Singapore in Motion: Insights on Public Transport Service Level Through Farecard and Mobile Data Analytics

Hasan Poonawala
IBM Research
9 Changi Business Park
Singapore
hasanp@sg.ibm.com

Vinay Kolar
Cisco Systems Inc.
170 West Tasman Dr.
San Jose, CA
vinkolar@cisco.com

Sebastien Blandin
IBM Research
9 Changi Business Park
Singapore
sblandin@sg.ibm.com

Laura Wynter
IBM Research
9, Changi Business Park
Singapore
lwynter@sg.ibm.com

Sambit Sahu
IBM Research
1101 Kitchawan Road
Yorktown Heights, NY
sambits@us.ibm.com

Abstract

Given the changing dynamics of mobility patterns and rapid growth of cities, transport agencies seek to respond more rapidly to needs of the public with the goal of offering an effective and competitive public transport system. A more data-centric approach for transport planning is part of the evolution of this process. In particular, the vast penetration of mobile phones provides an opportunity to monitor and derive insights on transport usage. Real time and historical analyses of such data can give a detailed understanding of mobility patterns of people and also suggest improvements to current transit systems. On its own, however, mobile geolocation data has a number of limitations. We thus propose a joint telco-and-farecard-based learning approach to understanding urban mobility. The approach enhances telecommunications data by leveraging it jointly with other sources of real-time data. The approach is illustrated on the first- and last-mile problem as well as route choice estimation within a densely-connected train network.

Keywords

big data; mobility ; public transport route choice; map-matching

1. INTRODUCTION

1.1 Background

Traditional transportation planning processes can be costly and resource intensive since they are typically based on survey data. However, given the changing dynamics of mobility patterns, it is important for cities to maintain a comprehensive and high-frequency knowledge of mobility patterns. A data-driven approach to the monitoring, analysis, and planning of transport operations is paramount for smart cities.

The deep penetration of mobile phones provides an opportunity for transport agencies to monitor and derive valuable insights from commuters, and across their entire trips. With the proper analytics, anonymized mobile geolocation data has the potential for providing a detailed understanding of mobility patterns, with visibility on the trip segments traditionally un-observed, such as the first and last miles part of the trip. This understanding could then trigger a data-driven feedback for public transit improvements.

On its own, however, mobile geolocation data has a number of limitations. For example, the spatial resolution of the mobile geolocation data, for records without GPS information, may lead to high uncertainty in derived movement patterns. Secondly, the uncertainty in the penetration of mobile devices means that quantitative assessments of densities or volumes cannot be known precisely. Lastly, the heterogeneity of those two sources of errors over space (and time, for the latter) adds further difficulty to the results obtained from mobile geolocation data.

We thus propose a farecard-based learning approach to understand urban mobility from telecommunications data. Fusion with farecard data allows for harnessing the power of the mobile geolocation data while compensating for its limitations. In particular farecard data provides highly accurate quantitative knowledge on station to station travel-times and volumes, which can help anchor end-to-end trajectory information provided by mobile geolocation data.

We are also interested in designing Big Data models for two main types of analyses, namely the so-called first and last mile of public transport users, and the route choice of public transport users in the public transport network. Both require trajectory analytics and data fusion, which are described in this article.

First and last mile analysis is of importance to transport authorities to aid in defining services to and from train stations, including feeder bus routes and on-demand minibus service. The route choice of train passengers in a train network such as Singapore's is not directly available from the farecard data and so must be deduced from other sources. It is valuable to transport authorities in understanding route

KDD '16, August 13-17, 2016, San Francisco, CA, USA
© 2016 ACM. ISBN 978-1-4503-4232-2/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939723

crowding within the network, and adjusting accordingly route information systems. Additionally, the understanding of explanatory factors for route choice (travel-time, fare, etc.) is fundamental for planning.

1.2 Related work

With the increase of mobile data, the concept of crowd and *community sensing* [20] has seen a growing popularity in the recent decades. With the availability of GPS chips and additional sensors indirectly providing location contexts on most smartphones, spatio-temporal analytics which used to rely on inaccurate low-sample cellphone data has seen renewed interest for this new *analytic superfood* [17]. Mobile geolocation data has been used for instance for real-time traffic monitoring [36], adaptive routing [6, 30] and path inference [15] in the context of GPS-based participatory sensing, combined with other sources for city planning [28], and even as a potential replacement for the fixed sensing infrastructure, in particular on expressways [25].

A considerable amount of literature has been focused on extracting spatio-temporal activity patterns from large cellular network datasets [1], in combination with social media [11], and in some cases leveraging models from statistical physics [10]. Remarkable structure and consistency has been exhibited within mobile datasets [4, 31], supporting the design of long-term mobility forecast models able to achieve unexpectedly high accuracy [29]. It has been shown that based on mobile data, it is possible to identify tourists from residents, identify hotspots within a city, and detect preferred activity sequences, see for instance [3, 8, 34].

In the recent years, a growing literature has been concerned with augmenting geospatial and mobility insights with semantics information [18], in particular using social media during events [27] and for prediction of level of service on public transport networks [26].

In Singapore, the large penetration of mobile phones and the relatively high population density have been conducive to a number of studies analyzing public mobility patterns [21], deriving insights on the network resilience [14] based on network connectivity indicators and crowding. In [13] a system was developed to derive passenger traffic and route recommendations for the train network, using train network specific cell tower information. These studies serve as a building block for nation-wide simulation engines [24, 32], which can in turn support enhanced response to incidents [12, 16, 33].

In parallel, Big Data and cloud computing platforms have emerged leveraging fast and efficient frameworks able to run spatio-temporal and statistical models at scale [19] on terabytes of data.

We focus on the study of travel patterns which usie the public transit network. This requires the development of novel analytics for travel mode detection in the context of spatially uncertain mobile geolocation data. We further derive new actionable insights by fusion of mobile geolocation data with farecard data, and show that such insights can support the deployment of new data-driven practices for real-time monitoring and transit planning.

