This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2009 proceedings.

SEER: Metropolitan-scale Traffic Perception Based on Lossy Sensory Data

Hongzi Zhu[1], Yanmin Zhu[1], Minglu Li[1] and Lionel M. Ni[1, 2]
[1] Shanghai Jiao Tong University
[2] Hong Kong University of Science and Technology
[1]{hongzi, yzhu, li-ml}@cs.sjtu.edu.cn, [2]ni@cse.ust.hk

***Abstract*—Intelligent transportation systems have become increasingly important for the public transportation in Shanghai. In response, ShanghaiGrid (SG) aims to provide abundant intelligent transportation services to improve the traffic condition. A challenging service in SG is to estimate the real-time traffic condition on surface streets. In this paper, we present an innovative approach SEER to tackle this problem. In SEER, we deploy a cost-effective system of taxi traffic sensors. These taxi sensory data are found to be noisy and very lossy in both time and space. By intensively mining the spatio-temporal correlations along with the evolution of traffic condition, SEER provides wealthy knowledge to setup statistical models for inferring traffic condition when they cannot be directly calculated. As an example, we demonstrate utilizing multi-channel singular spectrum analysis (MSSA) to iteratively produce estimates of traffic condition in a metropolitan scale. The optimal window width of MSSA is determined with the basic periodicity found in traffic condition. Moreover, we minimize the number of channels required by MSSA to estimate traffic condition at any location. Given a desired estimation granularity, we optimize the MSSA parameters to minimize the estimation error.**

Keywords-traffic inference; GPS system; mobile sensor networks; spatio-temporal correlation

I. INTRODUCTION

With the booming economy, traffic has been increasing in the past two decades in Shanghai. This has put a heavy burden on the city infrastructure of the road networks. The public suffers constant traffic congestion. In response to the increasing challenge in public transportation, we launched a research project called ShanghaiGrid [1][2] in 2005, with full support from the Shanghai Government. The goals of the project are two-fold. First, it tries to make the available transportation infrastructure to be used more efficiently. Second, it aims to provide the public with a wide spectrum of intelligent transportation system (ITS) applications, ranging from navigation, trip planning and optimal route selection to congestion avoidance and bus arrival prediction.

One of the most important tasks of the project is to determine traffic condition of the road networks. The significance of this task shows in two aspects. On one hand, it provides the foundation for infrastructure construction planning as well as design optimization of public transportation systems like bus network and metro system. On the other hand, it provides the public with valuable information to plan their travels and to reduce overhead on roads.

However, it is very challenging to determine traffic condition in a metropolis like Shanghai. First, traffic condition is time-varying. Moreover, the changing of traffic condition is often unpredictable as there are so many possible factors influencing the traffic such as incidents, infrastructure construction, weather and festivals. The provided information of traffic condition would be no use if the system needs a long period of time to make the estimation. Second, it is hard to determine traffic condition of the whole road networks. In Shanghai, there are more than 22,413 intersections which connect about 33,290 road segments. It is nontrivial to provide accurate traffic information on all of these road segments.

In industry, there are already a great number of efforts aiming to provide such valuable traffic information. One simple solution would be using radio to broadcast congestion information. In such a system, congestion can be detected by eyewitness reports from commuters or news organizations. Such reports can be very coarse-grained in terms of congestion locations and duration. Many ITSs have deployed considerable sensors such as closed-circuit cameras and vehicle loop detectors as infrastructure [3][4]. Unfortunately, the coverage of these systems is supremely limited due to the high deployment and maintenance costs. It is practically infeasible to install traffic monitoring systems densely enough to cover the entire road networks.

Instead, we propose a systematic approach, called SEER, to traffic perception on a metropolitan scale. Our approach is made up of several components. First, we define an expressive metric to reflect the traffic condition at a given site. It is not straightforward to define a good traffic condition because there are no obvious criteria. We define a metric, *transit velocity*, as the maximum speed at which vehicles can safely transit the site. Intuitively, a high transit velocity within the speed limitation implies a good traffic condition. Second, we deploy a cost-effective system of taxi traffic sensors. On each taxi, a GPS receiver is installed (as shown in Fig. 1). It periodically reports its instantaneous speed and location information to a data center. Therefore, while moving around in the city, taxies

*This research was supported in part by Hong Kong RGC Grants HKUST617908 and HKBU 1/05C, the Key Project of China NSFC Grant 60533110 and 90612018, STCSM Grant No. 05DZ15005 and the National Basic Research Program of China (973 Program) under Grant No. 2006CB303000.

Fig. 1. A taxi with a commercial GPS device installed, the highlight area in the inset shows such a device.

act as mobile sensors perceiving surrounding traffic condition. Third, with the availability of taxi sensory data collected throughout the city, we propose an efficient algorithm which can answer traffic condition queries at any site in the city at any time. In particular, it can even make accurate prediction of traffic in a short period.

It is difficult to determine traffic condition by directly using the sensory data. First, the taxi sensory data is erroneous. The GPS location data is often not accurate, and the error can be as large as 100 meters. As a result, it is difficult to map such a sensor data back to the road map. Second, sensory data may vary from taxi to taxi significantly even they are report at the same location and at the same time. In other words, each sensory data report is associated with a certain degree of noise. Third, the data is lossy and not uniformly distributed both in time and in space. For example, there are 90% of roads that do not have sensory data for more than 80% of all the 1,440 minutes in a day. The fraction would not be less than 50% when count the number of roads that are short for data for more than 12 hours in a day. In addition, we have observed that about 80% sensory data are collected from only about 20% roads.

