International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

Road Traffic Speed Prediction: A Probabilistic Model Fusing Multi-Source Data

SHOBANASHRI S [1], Mr. S.SENTHILKUMAR, [2], Mr.M.ANNAMALAI[3]

[1] Department of Computer Science and Engineering,
VMKV ENGINEERING COLLEGE,SALEM.
[2,3]Assistant Professor, Department of Computer Science and Engineering,
VMKV ENGINEERING COLLEGE,SALEM.

Abstract - *Road traffic speed prediction is a challenging problem in intelligent transportation system (ITS) and has gained increasing attentions. Existing works are mainly based on raw speed sensing data obtained from infrastructure sensors or probe vehicles, which, however, are limited by expensive cost of sensor deployment and maintenance. With sparse speed observations, traditional methods based only on speed sensing data are insufficient, especially when emergencies like traffic accidents occur. To address the issue, this paper aims to improve the road traffic speed prediction by fusing traditional speed sensing data with new-type "sensing" data from cross domain sources, such as tweet sensors from social media and trajectory sensors from map and traffic service platforms. Jointly modeling information from different datasets brings many challenges, including location uncertainty of low-resolution data, language ambiguity of traffic description in texts and heterogeneity of cross-domain data. In response to these challenges, we present a unified probabilistic framework, called Topic-Enhanced Gaussian Process Aggregation Model (TEGPAM), consisting of three components, i.e. location disaggregation model, traffic topic model and traffic speed Gaussian Process model, which integrate new-type data with traditional data. Experiments on real world data from two large cities in America validate the effectiveness and efficiency of our model*

1.INTRODUCTION

1.1 Background and Motivation

ROAD traffic monitoring is of great importance for urban transportation system. Traffic control agencies and drivers could benefit from timely and accurate road traffic prediction and make prompt, or even advance decisions possible for detecting and avoiding road congestions. Existing methods mainly focus on raw speed sensing data collected from cameras or road sensors, and suffer severe data scarcity issue because the installation and maintenance of sensors are very expensive [56]. At the same time, most existing techniques based only on past and current traffic conditions (e.g [9], [54], [25], [38]) do not fit well when real-world factors such as traffic accidents play a part.

To address the above issues, in this paper we introduce new-type traffic related data arising from public services: 1) Social media data, which is posted on social networking websites, e.g. Twitter and Facebook. With the popularization of mobile devices, people are more likely to exchange news and trifles in their life through social media services, where messages about traffic conditions, such as "Stuck in traffic on E 32nd St. Stay away!", are posted by drivers, passengers and pedestrians who can be viewed as sensors observing the ongoing traffic conditions near their physical locations. Meanwhile, traffic authorities register public accounts and post tweets to inform the public of the traffic status, such

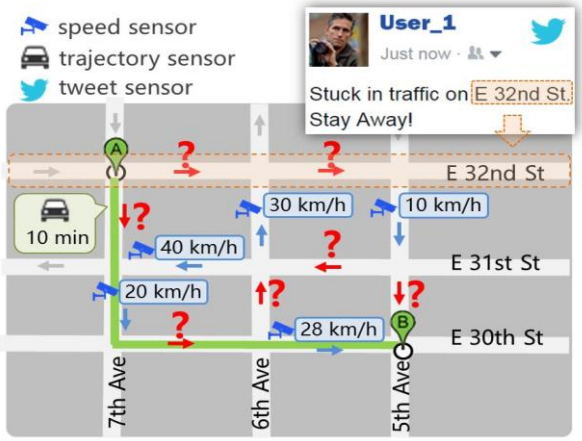


Fig. 1: Problem setting. Our goal is to predict the traffic speed of specific road links, as shown with the red question marks, given: 1) some speed observations collected by speed sensors, as shown in blue; 2) trajectory and travel time of OD pairs. Note that speeds of passed road links are either observed or to be predicted; 3) tweets describing traffic conditions. Note that the location mentioned by a tweet may be a street covering multiple road links. as "Slow traffic on I-95 SB from Girard Ave to Vine St." posted by local transportation bureau account. Such text messages describing traffic conditions and some of them tagged with location information are accessible by public and could be a complementary information source of raw speed sensing data.

