Forecasting Collector Road Speeds Under High Percentage of Missing Data

Xin Xin*
School of Computer Science
Beijing Institute of Tech.
xxin@bit.edu.cn

Chunwei Lu
Autopia Mobile
Tech Group Inc.
myles.lu@chetuobang.com

Yashen Wang
School of Computer Science
Beijing Institute of Tech.
yswang@bit.edu.cn

Heyan Huang
School of Computer Science
Beijing Institute of Tech.
hhy63@bit.edu.cn

Abstract

Accurate road speed predictions can help drivers in smart route planning. Although the issue has been studied previously, most existing work focus on arterial roads only, where sensors are configured closely for collecting complete real-time data. For collector roads where sensors sparsely cover, however, speed predictions are often ignored. With GPS-equipped floating car signals being available nowadays, we aim at forecasting collector road speeds by utilizing these signals. The main challenge compared with arterial roads comes from the missing data. In a time slot of the real case, over 90% of collector roads cannot be covered by enough floating cars. Thus most traditional approaches for arterial roads, relying on complete historical data, cannot be employed directly. Aiming at solving this problem, we propose a multi-view road speed prediction framework. In the first view, temporal patterns are modeled by a layered hidden Markov model; and in the second view, spatial patterns are modeled by a collective matrix factorization model. The two models are learned and inferred simultaneously in a co-regularized manner. Experiments conducted in the Beijing road network, based on 10K taxi signals in 2 years, have demonstrated that the approach outperforms traditional approaches by 10% in MAE and RMSE.

Introduction

With fast-paced lifestyles, people living in major cities are increasingly concerned about the road congestion problem. Spending hours daily on the road constantly compromises one's working efficiency and mood. As a result, forecasting road speeds has become a crucial urban service nowadays. It help drivers to avoid roads expected to be congested soon.

The problem of the road speed prediction has been extensively studied previously. But most previous work are limited in dealing with freeways and arterial roads, where sensors (such as loops, cameras) are configured closely to collect complete real-time road speeds. Due to the nontrivial cost, collector roads are often sparsely covered by these sensors, where most existing approaches will fail to work. Collector roads refer to secondary main roads that connect arterial roads in cities, as illustrated in Fig. 1(a). In route planning, forecasting collector road speeds usually has the same importance with arterial roads, as they always compose the alternate routes. Unfortunately, this task has not been thoroughly explored yet, and most predictions for collector roads are simply approximated by historical average values.

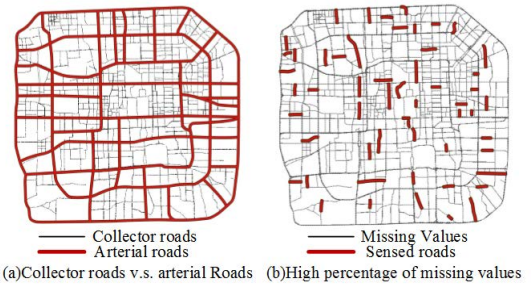


Figure 1: The task of forecasting collector road speeds

While configuring sensors for each collector road is very difficult, other ubiquitous data sources indicating collector road speeds are already available. A typical example is the signals from GPS-equipped floating cars, such as taxies and cargo vans. These vehicles travel both arterial and collector roads in cities, and their real-time speeds can sense the corresponding road speeds. Therefore, this brings alternative opportunities for collector road speeds prediction.

The high percentage of missing data is the main challenge, which differentiates the task in this paper from previous tasks. The challenge of data sparsity is also common in general urban computing tasks (Zheng et al. 2014). Traditional tasks are based on complete historical data, recorded by sensors. But in our case, the historical data is incomplete and sparse. Figure 1(b) shows a snapshot of Beijing collector roads covered in ten minutes by 10K taxis. Only 10% have reliable observed road speeds, while a large percent of the data is missing. Hence, most previous methods, relying on complete data, cannot be directly applied.

In this work, aiming at forecasting collector road speeds effectively and efficiently, we propose a multi-view learning framework based on the following assumptions:

1. *Temporal Patterns*. The future speed of the active collec-

*The corresponding author.

tor road depends on the current speeds of its neighbor roads, as well as the current overall speed of its region.

2. *Spatial Patterns*. The collector roads, which share common spatial contexts, such as the road level, the surrounded points of interest (e.g., schools, hospitals), and the corresponding region, are likely to have similar speeds.
3. *Patterns Agreement*. For the missing data, predictions based on accurate temporal patterns and accurate spatial patterns, will agree with each other.

The first assumption leads to a two-layer hidden Markov model (LHMM). In the first layer, we divide a city into regions, and utilize the Markov property to model the temporal patterns of region speeds. In the second layer, we model temporal patterns of collector road speeds. The second assumption leads to a collective matrix factorization model (CMF) for spatial pattern analysis. The third assumption leads to a multi-view learning framework of the LHMM and the CMF, where disagreements from the two views are penalized. Our work has three primary contributions.

- *Collector road speed prediction*. This previously ignored task has been thoroughly studied in this paper, serving as complementary information with traditional arterial road speed predictions for the route planning service.
- *Dealing with data sparsity*. The multi-view framework bridges the low-rank property of matrix factorization, being demonstrated effective in dealing with the missing data case, with the hidden Markov model. The joint model naturally combines spacial patterns and temporal patterns in alleviating the data sparsity problem.
- *Real evaluation*. By conducting experiments in Beijing road network, with trajectories of around 10K taxies in 2 years, we demonstrate that the proposed approach outperforms traditional methods by 10% in MAE and RMSE.