1.3 Contributions

In this work, we generate insights on the Singapore public transport mobility patterns through a fusion approach which includes mobile geolocation data as well as farecard data from the public transport network. The contributions of this work include:

- the development of trajectory analytics for mobile geolocation data calibrated automatically from farecard data, allowing for both high accuracy and high spatial coverage,
- application of the calibrated travel pattern model to the analysis of the first and last mile problem, with categorization of train stations according to their accessibility properties,
- the development and calibration of data-driven route choice models and the resulting analyses of public transport route choices, in particular explanatory features, by using the calibrated travel pattern model.

The remainder of the paper is organized as follows. Section 2 describes the data characteristics, Section 3 presents our Big Data platform, and Section 4 introduces the core analytics developed for the particular use cases of interest, as well as our calibration procedure. Section 5 presents the insights obtained on the problem of the *first and last mile*, and Section 6 addresses the problem of understanding the explanatory factors of public transport route choices. Finally, Section 7 provides concluding remarks.

2. DATASET

In this section, we describe our two main sources of movement data: the anonymized StarHub mobile geolocation data and the anonymized Singapore Land Transport Authority (LTA) farecard data, as well as the key characteristics of the public transport network.

2.1 Public transport network characteristics

The public rail network used in this study is composed of five *Mass Rapid Transit* (MRT) lines: the North-South, East-West, North-East, Circle and Downtown lines. There are 123 stations on the five lines, of which 15 are interchange stations connecting two or more lines. The network boasts a ridership of over 2 million passengers per day.

2.2 Farecard data

All train fares on the Singapore public transport network are paid using cashless smart cards which require a passenger to tap in to enter the network and tap out to exit the network, thus providing the origin, destination, time, and travel-time, of every journey within the train network. The anonymized farecard data thus has a unique entry for each origin-destination journey made by a commuter including the following data fields for each journey (1) anonymized card ID, (2) origin MRT station, (3) destination MRT station, (4) entry timestamp, (5) exit timestamp.

2.3 Mobile geolocation data

The StarHub mobile geolocation data format includes a unique anonymized entry for each geo-localized record generated from a StarHub anonymized customer mobile phone, including the following fields; (1) record time, (2), anonymized user ID, (3) latitude, (4) longitude, (5) accuracy, (7) network event type, where the various network events are further described in the following section.

3. BIG DATA PLATFORM

In this section, we describe the Big Data platform supporting the analytics components. We use IBM *City in Motion* (CiM) system. The system is built on a Hadoop-based platform with a custom spatio-temporal engine.

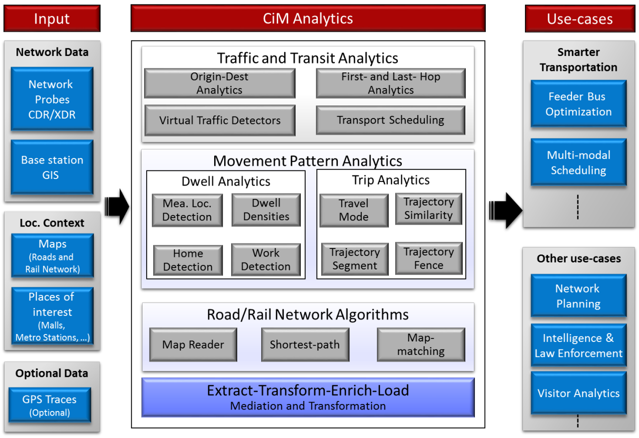


Figure 1: **IBM City in Motion:** Conceptual Architecture.

The conceptual architecture of CiM is shown in Figure 1. The CiM acts on mobile geolocation data to provide meaningful insights about movement patterns in a city. Core algorithms are generically constructed such that CiM can be used as a fundamental system for spatio-temporal analytics in various domains such as Smarter Transportation Planning, Cellular Network Planning, Visitor Analytics for different points-of-interest, and Law Enforcement. We now briefly explain the main components of the system.

3.1 Data and Ingestion

CiM ingests generically available mobile geolocation data including network events, and other forms of openly available data such as maps.

Generic mobile geolocation inputs.

Network event data consists of anonymized subscriber-level events that are recorded by the network. Network events are usually generated under the following conditions: (1) place or receive a call or SMS, (2) periodic log of mobile data usage, (3) location update (typically, when anonymized user moves from one region of the city to another), or (4) network event such as network congestion, call drop, etc.

In addition to the network events, CiM ingests location context and map data. The map data provides CiM with details about the road and the rail network, and is used by the analytics engine to reason about the mode of transport of the anonymized user. Point-of-interest data provides information about important locations to be analyzed, such as malls and train stations.

Extract, Transform, Enrich, Load.

One of the most challenging tasks in ingesting network events is handling the scale, which may range from hundreds of millions to over a billion records per day. CiM ingests network events using Apache Flume and Kafka thus allowing CiM to reach this scale.

Network event data is enriched and converted to a standard format called the *Denormalized Network Event* (DNE). Here, only selective fields required for analysis (anonymized user identifier, time-of-the-event, and event type) are used and the rest of the record is excluded. In this study, we ingest close to 3 billion records for around 2.6 million anonymized users.

The multi-dimensional DNE data is indexed and stored in Hive tables [7]. Indexing the data is not a trivial task. Each analytic component may query the data using different dimensions; certain analytics require all events of an anonymized user for a given time-frame, others may require all events within a certain space-time window. Thus there is no single optimal way to index and cluster the data. We partition the DNE first by time (at date granularity), and then cluster by the user id. This enables the core algorithms – the dwell and trip analytics – to compute meaningful locations and trips of an anonymized user in a given time period.

3.2 Road/Rail Network Algorithms

Several analytics require basic spatio-temporal algorithms on graphs, such as the road and rail network, to determine the location context of the user. Examples of such analytics in CiM include detection of mode of transportation used and computation of the shortest path on the rail network. The main graph-based spatio-temporal functionalities in CiM are: (1) ingestion of the road and rail network; (2) fast shortest-path algorithms on the graph between two points using constrained graph search; (3) *Hidden Markov Model* (HMM) based map-matching algorithm to identify the most likely trajectory on a road or rail network given mobile geolocation data [23]. On the Hadoop platform, the road and rail network are stored in Hive tables. The algorithms for shortest-path and map-matching are exposed as Hive User Defined Functions (UDFs).