(OD) pair on a map, such services can recommend optimal route from the origin to the destination with least time, and trajectories can be collected once drivers use the service to navigate. Here a trajectory is a sequence of links for a given OD pair, and a link is a road segment between neighboring

intersections. Correspondently, a trajectory travel time is an integration of link travel times, which are related to the real-time road traffic speeds. Longer trajectory travel time indicates that some involving road links may be congested with lower traffic speed. Trajectory data is useful for a wide range of transportation analyses and applications [49] [9].

Based on the above observations, where traditional traffic sensing data are limited while new-type data from social media and map service begin to spring up, our goal is to predict the road-level traffic speed by incorporating new-type data with traditional speed sensing data. To motivate this scenario, consider a road traffic prediction example depicted in Fig.1. Those links in red question marks are not covered by traditional speed sensors, but may be passed by trajectories attached with travel time information, or mentioned in tweets describing traffic conditions, so their speeds can be inferred fusing multiple cross-domain data.

1.2. Challenges

When integrating traditional traffic speed data (e.g. sensing data) with new-type data (e.g. Twitter data and trajectory data) to predict road traffic speed, technical challenges arise due to the characteristic of each data source:

location uncertainty of low-resolution data; tweet data and trajectory data are called low-resolution data because we cannot directly locate them into specific road links. Most tweets have no location tags, so geographic location language is the main clue, which however is vague. For example, expression like "Stuck in traffic on E 32nd St. Stay away!" covers the whole street without precise road locations. Meanwhile, travel time of a trajectory is an aggregate measure based on the speed of multiple links, which may vary widely. Thus a strategy is required to disaggregate the data to specific road links;

language ambiguity of traffic description in tweets; the expressions depicting traffic conditions are diverse, and may denote different speed values. An example is shown in Fig.2, which shows the frequency distribution over the degree of congestion when people use congestion-related words. Meanwhile some words not directly related to traffic may also have strong implication to link speed, such as words complaining bad weather. Thus a linguistic model is required to capture the patterns between discrete descriptive words and continuous speed values;

heterogeneity of multi-source data; the data sources have diverse properties and latent relations with the road traffic speed. For example, tweets possess latent topics which cluster based on speed levels, and negative correlation existed between trajectory travel time and traffic speed of involving links. Therefore a unified framework is re-quired to model these properties and aggregate the latent relations between heterogeneous data to predict speed synthetically.

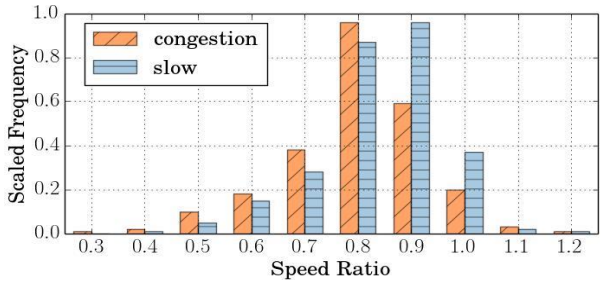


Fig. 2: The distribution of word frequencies when people use words "congestion" and "slow" to describe traffic, w.r.t the ratio between current speed and a reference speed, which is defined by INRIX as the "uncongested free flow speed" for each road segment. X-axis denotes the speed ratio and Y-axis denotes the frequency scaled w.r.t the biggest value.

1.3. Contributions

In spite of the good potential of these new-type data, to the best of our knowledge, the problem of road-level traffic speed prediction using multiple data sources has not been well explored before, especially with the aforementioned challenges. In this paper, we propose a unified statistical framework, entitled Topic Enhanced Gaussian Process Aggregation Model (TEGPAM) fusing multi-source data, which includes traditional speed sensing data, and new-type "sensing" data from social media and map services. The framework combines the location disaggregation model to decompose vague locations into specific links, the traffic topic model to handle the language ambiguity in tweets and the Gaussian Process model to capture the spatial correlation in traffic sensing data.