Related Work

Traffic Forecasting with Dense and Complete Data

Approaches for predicting traffic conditions with dense and complete data have been investigated in detail recently. Typical methods include random walk (Wang et al. 2006), autoregressive integrated moving average model (ARIMA) (Hamilton 1994; Haworth and Cheng 2012; Min and Wynter 2011; Smith, Williams, and Keith Oswald 2002; Williams and Hoel 2003), neural networks (Lint et al. 2005; Park and Rilett 1998; Van Lint, Hoogendoorn, and van Zuylen 2005), support vector regression (Castro-Neto et al. 2009), hidden Markov model (Qi and Ishak 2014; Thiagarajan et al. 2009) and conditional random fields (Djuric et al. 2011). Hidden Markov model is one of the most competitive approaches among them (Qi and Ishak 2014).

Differences. The above approaches are based on dense and complete data. Hence they only perform well for freeways and arterial roads, where sensors can cover. But in our case, we target at forecasting collector road speeds, where a large percent of historical records are missing. Most previous approaches cannot be employed directly.

Traffic Forecasting with Sparse Data

The traffic missing data imputation issue targets at completing the historical data, where failures occur in sensors. Approaches include neighbor-based heuristic rules (Ni et al. 2005; Smith, Scherer, and Conklin 2003), neural networks (Zhong, Lingras, and Sharma 2004), support vector machines (Zhang and Liu 2009), hidden Markov model (Herring et al. 2010), factorization-based methods (Qu et al. 2009; Tan et al. 2013), and etc (Bejan and Gibbens 2011; Yuan et al. 2011; Zhu et al. 2013). Recently, Zheng at el. (Shang et al. 2014; Zheng et al. 2010) utilized factorization techniques to complete the missing data for taxi related applications, and demonstrated its effectiveness for alleviating data sparsity.

Differences. The above approaches mainly focus on completing historical data, where observations exist in all time slots. But our work focuses on future data prediction, where no observations are acquired. Although one can utilize these data imputation approaches to recover the historical missing data first, and then utilize traditional methods to predict the future data, the error propagation will impair the performance significantly. Hence, at this point, our work can be seen as solving the missing historical data imputation and the future data prediction simultaneously to avoid error propagations.

Multi-view Learning

Multi-view learning has been widely used when large percent of the data is unlabeled/unobserved (Sridharan and Kakade 2008; Sun 2013; Yu et al. 2011; Zheng, Liu, and Hsieh 2013). It is usually composed of two independent features/models as views. The underlying assumption is, either view can make accurate predictions based on sufficient data. Thus hypothesis from either view, whose predictions tend to disagree with predictions based on the other view, is eliminated. The two views are co-regularized to avoid overfitting. In this paper, we build a multi-view framework between hidden Markov model and matrix factorization. The former models the temporal patterns, and the latter models the spatial patterns.

Prediction Framework

Problem Definition

The major task of this paper is to predict collector road speeds in cities. The dataset to be utilized contains the map data and the floating car data. ***The map data*** consists of road ids, road connection relations, and road contexts. Following the presentation in Shang et al. and Wang et al.'s work (Shang et al. 2014; Wang, Zheng, and Xue 2014), road contexts are pre-processed into the road-context matrix as shown in Fig 2(a), with three sub-categories: (1) road properties f_r, (2) points of interest (POIs) f_p, and (3) global positions f_g. f_r contains road lengths, road levels, numbers of connections, and directions, with all values normalized into $(0, 1)$. Ten types of POIs are selected to describe a road's surroundings, including hospitals, schools, and etc. f_p denotes the distribution of the numbers of each POI type near a road. The city map is divided into 16 grids,

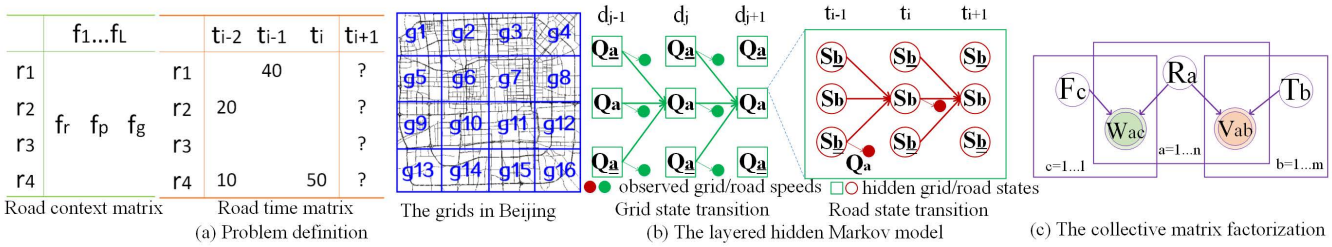


Figure 2: Problem definition & graphical models of LHMM and CMF

as shown in Fig. 2(a). f_g describes the grid of a road, indicating its global position in the city. It is denoted by the active grid's neighbors. Take roads in g_6 for example, it is denoted as $(1,1,1,0,1,0,1,0,1,1,1,0,0,0,0,0)$[1]. In such a notation, adjacent grids will share similar vectors. ***The floating car data*** has been pre-processed into the road-time matrix as shown in Fig. 2(a). The entry denotes the speed of a road (precisely defined in the next section) in a time slot of 10 minutes. t_i is the current time slot. Suppose there are totally n collector roads, m time slots (including the future time slot t_{i+1}), and l kinds of road features.

