City-Scale Traffic Estimation from a Roving Sensor Network

Javed Aslam
College of Computer and Information Science
Northeastern University
jaa@ccs.neu.edu

Sejoon Lim
CSAIL
Massachusetts Institute of Technology
sjlim@csail.mit.edu

Xinghao Pan
DSO National Laboratories
pxinghao@dso.org.sg

Daniela Rus
CSAIL
Massachusetts Institute of Technology
rus@csail.mit.edu

Abstract

Traffic congestion, volumes, origins, destinations, routes, and other road-network performance metrics are typically collected through survey data or via static sensors such as traffic cameras and loop detectors. This information is often out-of-date, difficult to collect and aggregate, difficult to analyze and quantify, or all of the above. In this paper we conduct a case study that demonstrates that it is possible to accurately infer traffic volume through data collected from a roving sensor network of taxi *probes* that log their locations and speeds at regular intervals. Our model and inference procedures can be used to analyze traffic patterns and conditions from historical data, as well as to infer current patterns and conditions from data collected in real-time. As such, our techniques provide a powerful new sensor network approach for traffic visualization, analysis, and urban planning.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Sensor network, prediction, estimation, traffic, GPS, Taxi, Inductor loop detector

1 Introduction

Understanding traffic conditions and patterns, such as origins and destinations or trips, car routes, and traffic volume and congestion are critical for urban planning. Traffic information is collected today through manually conducted surveys (for origins, destinations, routes), or using static sensors such as traffic cameras and loop detectors[1] (for volume, congestion). However, survey data is often incomplete, inaccurate, and out-of-date, and static sensor data is incomplete and often difficult to analyze and aggregate, especially in real-time.

In this paper we consider a third source of data: a vehicular sensor network that consists of a roving fleet of dynamic sensor "probes". Commercial vehicles are often outfitted with GPS devices that log their locations and speeds at regular intervals, as increasingly are federal, state, and municipal vehicles. These vehicles form as a mobile sensor network, providing real-time information on the state of the road network.

In a large-scale study conducted in the country of Singapore, we collected the travel data (including GPS, speed, and car status) from a fleet of 16,000 taxis for the month of August 2010, representing approximately 500 million individual data points. The taxis transmit the data using a cellular network. Thus, their data can be used in a real-time mode or in a historical mode. We provide intuition as to why a vehicular network consisting of taxis is well suited to the problem of describing traffic patterns: we demonstrate empirically and theoretically that taxis, no matter what their initial locations, tend to rapidly "spread out" within their allowed regions, thus providing good and consistent "coverage" of the road network, by showing that they move in a random walk across the city-state. We show how taxi volume data can be used to automatically determine distributions of origins and destinations. The taxi origin and destination study is a first step towards using dynamic probes for automatically estimating automatically detailed urban-scale mobility patterns.

This work builds on prior studies on traffic to estimate traffic volume and speed [9], and mobility to measure the origins, destinations, and trajectories of trips [17].

SenSys'12, November 6–9, 2012, Toronto, ON, Canada.

[1] Loop detectors are inductive loops installed in the road network, typically at intersections. They can detect metal and thus and count vehicles.

The state of the art in estimating traffic uses static sensors such as loop detectors or traffic cameras [3]. The static sensors are installed at fixed points and provide traffic estimates at that single location. They require significant effort to deploy and maintain and fail to capture traffic flow and trajectory information. Some recent studies have began to investigate the use of GPS devices as dynamic traffic probes for inferring traffic volume using existing mathematical models [11, 16], or they estimate traffic speed from GPS [24, 23] or from special-purpose static sensors [1, 2].

Measuring mobility patterns is more challenging than measuring traffic volume. The state-of-the-art methodology for recording the origins and destinations (OD) of trips is a manually-conducted survey [10, 22, 4, 5]. OD surveys include household, workplace, or roadside surveys and aim to identify where people start and end their trips. There are many shortcomings with this method. The surveys measure the average rather than actual travel behavior, they cover only a small subset of trips, and given the manual nature of the method, the information (e.g. travel time) is poorly estimated by the interviewee. More recently, estimating origin and destination has also been done using the observed link flow information [8, 12, 7, 6, 15, 13, 14]. However, the results of this method depends on an underlying model of traffic and the number of link flow measurement locations. Surveys have also been used to estimate route choices, although the results are not reliable due to the complexity and scale of the route selection problem in a dense network of roads. Truck fleets have been used to identify truck routes using loop detector counts [21]. To our knowledge there are no known studies on mobility details such as origin, destination, and routes using dynamic sensors.

The rest of this paper is organized as follows. Section 2.1 presents our method and data for inferring origins and destinations using dynamic probes installed in taxis. Section 2.4 provides an intuition for why taxis make for good dynamic probes, by showing that have a rapidly mixing property. Section 2.5 considers the dependency between the size of the taxi sensor network and its ability to provide road coverage and accurate traffic volume predictions. Section 3 describes several traffic applications for taxi networks, including traffic volume (i.e. congestion) analysis and visualization, hotspot analysis and visualization, and overall trip origins and destinations analysis and visualization.

2 Modeling and Predicting Traffic with Taxi Probes

Our hypothesis in this paper is that a small number of dynamic probes[2] are sufficient to characterize overall traffic at a city scale. Two natural questions arise: (1) How well does the data from a dynamic sensor network of taxis represent overall traffic? (2a) If it is representative, how much data is needed to infer a good model for traffic? and (2b) If it is not representative, is the bias consistent and correctable?

Our approach to these questions uses a sensor network with two types of sensor data: static and dynamic. Static sensors are placed at fixed locations to collect information about traffic as it passes by, for example loop detectors or traffic cameras. Dynamic sensors are attached to the vehicles themselves and collect information about individual vehicles as they move. Note that with this setup, the dynamic data is strictly richer than the static data in the sense that the static data can be *inferred* from the dynamic data, but not vice versa. Why? Suppose that every vehicle was outfitted with a dynamic sensor and every intersection was outfitted with a static sensor. The static sensors are effectively collecting *macro-level data*, such as the number of vehicles passing through a given intersection during a given time interval, and this information can be inferred from the *micro-level data* collected by the dynamic vehicle sensors, for example by simply counting the number of such vehicles whose sensed routes pass through the given intersection during the given time interval. However, the micro-level dynamic data cannot be inferred from the static sensor data, unless individual vehicles can be identified.

Now suppose that we did not have every vehicle outfitted with a dynamic sensor, but we did have a perfect random sample of such vehicles so outfitted. Then the static data inferred from the dynamic data would not be correct in *absolute* terms, but it would be correct in *relative* terms. In other words, if $1/3$ of the vehicles (chosen uniformly at random) were equipped with dynamic sensors, then the static traffic data inferred should be 1/3 of the actual data collected from the static sensors. The *distribution* of traffic should be correct.

This observation provides a mechanism for quantitatively testing how representative is the probe data: Given any set of static sensor measurements, infer those measurements from the dynamic probe data and compare the results in relative terms, e.g., by comparing the corresponding traffic distributions and/or traffic volumes. If the probe data is representative, these distributions and/or volumes should match; if not, then one can quantify the mismatch, identify specific areas of match and mismatch, correct for consistent biases, and so on. Note that one would naturally suspect that probe data is not generally representative of all traffic: for example taxis do not ply the same routes as trucks. But we can qualify and quantify this mismatch.

There may be areas of the city where the taxi data is quite representative, for example, in the downtown area. Here we can use the taxi data to quantify the *amount* of taxi data needed to infer good models of traffic: one can sub-sample the taxi data and see how the inferred models of traffic degrade vs. the gold-standard static sensor data.

2.1 Experimental Testbed: Taxis and Loop Detectors

We use two sources of sensor data: taxi data from a large fleet of taxis in Singapore, and loop detector data for the entire road network in Singapore. The loop detector data is used as ground truth for traffic volume. Our study uses four weeks of data (August 2010) from 16,000 taxis in Singapore which amounts to approximately 500 million data points (31GB). Each taxi record contains the car id, the driver id, the time stamp, the latitude, the longitude, and status of operation (represented by one of the following four attributes:

[2]In this paper we use taxis, probes, and dynamic sensor network interchangeably.

