transportation Research Part A: Policy and Practice 75, 1–15.

Hatipkarasulu, Y., Wolshon, B. and Quiroga, C. (2000), A GPS approach for the analysis of car following behavior, in ‘Transportation Research Board 79th Annual Meeting’.

Hellinga, B. and Fu, L. (1999), ‘Assessing expected accuracy of probe vehicle travel time reports’, Journal of Transportation engineering 125(6), 524–530.
Herrera, J. C. and Bayen, A. M. (2008), Traffic flow reconstruction using mobile sensors and loop detector data, in ‘Transportation Research Board 87th Annual Meeting’.

Hood, J., Sall, E. and Charlton, B. (2011), ‘A GPS-based bicycle route choice model for San Francisco, California’, Transportation letters 3(1), 63–75.

Ikeda, Y. and Inoue, M. (2016), ‘An evacuation route planning for safety route guidance system after natural disaster using multi-objective genetic algorithm’, Procedia Computer Science 96, 1323–1331.

Iqbal, M. S., Choudhury, C. F., Wang, P. and González, M. C. (2014), ‘Development of origin–destination matrices using mobile phone call data’, Transportation Research Part C: Emerging Technologies 40, 63–74.

Jensen, A. F. and Larsen, T. V. (2014), Travel-time estimation in road networks using GPS data. Unpublished.

Jiang, S., Ferreira Jr, J. and González, M. C. (2015), Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore, in ‘Int. Workshop on Urban Computing’.

Johansson, A. and Helbing, D. (2010), Analysis of empirical trajectory data of pedestrians, in ‘Pedestrian and Evacuation Dynamics 2008’, Springer, pp. 203–214.

Joo, S., Oh, C., Jeong, E. and Lee, G. (2015), ‘Categorizing bicycling environments using GPS-based public bicycle speed data’, Transportation Research Part C: Emerging Technologies 56, 239–250.

Karypis, G., Han, E.-H. and Kumar, V. (1999), ‘Chameleon: Hierarchical clustering using dynamic modeling’, Computer 32(8), 68–75.

Kesting, A. and Treiber, M. (2013), Traffic Flow Dynamics: Data, Models and Simulation, Springer: Berlin.

Khattak, A. J., Khattak, A. J. and Council, F. M. (2002), ‘Effects of work zone presence on injury and non-injury crashes’, Accident Analysis & Prevention 34(1), 19–29.

Kim, S. and Coifman, B. (2014), ‘Comparing INRIX speed data against concurrent loop detector stations over several months’, Transportation Research Part C: Emerging Technologies 49, 59–72.

Koshak, N. and Fouda, A. (2008), Analyzing pedestrian movement in Mataf using GPS and GIS to support space redesign, in ‘The 9th international conference on design and decision support systems in architecture and urban planning’.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. and Marra, M. A. (2009), ‘Circos: An information aesthetic for comparative genomics’, Genome research 19(9), 1639–1645.
Kusano, S. and Inoue, M. (2013), Safety route guidance system using participatory sensing, in ‘Consumer Electronics (GCCE), 2013 IEEE 2nd Global Conference on’, IEEE, pp. 363–364.

Kuwahara, M. and Sullivan, E. C. (1987), ‘Estimating origin-destination matrices from roadside survey data’, Transportation Research Part B: Methodological 21(3), 233–248.

Langford, B. C., Chen, J. and Cherry, C. R. (2015), ‘Risky riding: Naturalistic methods comparing safety behavior from conventional bicycle riders and electric bike riders’, Accident Analysis & Prevention 82, 220–226.

Lee, H. and Coifman, B. (2011), ‘Quantifying loop detector sensitivity and correcting detection problems on freeways’, Journal of Transportation Engineering 138(7), 871–881.

Lee, R. J., Sener, I. N. and Mullins, J. A. (2014), Emerging data collection techniques for travel demand: A literature review, Technical report, TTI/SRP/14/161407-2.

Li, X., Lord, D., Zhang, Y. and Xie, Y. (2008), ‘Predicting motor vehicle crashes using support vector machine models’, Accident Analysis & Prevention 40(4), 1611–1618.

Li, Y., Luo, J., Chow, C.-Y., Chan, K.-L., Ding, Y. and Zhang, F. (2015), Growing the charging station network for electric vehicles with trajectory data analytics, in ‘Data Engineering (ICDE), 2015 IEEE 31st International Conference on’, IEEE, pp. 1376–1387.

Lima, A., Stanojević, R., Papagiannaki, D., Rodriguez, P. and González, M. C. (2016), ‘Understanding individual routing behaviour’, Journal of The Royal Society Interface 13(116), 20160021.