The problem under investigation is essentially how to effectively and efficiently predict the missing values of the future time slot t_{i+1} in the road-time matrix, by utilizing a sample of recent speed records and the map data? Please note, if speeds in t_{i+1} are predicted, the ones in t_{i+k} can be iteratively predicted based on predictions in $[t_0, t_{i+k-1}]$.

Temporal Analysis

A two-layer hidden Markov model, as shown in Fig. 2 (b), is proposed for temporal analysis. The inside layer is defined on collector roads, and it models their speeds in a time slot. The outside layer is defined on the grid (see Fig. 2 (a)), and it models the overall speed of a region in a period. A ***period*** is defined as a time window of 6 continuous time slots. The elements of the model is defined as follows.

Definition 1 (Road (Grid) Speed V_{ab} (U_{xy})) *V_{ab} (or U_{xy}) is the average speed on a certain road r_a (or grid g_x) in a time slot t_b (or period d_y). When the number of floating car signals, which contain their instant speeds, is large than a threshold (10 in this paper), it is calculated as the average value; or otherwise, the road (or grid) speed data is missing.*

Definition 2 (Road (Grid) State S_{ab} (Q_{xy})) *In time slot t_b (or period d_y), road r_a (or grid g_x) has its hidden state $S_{ab} \in \{0,1\}$ (or $Q_{xy} \in \{0,1\}$). State 0 means the road (or grid) is crowed, and state 1 means the road (or grid) is free.*

Definition 3 (Road (Grid) Speed Distribution D_S (D_Q)) *When S_{ab} (or Q_{xy}) is known, V_{ab} (or U_{xy}) follows an unique Gaussian $\mathcal{N}(v_{ab}, \sigma^2_{ab}|S_{ab})$ (or $\mathcal{N}(u_{xy}, \sigma^2_{xy}|Q_{xy})$).*

Definition 4 (Road (Grid) State Trans. Probs. P_S (P_Q)) *We utilize 1-order Markov property assumption in modeling state transitions. The future state of the active grid depends only on its neighbors' current states, denoted*[2] *by $P(Q_{x,d_{j+1}}|history) = P(Q_{x,d_{j+1}}|n(Q_{x,d_j}))$. The future state of the active road depends on both its neighbors' current states and its corresponding grid's current state, denoted*[3] *by $P(S_{a,t_{i+1}}|history) = P(S_{a,t_{i+1}}|Q_{g(a),d(t_i)}, n(S_{a,t_i}))$. To simplify the model complexity, the neighbors of a road (or grid) contain itself, and up to only two closest neighbors*[4].

Definition 5 (Initial Road (Grid) State Dist., π_S (π_Q)) *π_S (π_Q) refers to initial distributions of road (grid) states.*

Under the above definitions, the parameters of the LHMM are $\vartheta_{hmm} = \{\pi_S, \pi_Q, D_S, D_Q, P_S, P_Q\}$, and its variables are $\{S, Q\}$. ϑ_{hmm} is learned through a training process, which will be discussed later. The inference process is to find $\{S, Q\}$ that can maximize the following likelihood.

$$
\begin{aligned}
&P(V, U, S, Q; \vartheta_{hmm}) = \\
&\prod_{x=1}^{16} P(Q_{x,d_1}) P(U_{x,d_1}|Q_{x,d_1})^{I^U_{x,d_1}} \\
&\cdot \prod_{j=2}^{d} \prod_{x=1}^{16} P(Q_{x,d_j}|n(Q_{x,d_{j-1}})) P(U_{x,d_j}|Q_{x,d_j})^{I^U_{x,d_j}} \\
&\cdot \prod_{a=1}^{n} P(S_{a,t_1}) P(V_{a,t_1}|S_{a,t_1})^{I^V_{a,t_1}} \\
&\cdot \prod_{i=2}^{m} \prod_{a=1}^{n} P(S_{a,t_i}|n(S_{a,t_{i-1}}), Q_{g(a),d(t_i)}) P(V_{a,t_i}|S_{a,t_i})^{I^V_{a,t_i}}
\end{aligned}
$$

When $\{S, Q\}$ are found, the imputation of missing data and the prediction of future road speeds are calculated as the expectation speeds over the two states, as

$$
\begin{aligned}
f^{lhmm}_{ab} = &E(V_{ab}|S_{ab}=0) P(S_{ab}=0|Q_{g(a),d(b-1)}, S - S_{a,b}) \\
&+ E(V_{ab}|S_{ab}=1) P(S_{ab}=1|Q_{g(a),d(b-1)}, S - S_{a,b}).
\end{aligned}
$$

Spatial Analysis

A collective probabilistic factorization model, as shown in Fig. 2 (c), is proposed for spatial analysis. Let V be an $n \times m$ data matrix, whose element V_{ab} is the speed of road r_a in time slot t_b. Let W be an $n \times l$ data matrix, whose element W_{ac} is the c^{th} context of road r_a. The matrix V is factorized into two latent feature matrices R and T, where R is an $n \times k$ matrix, T is an $m \times k$ matrix, and k is the dimensionality of latent feature vectors in factorization. V_{ab}

[1] The 1s refer to its neighbors, and the 0s refer to others.