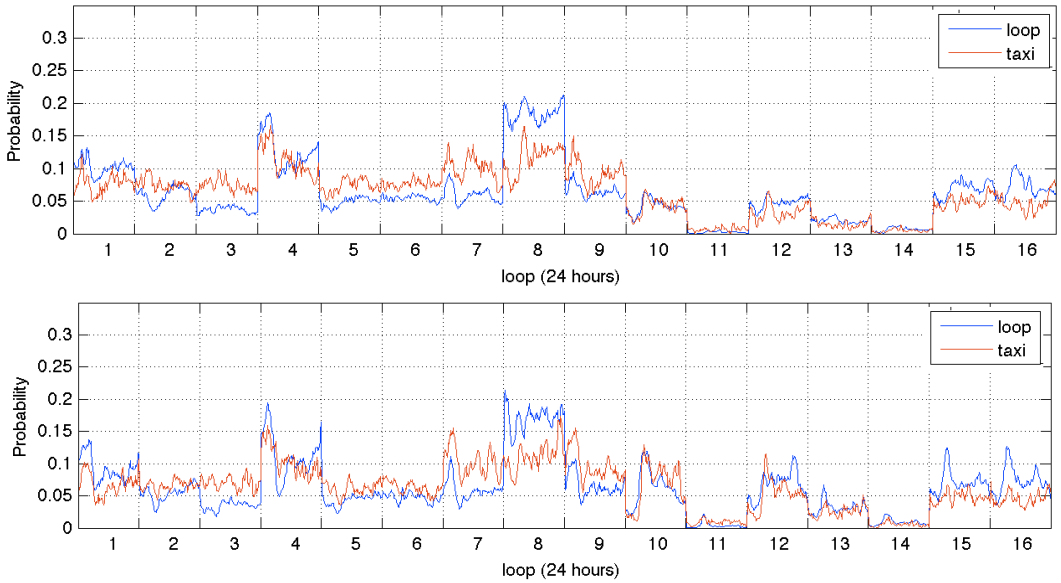


Figure 1. Distribution for Aug 1 and Aug 2 over selected 16 road segments. The x-axis includes 16 segments, one per road. Each of these 16 segments includes 96 points, one for each 15 minute time slot during a 24 hour day. The y-axis shows the fraction of traffic at that location and time. The two curves are not well matched, although we notice that within every hour time slot the offsets seem consistent.

free, person on board, busy, on break). Records are logged at interval between 30 seconds and 2 minutes, depending on network connectivity. Our study also uses the the loop count data we obtained from Land Transportation Authority (LTA) in Singapore for about 12,000 loop detectors in about 1,000 intersections in Singapore for the same period of time, August 2010. Each loop detector record gives the number of cars that pass over each loop detector during a 15-minute time slot. There are 2,688 time-slots for the first four weeks of August 2010. We used these time slots for our studies and map all taxi and loop detector data in these slots.

Processing the taxi data to get the taxi counts for the corresponding time windows for the loop detector data required several steps. First, the data is mapped to a time series of GPS points for each car. Next, we match the time series of GPS points to a sequence of road segments in the road network of Singapore. To overcome the noise and sparsity of the taxi GPS data, we used a map matching method based on the Viterbi algorithm [19]. Third, we count the number of taxis on road segments where loop detectors exist. From this process, we get the taxi counts for each location in the road, which is regarded as the sampled count for the probe traffic. The loop count serves as the ground truth for general traffic. Finally, we smoothed the count data by sliding averages over a sequence of time slots ordered in time[3].

Figure 1 shows the taxi and loop detector count data for 16 Singapore road segments we selected randomly. Note that the taxi distribution (in red) tends to overestimate the loop distribution (in blue) during much of the day, and that the overestimation varies. During the morning rush hour, the taxi and loop distribution values are nearly identical. Thus, while there exists a bias, this bias certainly changes throughout the day, though it appears relatively consistent across days.

2.2 Using Taxi Probes to Infer General Traffic

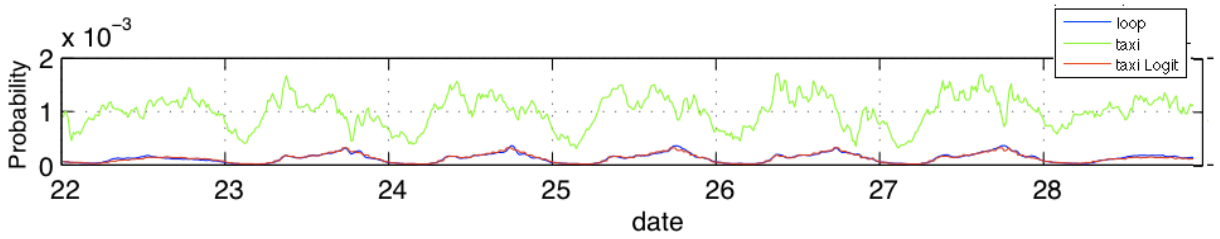
We employ machine learning, and in particular, a cross-validation study, to determine the simplest corrective model for inferring vehicle distribution as detected by loop sensors from vehicle distribution as detected by taxi sensors. We extend this result with a cross-validation study that shows that general traffic volume can also be inferred from the taxi data. Figure 2 shows the results of learning the corrective coefficients for taxi data and demonstrates that taxi data can indeed be used to predict general traffic. Our analysis shows that the best model for accurately inferring traffic distribution and volume uses (1) the hour of the day and (2) whether the day is a workday or non-workday.

Let R be the set of roads in our study. We selected the 1000 road segments from the Singapore road network whose lanes are equipped with loop detectors: $R = \{r_1, r_2, \cdots, r_{1000}\}$. For road r and time slot t we normalize the traffic value relative to the overall traffic, to determine the fraction of traffic at that location at that time. Specifically, if t_r and l_r are the vectors representing the taxi and loop detector counts for each 15 minute time slot for road r and $\tilde{t}_r$ and $\tilde{l}_r$ are the corresponding distribution of taxis and loop detector counts[4], then

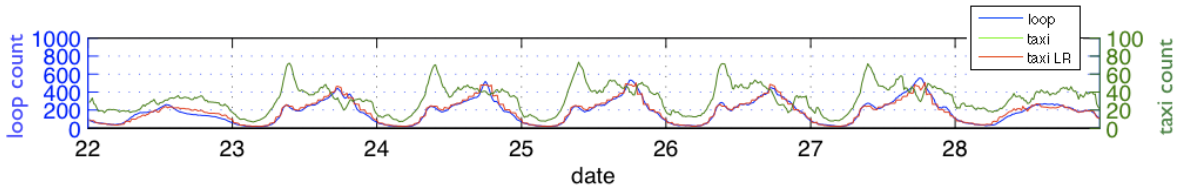
$$\tilde{t}_r = \frac{t_r}{\sum_{i=1}^{1000} t_i}, \quad \tilde{l}_r = \frac{l_r}{\sum_{i=1}^{1000} l_i} \tag{1}$$

[3]The window size of sliding averages was determined as the minimum value that makes the aggregate number of data points over the window is at least 100.

[4]Since we use 4 weeks of data, the length of t_r and l_r is 2688.



(a) Distribution from loop detector data (blue), taxi data (yellow), and after applying the regression for taxi data (red). The probability is found using (1). The x-axis represents the day of the month (in this case week 4 of August 2010). Each day is further divided into 96 15-minute intervals. The y-axis shows the distribution for each give day and 15 minute slot. The ground truth provided by loop detectors is shown in blue. The original taxi data is shown in yellow. The taxi data processed according to the learned parameters is shown in red. Notice that the red curve is a very good match to the blue curve. The large variation between loop detector data and taxi data is removed after applying the logistic regression.



(b) Count from loop detector data (blue), taxi data (green), and after applying the regression for taxi data (red). Whereas logistic regression was used for regression of distributions, linear regression was naturally used for count regression. Notice the red and blue curves are well matched.

Figure 2. Result of regression

Our goals are (1) to examine if $\tilde{t}_r$ can be used to infer $\tilde{l}_r$ and (2) to determine how well we can predict $\tilde{l}_r$ from $\tilde{t}_r$. To infer the relationship between $\tilde{l}_r$ and $\tilde{t}_r$, we used a logistic regression model as follows:

$$\tilde{l}_r = \frac{1}{1+e^{-(\beta_0+\beta_1\tilde{t}_r)}},$$

where β_0 and β_1 are the two regression parameters we learn.

We find the best categorization of time slots for learning β_0 and β_1 that result in the best prediction of $\tilde{l}_r$, using the day-of-week and time-of-day categorization.

The intuition behind this is from the observation of the traffic distribution pattern. We saw significant periodic pattern based on day and week. Specifically, our day-of-week categories include "Each Day of Week", "Workday/Non-workday", and "All Days together", where non-workday means Saturday, Sunday or holiday. Our time-of-day category varies from 15 minutes to 24 hours. For example, Workday/Non-workday as day-of-week category and 2 hours as time-of-day category results in $2 \times 12 = 24$ pairs of regression parameters.

We divided the 4 weeks (Aug 01- Aug 28) data into four one-week testing sets and four associated three-week training sets. We learn the regression parameters using the training set and apply the parameters to the left-out test set to find the test error. Figure 3 shows the leave-one-out cross validation training and testing RMSE errors after performing logistic regression for a road segment. As expected, the training error decreases when we use more complex models; it decreases as the number of time-of-day slots increase, and as the number of day-of-week slots increase. However, the testing error does not always decrease as the model complexity increases (generally known as over-fitting). Figure 3 shows that the best test error was achieved using Workday/Non-Workday with 15 minute slots, though substantially similar results were achieved with slots as long as 1 hour. Figure 2(a) shows that using taxi data and the Workday/Non-workday 1 hour time model we can predict the general traffic distribution with high accuracy.

A similar analysis was done to infer traffic volume, represented as counts, from the dynamic probes using linear regression. The best test error was achieved for the model Workday/Non-workday and 1 hour time slot, Figure 2(b), showing that using taxi data and the Workday/Non-Workday 1 hour time model we can predict general traffic counts with high accuracy.

Let us call RMSE as *absolute RMSE* and define the *relative RMSE* as the coefficient of variation of the RMSE as follows: relative RMSE $= \frac{\text{absolute RMSE}}{\text{mean of loop detector distribution}}$. To see how representative the taxi data is for the general traffic in Singapore, we computed the relative RMSE for all 1,000 road segments and observed that the relative RMSE is less than 10% for vast majority (80.3%) of the road segments.