Specifically, this paper makes the following contributions:

- **Integration of data from multiple cross-domain sources**. We implement the idea of improving traffic speed pre-diction by integrating speed sensing data with new-type traffic-related data, such as tweet and trajectory.

- **Formulation of the unified TEGPAM framework**. We propose a unified probabilistic framework TEGPAM that combines the disaggregation model, topic model with Gaussian Process model and is learned by variational methods and a stochastic EM algorithm.

- **Extensive experiments to validate the performance of the proposed method**. We validate our approach us-ing real-world data collected from two large American cities. The extensive experiments show the effectiveness of TEGPAM, as well as the model efficiency and reliability.

- **Elaborate analyses of introduced traffic-related data**. We explore the impacts of different data sources, By decomposing TEGPAM into sub models and changing the combination ratio of datasets. Comparative experiments demonstrate the effectiveness of each data source.

The rest of this paper is organized as follows. Section 2 reviews related works. Section 3 gives a preliminary to Gaussian Process. Section 4 defines the problem and presents the model design. Section 5 gives model inference. Section 6 analyzes the results of experiments on real data. Section 7 concludes the paper and suggests future directions.

2. RELATED WORKS

Traffic prediction problem can be broadly classified into short-term and long-term prediction [1], considering three main basic traffic measurements: traffic flow, an equivalent flow rate in vehicles; speed, mean of the observed vehicle speeds; lane occupancy, the percentage of time that the sensor is detecting vehicle presence. This paper focuses on the short-term traffic speed prediction combining multi-source heterogeneous data, which, as far as we know, has not been well explored before. This part gives a summary on short-term traffic speed prediction and the exploration on fusing multiple information sources.

Short-term Traffic Speed Prediction: The presented methods can be classified into two categories:

1) **parametric methods**, assume that traffic speed follows a probability distribution based on a fixed set of param-eters. Time series analysis technique is applied in traffic speed prediction based on the periodicity of traffic speed during a day or a week. Auto-Regressive Moving Average (ARMA) models are adopted in [46] and [38], where Mul-tivariate Spatial-Temporal Auto-Regressive (MSTAR) model is adopted to include dependency among observations from neighboring locations. A review about Auto-Regressive Integrated Moving Average (ARIMA) time series methods can be found in [55]. ARIMA and Winters exponential smoothing techniques are used to forecast urban freeway flow in [54]. [53] separate ARIMA models for a set of loop detectors that incorporate information from upstream measurement locations. A single Space-Time Auto-Regressive Integrated Moving Average (STARIMA) model is proposed to describe the spatiotemporal evolution of traffic flow in an urban network in [26], which is essentially a constrained Vector Autoregressive Moving Average (VARIMA) model [13] with constraints that reflect the topology of a spatial network and result in a drastic reduction in the number of parameters. A Generalized Space-Time ARIMA (GSTARIMA) method is proposed in [57], which extends ARIMA in spatial and temporal dimension and is more flexible because parameters are designed to vary per spacial location. Kalman filter-based approaches are used in [11] and [14], and show advantages for on-line estimation of traffic flows. Markov logic network is used to simultaneously predict the congestion state in [30]. A structured time series model is proposed in multivariate form for short-term traffic prediction in [12].

2) **non-parametric methods**, make no distribution assumptions and the number of parameters scales with the number of training data. K-nearest neighbor (KNN) non-parametric regression methods, e.g. [9], [21], [58], find the k-nearest neighbors using Euclidean distance and calculate the weight. Neutral Networks (NNs), e.g. [50], [27], are biologically-inspired systems and can be trained to approximate virtually any nonlinear function given adequate data and a proper network architecture. NNs have many derivatives for short-term prediction, such as back propagation neutral network with genetic algorithms [1] and wavelet networks [22]. Travel speed of each road segment is computed using the GPS trajectories by a context-aware matrix factorization approach in [45]. To adaptively route a fleet of cooperative vehicles under the uncertain and dynamic road congestion conditions in [33] and [34], a GP probabilistic model is proposed to capture the spatial and temporal relationships of travel speeds over road segments and temporal contexts, especially with estimating the mean and covariance of the GP prior from the historical data. Geostatistical interpolation techniques named Kriging are proposed to capture spatial and temporal evolutions of traffic flows in [48].