[2] $n(Q_{x,d_j}) = \{Q_{z,d_j} : z \in n(x)\}$, $n(x)$ is g_x's neighbors.

[3] $g(a)$ refers to the corresponding grid of r_a, and $d(t_i)$ refers to the corresponding period of t_i. $n(S_{a,t_i}) = \{S_{e,t_i} : e \in n(a)\}$.

[4] We utilize the number of consecutive occurrences in the floating car trajectories to justify how close two roads (or grids) are.

is modeled as a Gaussian distribution with the mean $R_a^T T_b$. Similarly, the matrix W is factorized into two matrices R and F, where R is the above $n \times k$ matrix, and F is an $l \times k$ matrix. In the factorization, the road latent feature matrix R is shared by both V and W. Finally, the elements of R, T and F are given the zero-mean Gaussian distributions as the empirical priors. Collectively modeling V and W has two advantages. First, contexts have been incorporated in imputations and predictions. In factorizing W, roads with similar contexts will generate similar latent vectors, propagated into the factorization of V. Second, since W is a dense matrix, and V is the target sparse matrix, collective factorization alleviates the data sparsity problem of factorizing the matrix V alone.

Under the above definitions, the parameters of the CMF are $\vartheta_{cmf} = \{\sigma_R^2, \sigma_T^2, \sigma_F^2, \sigma_V^2, \sigma_W^2\}$, and its variables are $\{R, T, F\}$. ϑ_{cmf} is set empirically. The inference process is to find $\{R, T, F\}$ that can maximize the likelihood, as

$$P(V, W, R, T, F; \vartheta_{cmf}) \propto \prod_{a=1}^{n} \prod_{b=1}^{m} \left[\mathcal{N}(V_{ab}; R_a^T T_b, \sigma_V^2)\right]^{I_{ab}^V} \cdot \prod_{a=1}^{n} \prod_{c=1}^{l} \mathcal{N}(W_{ac}; R_a^T F_c, \sigma_W^2) \cdot \prod_{a=1}^{n} \mathcal{N}(R_a; 0, \sigma_R^2) \cdot \prod_{b=1}^{m} \mathcal{N}(T_b; 0, \sigma_T^2) \cdot \prod_{c=1}^{l} \mathcal{N}(F_c; 0, \sigma_F^2)$$

When $\{R, T, F\}$ are found, the imputation of missing data and the prediction for future speeds are calculated[5] as $f_{ab}^{cmf} = R_a^T T_b$.

Multi-view Learning Framework

Multi-view learning is a typical semi-supervised learning paradigm. In our application, the labeled data is the road speeds sensed by floating cars, and the unlabeled data is the missing values. We provide two views above for collector road speed predictions, the temporal view and the spatial view. The assumption of multi-view learning in our task is, the missing value predictions from either view should agree with the other. In other words, if the models from the two views are trained simultaneously, it will be penalized when the predictions from the two views for the same missing value are dissimilar.

The joint likelihood to be maximized in our multi-view learning framework is defined as follows. The first two terms are the likelihoods of the proposed LHMM and CMF, respectively. The third term is proposed to model the probabilities of disagreements between the two models on the missing data[6]. It is defined by a zero-mean Gaussian, which penalizes the disagreements between two views.

$$\max_{S,Q,R,T,F} P(V, U, S, Q; \vartheta_{lhmm}) \cdot P(V, W, R, T, F; \vartheta_{cmf}) \cdot \prod_{a=1}^{n} \prod_{b=1}^{m} \left[\mathcal{N}(f_{ab}^{lhmm} - f_{ab}^{cmf}; 0, \sigma^2)\right]^{\bar{I}_{ab}^V} \quad (1)$$

Joint Inference Process The inference of the multi-view learning framework is to find $\{S, Q, R, T, F\}$ that can maximize the above joint objective function. An iterative process is conducted. In each iteration, we fix the variables of one model, and search the variables of the other.

When the variables of the LHMM $\{S, Q\}$ are fixed, the joint objective function can be converted to minimize

$$L(R, T, F) = \frac{1}{2\sigma^2} \sum_{a=1}^{n} \sum_{b=1}^{m} \bar{I}_{ab}^V (f_{ab}^{lhmm} - R_a^T T_b)^2 + \frac{1}{2\sigma_V^2} \sum_{a=1}^{n} \sum_{b=1}^{m} I_{ab}^V (V_{ab} - R_a^T T_b)^2 + \frac{1}{2\sigma_R^2} \|R\|_F^2 + \frac{1}{2\sigma_W^2} \sum_{a=1}^{n} \sum_{c=1}^{l} (W_{ac} - R_a^T F_c)^2 + \frac{1}{2\sigma_F^2} \|F\|_F^2 + \frac{1}{2\sigma_T^2} \|T\|_F^2.$$

A local minimum of the objective function can be found by performing gradient descent in $\{R, T, F\}$.