Fortunately, we have observed that there are strong correlations of traffic condition over both time and space. By using conditional entropy and mutual information, we find out that knowing the traffic condition in the past does help determine the current traffic condition. Moreover, the traffic condition evolves in a basic periodicity of one day. Along with the spatial dimension, we notice that the traffic condition at a site has strong correlation with the traffic condition of a limit number of other sites.

For making use of the natural spatio-temporal correlations, we employ multi-channel singular spectrum analysis (MSSA) as an integral part of our solution. MSSA is a nonparametric algorithm that can effectively eliminate noise from the real signal in a time series. In our problem, the real traffic condition is our signal, and each sensory report deviates from the real signal to a certain extent. Furthermore, it provides the facility to recover signals in face of missing data. However, there are two key questions need to be answered when using MSSA in our problem. The first is how to determine the number of dimensions of the vector space that MSSA embeds the time series into. We find the optimal parameters by setting the number of dimensions to the basic periodicity of one day. The second is how many channels are required for MSSA to estimate the traffic condition at a site. By the spatial correlation analysis, we identify the minimum number of channels and thus reduce the computation overhead in a great deal.

The rest of this paper is structured as follows. Section II compares SEER with related work. In Section III, we describe the characteristics of the taxi sensory data. Section IV presents the spatial and temporal correlations of traffic condition we have observed. We demonstrate utilizing MSSA to estimate traffic condition in Section V. Section VI describes the methodology to evaluate the performance of SEER and presents the results. We have some discussion in Section VII. Finally, we present concluding remarks and outline the directions for future work in Section VIII.

II. Related Work

Early in the 1930s, a number of work started to focus on empirical studies on the traffic flow theory in terms of speed, flow rate, vehicle density, and many other factors [7][8][9]. However, the attention was much paid to find the relationships among traffic factors.

As advanced ITSs evolve, there are a lot of efforts focusing on measuring traffic and incident detections by deploying sensors on roads [3][4]. These schemes have a limited coverage problem due to high deployment and maintenance costs. Recently, there are some ITSs starting to using GPS and/or cellular positioning systems to sample the traffic [5][6]. However, most of them focused on highways or freeways, where the traffic light delay is not an issue in these circumstances. In 2007, Jungkeun Yoon et al [10] proposed a scheme to estimate surface street traffic using GPS data. They concerned how to define an appropriate traffic metric to characterize traffic states whereas we focus on how to determine the traffic condition given certain traffic metric.

III. Taxi Sensory Data

In this section, we give a brief description of the taxi sensory data collected from SG. We then present the underlying characteristics found in the data.

A. Collecting Vehicular Sensory Data

In an initial pilot effort in SG, we deploy a cost-effective system of vehicular traffic sensors. Commercial GPS receivers are installed to certain public vehicles (around 6,850 taxies and 3,620 buses). In Fig. 1, a taxi with a GPS receiver is shown. A vehicle actively reports its current status information back to a data center through a wireless cell-phone data channel (i.e., GPRS). Due to the GPRS communication cost for transmitting the GPS location information back to the data centre, drivers prefer to choose relatively large intervals. The typical value is one minute. The information we can obtain directly from GPS

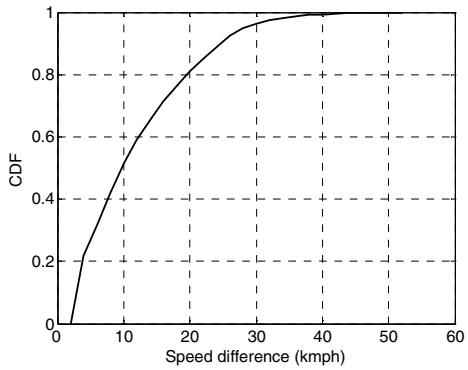

Fig. 2. CDF of speed difference at the same location at the same time

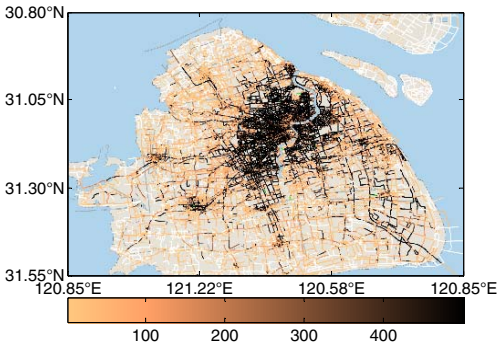

Fig. 3. Spatial distribution of GPS samples for one week from Dec. 15 to Dec. 22, 2006

reports is very limited: a vehicle's location coordinates, timestamp, and optional speed and heading. Despite of the long report intervals and coarse-grained information of the vehicular sensory data, the traffic information contained within the data is very valuable for study of traffic condition.

Currently, we use only taxi sensory data for traffic condition perception because of the sensitivity of taxies to the traffic condition and broad coverage of taxi traces in time as well as in space. For this purpose, we collect trace data of about 15 months from October in 2006 to December in 2007 (One-week demonstration taxi GPS data are available at *http://www.cse.ust.hk/dcrg*). Other heterogeneous traffic information sources obtained from SG will be considered in the future.