Traffic Modeling with Multi-Source Heterogeneous Data:

Some researchers attempted to combine traffic sensing data with other data sources, to handle external factors such as traffic accidents (e.g. [36], [42]), mobile sensors (e.g. [39], [40]) and weather (e.g. [37], [2]). [37] reviews the literature on the impact of weather on traffic demand, traffic safety, and traffic flow relationships. A trajectory-based community discovery method is proposed in [32], where the trajectory similarity is modeled by several types of kernels for different information markers (e.g. semantic properties of the locations and the movement velocity). [29] tackles the rents/returns bike number prediction prob-lem using multiple features, e.g. time and meteorology, as measures of similarity functions in multi-similarity-based inference model. While [32] and [29] introduce different information sources as new features for computing the similarity, our work assumes the latent relations between these informations, and constructs a Bayesian generative process. As crowdsourcing data from a crowd of online social platform become more available, researchers begin utilizing social content to estimate traffic conditions. Twitter data are matched to detect traffic incidents in [36]. In [39], traffic anomaly detection uses crowd sensing with two forms of data, human mobility and social media, and the detected anomalies are described by mining representative terms from the social media that people posted when the anomaly happened. Few methods incorporate social me-dia text data (e.g. Twitter data) to improve traffic speed prediction. [31] extends spatiotemporal GP in [34] to three dimensional topic-aware GP, where topics on road links are probabilistic modeled based on the user, space and time of tweets. [15] do not tackle the location uncertainty problem of tweets, because the inference of traffic status based on words of tweets only focuses on the average regional traffic flow, which is insufficient for predicting road speed.

3. GAUSSIAN PROCESS PRELIMINARIES

Gaussian Processes (GPs) have been widely studied in many fields, such as spatio-temporal modeling [34] [35]. Given a set of road segments S under a specified time stamp, we spatially model the traffic speed of road segments via a function $f : S \rightarrow R_+$, which outputs the traffic speed for a given road link s.

Assume that f is sampled from a Gaussian process prior: $f(s) \sim GP(\mu(s); k(s; s'))$, which is fully specified by the mean function and the covariance, or kernel, function:

$$\mu(s) = E[f(s)]$$

$$k(s, s') = E[(f(s) - (s))(f(s') - \mu(s'))]$$

An important property of GP is that if two sets of variables property of GP is that if two sets of variables are jointly Gaussian, the conditional distribution of one set conditioned on the other is Gaussian, that is the basis to compute the posterior analytically [41].

Suppose that there are currently observed links S S with speed observations $V = \{v_s, s \in S\}$, where the traffic speed v_s for each links $s \in S$ follows $v_s \sim N(f(s),\sigma^2)$, where σ^2 is i.i.d. Gaussian noise. Then we can calculate the posterior distribution given the prior distribution with mean and kernel function, and the current observations V, which is still a GP distribution:

$$v_s \,|V, \mu, k \sim GP(\mu^{post}, k^{post}) \quad (1)$$

where

$$\mu^{post}(s) = \mu(s) + k(S,s)^T [K + \sigma^2 I]^{-1} (V-\mu) \quad (2)$$

$$k^{post}(s,s') = k(s,s') - k(S,s)^T[K + \sigma^2 I]^{-1} k(s,s') \quad (3)$$

where is the mean vector and K is the kernel Gram matrix, which are generated through historical speed records at observed links S:

$$\mu = [\mu(s)]_{s,s' \in S} \in R^{|S|}$$

$$K = [k(s,s')]_{s,s' \in S} \in R^{|S|x|S|}$$

Column vector $k(S, s)$ is the kernel values between $s \in S$ and every current observations in S:

Eq.(2) implies that the posterior mean $\mu^{post}(s)$ is determined by its prior mean $\mu(s)$ and the deviation between the historical observations and their prior means. If the positive covariance k(s, s') between road links s and s0 is high, the current observation of s' will have more impacts on post(s) with $(v_{s'} - \mu(s))$. Eq.(3) presents the property that the posterior covariance $k^{post}(s, s')$ between s and s' will decrease if we have more current observations related to s and s'. Meanwhile, the posterior $k^{post}(s, s')$ decreases faster with high k(s; s0) between s and s0.