2.3 Generalization

Whereas travel time data for each road segment can be observed by taxis, the volume data is only available for the locations where loop detectors exist. In this section we describe how to estimate volume for every road segment. We develop a computational approach to estimating the volume for locations where the loop sensor (hence loop data) is not available. We estimate the traffic volume using the predicted loop count learned in Section 2.2.

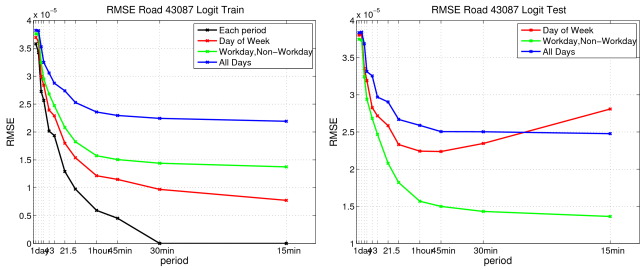


Figure 3. Training(left plot) and test(right plot) RMSE for probability of a road. The x-axis represents the time window considered. The y-axis represents the RMSE. We can observe that the training error decreases as the model gets more complex, but the test error shows over-fitting effect. The best category for this road is Workday/Non-workday and 15minutes.

Let O be the set of all roads with associated loop detectors. Suppose road i does not have an associated loop detector ($i \notin O$). Our goal is to estimate the volume for road i using the real-time volume data from all the roads with associated loop sensors O and taxi sensor information for the road set $O \cup \{i\}$. Intuitively, we will estimate the volume of road i by the weighted average of inferred volume using the method given in Section 2.2 applied to other roads with loop detectors, where the weight is defined by the similarity between road i and a road j in O.

We quantify the similarity between roads i and j by several measures:

- measure 1: $m_1(i,j) = \frac{1}{d(i,j)}$, where $d(i,j)$ is the Euclidean distance between road i and road j
- measure 2: $m_2(i,j) = \frac{1}{a(i,j)}$, where $a(i,j)$ is the angular difference between the orientation of road i and that of road j
- measure 3: $m_3(i,j) = \frac{1}{l(i,j)}$, where $l(i,j)$ is the difference between number of lanes of road i and that of road j
- measure 4: $m_4(i,j) = \frac{1}{t(i,j)}$, where $t(i,j)$ is the difference between taxi count for road i and that of road j

Each of these four measures is an inversely proportional relation. Measures 1, 2, and 3 are static. The information depends on neither time nor traffic conditions. Measure 4 captures the dynamic real-time information observed by taxi probes.

We define an aggregated similarity measure using the four measures. For a pair of roads i and j, (i,j), the aggregate similarity measure is given as follows:

$$s(i,j) = m_1(i,j)^{u_1} \times m_2(i,j)^{u_2} \times m_3(i,j)^{u_3} \times m_4(i,j)^{u_4} \quad (2)$$

where u_k is the indicator for whether measure m_k should be considered for deciding the aggregate similarity.

$$u_k = \begin{cases} 1 & \text{if measure } m_k \text{ should be considered} \\ 0 & \text{otherwise} \end{cases}$$

We choose the best u_k empirically.

Algorithm 1: Estimate-Volume

Data: $v(j)$: inferred loop count for road j where loop detector exists;
$v_t(i)$: taxi count data for road i;
euclidean distance between all pairs of roads;
angular difference of orientation between all pairs of roads;
difference in number of lanes between all pairs of roads;
difference in taxi count for all pairs of roads;
u_k: the indicators for each similarity measure;
Result: $\tilde{v}(i)$: estimated relative volume for road i without loop detector

```
O = a set of all roads where loop detectors exist ;
for j ∈ O do
    Find m1(i,j), m2(i,j), m3(i,j), and m4(i,j) ;
    s(i,j) = m1(i,j)^u1 × m2(i,j)^u2 × m3(i,j)^u3 × m4(i,j)^u4;
    w_i(j) = s(i,j) / Σ_{∀k∈O} s(i,k), ∀j ∈ O
end
ṽ(i) = (v_t(i)) Σ_{∀j∈O} w_i(j) v(j)/v_t(j);
```

Algorithm 1 describes the method for real-time estimation of traffic volume $\tilde{v}(i)$ for road $i \notin O$, using the real-time loop count estimation from all roads in O. Our algorithm estimates the volume by the weighted average of the available inferred loop detector counts for all the roads in O. The weight assigned to each loop detector on road j for the estimation of volume for road i, $w_i(j)$, is defined as the normalized aggregate similarity measure for all the road segments in O as follows:

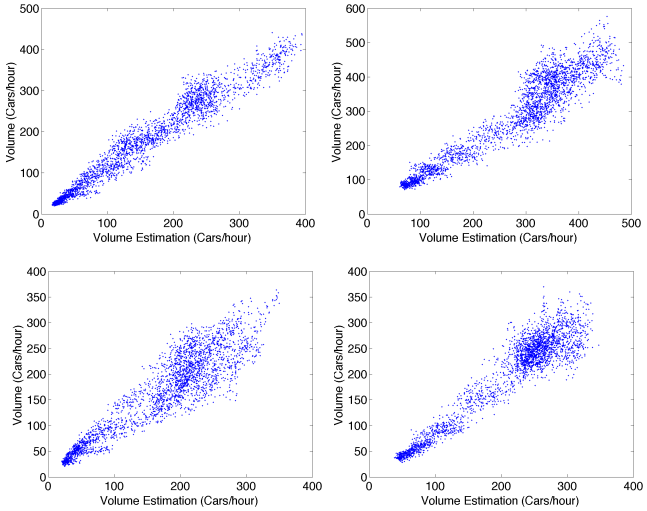
$$w_i(j) = \frac{s(i,j)}{\sum_{\forall k \in O} s(i,k)}, \exists j \in O \quad (3)$$

The estimation of volume for i, $\tilde{v}(i)$, is calculated as the weighted average of the other inferred loop counts using the weight $w_i(j), \forall j \in O$ as follows:

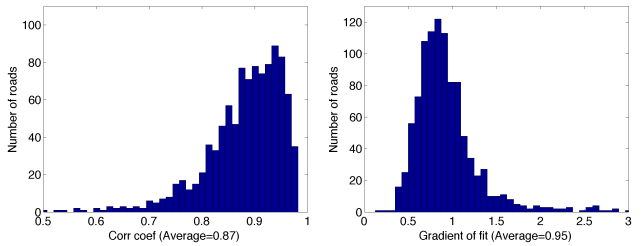
$$\tilde{v}(i) = (v_t(i)) \sum_{\forall j \in O, v_t(j) \neq 0} w_i(j) \frac{v(j)}{v_t(j)} \quad (4)$$

where $v(j)$ is the observed loop count of road j divided by the number of lanes of road j.

We evaluate performance using our data (Section 2.1) with a leave-one-out cross-validation approach. We selected 1,000 locations with loop detector data as the test set, O. We use each road segment $i \in O$ as a test case. For each i we define $O_i = O \setminus i$ to contain all the road segments with loops. We use the loop counts on i as ground truth to quantify our estimates. The data consists of loop detector data for August 2010 and taxi data for the same period, where the loop counts and taxi speeds are collected for each 15-minute time slot. Thus, we have the loop counts and taxi speeds for 1,000 locations with 31 days $\times$ 96 slots per day = 2,976 data



(a) Volume for four randomly selected roads for a test set of 1000 roads. The estimated volume (x-axis) is strongly linearly correlated with the real volume (y-axis), and the slope is very close to 1.



(b) (left) Correlation coefficient between estimated volume and real volume. The correlation coefficient is high around 0.9 (right). 'Gradient' is the slope of minimum-squared-error fit between real loop counts and the estimated loop counts. Most of the roads lie in the range of $0.7 \sim 1.3$. Thus, the gradient error is mostly in the 30% range. As such, the estimated volume approximates well the real volume.

Figure 4. Volume estimation quality

points. Using these data, we ran Algorithms 1 for 1,000 test cases.

Fig. 4 shows the quality of volume estimation. The estimated volume and the real volume have a high linear correlation, and the slope of the linear relationship is highly concentrated around one.

2.4 Taxi Distributions Are Rapidly Converging

Our thesis is that no matter where an individual taxi is located, or where it starts its day, its location will rapidly become indistinguishable from the overall taxi distribution. The reason for this is the stochastic nature of a taxi's passengers—unlike a mail delivery truck that likely follows a fixed route, a taxi will randomly visit locations based on the random nature of the passengers' destinations.

In effect, one can view a taxi's movement as a random walk, with transitions between regions governed by the conditional probabilities of passenger movements from origins to destinations. Although the driver does maintain full control of the taxi's location when there are no passengers on board, we show empirically that taxi distributions tend to converge regardless of the taxis' initial locations. Thus, the locations of the taxis can be viewed as having been randomly drawn from the overall taxi distribution at any point in time. As such, they are well suited to the problem at hand: Regardless of the initial locations of the dynamic taxi probes, we can expect them to quickly spread, providing good coverage in all regions.