$$\nabla_{R_a} L = \frac{1}{\sigma^2} \sum_{b=1}^{m} \bar{I}_{ab}^V (f_{ab}^{lhmm} - R_a^T T_b) T_b + \frac{1}{\sigma_R^2} R_a + \frac{1}{\sigma_V^2} \sum_{b=1}^{m} I_{ab}^V (V_{ab} - R_a^T T_b) T_b + \frac{1}{\sigma_W^2} \sum_{c=1}^{l} (W_{ac} - R_a^T F_c) F_c$$

$$\nabla_{T_b} L = \frac{1}{\sigma^2} \sum_{a=1}^{n} \bar{I}_{ab}^V (f_{ab}^{lhmm} - R_a^T T_b) R_a + \frac{1}{\sigma_T^2} T_b + \frac{1}{\sigma_V^2} \sum_{a=1}^{n} I_{ab}^V (V_{ab} - R_a^T T_b) R_a$$

$$\nabla_{F_c} L = \frac{1}{\sigma_W^2} \sum_{a=1}^{n} (W_{ac} - R_a^T F_c) R_a + \frac{1}{\sigma_F^2} F_c \quad (2)$$

When the variables of the CMF $\{R, T, F\}$ are fixed, the objective function can be converted to maximize

$$L(S, Q) = P(V, U, S, Q; \vartheta_{hmm}) \cdot \prod_{a=1}^{n} \prod_{b=1}^{m} \left[\mathcal{N}(f_{ab}^{lhmm} - f_{ab}^{cmf}; 0, \sigma^2)\right]^{\bar{I}_{ab}^V}.$$

We first search Q, and then search S. Since grids are much easier to be covered by floating cars than collector roads, we have $I_{xy}^U = 1$ always. Thus searching Q is independent with the CMF variables. The Gibss sampling approach with simulated annealing is employed in searching the appropriate Q. In each iteration, for each state, we sample Q_{x,d_j} as

$$P(Q_{x,d_j} = \rho | Q - Q_{x,d_j}, U) \propto P(Q_{x,d_j} = \rho | n(Q_{x,d_{j-1}}); P_Q) P(U_{x,d_j} | Q_{x,d_j} = \rho; D_Q) \cdot \prod_{z, x \in n(z)} P(Q_{z,d_{j+1}} | Q_{x,d_j} = \rho, Q_{n(z),d_j} - Q_{x,d_j}; P_Q). \quad (3)$$

The searching of S depends on both LHMM variables and CMF variables. Similarly, the Gibss sampling approach with simulated annealing is employed in searching the appropriate S. In each iteration, for each state, we sample S_{a,t_i} as

$$P(S_{a,t_i} = \tau | S - S_{a,t_i}, V, Q) \propto P(V_{a,t_i} | S_{a,t_i} = \tau; D_S)^{I_{a,t_i}^V} \cdot \prod_{c, a \in n(c)} \mathcal{N}(f_{c,t_i}^{cmf} - (f_{c,t_i}^{lhmm} | S_{a,t_i} = \tau, S - S_{a,t_i}); 0, \sigma^2)^{\bar{I}_{a,t_i}^V} \cdot \prod_{c, a \in n(c)} P(S_{c,t_{i+1}} | S_{a,t_i} = \tau, S_{n(c),t_i} - S_{a,t_i}, Q_{g(a),d(t_i)}) \cdot P(S_{a,t_i} = \tau | n(S_{a,t_{i-1}}), Q_{g(a),d(t_i)}). \quad (4)$$

Parameter Estimation The parameters of the CMF model, $\{\sigma_R^2, \sigma_T^2, \sigma_F^2, \sigma_V^2, \sigma_W^2\}$, are configured empirically. In this work, we set $1/\sigma_R^2 = 1/\sigma_T^2 = 1/\sigma_F^2 = 0.1$, $1/\sigma_V^2 = 1.0$, and $1/\sigma_W^2 = 0.5$. Directly estimating the parameters of the LHMM model, $\{D_S, D_Q, \pi_S, \pi_Q, P_S, P_Q\}$, is intractable. In this paper, we approximately set the parameters of $\{D_S, D_Q, \pi_S, \pi_Q\}$ by heuristic rules, and estimate P_S, P_Q through the maximum likelihood estimation (MLE). In setting D_S, for each road r_a in the time slot t_b, four

[5] Although when $b = t_{i+1}$, there is no observations, through the multi-view learning framework, this problem will be fixed.

[6] $\bar{I}_{ab}^V = 1$ when V_{ab} is missing; or otherwise, $\bar{I}_{ab}^V = 0$.

Table 1: Statistics of the Beijing collector road network

	Level 1	Level 2	Level 3	Level 4
Num. of roads	503	978	1073	261
Total length	588	829	828	203
% covered/time slot	8.26%	6.89%	3.25%	2.01%

parameters need to be configured, $\mathcal{N}(v_{ab}^{S_{ab}=0}, (\sigma_{ab}^{S_{ab}=0})^2)$ and $\mathcal{N}(v_{ab}^{S_{ab}=1}, (\sigma_{ab}^{S_{ab}=1})^2)$. We rank all the historical road speeds of r_a in t_b in ascending order. The average of the first half is utilized to set $v_{ab}^{S_{ab}=0}$, and the variance is utilized to set $(\sigma_{ab}^{S_{ab}=0})^2$; and the average of the second half is utilized to set $v_{ab}^{S_{ab}=1}$, and the variance is utilized to set $(\sigma_{ab}^{S_{ab}=1})^2$. We utilize similar heuristics to set D_Q. The values of π_S, π_Q are simply set uniformly. The estimation for $\{P_S, P_Q\}$ is to find the best $\{P_S, P_Q\}$ to maximize the joint likelihood defined in Eq. 1 in the training data. An iterative process is conducted. A detailed algorithm is shown in Algorithm 1.