Essentially, the kernel function k, generated from his-torical observations depicting the relation links, captures the spatial correlation of road network. If the covariance of two road links s and s' intuitively infer that they are close in the network structure.

4. MODEL DESIGN

This section begins by formalizing the speed prediction problem in Section 4.1. Then we introduce three models from Section 4.2 to 4.4 to tackle the challenges aforementioned in the introduction, i.e. a disaggregation model for location uncertainty in tweet and trajectory data, a traffic topic model for tweet language ambiguity and a GP model for capturing the spatial correlation of speed sensing data. Section 4.5 integrates three models dealing with different information source into a novel probabilistic model, named TEGPAM, under the Bayesian framework.

5.TEGPAM

Integrating the components introduced in the above subsections completes the design of the new probabilistic model, named the Topic-Enhanced Gaussian Process Aggregation Model (TEGPAM). Fig.3 gives the graphical representation of our model.

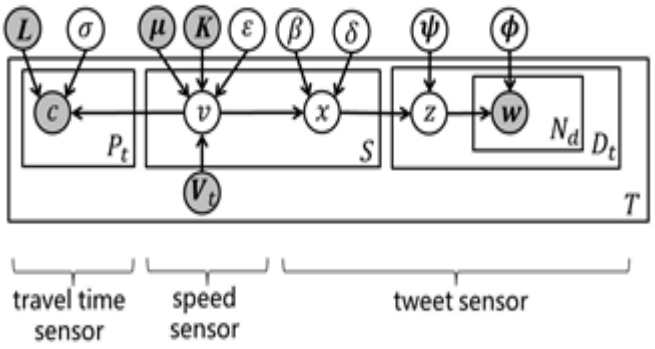


Fig. 3: Graphical model for TEGPAM, with three part dealing with the three aforementioned data sources.

6. EXPERIMENTS

Experiments on traffic speed prediction of two large Amer-ican cities are conducted to evaluate the following perfor-mance indicators: prediction accuracy, model efficiency, and model stability. This section is organized as follows: Section 6.1 introduces the experiment setting, including datasets, benchmark methods and predictive metrics. Section 6.2 val-idates our model of the overall performance regarding the prediction accuracy and efficiency. Section 6.3 provides an elaborate evaluation of the TEGPAM's effectiveness when applied to different data combinations. Section 6.4 discusses the model efficiency, and two factors of model stability: 1) sensitivity to parameters and 2) reliability on noisy tweets.

6.1 Experiment Setting

6.1.1 Datasets

We obtain three data sources for road traffic speed prediction: 1) **Traffic speed data**. INRIX database [20] provides

TABLE 2: Dataset size for each city.

City	Data	# total links	# trajectories	# tweets
Philadelphia		847,714	3,888,000,00 0	333,764
Washington D.C.		1,286,284	7,776,000,00 0	404,872

traffic speeds for each road link at a 5-minute rate, from June 1, 2013 to March 31, 2014, across two cities: Washington D.C. and Philadelphia. 2) Trajectory data. Trajectories are generated from INRIX database at a 5-minute rate. Given a random OD pair, we synthesize a trajectory by comput-ing the shortest path between them (i.e. using Johnson's algorithm [23]). With the length and speed information of links from INRIX, the travel time of this trajectory is obtained by adding the time of each link up and corrupting it with a Gaussian noise. 3) Twitter data. Tweets in the same time period and cities are collected via the Twitter REST search API. Traffic related tweets are preliminarily extracted by matching at least one term of a predefined vocabulary developed by domain experts, which included terms like "traffic", "accident", "stuck", "crash", etc, then further classified and filtered using an SVM classifier that was trained based on manually labeled 10,000 tweets (50% positive and 50% negative tweets). With road records containing the geo-coordinates, names and aliases, we geocode tweets to road links by matching their geo-tag and text content to the front end of those links, which corresponds to the driving out direction and is denoted as Head. Different driving directions are denoted as different road links. After geocoding, there are 5 major roads with 35 road links mentioned in the Philadelphia twitter data, and 8 major roads with 44 links in Washington D.C. respectively. Details of each data source are show in TABLE 2.