Lindsey, G., Hankey, S., Wang, X., Chen, J. and Gorjestani, A. (2013), ‘Feasibility of using GPS to track bicycle lane positioning’. Intelligent Transportation Systems Institute, Center for Transportation Studies, University of Minnesota.

Liu, Y., Jia, G., Tao, X., Xu, X. and Dou, W. (2014), A stop planning method over big traffic data for airport shuttle bus, in ‘Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on’, IEEE, pp. 63–70.

Lo, H., Zhang, N. and Lam, W. H. (1996), ‘Estimation of an origin-destination matrix with random link choice proportions: A statistical approach’, Transportation Research Part B: Methodological 30(4), 309–324.

Lu, Y., Yang, X. and Chang, G.-L. (2014), ‘Algorithm for detector-error screening on basis of temporal and spatial information’, Transportation Research Record: Journal of the Transportation Research Board (2443), 40–48.

Luo, D. and Ma, X. (2014), Analysis of cyclist behavior using naturalistic data: Data processing for model development, in ‘International Cycling Safety Conference (ICSC2014), 3rd, 2014, Gothenburg, Sweden’.
Lyu, Y., Chow, C.-Y., Lee, V. C., Li, Y. and Zeng, J. (2016), T2CBS: Mining taxi trajectories for customized bus systems, in ‘Computer Communications Workshops (INFOCOM WKSHPs), 2016 IEEE Conference on’, IEEE, pp. 441–446.

Ma, J. and Kockelman, K. (2006), Crash frequency and severity modeling using clustered data from Washington state, in ‘Intelligent Transportation Systems Conference, 2006. ITSC’06. IEEE’, IEEE, pp. 1621–1626.

Mahapatra, S., Wolniak, M., Beckett, E. and Farokhi Sadabadi, K. (2015), 2015 Maryland state highway mobility report, Technical report, Maryland State Highway Administration.

MAP-21 (2016), [https://www.fhwa.dot.gov/map21/factsheets/pm.cfm](https://www.fhwa.dot.gov/map21/factsheets/pm.cfm) Accessed: 2016-12-23.

Marković, N., Ryzhov, I. O. and Schonfeld, P. (2015), ‘Evasive flow capture: Optimal location of weigh-in-motion systems, tollbooths, and security checkpoints’, *Networks* **65**(1), 22–42.

Marković, N., Ryzhov, I. O. and Schonfeld, P. (2017), ‘Evasive flow capture: A multi-period stochastic facility location problem with independent demand’, *European Journal of Operational Research* **257**(2), 687–703.

McArdle, G., Demšar, U., van der Spek, S. and McLoone, S. (2014), ‘Classifying pedestrian movement behaviour from GPS trajectories using visualization and clustering’, *Annals of GIS* **20**(2), 85–98.

McNally, M. G. (2007), ‘The four-step model’, *Handbook of transport modelling* **1**, 35–41.

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J. and Damas, L. (2016), ‘Time-evolving OD matrix estimation using high-speed GPS data streams’, *Expert Systems with Applications* **44**, 275–288.

Mudge, R., Mahmassani, H., Haas, R., Talebpour, A. and Carroll, L. (2013), Work zone performance measurement using probe data, Technical report, FHWA-HOP-13-043.

Neis, P. and Zipf, A. (2008), ‘Openrouteservice. org is three times “open”: Combining opensource, opens and openstreetmaps’, *GIS Research UK (GISRUK 08). Manchester*.

Newson, P. and Krumm, J. (2009), Hidden Markov map matching through noise and sparseness, in ‘Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems’, ACM, pp. 336–343.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011), ‘Scikit-learn: Machine learning in Python’, *Journal of Machine Learning Research* **12**(Oct), 2825–2830.
Pelekis, N., Andrienko, G., Andrienko, N., Kopanakis, I., Marketos, G. and Theodoridis, Y. (2012), ‘Visually exploring movement data via similarity-based analysis’, *Journal of Intelligent Information Systems* **38**(2), 343–391.

Pinelli, F., Nair, R., Calabrese, F., Berlingerio, M., Di Lorenzo, G. and Sbodio, M. L. (2016), ‘Data-driven transit network design from mobile phone trajectories’, *IEEE Transactions on Intelligent Transportation Systems* **17**(6), 1724–1733.