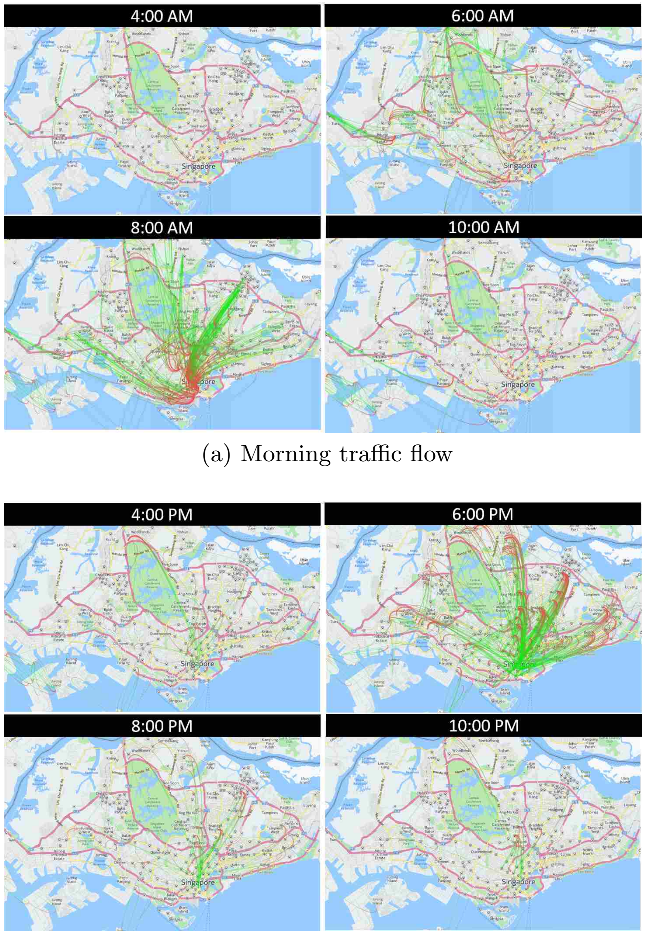
3.3 Visualization

For visualization purposes, we aggregate trajectories by origin zone-destination zone (OD) pairs, where zones are 250m x 250m squares. The start and the end point of the trip computed by the *trajectory cut* algorithm, described in the following section, are mapped to their respective zones. We aggregate the number of trajectories at 1 hour interval periods for each day and the estimated aggregate origin destination flow is used by the application-specific modules. We also provide the mean and median measures for the trip distance and duration, along with the number of trips for an OD pair. Figure 2 corresponds to the OD visualization described, for various times of day.

4. MOVEMENT PATTERN ANALYTICS

CiM consists of two types of fundamental spatio-temporal analytics: (1) Dwell analytics determine the meaningful locations where people spend time, and (2) Trip Analytics determine how people move. Properties of mobile geolocation data render these analytics non-trivial.

First, mobile geolocation data is very sparse (around 40 events per person per day) compared to GPS location traces, typically sampled around 1 Hz. In addition, geolocation uncertainty in cellular data typically ranges in the hundreds of meters, whereas GPS data uncertainty is less than 10 meters. Hence, our cellular spatio-temporal analytics compo-



(a) Morning traffic flow

(b) Evening traffic flow

Figure 2: **Origin Destination weekday traffic**: represented by a green arrow from the origin with red head at the destination. The morning traffic toward the Central Business District and evening traffic from the Central Business District on the South can be observed at peak times.

nents are tailored to spatially uncertain and under-sampled geolocation data.

4.1 Dwell analytics

Dwell analytics estimate user home and work location based on the duration of stay at different times of the day. Dwell analytics also estimate additional anonymized user-specific *meaningful locations*, where an anonymized user spends significant time during weekdays and weekends. Home, work and meaningful locations can be aggregated and visualized as a heatmap of where people live, work or spend time. Important places of an anonymized user are stored into HDFS as Hive tables. They are also stored in Elastic-Search, which is an indexed storage for fast retrieval [5]. Home, work and meaningful locations are indexed by the anonymized user identifier and the spatial zone coordinate. In addition, meaningful locations are also indexed by type of day (weekday/weekend) and 90th percentile of the time range of visits to the given location.

4.2 Trip analytics

The Trip Analytics component estimates anonymized user trajectories from raw DNE data. We describe some of the core trajectory algorithms below, focusing on the scalability and Big Data aspects.

Trajectory segmentation.

We utilize stay-point detection algorithms [37] to detect the start and end of a meaningful trip of a user. All the trips of a user are stored into the Hive table. This table serves as a foundation for all trajectory-related algorithms in CiM, such as origin-destination analysis.

Trajectory fencing.

A common requirement consists of assessing whether a trajectory intersects a point of interest. For example, to determine if a user has possibly traveled in a train, it may be useful to compute if the user trip intersects a train station polygon. We have developed *trajectory fence* algorithms to determine if a trajectory cuts a polygon. The polygon of interest is divided into disjoint triangles and a trajectory into a set of lines between hops. The algorithm checks if any line of a trajectory cuts at least one triangle using a fast line cutting polygon algorithm [22].

Trajectory similarity.

It is often required to estimate the distance between two trajectories, for instance a user trip and a train route. We have developed trajectory similarity algorithms based on *least common substring* (LCSS) [35].

4.3 Travel mode detection

In order to provide public transit specific travel patterns insights from mobile geolocation data, it is necessary to identify the subset of the total daily trips including a public transit network segment. To this end, we propose a four-phase approach to process the time-series of sparse mobile geolocation data network events and leveraging the farecard and potentially other sources of real-time data.