6.2 Benchmark Methods

To validate the performance of our approach fusing multiple data sources, particularly, to explore the impact of each data source, this subsection designs several comparative methods, which are based on the decomposition of our approach TEGPAM.

To make the expression clear, label data sources ftraffic speed, trajectory, twitterg as f1; 2; 3g. Denote M-i as the model excluding the data sources i f1; 2; 3g. Then we design sub models in terms of different data combinations:

TEGPAM: our full model introduced in section 5 and 6, using traffic speed, travel time and twitter data. The model is learned by variational inference.
M-1: trajectory and twitter based model, without incorporating traffic speed sensing data.

M-2: traffic speed and twitter based model, without handling trajectory data.

M-3: traffic speed and trajectory based model, without handling twitter data.

M-13: trajectory based model.

M-23: traffic speed based model, which is essentially a simplification of Gaussian Process Dynamic Congestion Model (GPDCM) in [33].

We infer the parameters of those models based on the same distribution assumptions, and we train parameters under the same settings.

Baseline methods: We also compare our approach with three baseline methods: K-nearest neighbor model (KNN) [58], GSTARIMA [57] and the tweet semantic based method(TwiSemantic) [15]. KNN and GSTARIMA are based on recent speed observations with considering the road network topology. TwiSemantic combines recent traffic speed with tweets semantics using linear regression.

In KNN, we use non-weighted algorithm and the neigh-bor number is 5 with the best result here. In GSTARIMA, we set the spatial weighted matrix following the paper. In TwiSemantic, tweet semantics are mapped into the same vocabulary as our model used, which contains 1857 words and is obtained by removing stop words and words with frequencies lower than 10 from traffic related tweets. Our model is initialized by pre-analyzing a small fraction of data, with = 1; equal to the opposite value of the speed median in the fraction, $_{i;j} = _K^1$; $_{j;k} = _M^1$; and the topic number K is 2, denoting congested and normal. The dataset is divided into training and testing data by time stamps. In the training stage, the speed variables $v_{t;s}$ are observed to learn the model parameters; in the testing stage, the speeds are latent, the posterior distributions of which are inferred with fitted model parameters.

6.3 Overall Comparison

To show the improvement of fusing more data, we compare the sub-models M-1, M-2, M-3 fusing two data sources and M-23, M-13 based on one data source in the baseline methods, note that M-12 is not added to this set of experiments because of the insufficiency of only using twitter data. The percentage of speed sensor and trajectory are all set as 50%, and the fraction of testing data ranges from 1=6 to 5=6.

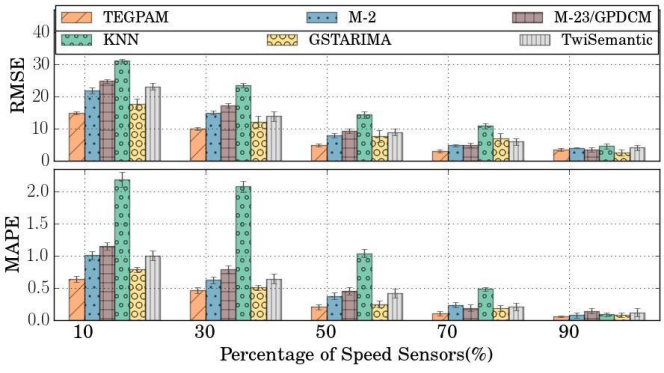
The result is shown in Fig. 8. We observe that TEGPAM using 3 data sources performs steadily the best, while M-1, M-2

and M-3 fusing 2 data sources take second place and M-23 and M-13 using only one data source performs the worst, which validates the intuition that speed prediction combining more information can improve the accuracy. Meanwhile, comparing the error of 2 data based models, M-1 is the worst, so excluding speed sensing data impacts the prediction most, which implies that speed sensing data might be more affective than trajectory data while trajectory is better than twitter data. Observing M-23 better than M-13 also proves the indication.