B. Characteristics of Taxi Sensory Data

Before we start to determine traffic condition based on taxi sensory data, it is helpful to understand the unique characteristics of the data.

In the city setting with dense high buildings and viaducts, the GPS reports from taxies can be very erroneous. The error of reported locations can be as large as 100 meters. To tell which road a taxi is actually monitoring, we need to recover each sample back on track. We deal with this problem using map-matching [11]. To the best, we can accurately recover about 90% of the data with the left regarded as an inevitable source of noise.

In addition, we also find that individual reports vary significantly even they are collected from the same location at the same time. Figure 2 shows the cumulative distribution function (CDF) of speed difference derived from reports at the same location at the same time. It can be seen that the CDF increases slowly with a relatively long tail, which implies the individual reports can vary largely. The derivation of this variance may be ascribed to individual driving behaviour. For example, a taxi may stop arbitrarily to pick up or drop passengers. In other words, each sensory data report is associated with a certain degree of noise.

Further, we consider the spatial distribution of taxi sensory data. Figure 3 shows the number of samples on each road in a week from December 15 to December 22 in 2006. Totally, there are 42,722 road segments covered by samples of 4,450 taxi traces. It is clear to see that most of the GPS samples are scattered in the downtown area where taxies congregate more densely than in suburbs. The CDFs of sample density on each road are shown in Fig. 4. The data are taken on a weekend, on a workday and for a whole week, respectively. We observe an obvious Pareto distribution in which the "80-20 rule" [12] stands, i.e., 20% of the road segments owns 80% of the data.

Next, we concern the distribution of taxi sensory data in time. We are interested in the probability distribution of the *inter-report times*, which refers to the time intervals between any two consecutive reports received from a location over time. Figure 5 shows the complementary cumulative distribution function (CCDF) of inter-report times. It can be seen that the middle part of the CCDF is almost linear in log-log scale, which indicates a power law. This means a location may frequently has no sensory data for a long time. Figure 6 shows the CCDFs of the proportion of time with no sensory data in a day in different observation granularities. The time windows used to collect sensory reports are one minute, 30 minutes and 60 minutes, respectively. It shows that about 90% of roads have no samples in 80% of the 1,440 minutes in a day. The fraction is about 50% when counting the number of

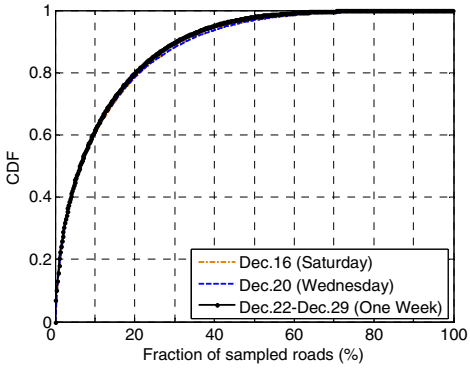

Fig. 4. CDFs of sample density at each road

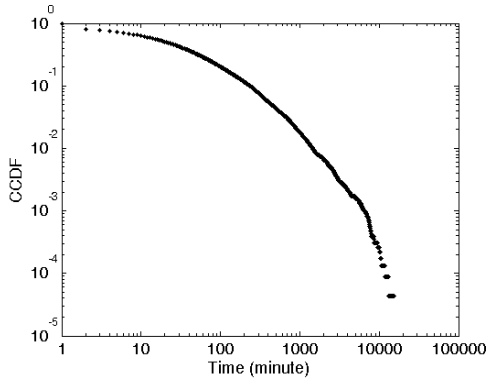

Fig. 5. CCDF of inter-report times

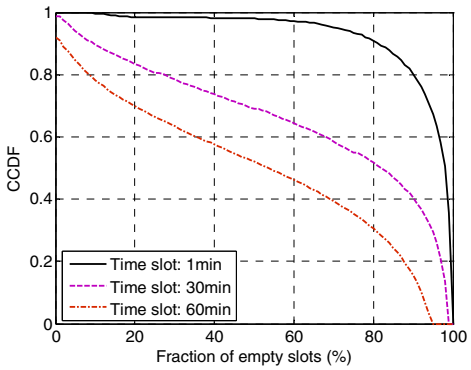

Fig. 6. CCDFs of the proportion of time with no sensory data

road segments that are short of samples for 12 hours in a day.

In summary, there are three crucial characteristics of the taxi sensory data with respect to using these data to determine traffic condition. First, the sensory data are erroneous in terms of large location deviation. Second, individual driving behavior introduces noise in the sensory data. Last, the distribution of the taxi sensory data is very lossy and nonuniform in time and space.

IV. Unveiling Spatio-temporal Correlation

According to the above study, it is not feasible to directly derive traffic condition simply from the trace data due to noise and the sparseness of sampling. In this section, we first model the problem. Then we examine the spatio-temporal correlations of traffic condition.

A. Problem Modelling

Speaking of traffic condition, we are concerning about the transportation capability of the road networks. It is not straightforward to define a good traffic condition because there are no obvious criteria. We define a metric, *transit velocity*, as the maximum speed at which vehicles can safely transit a location. Intuitively, a high transit velocity within the speed limitation implies a good traffic condition.