The RMMC assumption is backed by the data analysis from a fleet of 16,000 taxis in Singapore. We do not know if the assumption holds for other urban environments. We believe that for cities where the number of taxis and the frequency of trips are high enough relative to the area of the city, we will observe the RMMC assumption. However, more research and experimental studies are needed to verify the generality of the RMMC assumption.

In the remainder of this section, we analyse taxi movement over 27 regions (see Section 2.4.1, less one outlier region) at 15 minute intervals. The transition probabilities are empirically derived from the taxi data.

2.4.1 *Partitioning of Singapore into Traffic Zones*

To make the analysis more tractable, we partitioned the area of Singapore into a number of regions. This was done in a completely data-driven fashion, by applying the standard K-means clustering on a subset of the origins and destinations extracted from the taxi data. (198721 origins and destinations from trips of 500 taxis over the first week were used for the K means clustering.) The K-means clustering algorithm iteratively assigns each origin or destination to one of K centroids to which the origin or destination is closest (as measured by Euclidean distance), and then adjusts each of the K centroids to the means of the origins and destinations assigned to it. This results in a partition of the data space into Voronoi regions, each represented by one of the K means or centroids.

In choosing the value of K, we followed [18] in the use of a criterion, comprised of a linear combination of the *distortion* (or encoding error) and regularization term:

$$\sum_{i=1}^{N} (\vec{x}_i - \vec{c}_{\text{ENCODE}(\vec{x}_i)})^2 + 2\lambda K \log N$$

where N is the number of origins and destinations, $\vec{x}_i$ are the GPS locations of the origins and destinations, $\vec{c}_{\text{ENCODE}(\vec{x}_i)}$ is the nearest centroid to $\vec{x}_i$, and λ is the regularization factor. In this above criteria, the distortion $\sum_{i=1}^{N} (\vec{x}_i - \vec{c}_{\text{ENCODE}(\vec{x}_i)})^2$ is the sum of Euclidean distances of each origin or destination from its nearest centroid, and $2\lambda K \log N$ is the regularization term. The "optimal" K^* is chosen to minimize this criteria.

For our analysis, we have chosen $\lambda = 35$, which results in $K^* = 28$. Although the "optimal" K is sensitive to the choice of λ, we believe that our choice is validated for the following reasons:

1. By visually examining the graph of distortion against K, we find an "elbow" at around $15 \leq K \leq 35$;
2. There happens to be 28 postal districts in Singapore [20];

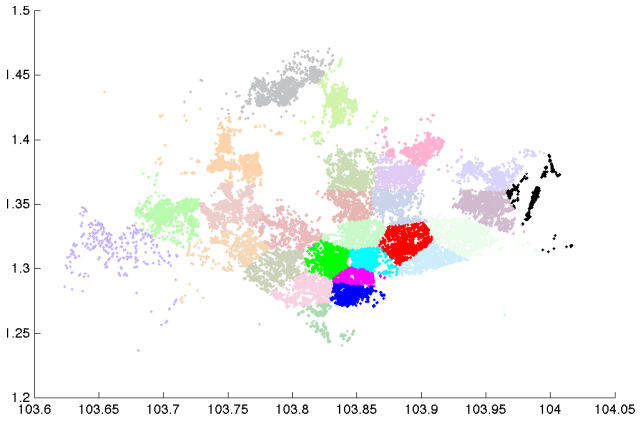


Figure 5. Voronoi regions of Singapore, extracted through K means clustering on origins and destinations of taxi trips. A total of 27 regions are shown in this map, with one outlier region excluded. The highlighted 6 regions will be discussed in detail in Section 3.3.2.

3. The resultant Voronoi regions appear to correspond to semantically significant areas in Singapore

We also verified that observations presented in this paper are robust to a wide range of $K = 15, \ldots, 35$.

Figure 5 shows the Voronoi regions extracted through K means clustering (27 regions are shown, with an outlier region excluded). Each dot in the plot represents a single origin or destination of a taxi trip, and is color-coded to represent the region to which it belongs. Hence, each region is represented by a continuous patch of color-coded dots. Although no additional map information was used in the creation of the plot, the map of Singapore clearly emerges from the data, with the empty patches corresponding to inaccessible areas (e.g. reservoirs or forests). We observe that semantically significant regions such as the Airport in the East, the Central Business District and Orchard regions in the South; Tuas in the West, and Woodlands in the North, are also visible.

2.4.2 *Taxis Are Rapidly Mixing*

To quantify the rate of convergence, we compute the maximum difference (i.e. the L^∞ distance[5]) between any initial taxi distribution and the observed overall taxi distribution after 15 minutes, 30 mintues, and 1 to 8 hours, when following the empirical transition probabilities. In effect, this is equivalent to the following: (1) consider a single taxi located at any initial position, (2) compute the probabilities that this taxi will be located at any given position after a set period of "mixing", and (3) compare this distribution to the overall taxi distribution at that time. Note that we consider the worst-case possible initial conditions (and thus our results apply to any initial taxi distribution or individual taxi location).

Let $\vec{\pi}_t = (\pi_1, \pi_2, \ldots, \pi_r)^T$ be the overall taxi distribution at time t, and $\vec{x} = (x_1, x_2, \ldots, x_r)^T$ be any initial distribution of taxis, where r is the number of regions used for analysis. Let $A^*_{u \to v}$ be the transition probabilities matrix from time u to v. Note that the transition probabilities matrix is estimated using taxi data and that we do not make any Markovian assumptions in its construction.

Hence, $\vec{\pi}_t^T = \vec{\pi}_0^T A^*_{0 \to t}$ is the overall distribution at time t. The time-t distribution obtained from an initial distribution of $\vec{x}$ is $\vec{x}^T A^*_{0 \to t}$.

We now present an upper bound on the L^∞ distance between any initial taxi distribution and the observed overall taxi distribution after a given time t.

Claim: $||\vec{x}^T A^*_{0\to t} - \vec{\pi}_t^T||_\infty \leq \max_{i,j} ||(\vec{e}_i - \vec{e}_j) A^*_{0\to t}||_\infty$, where $\vec{e}_i$ is the standard basis vector $(0, \ldots, 0, 1, 0, \ldots, 0)$, and $||\cdot||_\infty$ is the L_∞ norm, returning the maximum absolute element of a vector.

Proof: Let $[\vec{y}]_k$ denote the k element of any given vector $\vec{y}$. Note that $\vec{x} = \sum_i x_i \vec{e}_i$, so:

$$\begin{aligned}
\left[\vec{x}^T A^*_{0\to t}\right]_k &= \left[\sum_i x_i \vec{e}_i^T A^*_{0\to t}\right]_k = \sum_i x_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k \\
&\leq \sum_i x_i \max_j \left[\vec{e}_j^T A^*_{0\to t}\right]_k = \left(\sum_i x_i\right) \max_j \left[\vec{e}_j^T A^*_{0\to t}\right]_k \\
&= \max_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k
\end{aligned}$$

$$\begin{aligned}
\left[\vec{x}^T A^*_{0\to t}\right]_k &= \left[\sum_i x_i \vec{e}_i^T A^*_{0\to t}\right]_k = \sum_i x_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k \\
&\geq \sum_i x_i \min_j \left[\vec{e}_j^T A^*_{0\to t}\right]_k = \left(\sum_i x_i\right) \min_j \left[\vec{e}_j^T A^*_{0\to t}\right]_k \\
&= \min_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k
\end{aligned}$$

which bounds

$$\min_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k \leq \left[\vec{x}^T A^*_{0\to t}\right]_k \leq \max_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k$$

Similarly, we can show for $\vec{\pi}_t^T = \vec{\pi}_0^T A^*_{0\to t}$:

$$\min_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k \leq \left[\vec{\pi}_t^T\right]_k \leq \max_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k$$

Thus:

$$\begin{aligned}
&\min_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k - \max_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k \\
\leq\ & \left[\vec{x}^T A^*_{0\to t} - \vec{\pi}_t^T\right]_k \\
\leq\ & \max_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k - \min_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k
\end{aligned}$$

$$\begin{aligned}
\implies \quad & \left|\left[\vec{x}^T A^*_{0\to t} - \vec{\pi}_t^T\right]_k\right| \\
\leq\ & \left|\max_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k - \min_i \left[\vec{e}_i^T A^*_{0\to t}\right]_k\right| \\
=\ & \max_{i,j} \left|\left[(\vec{e}_i - \vec{e}_j)^T A^*_{0\to t}\right]_k\right|
\end{aligned}$$

$$\implies ||\vec{x}^T A^*_{0\to t} - \vec{\pi}_t^T||_\infty \leq \max_{i,j} ||(\vec{e}_i - \vec{e}_j)^T A^*_{0\to t}||_\infty \ \square$$

Note the above upper bound $\max_{i,j} ||(\vec{e}_i - \vec{e}_j)^T A^*_{0\to t}||_\infty$ holds for any initial distributions $\vec{x}$ and $\vec{\pi}_0$. In other words, ***any*** two intial distribution will converge to within a L_∞ distance of $\max_{i,j} ||(\vec{e}_i - \vec{e}_j)^T A^*_{0\to t}||_\infty$ of each other within t

[5] A low L^∞ distance indicates that the L^1 and L^2 distances must necessarily be low as well.

time-steps. Hence, to show that the taxi distributions converge regardless of initial locations, we need to show that this upper bound decreases quickly as t increases.