In this work, we focus on identifying trajectories where a significant sub-trajectory aligns with a train network segment. Applying a map-matching algorithm on all the trips being prohibitively expensive from a computational standpoint, we propose instead a fast two-step heuristic to efficiently filter out trajectories that are unlikely to correspond to a trip on the train network. A map-matching algorithm is then applied to these filtered trajectories. Lastly, we join the estimated train network trips with patterns estimated from the farecard data, and iterate until a target accuracy has been achieved.

Step 1*: determine start and end train stations.*

We recognize that a sampled trajectory trace can be a trajectory on a train only if it cuts through at least two train stations. We use the *trajectory fence* algorithm described above to check if the trip cuts train station polygons. For the filtered trajectories, we determine the start and the end train stations in this step. Next steps are executed only for trajectories filtered in this stage.

Step 2*: filter based on intermediate points.*

We use a heuristic that all points of a trip between the start and end train stations should be *close* to the rail network. We store all line segments of the rail network within a spatial index, such as r-Tree [9]. For each trajectory filtered in step 1, we determine if any point between the start and end train stations is farther than a threshold (1km in this study) from the rail network segments. It that is the case, we conclude that the trip does not correspond to a train trip since at least one point between the candidate start and end train stations is far from the rail network. Step 3 is only applied to the trajectory successfully filtered by both step 1 and step 2.

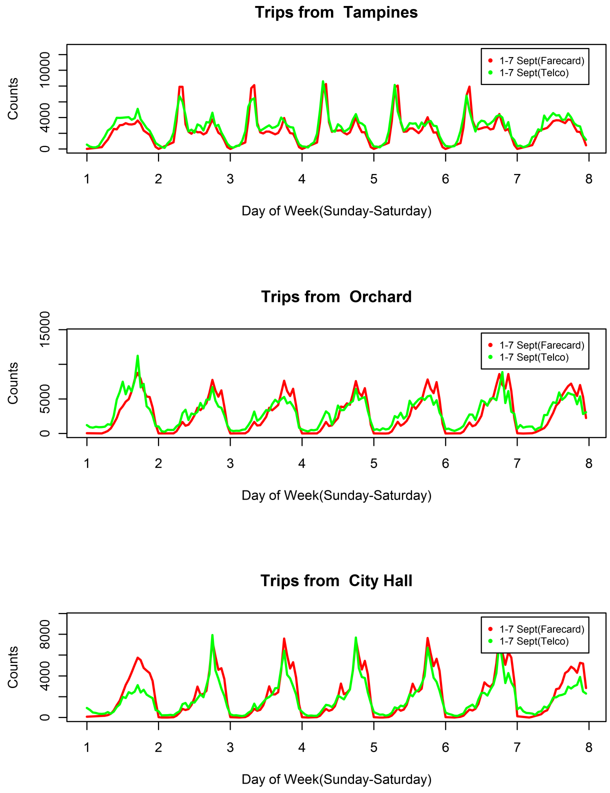


Figure 3: **Farecard-based learning:** for a popular residential station (Tampines), a popular commercial/shopping zone (Orchard) and a busy commercial station in the Central Business District (City Hall).

Step 3*: map-matching and similarity test.*

We then *snap* the candidate trajectory on the rail-network using the map-matching algorithm described in Section 3.2. The HMM-based map-matching algorithm uses an adequate observation noise model with a 1km variance. This relatively large observation noise also causes scenarios where trips corresponding to nearby roads may be falsely categorized as train trips by the map-matching module. Hence, we utilize a heuristic wherein we compare the sampled sub-trajectory between the start and the end train station with the map-matched sub-trajectory using a *trajectory similarity* algorithm. If the similarity is sufficiently high, we accept the trajectory as a train trip.

Step 4*: farecard-based learning.*

The output from step 3 is then appropriately scaled based on estimated device penetration rate, and is compared against similar observed quantities from farecard data. CiM iterates if the error metric is too high. This process is illustrated in Figure 3, representing a week of farecard tap-in data and the corresponding tap-in as per step 3. Travel mode detection parameters are updated via local search based on the observed discrepancy until convergence.

For each trip including a train network leg, we eventually obtain from the travel mode detection component both a sub-trajectory corresponding to the train segment, and an entire end-to-end trajectory of the anonymized user trip as per the DNE data. Both trajectories are stored in Hive tables clustered by start and end train stations.

4.4 Evaluation

The general statistics of the dataset and movement patterns are presented in Table 1 and Table 2.

Number of records	3089 M
Number of subscribers	1.7 M
Number of records per user	1800

Table 1: **General Statistics for 15 days of data.**

Number of Trips	33 M
Number of Trips after MRT Filter	27 M
Number of MRT Trips after Map-matching	21.1 M
Number of MRT Trips after Calibration	14 M
Number of MRT Trips after scaling by penetration factor	46.6 M
Number of MRT Trips as per Farecard Data	40 M

Table 2: **Movement Patterns (15 days)**

The evaluation of the performance of the trajectory analytics can be measured quantitatively on a network level by computation of standard rank correlation, such as Spearman rank correlation coefficient or Kendall rank correlation coefficient, presented in Table 3. Specifically, we compare the ranked correlation coefficient for the origin-destination flows, estimated on one hand from the farecard data, and on the other hand from the mobile geolocation data.

Spearman rank correlation coefficient	Kendall rank correlation coefficient
0.60	0.44

Table 3: **Top ODs matching:** between farecard-based top ODs and mobile geolocation-based top ODs.

For completeness, we also present in Table 4 typical values for the qualitative similarity between origin-destination flow rankings from mobile geolocation and farecard data.
In the following section, we present applications of the cal-

Top ODs	Matching between cellular and farecard
100	40
500	225
1000	526
2000	1185

Table 4: **Top ODs matching:** qualitative metric.

ibrated mobile geolocation-based public transit travel patterns model.