Let four-tuples (*id*, *location*, *time*, *speed*) denote the GPS reports, where *id* is the identifier of a taxi, *location* is the current coordinates of the taxi, *time* is the report time and *speed* is the instantaneous speed of the taxi. We collect all the reports from each taxi during a time window, denoted by T, and get the set of sensory data, denoted by D. We say a road is *covered* if there is a report that is issued from this road. Let R denote the set of roads that D has covered. With the metric of transit velocity, we define our problem of traffic perception as: *Given the set of sensory data D, how to determine the transit velocity at any location in the road set R at any time in the time window T?*

There is no instant answer to this problem because of the innate characteristics of the sensory data. Even there are sufficient reports obtained at the queried location and time, it is hard to determine what the transit velocity is due to the existence of noise. It can be more difficult to answer the problem when there are no reports available.

In the following subsections, we first measure transit velocity using average speed of reports. We then try to characterize the correlations of average speeds over time and space.

B. Characterizing Temporal Traffic Correlations

To measure the transit velocity at location l at time t, we calculate the *average speed* of reports which are obtained from a distance interval centered at l and a time interval centered at t. We refer to the length of the distance interval and that of the time interval as the calculation granularity, denoted by (Δs, Δt). With a calculation granularity, we can establish a neighborhood of a location and divide continuous time into separate time slots. Formally, the average speed at location l at time t can be calculated as:

$$T_t(l) = \begin{cases} \sum_{i=1}^{n} v_i / n & if\ n \geq m \\ NaN & otherwise \end{cases}, \quad (1)$$

where n is the total number of reports collected from the neighborhood of l in the time slot of t, v_i is the speed of the i^{th} report and m is the minimum number of reports to calculate the average. We set m larger than one (e.g., at least 3 reports) to reduce the impact of individual driving behavior. If there is no sufficient reports available, i.e. $n<m$, the average speed is left blank with no value assigned.

Let us look at a simple case where we relax the spatial calculation granularity Δs. We consider average speeds on a *road segment*, which refers to the part between two neighboring intersections of a road in one direction.

For ease of computation, we further discretize the continuous average speed values into Q disjoint sub-intervals without losing generalization. We hereafter use sub-intervals

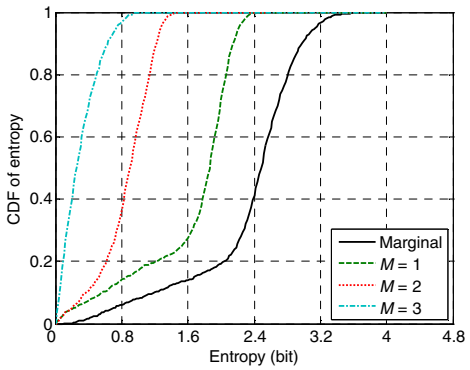

Fig. 7. CDFs of marginal entropy and conditional entropy of average speeds

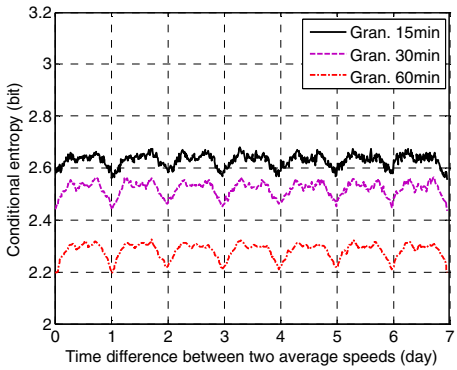

Fig. 8. Conditional entropy given average speed in each time slot in the last week

to represent corresponding average speeds. For example, if the average speed value is 48kmph and the speed values are separated by 10kmph, we say the average speed is four.

In order to understand how traffic condition evolves over time, we need to answer two specific questions, namely 1) how historical information is related to the current traffic condition and 2) how much historical information are related to determine the current traffic condition.

We first examine whether or not knowing the traffic condition on a road segment in the past can help us determine the current traffic condition on that road segment. We do this by computing the entropy of average speeds on each road segment and the conditional entropy of the average speed on a road segment given previous M average speeds. Let X be the random variable representing the average speeds on a road segment r. If we have observed the road segment for N time slots, the time series of average speeds can be denoted by a vector $V_r = (k_0, k_1, \ldots, k_{N-1})$ where $k_i \in [0, Q-1], 0 \le i \le N-1$ is the average speed in time slot i. Assume each of these Q average speeds appeared s_j times in V_r, $0 \le j \le Q-1$. Thus, the probability of the average speed on the road segment being j can be computed as s_j/N. Therefore, the entropy of X is:

$$H(X) = \sum_{j=0}^{Q-1} (s_j / N) \log_2 \frac{1}{s_j / N}. \quad (2)$$

When M=1, let X' be the random variable for the immediately previous average speed on the road segment given the average speed X. X' and X have the same distribution when N is large enough. The vector V_r can be written as $W_r = \{(k_i, k_{i+1}) : 0 \le i \le N-2\}$. Therefore, the joint entropy of X' and X can be computed as:

$$H(X', X) = \sum_{(x', x) \in W} P(x', x) \log_2 \frac{1}{P(x', x)}, \quad (3)$$

where $P(X', X)$ is the number of times (x', x) appears in W_r divided by the total umber of elements in W_r. With $H(X)$ and $H(X', X)$, the conditional entropy of X given X' is:

$$H(X \mid X') = H(X', X) - H(X') = H(X', X) - H(X). \quad (4)$$

When M=2, let X'' denote the random variable representing the distribution of the previous *two* average speeds given X. Similarly, the conditional entropy $H(X \mid X'')$ is:

$$\begin{aligned} H(X \mid X'') &= H(X'', X) - H(X'') \\ &= H(X'', X) - H(X', X), \end{aligned} \quad (5)$$

The joint entropy $H(X'', X)$ can be calculated similarly. We can continue the process and get the joint entropy when M is larger than two.