Algorithm 1 Parameter Estimation for LHMM

Input: The historical floating car data and the map data
Outputs: P_S, P_Q

1: Initialize S, Q, R, T, F, P_S, P_Q
2: **for** each iteration **do**
3: Update R, T, F by Eq. 2
4: Update S, Q by Eq. 3 and Eq. 4
5: Update P_S, P_Q by
$P(Q_{x,y+1} = \rho | n(Q_{xy})) = \frac{\#(Q_{x,y+1}=\rho | n(Q_{xy}))}{\#(n(Q_{xy}))}$
$P(S_{a,b+1} = \tau | n(S_{ab}), Q_{g(a),d(b)}) = \frac{\#(S_{a,b+1}=\tau | n(S_{ab}), Q_{g(a),d(b)})}{\#(n(S_{ab}), Q_{g(a),d(b)})}$
6: **end for**

Complexity Analysis In the inference step, the complexity for the gradient descent method in each iteration is $O(|V|^+K + NMK)$, where $|V|^+$ is the number of non-zero entries in V; and the complexity for the Gibss sampling method in each iteration is $O(NM)$. In the training step, the complexity of the iterative method in each iteration is $O((|V|^+K + NMK)H + NMH)$, where H is the number of training samples. Since our algorithm will converge after 5 to 10 iterations, this complexity analysis shows that the proposed framework is very efficient and can scale up with respect to very large data.

Experiments

Experimental Setup

Experimental verifications are conducted in the road network of Beijing, China. As shown in Fig. 1, all the main collector roads inside the 4^{th} ring are selected, who cover a $16km \times 16km$ spatial range, with a total length of $2,448$ km. There totally 2,815 roads, divided into four levels (L1-L4) according to the road width. The smaller the level is, the wider the road is. We use GPS trajectories generated by 8,126 taxis over a period of 2 years, with 2,036,792,043 GPS points and a length of 782,523,971km in total, sampled at a rate of 79 second/point in average. Table 1 shows the coverage of different road levels, after the GPS points are converted into the road-time matrix. The statistics can demonstrate the high percentage of missing data in our task. We utilize the data from Sep. 2013 to Nov. 2013 for evaluation, and the data from Nov. 2011 to Aug. 2013 for learning the parameters. In the evaluation, we extract the data from 7am to 23pm, and divide them into 8 groups, with each group containing 2 periods and 12 time slots. In each group, the first 11 time slots are utilized to infer the variables, and the last time slot is utilized for future speed predictions and evaluations. In the experiments, we set $\sigma = 20$ in Eq. (1), and the dimensionality of CMF $K = 20$. We utilize two metrics, the Mean Absolute Error (MAE), and the Root Mean Square Error (RMSE) for evaluations. The speeds are measured by km/h. A smaller MAE or RMSE value indicates a better performance. They are defined[7] as

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}, RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}.$$

We compare our framework with the following baselines.

1. AVG: The average speed of a road (RAVG) and the average speed of a road in the same time slot (RTAVG).
2. KNN: We employ the logistic regression from the top 5 nearest roads (including the active road), whose current speed is available, to predict the future speed.
3. Cascade: We first employ matrix factorization techniques to complete the missing historical data, and then predict the active road's speed by the neural networks model.
4. HMM: A simplification of the LHMM with the grid layer removed, similar to the method in (Herring et al. 2010).
5. LHMM: The proposed Layered HMM without factorization techniques from spatial analysis incorporated.
6. LHMM+MF: The multi-view learning framework of the LHMM and the factorization without road contexts.

The last two baselines are simplified versions of the propose framework. Our proposed model is denoted as LHMM+CMF. In LHMM+CMF, both LHMM and CMF have their own predictions. The predictions from the LHMM are finally selected in evaluation, as its performances are better.

Overall Performances

Table 2 shows the performances of different methods. The bottom bolded three methods are the ones proposed in this paper, and the percentages are the improvements from the best method. From the comparisons of the LHMM+CMF and the first five baselines, it is observed that the proposed multi-view framework outperform previous methods by more than 10% in both MAE and RMSE. The AVG methods perform the worst. Since in the Cascade method, errors in the matrix factorization are propagated to the neural

[7] $\hat{y}_i$ is the prediction, and y_i is the ground truth.