This upper bound has a very natural interpretation. The distribution $\vec{e}_i^T A^*_{0\to t}$ is obtained when a subset of taxis concentrated in a single region is tracked for t time-steps. This initial configuration is arguably the worst case scenario, which we intuitively expect to take the longest time to converge. The upper bound is thus a measure of the worst case scenario, where two extreme initial distributions are allowed to converge over t time-steps.

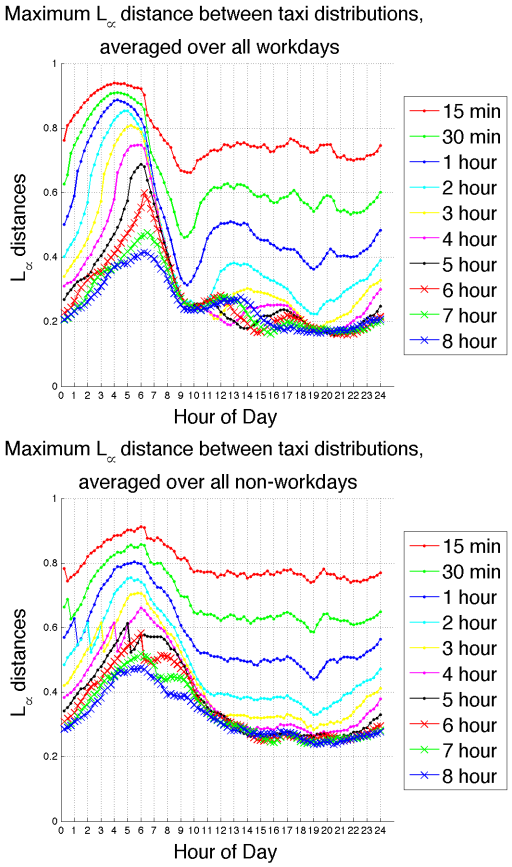


Figure 6. Upper bound on L^∞ distance between distributions of taxis with different initial distributions, after a given period of time. Within a short amount of time, low upper bounds on the L^∞ distances are attained, indicating that the taxi distributions have converged, so the taxi probes provide good coverage regardless of the initial distribution.

Figure 6 shows the results of our analysis for workdays and non-workdays. Each curve represents a different time period such as "6 hours" (red curve with crosses); this corresponds to the amount of time any initial taxi distribution (or individual taxi) is allowed to "randomly walk" before its posterior distribution is compared to the overall taxi distribution. The y-axis gives the upper bound on the L^∞ distance between the "random-walked" taxi distribution and the empirical overall taxi distribution. For instance, at 09:30 on a workday, the upper bound on the L^∞ distance after two hours of "random-walks" is approximately 0.25, indicating that starting from 07:30, any two initial taxi distributions will come within an L^∞ distance of 0.25 of each other by 09:30.[6] Note that the convergence rate is considerably worse in the early morning hours, as there is far less taxi traffic in the middle of the night. As can be seen, we have fairly rapid convergence during most all waking hours.

In our analysis, these upper bounds represent extreme, worst-case scenarios: the taxis are concentrated in a single region, thus providing little coverage of the country. The low upper bound values demonstrate that any initial distribution of taxis will quickly disperse over the country to provide coverage on all regions.

We next show empirically that accuracy estimates of the general traffic volumes are possible using only small number of probes.

2.5 How Many Taxis Sensors Are Needed?

Since taxi sensor data can be used to infer general traffic volume, we conclude that a taxi sensor network is a good proxy for general traffic and would like to know how many nodes are needed to infer traffic within some desired accuracy. There are several attributes that trade-off (1) the number of nodes; (2) the amount of time the nodes roam to collect historical data; (3) the fraction of roads the nodes cover, and (4) the accuracy of the general traffic prediction from the taxi sensor network nodes. In this section we show that a relatively small number of taxis is sufficient to get good road coverage, and that if the taxis roam for a relatively small amount of time, general traffic can be inferred with high accuracy as measured by RMSE.

2.5.1 Coverage vs. Number of Taxis

Section 2.4 shows that taxis travel randomly and therefore they cover Singapore broadly. In this section we quantify this property of taxi movement by investigating the fraction of road segments they cover. We measure *coverage* as the fraction of roads that were driven on at least k times[7] by any taxi. We examined the coverage for different time windows and different k, using sets of taxis of size N that were randomly selected using the following procedure. We generate a random permutation of 16,000 taxis and use the first N taxis from the permutation sequence.

Figure 7 was drawn for a single permutation of 16,000 taxis. The plots show the fraction of the 1,000 road segments driven on least 30 times during 2 different time windows. The number of taxis used for the coverage is given by x-axis. Different curves correspond to different times of day, as denoted in the legend. The time window is 15 minutes (left) and 1 hour (right). For workday coverage during 15 minutes (left plot), we conclude that 700 taxis are enough to cover 70% of the roads for most of the day's 1-hour time windows except those in the middle of the night when the number of vehicles on the road is sparse. Figure 8 shows cumulative histograms of 300 random permutations. The steep slopes

[6] These values are the *average*, over all work or non-work days, of the *worst-case*, over all initial taxi distributions, for the time period of interest.

[7] k is a parameter we choose to support the methodology of inferring general traffic from the taxi counts.

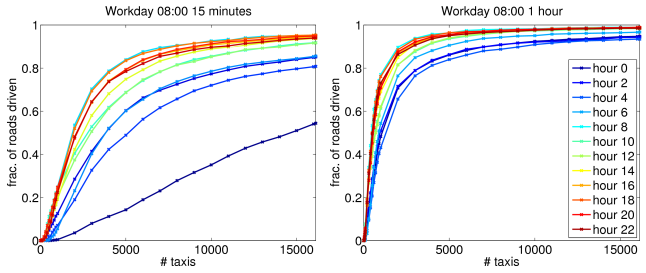


Figure 7. Coverage according to the number of probes. The Left and right plots show what fraction of roads are covered 30 times during 15 minutes and 1 hour respectively (y-axis) as the number of probes grows (x-axis). Color encodes the time of day for the coverage measurement.

indicate that the coverage result is not very different from one permutation sequence to another. From the right plot, we can see that 2000 taxis are enough to cover 90 % of the total loop detector locations during 08:00 $\sim$ 08:15 on all the workdays.

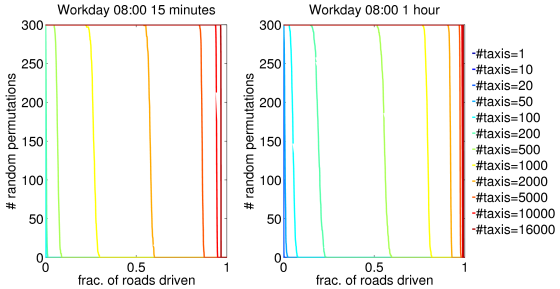


Figure 8. Cumulative histogram of coverage by 30 samples for 300 random permutations for 8am. The steep slope of each line means that the variation from one random order selection of 16000 taxis are not very different from the others. Thus, we don't need to worry too much about which set of taxis we should choose

2.5.2 RMSE vs. Number of Taxis

Next, we consider how many taxis are needed in order to estimate traffic on the Singapore road network within a desired RMSE. We use as base case for traffic estimation the RMSE derived in Figure 2(a) using the entire fleet of 16,000 taxi probes and ask, how much will the RMSE degrade if we use a subset of size given by parameter k of the the probes selected randomly. The RMSE decreases as the number of taxi probes increases. However, the RMSE bottoms out at a certain point so that increasing number of probes has no effect on improving the RMSE value.

Let $RMSE_T$ denote the RMSE achievable with our entire fleet of 16,000 dynamic probes and let $n_\lambda(c)$ be defined as the minimum number of taxis that can be used to infer general traffic with RMSE at most $(1+\lambda)RMSE_T$ for a given traffic model c. Figure 9 shows a cumulative histogram of $n_\lambda(c)$ for various λ values where c is the model workday/non-workday with a 1 hour time slot. The y-axis of both plots represents the fraction of road segments among the total 1,000 road segments. The left plot is drawn over *# probes* for various λ, and the right one is drawn over λ for various *# taxis*. In Figure 9 the increase in percentage is steep from 0 to 1000, which shows that increasing the number of probes from 0 to 1,000 improves estimation. Given 2,000 probes we can estimate the general traffic counts as given by the loop detectors for 65% of the road network with precision at least $1.1RMSE_T$.