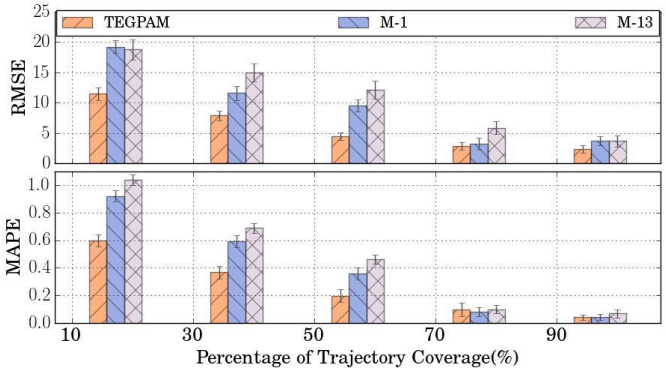
6.4 Effectiveness of Traffic Speed Data

We firstly compare the performance of models with or without using speed sensors, to demonstrate the effectiveness of speed data. Fix the percentage of speed sensor p_v and the percentage of trajectory coverage p_p as 50%, then test integrated TEGPAM, speed based M-23(GPDCM) and speed excluded M-1 with fraction of testing data as f1=6; : : : ; 5=6g. The results are shown in Fig. 9(a). The performance of TEGPAM is steadily better than the others, while only using speed sensors (M-23) is insufficient and limited, which again demonstrates the benefit of multi-source data.

To answer the questions: with other data sources, how sparse the historical traffic speed data can be to predict current traffic speed? We set the percentage of speed sensors as p_v = 10; 30; : : : ; 90%, under the fraction of testing data and the percentage of path coverage p_p as 50%. The traffic speed based models, TEGPAM, M-2 and M-23 (GPDCM) of our approach, and KNN, GSTARIMA, are applied on the training set. The results are shown in Fig.10(a). The score decrease trend of each model shows that with more current or recent observations, the missing speeds will be better predicted. However, when fewer than 70% speed sensors, TEGPAM fusing multi-source data performs better than the traffic speed based model (M-23/GPDCM, KNN and GSTARIMA), especially, the RMSE of TEGPAM is nearly 40%; 50% and 15% less than that of M-23/GPDCM, KNN and GSTARIMA when only 10% links are observed. The results show the impact of traffic speed data and prove the effectiveness of TEGPAM when speed sensors are largely unavailable.



(a) Speed and twitter based models.



(b) Trajectory based methods

Fig. 10: Comparisons with changing the percentage of data.

6.5 Effectiveness of Trajectory Data

Trajectory data, with time and link information, has a direct relationship with road speeds. In a trajectory with available travel time, more information about the speed of those road links in this path is contained. If the travel time is big, we will be more confident to infer that some road links in the trajectory are congested and the speeds of them must be low. This section validate the effectiveness of trajectory data in predicting unobserved traffic speeds.

We firstly validate the effectiveness of trajectory data by comparing models with or without using trajectory information. Three models are applied on the fraction of testing data as f1=6; : : : ; 5=6g: integrate TEGPAM, trajectory based M-13 and trajectory excluded M-2. The percentages of speed p_v and trajectory coverage p_p remain 50%. The RMSE and MAPE scores are shown in Fig. 9(b). The performance of M-2 is better than M-13, which implies that trajectory data alone is also not good enough to predict traffic speeds. Meanwhile, comparing M-2 here with M-1, and M-13 with M-23 in Fig.9(a), we observe that speed based model M-23 has a slight advantage to trajectory based model M-13, which validates the effectiveness of speed data in some degree.