Table 2: Overall Performance of different methods

Methods	MAE (km/h)	RMSE	Time (s)
RAVG	5.695 (45.3%)	9.301 (50.0%)	< 1
RTAVG	4.167 (25.2%)	6.632 (29.8%)	< 1
KNN	3.605 (13.5%)	5.770 (19.3%)	< 1
HMM	3.498 (10.9%)	5.383 (13.5%)	10
Cascade	3.478 (10.4%)	5.253 (11.4%)	18
LHMM	3.364 (7.27%)	5.123 (9.15%)	11
LHMM+MF	3.148 (1.02%)	4.786 (2.76%)	17
LHMM+CMF	3.116	4.654	19

network, performances are also unsatisfactory. The comparisons between the LHMM and the LHMM+CMF demonstrate the effectiveness of the proposed multi-view learning framework. When the spatial view, modeled by the CMF, is integrated into the temporal view, modeled by the LHMM, the performance has been improved by 7.27% in MAE and 9.15% in RMSE, which demonstrates that the patterns agreement assumption is effective in improving performances. The improvement from the LHMM compared with the HMM verifies the effectiveness of the grid layer. Since the data is very sparse, modeling the average speed of a grid will obtain an accurate coarse-grained information, which helps to alleviate the sparsity problem, and further improves the prediction for fine-grained road speeds. The improvements from the LHMM+CMF compared with the LHMM+MF verify the effectiveness of road contexts, which also help to alleviate the sparsity problem. The third column in the table shows the total time of predictions for all road speeds in a time slot. By only using one server (with 8-core, 3.6GHz CPU and 32GB RAM), all the proposed approaches can make road predictions within 20 seconds.

The top two sub-figures in Fig. 3 demonstrate the overall prediction performances in different road levels. From the figures, it is observed that wider roads are more difficult to predict than narrower roads. This satisfies the common sense that narrow roads usually have less cars, thus the road speeds are more stable; while wide roads are capable for more cars, thus the variances are larger. The bottom two sub-figures in Fig. 3 demonstrate the overall prediction performances in different time of a day. It is observed that the speeds in the noon time and the evening time are easier to predict than other time slots. This also satisfies the common sense that in these time slots, people are working inside buildings or staying at homes, when less cars occur on roads. In rush hours, the variances become larger, making predictions harder.

Convergence and Impact of σ

Figure 4(a) shows the convergence speed of the proposed multi-view framework. It is observed that the algorithm converges in around 8 iterations. Thus the algorithm is efficient to find the local optimized point. Figure 4(b) shows the impact of the parameter σ in Eq. (1). When σ is large, the penalties for the disagreements from the two views will be weaker. Thus when σ is larger than $1K$, the performance is equivalent to methods based on single views. When σ is small, the penalties for the disagreements will be stronger.

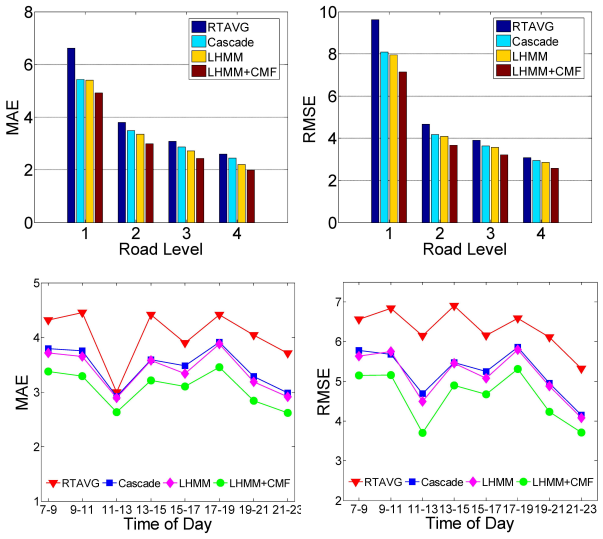
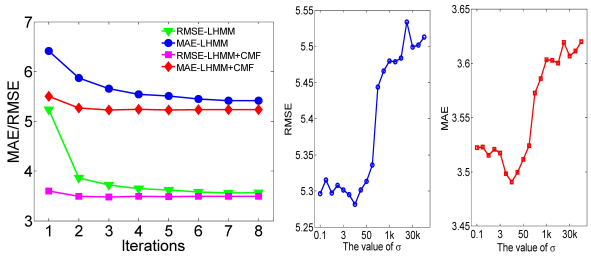


Figure 3: Performance on Different Road Levels and Time



(a) Convergence (b) Impact of σ

Figure 4: Convergence and impact of σ

When $\sigma = 20$, we obtain the best performance.

Conclusion

In this paper, we investigate a novel task, collector road speed predictions, which has been ignored due to the limitation of the available data. By proposing a multi-view framework to integrate the temporal and spatial patterns, we alleviate the data sparsity challenge in utilizing the floating car signals. Experiments in real data demonstrate the proposed algorithm can forecast collector road speeds in Beijing within 20 seconds, and obtain 10% improvements in MAE and RMSE, compared with traditional methods.

Acknowledgements

The work described in this paper was mainly supported by the National Basic Research Program of China (973 Program, Grant No. 2013CB329605), National Natural Science Foundation of China (No. 61300076), and Ph.D. Programs Foundation of Ministry of Education of China (No. 20131101120035).

References

Bejan, A. I., and Gibbens, R. J. 2011. Evaluation of velocity fields via sparse bus probe data in urban areas. In *Proc. of ITSC'11*, 746–753. IEEE.

Castro-Neto, M.; Jeong, Y.-S.; Jeong, M.-K.; and Han, L. D. 2009. Online-svr for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert systems with applications* 36(3):6164–6173.