QGIS Development Team (2015), *QGIS Geographic Information System*, Open Source Geospatial Foundation.  
**URL:** [http://qgis.osgeo.org](http://qgis.osgeo.org)

Rakha, H., Ahn, K. and Trani, A. (2004), ‘Development of VT-Micro model for estimating hot stabilized light duty vehicle and truck emissions’, *Transportation Research Part D: Transport and Environment* **9**(1), 49–74.

Rauh, N., Franke, T. and Krems, J. F. (2015), ‘Understanding the impact of electric vehicle driving experience on range anxiety’, *Human factors* **57**(1), 177–187.

RITIS (2016), ‘The regional integrated transportation information system (RITIS)’, [http://www.cattlab.umd.edu/?portfolio=ritis](http://www.cattlab.umd.edu/?portfolio=ritis) Accessed: 2016-10-14.

Santero, N. J., Nokes, W. and Harvey, J. T. (2005), ‘Virtual weigh stations: The business case’. Technical Memorandum TM-UCB-PRC-2005-3.

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z. and González, M. C. (2013), ‘Unravelling daily human mobility motifs’, *Journal of The Royal Society Interface* **10**(84), 20130246.

Sekula, P., Marković, N., Vander Laan, Z., Farokhi Sadabadi, K. and Skibniewski, M. J. (2017), Predicting work zone collision probabilities via clustering: Application in optimal deployment of highway response teams. Submitted for publication.

Sellmair, R. and Hamacher, T. (2014), ‘Optimization of charging infrastructure for electric taxis’, *Transportation Research Record: Journal of the Transportation Research Board* (2416), 82–91.

Sharman, B. and Roorda, M. (2010), Analysis tool to process passively collected GPS data for commercial vehicle demand modeling applications, in ‘Innovations in Freight Demand Modeling and Data Symposium, Washington, DC’.

Siša-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U. and Fotheringham, A. S. (2016), ‘Analysis of human mobility patterns from GPS trajectories and contextual information’, *International Journal of Geographical Information Science* **30**(5), 881–906.
Song, X., Zhang, Q., Sekimoto, Y. and Shibasaki, R. (2014), Intelligent system for urban emergency management during large-scale disaster, in ‘Proceedings of the Conference on Artificial Intelligence (AAAI14)’, pp. 458–464.

Stenneth, L., Wolfson, O., Yu, P. S. and Xu, B. (2011), Transportation mode detection using mobile phones and GIS information, in ‘Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems’, ACM, pp. 54–63.

Strauss, J., Miranda-Moreno, L. F. and Morency, P. (2015), ‘Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts’, Accident Analysis & Prevention 83, 132–142.

Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A. and González, M. C. (2015), ‘The path most traveled: Travel demand estimation using big data resources’, Transportation Research Part C: Emerging Technologies 58, 162–177.

Treiber, M., Kesting, A. and Thiemann, C. (2008), How much does traffic congestion increase fuel consumption and emissions? Applying a fuel consumption model to the NGSIM trajectory data, in ‘87th Annual Meeting of the Transportation Research Board, Washington, DC’.

Van der Spek, S., Van Schaick, J., De Bois, P. and De Haan, R. (2009), ‘Sensing human activity: GPS tracking’, Sensors 9(4), 3033–3055.

van Vuren, T., Carey, C. and MacDonald, M. (2011), Building practical origin-destination (OD/trip) matrices from automatically collected GPS data, in ‘European Transport Conference, Glasgow’.

van Zuylen, H. J. and Willumsen, L. G. (1980), ‘The most likely trip matrix estimated from traffic counts’, Transportation Research Part B: Methodological 14(3), 281–293.

Venugopal, S. and Tarko, A. (2000), ‘Safety models for rural freeway work zones’, Transportation Research Record: Journal of the Transportation Research Board (1715), 1–9.

Wang, P., Hunter, T., Bayen, A. M., Schechtner, K. and González, M. C. (2012), ‘Understanding road usage patterns in urban areas’, Scientific reports 2.

Wang, Z., Jin, B., Zhang, F., Yang, R. and Ji, Q. (2016), Discovering trip patterns from incomplete passenger trajectories for inter-zonal bus line planning, in ‘IFIP International Conference on Network and Parallel Computing’, Springer, pp. 160–171.

Wedin, D. (2015), Travel time estimation in Stockholm using historical GPS data, Master’s thesis, Uppsala University, Uppsala, Sweden.

Wei, W., Xiucheng, G., Jing, J. and Bin, R. (2010), GPS probe based freeway real-time travel speed estimation using Kalman filter, in ‘Intelligent System Design and Engineering Application (ISDEA), 2010 International Conference on’, Vol. 1, IEEE, pp. 797–800.
Widhalm, P., Yang, Y., Ulm, M., Athavale, S. and González, M. C. (2015), ‘Discovering urban activity patterns in cell phone data’, *Transportation* **42**(4), 597–623.