5. FIRST AND LAST MILE

5.1 Motivation

Using the results of the trajectory analytics, we analyze the spatial distribution of the initial (first mile) and final (last mile) segment of user trajectories before and after completing a probable train journey. The first and last mile are key quantities for public transit planners, since (1) a significant part of the trip travel-time can be associated with the first and last mile travel-time, and hence improvements to first or last-mile segments can have a large impact to public transport quality of service, and (2) there is currently very little data about these trip segments. In this section, we present our results for the spatial distribution of first and last mile segments as well as for the average first and last mile travel distances.

5.2 Residential vs shopping stations

The distribution of first mile locations of passengers boarding at Tampines MRT, a highly residential area, as estimated from trajectory analytics applied to geolocation data, is presented as a heatmap in Figure 4. Using the same visualization format, we present in Figure 5 the estimated distribution of first mile locations of passengers boarding at Bugis MRT, a popular commercial and shopping area. One can observe that the origin for the shopping area are qualitatively much closer to the train station. This is confirmed by the actual distances to the station for the first mile segments, presented in Table 5. Such insights can be derived in quasi real-time via fusion of mobile geolocation data with farecard, and thus supports frequent analyses, relevant for a fast-paced nation such as Singapore.

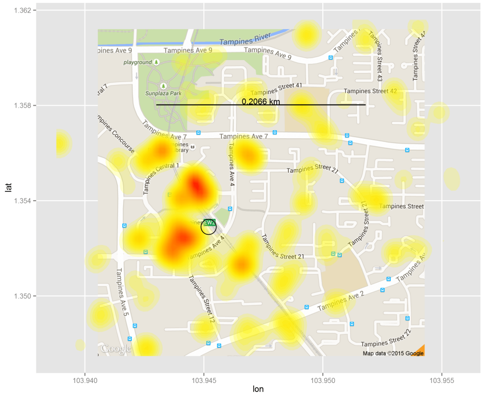


Figure 4: **First mile origin points at residential station:** Tampines MRT (represented as a circle on the map).

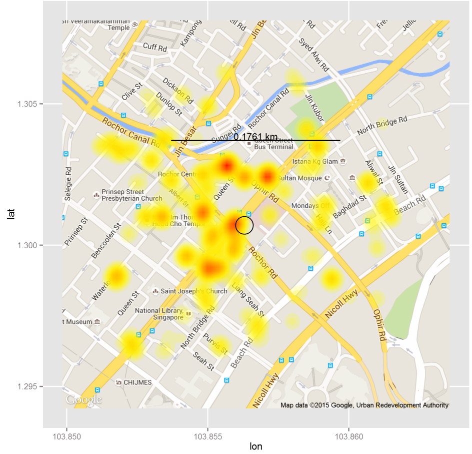


Figure 5: **First mile origin points at a shopping area station:** Bugis MRT (represented as a circle on the map).

Distance (kms)	Percentage Trips Tampines MRT	Percentage Trips Bugis MRT
0-1	39	75
1-2	51	24
2-3	7	1

Table 5: **First mile length:** for residential (Tampines) and shopping (Bugis) stations.

5.3 Analysis of first and last mile

The insights presented in the previous section for two stations can be reproduced for the entire train network. In Figure 6, we present a histogram of the percentage of commuters at each train station for which the trip origin is located within 1 km of the station. The left side of the histogram corresponds to stations with relatively low accessibility (percentage of commuters starting their trip within 1 km of the station is low), whereas the right hand side of the histogram corresponds to stations with relatively high accessibility (percentage of commuters starting their trip within 1 km of the station is higher). Figure 7 presents the analogous histogram for the last mile of the trip. The breakdown of the first mile histogram over a larger set of 42 train stations is provided in Table 6. In Figure 8, we illustrate the impact

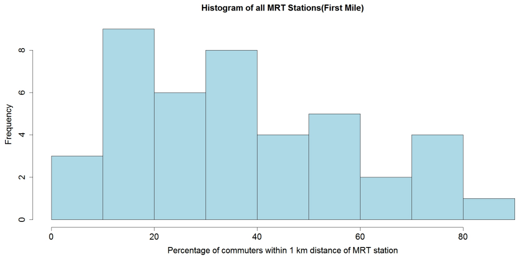


Figure 6: **Distribution of first mile shorter than** 1 **km:** across 42 MRT stations.

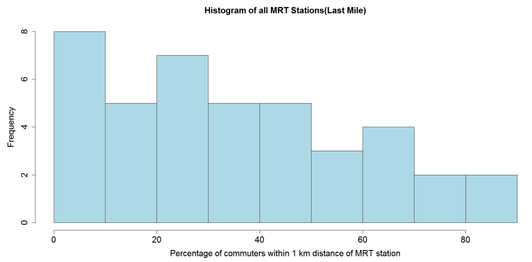


Figure 7: **Distribution of last mile shorter than 1 km:** across 42 MRT stations.

Station	Percent trips within 1 km	Station	Percent trips within 1 km
Admiralty	84	Ang Mo Kio	33
Redhill	78	Orchard	30
Bugis	74	Bishan	30
Bukit Gombak	71	Bukit Batok	27
Lavender	71	City Hall	24
Yio Chu Kang	66	Pasir Ris	23
Commonwealth	63	Buona Vista	23
Tanjong Pagar	59	Jurong East	23
Aljuned	56	Yishun	21
Novena	55	Marina Bay	19
Dhoby Ghaut	53	Clementi	18
Somerset	50	Tampines	17
Braddell	42	Lakeside	15
Boon Lay	40	Toa Payoh	13
Eunos	40	Bedok	12
Queenstown	40	Paya Lebar	11
Kembangan	39	Newton	11
Kallang	38	Chinese Garden	7

Table 6: **Percentage of First Mile trips** within 1 km of MRT station.

of advantageous vs disadvantageous first (and last) mile on the population of Singapore commuters. An origin station is "advantageous" if the true origin is within 1 km for the top quartile, and disadvantageous if the station is in the bottom quartile. We obtain the affected population of commuters from the farecard data. Note that perhaps surprisingly, the two subsets are nearly the same size, at about 24% of the total commuter trip population.