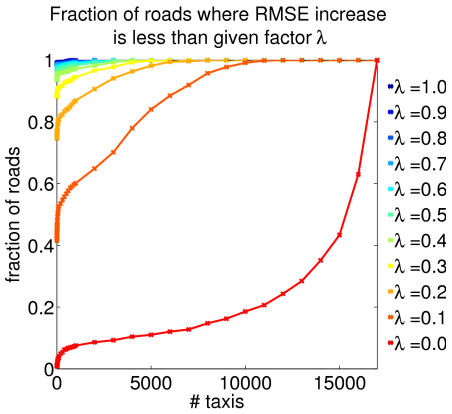


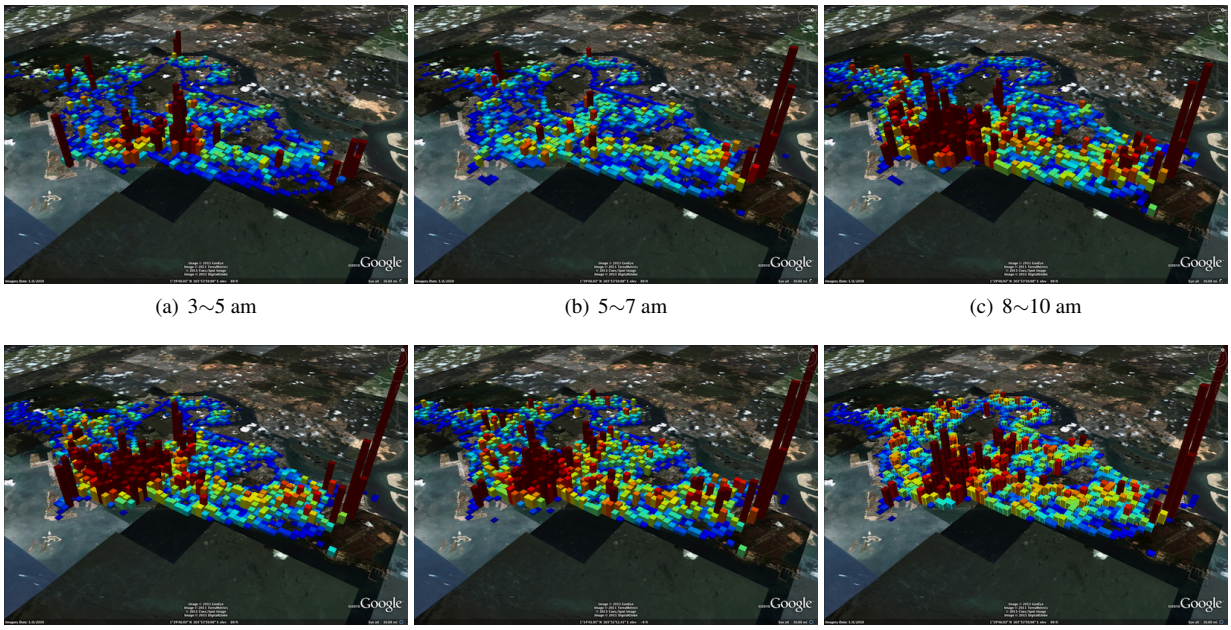
Figure 9. The dependence of the RMSE for inferred traffic volumes on the number of dynamic probes used. Cumulative histogram of $n_\lambda(c)$ Each curve in the plot corresponds to a different λ. The x-axis plots the error tolerance, and the y-axis plots the percentage of roads With 2000 probes and 10% additional error over $RMSE_T$ (the orange curve), we can predict the traffic volumes for 65% of the roads

3 City-scale Applications

The data collected from the roving sensor network of taxis can be analyzed to extract global trends about the environment, for example the location of hot spots. Sections 2.2, 2.4, and 2.5 provide a method for inferring traffic volumes from taxi probes, and an analysis of how many devices are needed to achieve the inference within some desired error limits. Many interesting traffic and mobility analyses can be performed using dynamic probes. In this section we give three examples of traffic analyses using dynamic probes: estimating taxi volume, estimating hotspots, and estimating the distribution of origins and destinations for taxis, which point to future directions and opportunities for using taxi probes to understand urban-scale mobility.

3.1 Volume

Figure 10 shows a snapshot of taxi volume for different hours of day on August 2nd (Monday) 2010 for Singapore. The volume is defined as the number of taxis observed in the square block over a given time and is plotted for a regional block size of 400meters $\times$ 400meters and a time size of 2 hours. Color also encodes a qualitative measure of volume according the the color wavelength, with red denoting the highest volume and blue denoting the lowest volume. Each bar height measure the volume. While the volume distribution across the country changes according to different times of day, some parts of the country retain the largest fraction of the traffic volume. We discuss this phenomenon in the next



(a) 3∼5 am (b) 5∼7 am (c) 8∼10 am

(d) 1∼3 pm (e) 5∼7 pm (f) 9∼11 pm

Figure 10. Snapshots of traffic volume measured by taxi probes for August 2nd 2010 (Monday) plotted on the map of Singapore. Bar height and color combine to encode traffic volumes. Note the large variation in volumes in different regions of the city.

section.

3.2 Hotspots

The red bars in Figure 10 visually indicate traffic hotspots. Some regions of the country remain hotspots regardless of the time of day (e.g. the Changi airport). Other regions (e.g. Orchard Rd) are hotspots during certain time periods. Figure 11(a) shows the top 9 hotspots, while Figure 11(b) shows the traffic variation during the day at each of these 9 hotspots (each location plotted as its own curve). We can observe different detailed volume variation over a day for those hotspots in Figure 11(b).

Hotspots such as A in Figure 11(a) show excess traffic volume in the early morning, and Hotspots such as F show excess volume during the morning rush hour. B shows excess traffic both in the morning and evening rush hours. C (airport) show all-time high volume. Hotspots such as D, E and much of northern area show that the traffic volume is relatively higher in the evening and early morning than morning rush hour.

3.3 Origins / Destinations

The streaming nature of the information from the taxi probes enable the aggregation of higher-order information for mobility analysis, for example where do trips originate and end and which trajectories were followed. In this section we give results on how dynamic probes can be used to analyze origins and destinations. In our future work we will describe how we can use dynamic probes to infer general traffic mobility patterns.

3.3.1 Number of Taxi Trips

A total of approximately 12 million trips were extracted from the taxi data, and the average number of trips starting (origins) and ending (destinations) at different times-of-day are shown in Figure 12, together with the number of taxis plying the road.

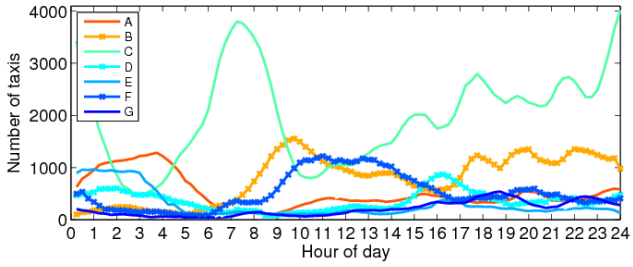
The number of taxis decreases from 8pm to a daily low at 5am, before dramatically increasing through the morning rush hours until 10am. The supply of taxis on the roads then remains nearly constant until the 8pm in the evening. The slight dip in taxi numbers at 5pm is possibly due to taxi shift changes.

The pattern for the demand of taxis (as represented by the number of origins / destinations) is similar to that of the supply of taxis, where fewer trips are taken between 8pm to 5am, before again increasing until 10am. On workdays, however, the peak demand is experienced at 9am. Even though the demand decreases after 9am, taxi drivers have started their own working hours, and continue to ply the roads, resulting in an over-supply during this period.

3.3.2 Distribution of Origins and Destinations

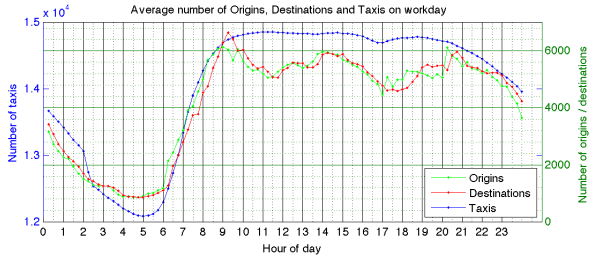
Each origin and destination from all 12 million trips were classified as belonging to one of the 28 regions. For each fifteen minute block starting from midnight on workdays, we calculated the empirical distributions of the origins and destinations over the 28 regions. This procedure was also repeated for trips on non-workdays. In total, 384 empirical distributions were obtained (one distribution for origins on workdays, origins on non-workdays, destinations on workdays, and destinations on non-workdays for each of the 96 fifteen minute blocks in a day).

(a) Hotspots with different daily volume patterns

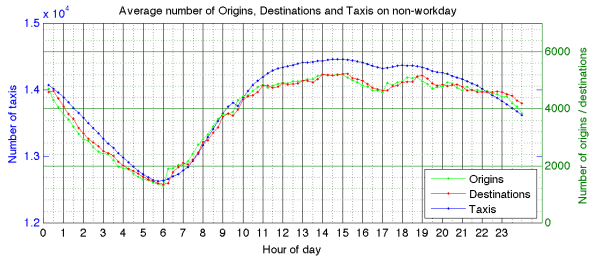


(b) Volume variation of hotspots over time

Figure 11. Hotspots according to time of day. The labels on the left map show the locations of the hotdpots. The curves show volume vs time of day for each of the 9 identified hot spots.



(a) Workdays



(b) Non-workdays

Figure 12. Number of taxis and trips (in terms of origins and destinations) at different times of days, on both workdays and non-workdays.

The randomness of each of these distributions were measured using *perplexity*.

DEFINITION 1. *Perplexity for a distribution p is defined as*

$$2^{H(p)} = 2^{-\sum_z p(z)\log_2 p(z)}$$

where $H(p) = -\sum_z p(z)\log_2 p(z)$ is the entropy of the distribution.