Xia, C., Cochrane, C., DeGuire, J., Fan, G., Holmes, E., McGuirl, M., Murphy, P., Palmer, J., Carter, P., Slivinski, L. et al. (2016), Lagrangian data assimilation in traffic-flow models. Unpublished.

Xie, Y., Lord, D. and Zhang, Y. (2007), ‘Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis’, *Accident Analysis & Prevention* **39**(5), 922–933.

Xu, Y. and González, M. C. (2016), Collective benefits in traffic during mega events via the use of information technologies. Unpublished.

Yamamoto, T., Liu, K. and Morikawa, T. (2006), Variability of travel time estimates using probe vehicle data, in ‘Proceedings of the Fourth International Conference on Traffic and Transportation Studies (ICTTS)’, pp. 278–287.

Yang, F., Jin, P. J., Wan, X., Li, R. and Ran, B. (2014), Dynamic origin-destination travel demand estimation using location based social networking data, in ‘Transportation Research Board 93rd Annual Meeting’.

Yang, H., Ozbay, K., Ozturk, O. and Xie, K. (2015), ‘Work zone safety analysis and modeling: A state-of-the-art review’, *Traffic injury prevention* **16**(4), 387–396.

Yang, H., Ozbay, K., Ozturk, O. and Yildirimoglu, M. (2013), ‘Modeling work zone crash frequency by quantifying measurement errors in work zone length’, *Accident Analysis & Prevention* **55**, 192–201.

Yang, J., Dong, J. and Hu, L. (2017), ‘A data-driven optimization-based approach for siting and sizing of electric taxi charging stations’, *Transportation Research Part C: Emerging Technologies* **77**, 462–477.

Yang, J.-S. (2005), Travel time prediction using the GPS test vehicle and Kalman filtering techniques, in ‘Proceedings of the 2005, American Control Conference, 2005’, IEEE, pp. 2128–2133.

Yang, Y., Widhalm, P., Athavale, S. and González, M. C. (2016), Mobility sequence extraction and labeling using sparse cell phone data, in ‘Proceedings of the Conference on Artificial Intelligence (AAAI16)’.

Yao, Z., Wei, H., Liu, H. and Li, Z. (2013), ‘Statistical vehicle specific power profiling for urban freeways’, *Procedia-Social and Behavioral Sciences* **96**, 2927–2938.

Yim, Y. B. and Cayford, R. (2001), Investigation of vehicles as probes using global positioning system and cellular phone tracking: Field operational test, Technical report, California Partners for Advanced Trans. and Highways (PATH).
Young, S. (2014), Use of probe and bluetooth data for arterial performance measures in the I-95 corridor coalition, *in* ‘NATMEC Workshop on Arterial Data and Performance Monitoring, Chicago, IL’.

Zanjani, A. B., Pinjari, A. R., Kamali, M., Thakur, A., Short, J., Mysore, V. and Tabatabaee, S. F. (2015), ‘Estimation of statewide origin-destination truck flows from large streams of GPS data: Application for Florida statewide model’, *Transportation Research Record: Journal of the Transportation Research Board* **2(2494)**, 87–96.

Zegeye, S. K., De Schutter, B., Hellendoorn, H. and Breunesse, E. (2009), ‘Model-based traffic control for balanced reduction of fuel consumption, emissions, and travel time’, *IFAC Proceedings Volumes* **42**(15), 149–154.

Zhang, L., Cirillo, C., Xiong, C. and Hetrakul, P. (2011), Feasibility and benefits of advanced four-step and activity-based travel demand models for Maryland, Technical report, MD-11-SP009B4S.

Zheng, Y. (2015), ‘Trajectory data mining: An overview’, *ACM Transactions on Intelligent Systems and Technology (TIST)* **6**(3), 29.

Zheng, Y., Li, Q., Chen, Y., Xie, X. and Ma, W.-Y. (2008), Understanding mobility based on GPS data, *in* ‘Proceedings of the 10th international conference on Ubiquitous computing’, ACM, pp. 312–321.

Zheng, Y., Liu, L., Wang, L. and Xie, X. (2008), Learning transportation mode from raw GPS data for geographic applications on the web, *in* ‘Proceedings of the 17th international conference on World Wide Web’, ACM, pp. 247–256.

Zheng, Y. and Xie, X. (2011), ‘Learning travel recommendations from user-generated GPS traces’, *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(1), 2.