Figure 9 shows interestingly that the impact to Singapore commuters in terms of both origin and destination being advantageous is again nearly the same, since those commuters whose O and D are both disadvantageous correspond to about 7% of all commuters for both groups.

Such insights are of use in new transit initiatives such as Beeline [2] which aims to dynamically adapt direct bus routes for high demand and high travel time origin-destination pairs. Similarly, first and last mile results can also help identify suitable locations to target pilot green and light modes deployment like bike-share and car-share facilities.

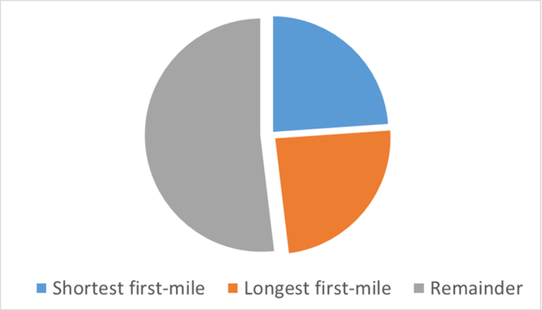


Figure 8: **Roughly equal proportion of commuters have advantageous vs disadvantageous origin stations**, in terms of first mile.

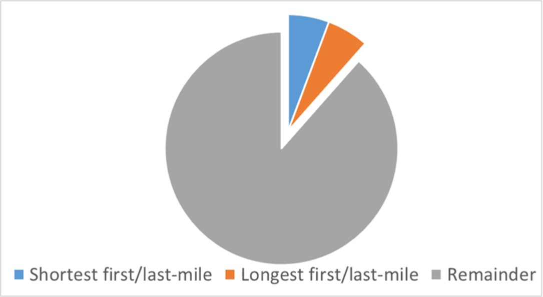


Figure 9: **Roughly equal proportion of commuters have both advantageous vs both disadvantageous Origin and Destination**, in terms of first and last mile.

6. REVEALING ROUTE CHOICE FACTORS

6.1 Motivation

Commuters often decide between several possible train routes from their origin to destination. The route choice may be motivated by a number of factors: distance, travel-time, comfort, crowdedness, cost. Estimating the sensitivity of commuters to specific factors is a valuable aspect of public transit planning, which can inform fare policy, service augmentation decisions, and network extensions plans. Additionally, a proper understanding of route choice is useful in the real-time management of incidents and events on the public transit network.

However, route choices and explanatory factors are not directly observable from farecard data which provides origin and destination, but does not identify the route taken. We leverage our trajectory analytics on mobile geolocation data, described in Section 4, to provide complete estimates of public transit trajectories, as a means to model the route choice process.

For a given origin-destination pair, the probability of a route within the public network being selected by commuters is an output of the trajectory analytics component. Figure 10 shows the distribution of likely number of used routes across all OD pairs in the network, where a route was assumed likely to be used in practice if the route choice probability was at least 5%. In this section, we present a discrete choice model for revealing the important explanatory features (and their relative importance) governing commuter decision of route choice in the train network.

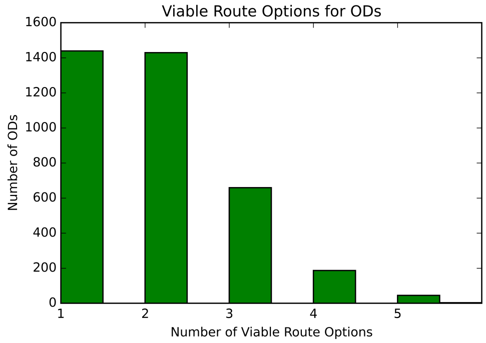


Figure 10: **Probability of route options set size** across all origin-destination pairs, as estimated from mobile geolocation data.

6.2 Discrete choice model

Let the graph $G = (V, E)$ represent the train network, where $V = [v_1, \cdots, v_n]$ are the stations and edges $E = (v_i, v_j)$ exist between connected stations s_i, s_j. Edge attributes include line id, distance, and other available features such as crowd level.

For each station pair, we compute a set P_{v_i,v_j} of at most $p_{\max}$ candidate paths which inherit the features of its edges. We set an upper bound for the maximum path length in order to avoid unreasonably long paths. For each station pair v_i, v_j, the discrete choice set consists of the path set P_{v_i,v_j}. The utility $U_{i,j,k}$ of using path k is a function of the path features and the origin-destination pair. The probability of choosing path k reads $p_{ijk} = Pr(U_{ijk} > U_{ijl,l\neq k}) = e^{U_{ijk}} / \sum_{l=1\dots p_{\max}} e^{U_{ijl}}$.

We use the following network features: (1) path length, (2) number of interchanges, (3) mean frequency of the trains, and (4) number of crowded interchanges, the latter two features being estimated from farecard data. Specifically, an interchange is labelled crowded for a candidate path if the number of commuters for the intended station-line combination exceeds a threshold at that time of the day.

For consistency, for each OD pair, we normalize the features in the interval $[0, 1]$. The model parameters are the weights associated with each feature in the utility function. The model is calibrated using the number of passengers for a given origin-destination pair from the farecard data, while the candidate paths and the proportion of commuters using a candidate path are provided by the trajectory analytics described in Section 4.

Given the expected distinct route choices at different times, we calibrate two distinct models for (1) morning peak on a non-holiday week day, (2) off-peak on a non-holiday week day. The morning peak model uses 13000 data points, and we obtain $R^2 = 0.56$. The off-peak model uses 28000 data points, and we obtain $R^2 = 0.55$.

We generate confidence intervals via bootstrapping of the model coefficients using a normal distribution with mean equal to the coefficient value and standard deviation equal to the standard error of the coefficient estimation. We bootstrap 1000 model executions and output a distribution of probabilities for every route and origin-destination pair. Figure 11 shows an example bootstrap simulation for a given route with choice probability 0.26.