Figure 7 shows the CDFs of the mean entropy and the mean conditional entropy, for M=1, 2, and 3, over all road segments. It can be seen that the conditional entropy when M=1 is much smaller than the marginal entropy and that the conditional entropy when M=3 is smaller than that when M=2 which is smaller then when M=1. This implies that the uncertainty about the average speed decreases when the previous average speeds on the road segment are known.

To answer the second question, we examine the correlation between the average speed in time slot t and that in time slot t-n and vary n from one to a large number. Let Y_n denote the random variable for the average speed in the previous n^{th} time slot given the average speed X. Then the condition entropy of X given Y_n is:

$$H(X \mid Y_n) = H(Y_n, X) - H(Y_n) = H(Y_n, X) - H(X). \quad (6)$$

Figure 8 shows the conditional entropy for each time slot in previous week. The average speeds are computed with temporal granularity of 15 minutes, 30 minutes and 60 minutes, respectively. In each case, we observe that the conditional entropy reaches a minimum when the value of n is times of 24 hours. This means that the uncertainty about the average speed on a road segment is least when we know the average speeds at the same time on past days. Easily, we can identify a periodicity of one day

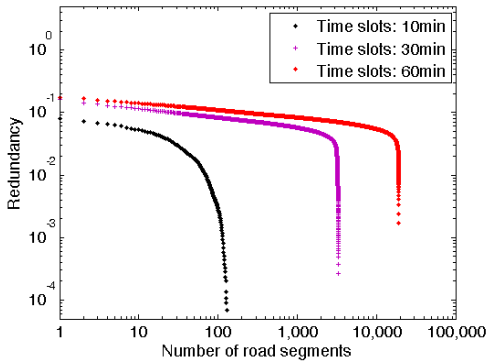

Fig. 9. Mean redundancy over all road segments, sorted in descending order

C. Characterizing Spatial Traffic Correlations

In this subsection, we examine whether or not knowing the traffic condition in the neighborhood of a road segment can help us determine the traffic condition on that road segment.

We quantify the correlation between average speeds on two different road segments as follows. Recall that each road segment has a time series of average speeds. Let X_{r1} and X_{r2} denote the random variables for the average speeds on road segment $r1$ and $r2$, respectively. We can obtain the mutual information of X_{r1} and X_{r2}, $I(X_{r1}, X_{r2})$, via the joint entropy $H(X_{r1}, X_{r2})$ and the marginal entropy $H(X_{r1})$ and $H(X_{r2})$ as follows:

$$I(X_{r1}, X_{r2}) = H(X_{r1}) + H(X_{r2}) - H(X_{r1}, X_{r2}). \quad (7)$$

We define the redundancy of X_{r1} and X_{r2} by

$$R(X_{r1}, X_{r2}) = \frac{I(X_{r1}, X_{r2})}{H(X_{r1}) + H(X_{r2})}. \quad (8)$$

Figure 9 shows the mean redundancy for all road segments under three temporal calculation granularities. It can be seen that the redundancy drops dramatically at certain number of road segments at all granularities. This implies that the average speeds on a road segment are only related to a limit number of road segments. Moreover, when the granularity decreases, this number also decreases rapidly. This result is valuable when estimating average speeds leveraging spatial correlations among different road segments. We only need to consider a modest number of related road segments to estimate a road segment. Notice that these road segments do not necessarily need to be geographically near from the road segment.

In summary, we make the following observations regarding traffic perception using taxi sensory data:

- The sampling GPS data is rather lossy in terms of spatio-temporal distribution. This results frequent gaps without sufficient samples available for estimating traffic condition. We should have confidence to reconstruct the traffic condition at missing points.
- There are various sources of noise involved. This may arise from the measurement errors of GPS devices as well as from the inaccuracy of map-matching algorithms, or from the individual driving behavior. Although we can reduce this impact by aggregating multiple samples, there is impossible to remove all the noise. We should have the capability to distinguish signal from noise.
- Traffic condition has apparent spatio-temporal correlations. We find out a basic periodicity of one day. This periodicity has nothing to do with the calculation granularity. Moreover, we find out that a road segment has much more correlation with only a small number of road segments when the calculation granularity is small.

V. MSSA Based Traffic Perception

In this section, we first give an overview of our traffic perception algorithm based on *multi-channel singular spectrum analysis* (MSSA), introducing its basic rational. Next, we give a brief review to MSSA. Then we describe the iterative procedure used to estimate unknown traffic condition. Finally, we discuss the optimal configuration of the algorithm parameters.

A. Overview

Based on our above observations, we hope to extract useful information from the noisy time series of traffic condition and thus provide insight into the unknown dynamics of the underlying transportation system that generated the series.