Djuric, N.; Radosavljevic, V.; Coric, V.; and Vucetic, S. 2011. Travel speed forecasting by means of continuous conditional random fields. *Journal of the Transportation Research Board* 2263(1):131–139.

Hamilton, J. D. 1994. *Time series analysis*, volume 2. Princeton university press Princeton.

Haworth, J., and Cheng, T. 2012. Non-parametric regression for space–time forecasting under missing data. *Computers, Environment and Urban Systems* 36(6):538–550.

Herring, R.; Hofleitner, A.; Abbeel, P.; and Bayen, A. 2010. Estimating arterial traffic conditions using sparse probe data. In *Proc. of ITSC'10*, 929–936. IEEE.

Lint, V.; Hoogendoorn, J. W. C. S. P.; J., H.; and van Zuylen. 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C* 347–369.

Min, W., and Wynter, L. 2011. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies* 19(4):606–616.

Ni, D.; Leonard, J. D.; Guin, A.; and Feng, C. 2005. Multiple imputation scheme for overcoming the missing values and variability issues in its data. *Journal of transportation engineering* 131(12):931–938.

Park, D., and Rilett, L. R. 1998. Forecasting multiple-period freeway link travel times using modular neural networks. *Transportation Research Record: Journal of the Transportation Research Board* 1617(1):163–170.

Qi, Y., and Ishak, S. 2014. A hidden markov model for short term prediction of traffic conditions on freeways. *Transportation Research Part C: Emerging Technologies*.

Qu, L.; Li, L.; Zhang, Y.; and Hu, J. 2009. Ppca-based missing data imputation for traffic flow volume: a systematical approach. *IEEE Transactions on Intelligent Transportation Systems* 10(3):512–522.

Shang, J.; Zheng, Y.; Tong, W.; Chang, E.; and Yu, Y. 2014. Inferring gas consumption and pollution emission of vehicles throughout a city. In *Proc. of SIGKDD'14*, 1027–1036. ACM.

Smith, B. L.; Scherer, W. T.; and Conklin, J. H. 2003. Exploring imputation techniques for missing data in transportation management systems. *Transportation Research Record: Journal of the Transportation Research Board* 1836(1):132–142.

Smith, B. L.; Williams, B. M.; and Keith Oswald, R. 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies* 10(4):303–321.

Sridharan, K., and Kakade, S. M. 2008. An information theoretic framework for multi-view learning. In *Proc. of COLT'08*, 403–414. Citeseer.

Sun, S. 2013. A survey of multi-view machine learning. *Neural Computing and Applications* 23(7-8):2031–2038.

Tan, H.; Feng, G.; Feng, J.; Wang, W.; Zhang, Y.-J.; and Li, F. 2013. A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies* 28:15–27.

Thiagarajan, A.; Ravindranath, L.; LaCurts, K.; Madden, S.; Balakrishnan, H.; Toledo, S.; and Eriksson, J. 2009. Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proc. of ENSS'09*, 85–98. ACM.

Van Lint, J.; Hoogendoorn, S.; and van Zuylen, H. J. 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C: Emerging Technologies* 13(5):347–369.

Wang, W.-X.; Wang, B.-H.; Yin, C.-Y.; Xie, Y.-B.; and Zhou, T. 2006. Traffic dynamics based on local routing protocol on a scale-free network. *Physical Review E* 73(2):026111.

Wang, Y.; Zheng, Y.; and Xue, Y. 2014. Travel time estimation of a path using sparse trajectories. In *Proceeding of the 20th SIGKDD conference on Knowledge Discovery and Data Mining*.

Williams, B. M., and Hoel, L. A. 2003. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of transportation engineering* 129(6):664–672.

Yu, S.; Krishnapuram, B.; Rosales, R.; and Rao, R. B. 2011. Bayesian co-training. *The Journal of Machine Learning Research* 12:2649–2680.

Yuan, J.; Zheng, Y.; Xie, X.; and Sun, G. 2011. Driving with knowledge from the physical world. In *Proc. of SIGKDD'11*, 316–324. ACM.

Zhang, Y., and Liu, Y. 2009. Missing traffic flow data prediction using least squares support vector machines in urban arterial streets. In *Proc. of CIDM'09*, 76–83. IEEE.

Zheng, V. W.; Zheng, Y.; Xie, X.; and Yang, Q. 2010. Collaborative location and activity recommendations with gps history data. In *Proc. of WWW'10*, 1029–1038. ACM.

Zheng, Y.; Capra, L.; Wolfson, O.; and Yang, H. 2014. Urban computing: concepts, methodologies, and applications. *ACM Transaction on Intelligent Systems and Technology (ACM TIST)*.

Zheng, Y.; Liu, F.; and Hsieh, H.-P. 2013. U-air: When urban air quality inference meets big data. In *Proc. of SIGKDD'13*, 1436–1444. ACM.

Zhong, M.; Lingras, P.; and Sharma, S. 2004. Estimation of missing traffic counts using factor, genetic, neural, and regression techniques. *Transportation Research Part C: Emerging Technologies* 12(2):139–166.

Zhu, Y.; Li, Z.; Zhu, H.; Li, M.; and Zhang, Q. 2013. A compressive sensing approach to urban traffic estimation with probe vehicles. *IEEE Transactions on Mobile Computing* 12(11):2289–2302.