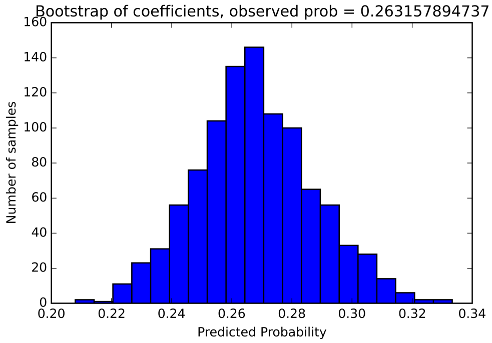


Figure 11: **Bootstrap simulation:** for a given route.

6.3 Example: route choice set

We illustrate one set of results from the route choice model on high occupancy routes from Ang Mo Kio, a popular residential station to Tanjong Pagar, a popular office location in the Central Business District of Singapore. Figure 12 depicts the route 1, with an estimated 52% of the commuters; Figure 13 shows route 2 with an estimated 24% of the commuters having one extra line change. Figure 14 shows route 3, with an estimated 15 % of the commuters and two extra line changes.

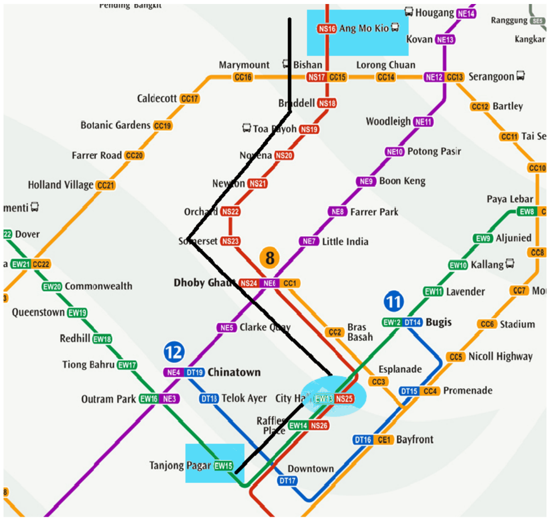


Figure 12: **Route 1:** with 52 percent Ang Mo Kio to Tanjong Pagar.

6.4 Explanatory features

The model feature coefficients and standard errors are listed in Table 7. The results illustrate that path distance remains the most important explanatory factor of the route choice, both during the morning peak and off-peak. One can also observe the relative increase of the sensitivity to crowd level during peak times, indicating the preference for low crowd level. The route choice model can be further improved by considering specific commuter types, regular commuters who know the network and would evaluate more advanced features (such as interchange distance, queues at train doors,

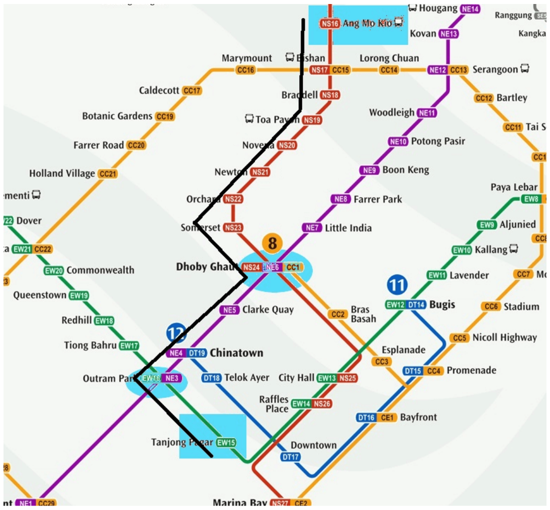


Figure 13: **Route 2:** with 24 percent Ang Mo Kio to Tanjong Pagar.

Feature	Peak coefficient	Off-peak coefficient
Intercept	2.91	3.00
# Interchanges	-1.22	-2.08
Distance	-3.08	-3.07
Train frequency	-0.64	-0.61
Crowd level	-0.59	-0.26

Table 7: **Feature importance:** for both the peak and off-peak models.

etc.) and non-regular commuters such as tourists with additional activity-specific considerations (hotspots on route, station attractivity, etc.)

7. CONCLUSIONS

We use mobile geolocation data and public transit data for generating complete insights on public transit travel patterns. We applied trajectory analytics on mobile geolocation data and showed that the limitations of mobile geolocation data can be addressed by leveraging the complementary strengths of public transit data via appropriate calibration and learning. We have shown that combining these data sources helps provide an accurate and complete picture of public transit trips, including first and last mile. The value of these insights was illustrated on two typical transport applications. Our conclusions on the estimation of first and last mile travel patterns show that the output of our system can be used for the design of on-demand public transit feeders and main public transit lines. We have also shown that parameters that are critical for optimal public transit planning, such as explanatory covariates for route choice, can be estimated from the adequate combination of mobile geolocation and public transit data via advanced learning analytics.

8. ACKNOWLEDGMENTS

We are grateful to the Singapore *Land Transport Authority* for sharing the anonymized farecard data and to *StarHub* for providing the anonymized mobile geolocation data used in this work. We acknowledge the support of the Singapore *Economic Development Board*.

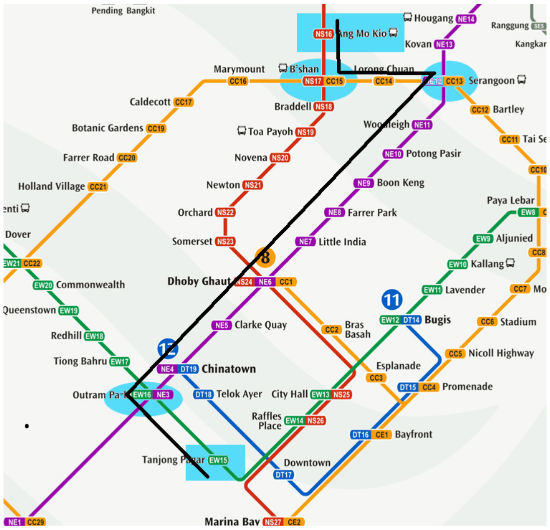


Figure 14: **Route 3:** with 15 percent Ang Mo Kio to Tanjong Pagar.

9. REFERENCES

[1] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82, 2013.