The perplexity has a natural interpretation for the OD distributions: if a trip has equal probability of starting in any region, then the perplexity of the empirical distributions of origins is exactly equal to 28, the number of regions. On the other extreme, if all trips begin in the same region, then the perplexity is equal to 1. A distribution with a perplexity value of X is as random as a uniform distribution over X regions.

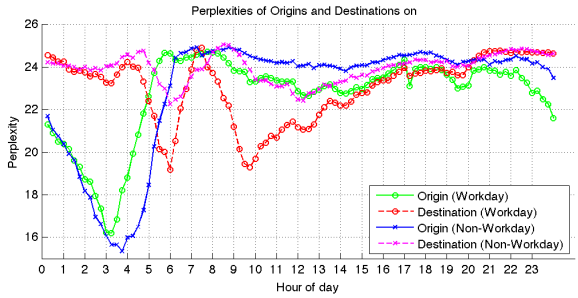


Figure 13. Perplexities of empirical distributions of origins and destinations

The lines marked with an 'O' in Figure 13 show the perplexities of the empirical distributions of origins and destinations at different times during workdays. The peak in the perplexities of origins is found at the morning rush hours as passengers leave their homes across Singapore. The distribution of origins stabilizes in the afternoon, before undergoing a drop from 10pm to a trough at 3am. During this period, trips largely originate from a small number of regions near the city centre that are dominated by offices and popular shopping malls, indicating that passengers are reversing their morning trips and heading home.

Conversely, a trough in the perplexities of destinations occurs at the morning rush hours, when passengers are traveling to the small number of city regions. Throughout the rest of the day, the perplexity increases as passengers travel to more diverse regions. Between 8pm - 12pm, perplexity of destinations reaches its nightly peak as passengers head to their homes in a large number of residential regions. A second trough occurs at 6am – this phenomenon will be explained below.

Similar patterns can be seen on non-workdays as well, as shown by the lines marked with an 'X' in Figure 13. In contrast to workdays, however, perplexities on non-workdays tend to be higher, indicating that the population's movements are less synchronized. Furthermore, the trough in perplexities of destinations, has shifted from 9:45am on workdays to 12:15pm on non-workdays. This suggests that a shift in the behavior of Singapore residents has occurred between workdays and non-workdays.

Perplexities are coarse-grained summary statistics that do not fully capture the richness of behavior encoded in the full

distributions. To better understand the evolution of OD distributions across time, we examined the probabilities of origins and destinations occurring in each region.

We have highlighted 6 regions of interest for discussion, 5 of which fall within or near the city center. The labels we have used for naming the regions are purely descriptive, and do not necessarily correspond to any official naming or demarcations. Nevertheless, a Singapore resident should be able to readily identify the regions and common activities associated with them.

The 6 regions of interest (as indicated by the corresponding bright colors in Figure 5) are:

1. Airport (Black): The easternmost region of Singapore holds Changi Airport, by far the most important civilian airport.

2. Geylang (Red): Although a largely residential area, Geylang is also well- known for some of its nocturnal activities.

3. Orchard (Green): Singapore's famous shopping strip, but also houses a substantial number of offices.

4. CBD (Blue): The Central Business District is dominated by high-rise offices, but also caters to party-goers with its pubs and clubs.

5. Fort Canning (Magenta): Some of Singapore's most popular clubs are found here. Being adjacent to Orchard and CBD, there are also shopping malls and offices in the area as well.

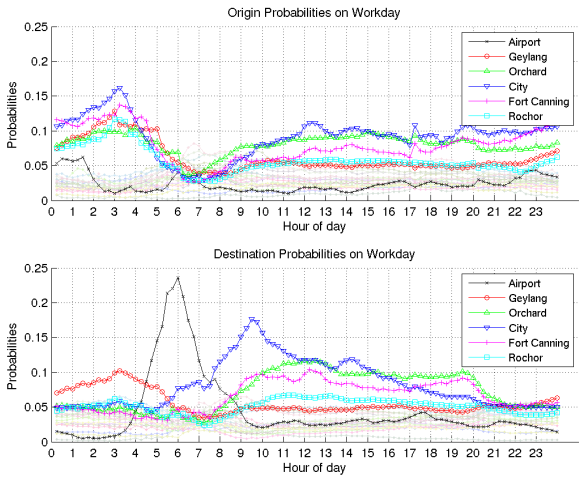
6. Rochor (Cyan): At the fringes of the city area, Rochor is home to a few high-rise offices and shopping malls.

Figure 14 shows the probabilities of trips originating and terminating in the 28 Voronoi regions. The 6 regions of interest are highlighted while the remaining 22 (mainly residential) regions are faded into the background. We discuss some of the patterns that emerge from these plots.

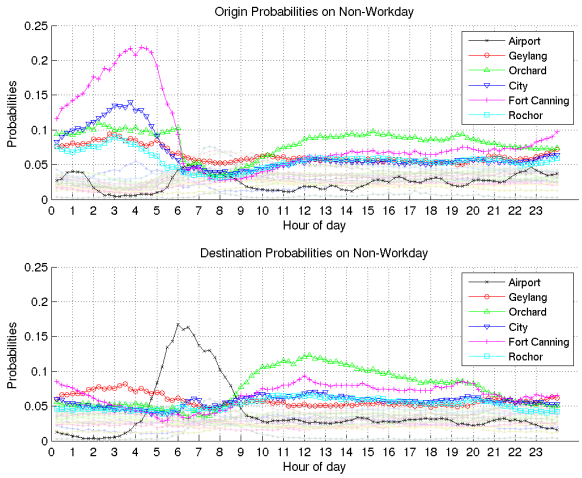
The city regions (CBD, Orchard, Fort Canning, and to a smaller extent, Rochor and Geylang) are the most popular regions for taxi trips to begin and end at most times of the days. The prominence of CBD as a destination from 8am to 10am reflects the arrival of people at their offices. As the day wears on, the CBD is then overtaken by other city areas such as Orchard and Fort Canning, where more leisure activities can be found.

One exception to the dominance of the city regions is the morning hours from 6am to 8am. During this time, one is more likely to find trips originating from residential areas as people leave their homes to begin their day. Conversely, the popularity of city regions as origins increases during the evening, since taxi passengers are now leaving the city for home. The same effect is observed with the destination probabilities when the city regions become unpopular after 8pm.

Another pattern that can be clearly seen from Figure 14 is the popularity of the Airport region as a destination in the early morning hours. We hypothesize that a large number of flights are scheduled to leave Singapore in the morning. This is further compounded by the lack of public trains in the early hours, leaving taxis as the major means of public transportation for air travelers. This spike in the popularity of the Airport region also accounts for the trough in perplexities of destinations at 6am seen in Figure 13.



(a) Workdays



(b) Non-workdays

Figure 14. Probabilities of empirical distributions of origins and destinations. Taxi trips' origins and destinations are mostly concentrated in the city regions, except for the morning rush hours where origins are dominated by residential regions. The airport is also a popular destination in the early morning hours.

The patterns on non-workdays (Figure 14(b)) are similar to the patterns on workdays (Figure 14(a)). A notable exception is the CBD, whose popularity as a destination drops drastically due to fewer people returning to offices on non-workdays.

On the other hand, during the late night hours 3am - 5am on non-workdays, trips are likely to originate from the Fort Canning region. This is likely to be due to the presence of nightspots that are typically frequented only on non-workdays when one does not have to report to work early the next morning.

3.3.3 O-D Relationship

In addition to studying origins and destinations separately, we also investigated the relationship between the two.

Specifically, we examined if the knowledge of a trip's origin could provide us with information about its destination, and vice versa. Such questions would be of particular interest to taxi drivers, whose expectations of their next task could be altered according to the location where the passenger is picked up. OD transitions are also important for transport planners as a first step towards understanding the load on the traffic network.

In Figure 15, we compare the perplexities of origins and destinations with the *conditional perplexities* of origins given destinations and of destinations given origins respectively.

DEFINITION 2. *The conditional perplexity of origins given destinations is defined as* $2^{H(O|D)}$, *where* $H(O|D) = -\sum_o \sum_d p(o,d) \log_2 p(o|d)$ *is the conditional entropy of origins given destinations.*

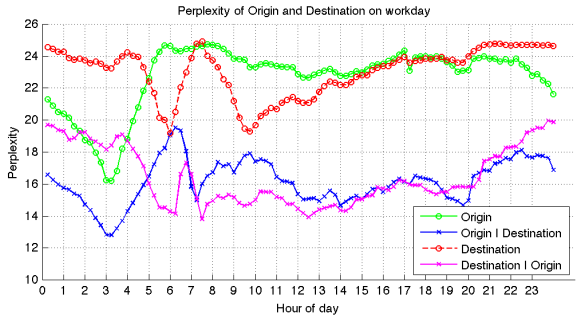
The conditional perplexity of destinations given origins is similarly defined.

The conditional perplexity of origins given destinations is the randomness or uncertainty about origins that remains after the destination has been made known. Thus, if the origins and destinations were independent, then we gain no information about origins through the knowledge of destinations, and the conditional perplexity is equals to the perplexity itself. Such a situation may arise if all trips converged on a single region (so the destination provides no additional information about the origin), or if all trips originated from the same region (so the origin provides no additional information about the destination). On the other hand, if the two are perfectly correlated, then knowledge of destinations leaves no uncertainty about the origins, so the conditional perplexity is 1. The difference between the perplexities and the conditional perplexities quantifies the information that one provides about the other.