First, we leverage the capability of MSSA to distinguish signal from noise contained in the traffic condition. MSSA takes advantage of both spatial and temporal correlations and decomposes the original time series into trends, oscillatory patterns and noise. A number of heuristic methods have been devised for signal-to-noise separation. With the trends and significant oscillatory patterns, we can reconstruct the signal.

Second, we use an iterative algorithm to deal with missing points in the time series. The algorithm iteratively produces estimates of unknown traffic condition using MSSA. Then the new estimates are used to compute a self-consistent lag-covariance matrix. The optimal window width and the minimal number of channels of MSSA are determined based on our observations on spatio-temporal correlations.

B. MSSA Review

MSSA is an ingenious application of the Karhumen-loève expansion for random processes [13]. It provides qualitative and quantitative information about the deterministic and

stochastic parts of system behavior recorded in a stationary time series even when the time series is short and noisy.

The MSSA method is data-adaptive and nonparametric based on embedding an L-channel time series with N data points $\{X_l(t):l=1,\ldots,L;t=1,\ldots,N\}$ in a vector space of dimension M. A multi-channel *trajectory matrix* $\tilde{X}=\left(\tilde{X}_1;\tilde{X}_2;\cdots;\tilde{X}_L\right)$ of X with M lagged copies can be formed by first augmenting each channel:

$$\tilde{X}_l=\begin{pmatrix} X_{l,1} & \cdots X_{l,M} \\ X_{l,2} & \cdots X_{l,N'+1} \\ & \cdots \\ X_{l,N'} & \cdots X_{l,N} \end{pmatrix}, 1\le l\le L \tag{9}$$

Thereafter, both spatial correlations between any two of L channels and temporal correlations in each channel can be obtain by computing the grand covariance matrix C_X:

$$C_X=\frac{1}{N'}\tilde{X}\tilde{X}^t=\left(C_{l,l'}\right)_{L\times L} \tag{10}$$

By diagonalizing the $LM\times LM$ matrix C_X, spectral information on the time series can be obtained. The eigenvectors E^k, $1\le k\le LM$, of grand covariance matrix C_X are called temporal extended empirical orthogonal functions (EEOFs). Each E^k consists of L consecutive M-long segments, with its elements denoted by $E^k_{l,m}$. The eigenvalues λ_k of C_X account for the partial variance of the original time series $X_l(t)$ in the direction of E^k. Corresponding to each EEOF, space-time principal components (PCs) A^k can be computed as:

$$A^k_n(t)=\sum_{m=1}^{M}\sum_{l=1}^{L}X_{l,n+m-1}E^k_{l,m}, \tag{11}$$

where n varies from 1 to N'.

Trends, oscillatory modes and noise contained in the entire time series can then be reconstructed by using linear combinations of these PCs and EEOFs. Specifically, the k^{th} reconstructed component (RC) at time n for channel l is:

$$R^k_{l,n}=\frac{1}{M_n}\sum_{m=L_n}^{U_n}A^k_{n-m+1}E^k_{l,m}, \tag{12}$$

The values of the normalization factor M_n, as well as of the lower and upper bound of summation L_n and U_n can be determined as,

$$M_n,L_n,U_n=\begin{cases}1,1,n, & 1\le n\le M-1\\ M,1,M, & M\le n\le N'\\ N-n+1,n-N+M,M, & N'+1\le n\le N\end{cases} \tag{13}$$

Generally, there are two main problems in using MSSA. One is how to determine the time window M (embedding dimension). The window size M should be larger than the longest periodicity that we are interested in. The other one is to determine what parts of the EEOFs are corresponding to significant oscillatory models. An oscillatory mode can be characterized by a pair of nearly equal eigenvalues and periodic eigenvectors that correspond to the same frequency.

C. Dealing with Missing Data

We adopt an iterative procedure as proposed in [14] [15] to utilize spatio-temporal correlations of traffic condition to estimate the missing points. Generally, this procedure iteratively produces estimates of missing points, which are then used to compute a self-consistent covariance matrix C_X and its EEOFs E^k. In the proposed methods, a brute-force cross-validation is required to optimize the window width M and the number of EEOFs that corresponds to significant oscillatory modes. Instead, we skip the cross-validation with confidence built on the foundation of our observations on temporal correlation. In addition, we minimize the number of channels required to reconstruct traffic condition based on our observation on spatial correlation.

Specifically, we first calculate average speeds according to a given temporal granularity. Points with no sufficient reports are regarded as missing points.

Then we set the window width of MSSA to the basic periodicity of one day. We center each channel of the original average speeds by computing the unbiased value of the mean and set the values of missing points to zero. A fraction of average speeds is left out for the purpose of validation.

Next, we start the inner-loop iteration by computing the leading EEOF E^1 and estimate the missing points using only R^1. Thus, we get a new time series with missing points estimated by R^1 and correct the mean. We then perform the MSSA algorithm again on the new series. Each estimate of the missing points is tested against the previous one until a convergence test is satisfied. Next, we perform outer-loop iterations by adding a second EEOF E^2 for estimation and repeat the inner-loop iteration. For each outer-loop iteration, we test the root-mean-square (RMS) deviation of the estimated average speeds with the reserved values. The outer-loop iterations are stopped when the minimum RMS deviations is found.

Finally, we take the parameters, K^* and M^*, that minimize the RMS deviation as the required optimum. To obtain the actual reconstruction, we repeat the inner and outer-loop iterations, using K^* and M^*, but with all available average speeds being included in the process.