[2] Beeline. Beeline. https://www.beeline.sg/.

[3] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *Intelligent Transportation Systems, IEEE Transactions on*, 12(1):141–151, 2011.

[4] N. Eagle and A. S. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[5] Elastic. Elastic search. https://www.elastic.co/products/elasticsearch.

[6] A. Flajolet, S. Blandin, and P. Jaillet. Robust Adaptive Routing Under Uncertainty. *submitted to Operations Research*, 2016.

[7] A. Foundation. Apache hive. https://hive.apache.org.

[8] F. Girardin, F. Calabrese, F. D. Fiore, C. Ratti, and J. Blat. Digital footprinting: Uncovering tourists with user-generated content. *Pervasive Computing, IEEE*, 7(4):36–43, 2008.

[9] A. Guttman. R-trees: A dynamic index structure for spatial searching. *SIGMOD Rec.*, 14(2):47–57, June 1984.

[10] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151(1-2):304–318, 2013.

[11] S. Hasan, X. Zhan, and S. V. Ukkusuri. Understanding urban human activity and mobility

patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, page 6. ACM, 2013.

[12] Y. He, S. Blandin, L. Wynter, and B. Trager. Analysis and real-time prediction of local incident impact on transportation networks. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pages 158–166. IEEE, 2014.

[13] T. Holleczek, S. Yin, Y. Jin, S. Antonatos, H. L. Goh, S. Low, A. Shi-Nash, et al. Traffic measurement and route recommendation system for mass rapid transit (mrt). In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1859–1868. ACM, 2015.

[14] T. Holleczek, L. Yu, J. K. Lee, O. Senn, C. Ratti, and P. Jaillet. Detecting weak public transport connections from cellphone and public transport data. In *Proceedings of the 2014 International Conference on Big Data Science and Computing*, page 9. ACM, 2014.

[15] T. Hunter, P. Abbeel, and A. Bayen. The path inference filter: model-based low-latency map matching of probe vehicle data. *Intelligent Transportation Systems, IEEE Transactions on*, 15(2):507–529, 2014.

[16] J. G. Jin, K. M. Teo, and L. Sun. Disruption response planning for an urban mass rapid transit network. In *transportation research board 92nd annual meeting, Washington DC*, 2013.

[17] J. Jonas. Analytic superfood. http://jeffjonas.typepad.com/jeff_jonas/2009/08/your-movements-speak-for-themselves-spacetime\\-travel-data-is-analytic-superfood.html.

[18] F. Kling and A. Pozdnoukhov. When a city tells a story: urban topic analysis. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 482–485. ACM, 2012.

[19] V. Kolar, S. Ranu, A. P. Subramainan, Y. Shrinivasan, A. Telang, R. Kokku, and S. Raghavan. People in motion: Spatio-temporal analytics on call detail records. In *Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on*, pages 1–4. IEEE, 2014.

[20] A. Krause, E. Horvitz, A. Kansal, and F. Zhao. Toward community sensing. In *Proceedings of the 7th international conference on Information processing in sensor networks*, pages 481–492. IEEE Computer Society, 2008.

[21] R. K.-W. Lee, T. S. Kam, et al. Time-series data mining in transportation: A case study on singapore public train commuter travel patterns. *International Journal of Engineering and Technology*, 6(5):431, 2014.

[22] T. Möller and B. Trumbore. Fast, minimum storage ray/triangle intersection. In *ACM SIGGRAPH 2005 Courses*, SIGGRAPH '05, New York, NY, USA, 2005. ACM.

[23] P. Newson and J. Krumm. Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 336–343, New York, NY, USA, 2009. ACM.

[24] N. B. Othman, E. F. Legara, V. Selvam, and C. Monterola. Simulating congestion dynamics of train rapid transit using smart card data. *Procedia Computer Science*, 29:1610–1620, 2014.

[25] A. D. Patire, M. Wright, B. Prodhomme, and A. M. Bayen. How much gps data do we need? *Transportation Research Part C: Emerging Technologies*, 2015.

[26] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva. Using data from the web to predict public transport arrivals under special events scenarios. *Journal of Intelligent Transportation Systems*, 19(3):273–288, 2015.

[27] A. Pozdnoukhov and C. Kaiser. Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks*, pages 1–8. ACM, 2011.

[28] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *Pervasive Computing, IEEE*, 6(3):30–38, 2007.

[29] A. Sadilek and J. Krumm. Far out: Predicting long-term human mobility. In *AAAI*, 2012.

[30] S. Samaranayake, S. Blandin, and A. Bayen. A tractable class of algorithms for reliable routing in stochastic networks. *International Symposium on Transportation and Trafic Theory (ISTTT), Procedia Social and Behavioral Sciences*, 17:341–363, 2011, doi:10.1016/j.sbspro.2011.04.521.

[31] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[32] L. Sun, D.-H. Lee, A. Erath, and X. Huang. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of mrt system. In *Proceedings of the ACM SIGKDD international workshop on urban computing*, pages 142–148. ACM, 2012.

[33] X. Tang, S. Blandin, and L. Wynter. A fast decomposition approach for transportation network optimization. In *World Congress*, volume 19, pages 5109–5114, 2014.

[34] A. Vaccari, L. Liu, A. Biderman, C. Ratti, F. Pereira, J. Oliveirinha, and A. Gerber. A holistic framework for the study of urban traces and the profiling of urban processes and dynamics. In *Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on*, pages 1–6. IEEE, 2009.

[35] M. Vlachos, G. Kollios, and D. Gunopulos. Discovering similar multidimensional trajectories. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 673–684, 2002.

[36] D. B. Work, O.-P. Tossavainen, S. Blandin, A. M. Bayen, T. Iwuchukwu, and K. Tracton. An ensemble kalman filtering approach to highway traffic estimation using gps enabled mobile devices. In *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, pages 5062–5068. IEEE, 2008.

[37] Y. Zheng. Trajectory data mining: An overview. *ACM Transaction on Intelligent Systems and Technology*, September 2015.