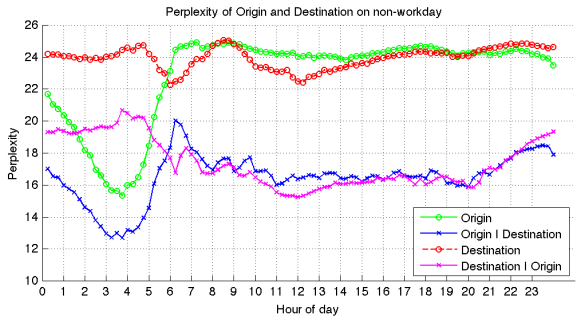
As Figure 15 shows, knowledge about destinations reduces the perplexities of origins by 3.5-9.5 regions on workdays and 2.5-8.5 regions on non-workdays. Knowledge about origins reduces the perplexities of destinations by 4.5-11 regions on workdays and 4-8.5 regions on non-workdays. The differences between perplexities and conditional perplexities are, however, not uniform across different times-of-day and workday / non-workday. Figure 16 shows the variation of the ratios of perplexities to conditional perplexities across the day.

The mutual information $I(O;D) = H(O) - H(O|D) = H(D) - H(D|O)$ is a symmetric measure of dependence between the origins and destinations. The ratio between the perplexities and conditional perplexities works out to be exactly $2^{I(O;D)}$, which is shown in Figure 16. We see that mutual information tends to be higher in the day than at night. Furthermore, a spike between 7am to 8am, followed by a depression at 10am, is observed on weekdays. To explain these observations, we examined the proportion of intra-regional trips, i.e. trips which originated and terminated within the same region.

As shown in Figure 17, the proportion of intra-regional trips is correlated with the mutual information between origins and destinations. At 7am-8am on workdays, many taxi trips are made within the regions around the perimeter of



(a) Workdays



(b) Non-workdays

Figure 15. Conditional perplexities of origins given destinations, and vice versa.

Singapore. We believe that this is caused by people traveling to work locations near their homes, and hence knowledge of the trip origin provides useful information on the probably nearby destination. This behavior is then overtaken by the movement of office workers into the city regions at 10am; since all trips likely end in the city, the knowledge of the trip's origin provides little additional information.

The proportion of intra-regional trips increases after 10am as people do not move far from their workplace. This allows us to again make better guesses about the destination if the origin is known. After work, however, trips are taken out of the city into the residential areas; since all trips likely originated from then city, the knowledge of the trip's destination provides little additional information. The low proportion of

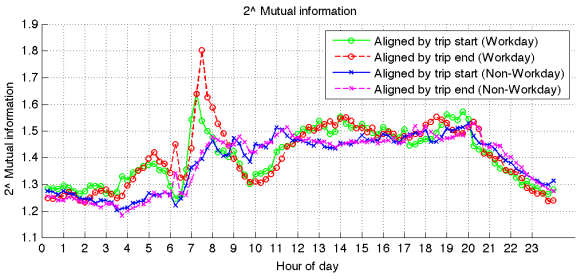


Figure 16. Mutual information between origins and destinations. A clear spike occurs between 7am to 8am on workdays, caused by an increase in the probability of intra-regional trips (see Figure 17).

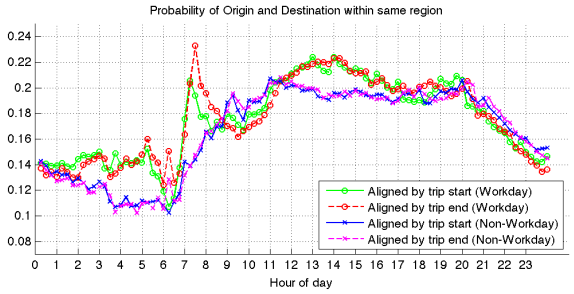


Figure 17. Probabilities of a trip originating and terminating within the same region, at different times of day, on both workdays and non-workdays. The spike between 7am to 8am on workdays corresponds to the increased mutual information between origins and destinations at the same time.

intra-regional trips (and low mutual information) then persists until the next morning.

Similar patterns are observed on non-workdays as people make short intra- regional trips in the day but disperse at night. The spike and depression are not as prominent, however, since there is less work-related movement.

4 Conclusion

In this paper we have shown that a vehicular taxi network can be used to infer traffic patterns such as congestion patterns, traffic hotspots, and historical traffic volumes. Because the movement patterns of taxis is similar to a random walk, a fleet of taxis can cover a city-scale road network quickly. Using data from a study conducted in Singapore, we show that a relatively small number of taxis (700) is needed to cover about 70% of the road network in order to provide sufficiently accurate traffic models. Traffic, as measured by taxis, provides a biased representation of general traffic in city-scale road networks. However, our case study shows this bias is consistent and we can learn the corrective parameters. Thus, we conclude that taxi probes can be used to provide accurate traffic forecasting using historical data from their past drives, and real-time traffic snapshots from their current drives.

The results in this paper are based on a case study done with a taxi network in Singapore. However, we believe that the nature of the results is general and applicable to other cities around the world where urban planners could turn taxis into a vehicular sensor network that will be able to capture data to characterize historical and real-time traffic patterns.

5 Acknowledgements

This research is funded by the Singapore-MIT Alliance for Research and Technology (The Future of Urban Mobility project), the NSF (grant numbers CPS-0931550, 0735953), and the ONR (grant numbers N00014-09-1-105, N00014-09-1- 1031). We are grateful for this support. We thank Land Transportation Authority (LTA) of Singapore for providing us with the inductor loop detector data, and we thank the anonymous Taxi company for providing us with the taxi data.

6 References

[1] Highway performance monitoring system, federal highway administration, http://www.fhwa.dot.gov/policyinformation/hpms.cfm.

[2] National average speed database, INRIX, www.inrix.com.

[3] Traffic detector handbook: Third edition, fhwa-hrt-06-108, october 2006, http://www.fhwa.dot.gov/publications/research/operations/its/06108/index.cfm.

[4] The 1995 national personal transportation survey (NPTS), http://npts.ornl.gov/npts/1995/Doc/publications.shtml. 1995.

[5] the new 2000 national household travel survey (NHTS), http://www.bts.gov/programs/national_household_travel_survey/. 2000.

[6] S. Baek, H. Kim, and Y. Lim. Multiple-Vehicle Origin–Destination matrix estimation from traffic counts using genetic algorithm. *Journal of Transportation Engineering*, 130(3):339–347, May 2004.

[7] M. Bierlaire and F. Crittin. An efficient algorithm for Real-Time estimation and prediction of dynamic OD tables. *Operations Research*, 52(1), Jan. 2004.

[8] E. Cascetta and S. Nguyen. A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research Part B: Methodological*, 22(6):437–455, Dec. 1988.

[9] T. L. David Schrank and S. Turner. TTI's 2010 urban mobility report, texas transportation institute, the texas a&m university system, http://mobility.tamu.edu. 2010.

[10] J. de Dios Ortzar and L. G. Willumsen. Modelling transport, third edition. *John Wiley & Sons*, 2001.

[11] M. González, C. Hidalgo, and A. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[12] M. L. Hazelton. Estimation of origin-destination matrices from link flows on uncongested networks. *Transportation Research Part B: Methodological*, 34(7):549–566, Sept. 2000.

[13] M. L. Hazelton. Inference for origin-destination matrices: estimation, prediction and reconstruction. *Transportation Research Part B: Methodological*, 35(7):667–676, Aug. 2001.

[14] M. L. Hazelton. Statistical inference for time varying origin-destination matrices. *Transportation Research Part B: Methodological*, 42(6):542–552, July 2008.

[15] M. L. Hazelton. Statistical inference for transit system Origin-Destination matrices. *Technometrics*, 52(2):221–230, May 2010.

[16] J. C. Herrera and A. M. Bayen. Traffic flow reconstruction using mobile sensors and loop detector data. *University of California, Berkeley*, 2007.

[17] T. Litman. Measuring transportation: traffic, mobility and accessibility. 2003.

[18] A. Moore. K-means and hierarchical clustering. http://www.autonlab.org/tutorials/kmens11.pdf, Nov 2001. Accessed Mar 30, 2011.

[19] P. Newson and J. Krumm. Hidden markov map matching through noise and sparseness. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 336–343, 2009. ACM ID: 1653818.

[20] U. R. A. of Singapore. List of postal districts. http://www.ura.gov.sg/realEstateWeb/resources/misc/list_of_postal_districts.htm. Accessed Mar 30, 2011.

[21] A. Recchia and J. C. Hadfield. Regional truck route study. *Southeastern Regional Planning and Economic Development District*, 2009.

[22] E. Richardson A.J. Ampt and A. Meyburg. Survey methods for transport planning. *Eucalyptus Press*, 1995.

[23] D. B. Work, S. Blandin, O. P. Tossavainen, B. Piccoli, and A. M. Bayen. A traffic model for velocity data assimilation. *Applied Mathematics Research eXpress*, 2010(1):1, 2010.

[24] 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 *47th IEEE Conference on Decision and Control, 2008 (CDC '08)*, pages 5062–5068, 2008.