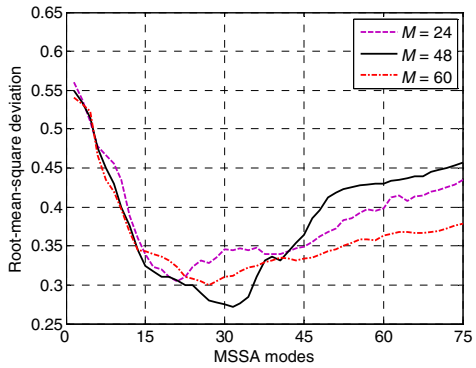

Fig. 10. RMS deviation as a function of window width *M* and number of MSSA modes

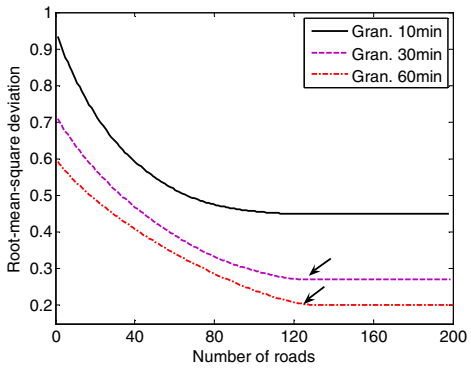

Fig. 11. RMS deviation as a function of number of roads

D. *Optimal Parameter Configuration*

As described above, we establish the optimal window width of MSSA based on our previous observations on temporal correlation. This decision can greatly accelerate the search for the optimal set of MSSA parameters. Besides the window width, we also minimize the number of channels needed to estimate average speeds on a certain road segment. We do this by leverage the observation that a road segment has much more correlation with only a small number of road segments when the calculation granularity is small. Therefore, it is not necessary to build up a large grand covariance matrix C_X and enormously reduce the computation overhead.

In the next section, we validate our iterative estimation process and examine the performance on using MSSA to determine traffic condition.

VI. Performance Evaluation

A. *Methodology and Metric Design*

In this section, we apply the MSSA-based traffic perception algorithm to the sensory data collected from a region in the Pudong district from December 1 to December 31 in 2006. Totally, there are totally 3,135 taxies involved in reporting traffic information on 235 roads.

We define the RMS deviation of two vectors of average speeds $V_1=[v_{1,1}, v_{1,2},\ldots,v_{1,n}]$ and $V_2=[v_{2,1}, v_{2,2},\ldots,v_{2,n}]$ as:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(v_{1,i}-v_{2,i})^2}{n}} \qquad (14)$$

B. *Impact of Window Width and MSSA Mode Quantity*

In this experiment, we investigate the impact of the time window employed in the iterative procedure on the performance of traffic perception. We set the temporal granularity to 30 minutes, and calculate average speeds on each road segment. We randomly choose 5% of average speed results for validation and carry out the iterative algorithm for 20 times for window widths of 24, 48 and 60. In each run of the experiment, all of the 235 roads are employed as MSSA channels.

Figure 10 shows the mean RMS deviation as a function of the window width *M* and the number of MSSA modes. We find out that the reconstruction error drops rapidly as the number of regular oscillatory modes increases. Nevertheless, the error gradually starts to increase as the number of MSSA modes keeps growing. This can be easily understood because more PCs corresponding to noise has joined the reconstruction. We also notice that the RMS deviation reaches the globally minimum when the time window width is 48, which is one day in time. This result validates our suggestion of using the basic periodicity found in traffic condition as the optimum time window width in MSSA.

C. *Impact of Channel Quantity*

In this experiment, we examine whether we can reduce the number of channels to be used in MSSA to reconstruct traffic condition at a road segment. We randomly choose 20 road segments for traffic perception. For each road segment, we rank all road segments according to the redundancy calculated using the method mentioned in Section IV. Then we gradually add the number of road segments to be involved in the iterative algorithm. Different temporal granularities are used to calculate average speeds.

Figure 11 shows the RMS deviation as a function of the number of road segments involved. It can be seen that, in general, the reconstruction error deceases as the number of road segments increases. Particular, when a small temporal granularity is used, only a small number of road segments can help decrease the RMD error. We also notice there are slop breaks in the dash lines in Fig. 11 as pointed out by the arrows. The reason may be that we only choose a small region to conduct the experiment and omit some related road segments. In addition, we can also find that using the road segments after the slope breaks does not provide any more benefit but computation overhead. This result strongly agrees with our analysis on spatial correlation of traffic condition.

D. *Impact of Temporal Granularity*

In the above experiment, we can find out that the reconstruction error increases when the temporal granularity

gets small. Reducing the granularity will cause more missing points and low signal-to-noise ratio. As the amount of mission points and noise increases, the significant PCs are "polluted" more, making it more difficult to remove the noise contributions. Even in this case, we find that the regular oscillatory modes can be determined correctly as long as the gap of missing data is not larger than any significant spatio-temporal correlations of traffic condition.

VII. Discussion

In this paper, the main coverage has focused on the innovation of a traffic perception system using pervasive taxi traffic sensors. On the one hand, we have put major efforts on establishing the prototype system of the taxi traffic sensors, which has been proved to be cost effective and successfully laid the foundation for our traffic perception algorithm. On the other hand, we have managed to extract traffic condition information from loss sensory data which by nature is erroneous and non-uniform. However, as a pioneering effort, the proposed approach is not perfect. Several problems need to be further investigated in future.