6.6 Effectiveness of Twitter Data

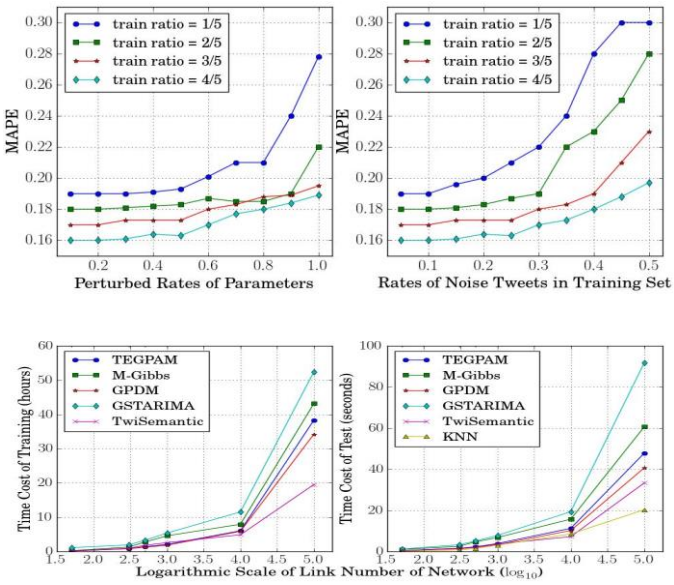
The traffic related information of twitter data is very dynamic, this subsection is designed to answer the question: what role does twitter data play in predicting current traffic speeds, a strong predictor or a good supplement to other data sources? To answer the question, we apply twitter based models, TEGPAM, M-1, M-2 of our approach and TwiSemantic on the settings of 50% speed sensor percentage and trajectory coverage. From Fig. 9(c), we observe that the integrate model TEGPAM performs steadily good, and with the same data source of speed and twitter, M-2 gains less error than TwiSemantic, which demonstrate the effectiveness of our model, especially the Traffic Topic model. Meanwhile, comparing model with and without using twitter, e.g. M-1 and M-13 in Fig. 10(b), M-2 and M-23 in

Fig.10(a), we observe that when the percentage is less than 50%, the models including twitter data (M-1 and M-2) perform better than those excluding twitter (M-13 and M-23). The results indicate that when observed speed percentage is low, Twit-ter data is a strong complement to speed sensing data.

6.7 Model Efficiency and Stability

Model efficiency is shown in Section 6.4.1. Then Section 6.4.2 and 6.4.3 validates two factors of model stability: 1)sensitivity to parameters and 2) reliability on noisy tweets.

Topic model and Traffic Speed Gaussian Process Model. Experiments on real data demonstrate the effectiveness and efficiency of our model. For Future work, we plan to implement kernel-based and distributive GP, so the traffic prediction framework can be applied into a real-time large traffic network.



7. CONCLUSIONS

This paper proposes a novel probabilistic framework to pre-dict road traffic speed with multiple cross-domain data. Ex-isting works are mainly based on speed sensing data, which suffers data sparsity and low coverage. In our work, we handle the challenges arising from fusing multi-source data, including location uncertainty, language ambiguity and data heterogeneity, using Location Disaggregation Model, Traffic.

ACKNOWLEDGEMENT

This work is supported by China 973 Fundamental R&D Program (No.2014CB340300), NSFC program (No.61472022, 61421003), SKLSDE-2016ZX-11, and Beijing Advanced Inno-vation Center for Big Data and Brain Computing.

REFERENCES

[1] B. Abdulhai, H. Porwal, and W. Recker. Short-term traffic flow prediction using neuro-genetic algorithms. ITS Journal-Intelligent Transportation Systems Journal, 7(1):3–41, 2002.

[2] R. Alfelor, H. S. Mahmassani, and J. Dong. Incorporating weather impacts in traffic estimation and prediction systems. Technical report, US Department of Transportation, 2009.

[3] M. T. Asif, N. Mitrovic, L. Garg, and J. Dauwels. Low-dimensional models for missing data imputation in road networks. 32(3):3527– 3531, 2013.

[4] C. M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., 2006. D. M. Blei and J. D. Lafferty. Correlated topic models. In NIPS, pages 147–154, 2005.

[5] Y. Gal, M. V. D. Wilk, and C. E. Rasmussen. Distributed variational inference in sparse gaussian process regression and latent variable models. Advances in Neural Information Processing Systems, 4:3257– 3265, 2014.

[6] D. B. Johnson. Efficient algorithms for shortest paths in sparse networks. Journal of the Acm, 24(1):1–13, 1977.

[7] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. Introduction to variational methods for graphical models. Machine Learning, 37:183–233, 1999.