Due to the discrete nature of sensory data of traffic reports from taxis, it is impossible to acquire the exact sensor data at a given point. Thus, we have proposed the approximating method by using sensor reports in the neighborhood. It is worth more careful study on how much the neighborhood size could be. Note that a larger neighborhood results in a larger set of sensor reports. Meanwhile, however, a sensor report further from the given point provides data of lower quality. Given a certain scenario of traffic condition and density of sensory reports, we believe there should be an optimal neighborhood size. Nevertheless, it is not trivial to determine the optimal value.

It is apparent that we could derive better traffic condition information if more sensory data were available. However, more sensory data also imply higher investment on recruiting more taxis and conveying a larger volume of sensory data back to the information center. In addition, it would introduce higher computation cost for processing raw data and executing the algorithm. Nevertheless, it is important to study the tradeoff between the volume of sensory data and the ability of answering queries on traffic condition. Such tradeoff study will allow us to determine how much sensory data should be acquired given a certain use requirement on query quality.

VIII. Conclusion and Future Work

In this paper we have presented the systematic approach to perceiving metropolitan traffic using a cost-effective system of taxi sensors. Although the raw sensory data collected by taxis are error-prone and non-uniform, our MSSA based algorithm can still effectively produce traffic condition of high quality. It also solves the lossy problem of sensory data in the sense that a given location may have a very limited number of sensor reports. As a result, this system overcomes many of the limitations for existing approaches, such as high cost and requirements on manpower. This system can be quickly implemented and serve the general public. The prototype system has been working and feeding valuable traffic condition information to the Transport Office of Shanghai.

With SEER having much space to improve, we will carry on our research in several directions. First, the problems discussed in Section VII will be carefully investigated and constant improvements will be made accordingly. Second, we will expand our system to include a richer set of sensory data input, by employing buses and volunteer cars. Moreover, we will also incorporate the distribution information of traffic lights in the city, which may have a non-negligible impact on traffic condition.

References

[1] Hongzi Zhu, Yanmin Zhu, Minglu Li, and Lionel M. Li, "ANTS: Efficient Vehicle Locating Based on Ant Search in Shanghai Tranportation Grid", *Proc. Int. Conf. Parallel Processing*, 2007.

[2] Hongzi Zhu, Yanmin Zhu, Minglu Li, and Lionel M. Li, "HERO: Online Real-time Vehicle Tracking in Shanghai", *Proc. IEEE INFOCOM*, 2008.

[3] B. Coifman, "Identifying the onset of congestion rapidly with existing traffic detectors", *In Transportation Research*, volume 37 of Part A, pages 277–291. 2003.

[4] W. Lin and C. Daganzo, "A simple detection scheme for delay-inducing freeway incidents", *In Transportation Research*, volume 31A of Part A, pages 141–155. 1997.

[5] M. McNally, J. Marca, C. Rindt, and A. Koos, "Tracer: In-vehicle, gps-based, wireless technology for traffic surveillance and management", Technical Report UCB-ITS-PRR-2003-23, California Partners for Advanced Transit and Highways (PATH), July 2003.

[6] J. Ygnace, C. Drane, Y. Yim, and R. Lacvivier, "Travel time estimation on the san francisco bay area network using cellular phones as probes", Technical Report UCB-ITS-PWP-2000-18, California Partners for Advanced Transit and Highways (PATH), September 2000.

[7] N. H. Gartner, C. Messer, A. K. Rathi, F. L. Hall, R. J. Koppa, R. W. Rothery, R. Kuhne, P. Michalopoulos, J. C. Williams, S. Ardekani, E. Hauer, B. Jamei, R. J. Troutbeck, W. Brilon, N. Rouphail, A. Tarko, J. Li, and E. Lieberman, "Traffic flow theory", Revised special report, Transportation Research Board (TRB), 2005.

[8] B. S. Kerner, "The Physics of Traffic", Springer-Verlag, 2004

[9] W. Leutzbach, "Introduction to the Theory of Traffic Flow", Springer-Verlag, 1988.

[10] Jungkeun Yoon, Brian Noble, and Mingyan Liu, "Surface Street Traffic Estimation", *Proc. ACM MobiCom*, 2007.

[11] Xu Li, Wei Shu, Minglu Li, Pei-En Luo, Hongyu Huang and Min-You Wu, "Performance Evaluation of Vehicle-based Mobile Sensor Networks for Traffic Monitoring", accept by *IEEE Transaction on Vehicular Technology*, 2008.

[12] Bookstein, and Abraham, "Informetric distributions, part I: Unified overview", *Journal of the American Society for Information Science* 41: 368–75, 1990.

[13] K. Fukunaga, "Introduction to Statistical Pattern Recognition", Academic Press, New York, 1970.

[14] Beckers, J. and Rixen, M., "EOF calculations and data filling from incomplete oceanographic data sets", J. Atmos. Ocean. Technol., 20, 1839–1856, 2003.

[15] D. Kondrashov and M. Ghil, "Spatio-temporal filling of missing points in geophysical data sets", *Nonlin. Processes Geophys*., 13, 151–159, 2006.