A Network Analysis of Road Traffic with Vehicle Tracking Data

Wen Dong and **Alex Pentland**
E15-383, 20 Ames Street
The MIT Media Laboratory
Cambridge, MA, 02142-4307
{wdong,sandy}@media.mit.edu

Abstract

The high resolution tracking data for hundreds to thousands of urban vehicles, as well as the availability of digitized map data, provide urban planners unprecedented opportunities for better understanding urban motor vehicle transportation and for better exploiting the knowledge thereof. This paper combines the domain knowledge of traffic engineering with machine learning techniques, and gives a new approach to the problem of traffic speed estimation and travel time prediction.

Introduction

Human group activities are diverse. They normally involve large terabytes data sets. The mechanism explaining (the data of) a specific type of human group behavior may be very different from the mechanism explaining another type. So do the purposes of our investigations. On the one hand, we can often find out some good patterns in the data sets related to human group activities and make good use of them without understanding the underlying mechanisms. On the other hand, a good understanding of the mechanisms may enable us to get better results and to estimate the hidden variables. In other words, model-driven parametric approaches are more natural to encode the domain knowledge and to "regularize" the target functions than data-driven non-parametric approaches. Both types of approaches use statistical learning methods such as support vector methods and Bayesian networks to model human group activities.

Road traffic is an important type of human behavior. This is reflected by the large number of publications and people's everyday concerns about the road traffic. On the other hand, there does not exist a published study of nation-wide and year-long road traffic based on fine-precision tracking records (consisting of longitudes, latitudes and timestamps) for a large number of moving vehicles according to our knowledge. Studying road traffic with terabyte vehicle tracking data could naturally be benefited by distributed sensor network technologies, machine learning methods, and the physics of road traffic.

This paper gives a case study of applying the statistical learning methods and the traffic theory to the problem of understanding the human behavior related to the road traffic. Our paper is the first that we know of to estimate and predict city-wide traffic speeds from vehicle tracking data.

The tracking data and the map data involved for this case study are kindly provided by NavSatCR© for our research. The tracking data contains the geographical positions, the time stamps, the speeds, and the headings of around 200 delivery vehicles operating over Costa Rica from the September of 2007 to the March of 2008 inclusive. The tracking is on a 10 second basis recorded by vehicle-mounted hardware when a vehicle is in operation (normally from 7:00 to 23:00). The heading of a vehicle ranges clockwise from 0° when the vehicle goes northward, to 90° when the vehicle goes eastward, and to 360° when the vehicle goes northward again. The speed of a vehicle is in km/h. It normally ranges from 5 km/h to 40 km/h on local roads, from 30 km/h to 60 km/h on arterial roads, and from 40 km/h to 100 km/h on major highways. The map data is used for GPS navigation in Costa Rica and is converted into separate ESRI Shape files and a dBase file for our data analysis. The generated shape files correspond respectively to (1) the POIs (points of interest, locations with names and attributes, such as restaurants, hotels, schools, and parking lots), (2) the unnamed points, (3) the lines defining the roads, (4) the polygons defining the lakes, parks, forests, etc, (5) the marine POIs, (6) the marine points, (7) the marine lines, and (8) the marine polygons. The generated dBase file contains the routing information, which describes how the road links are connected to one another. The map data contains 22,747 POIs and 35,369 roads. The roads form a network with 61,560 vertices (road intersections) and 78,507 edges (road links).

In the following sections, we will review the literature related to the traffic theories and related to traffic speed prediction. We will then give our Viterbi decoding algorithm to map the *(longitude, latitude)* sequences in the tracking data into the corresponding *(road segment, offset)* sequences. This "map matching" has long been identified (Smith *et al.* 2003) as an import problem to be solved for a study like ours. We will proceed to discuss the traffic patterns at the road link level in the tracking data set. Motivated by the traffic patterns in the data set and equipped with the knowledge of road traffic physics, we will give our algorithms for estimating/predicting traffic speed and predicting travel time. We will conclude with our general opinion on understanding human or human group behavior.

Road Traffic Theories

The physics of road traffic has a long research history and a rich collection of bibliographies. The fundamental diagram and the conservation equation of traffic flow are of our special interest, since they establish a relationship among the traffic speeds on different road links in terms of the relationship of the corresponding traffic flow rates. While traffic speeds can be much more reliably and easily estimated from individual vehicle speeds, the traffic in different parts of a road network is more directly related to each other via traffic flow rates. The implicit traffic speed relationship gives us a tool for the prediction of traffic speeds, the estimation/prediction of traffic flow rates, and the assessment of the feasibility of data-driven non-parametric approaches. The following two paragraphs give more details on the fundamental diagram and the conservation equation respectively.

The variables of the traffic on a road segment are the traffic **speed** v (km/h), **density** k (vehicles/kilometer), and **flow rate** q (vehicles/hour). Since $q = k \times v$, the knowledge of any two variables gives the third one. The **fundamental diagram of traffic flow** (Greenshields 1935) postulates that there is a (statistically) linear relationship between the traffic speed and the traffic density (with a negative slope) in a one-lane traffic flow, i.e., $k = k\text{max} - (k\text{max}/v\text{max}) \cdot v$ where $v \in [0, v\text{max}]$ and $k \in [0, k\text{max}]$. It can be extended to model the equilibrium state of multi-lane traffic flow. The fundamental diagram can be explained by a microscopic **car following model** that describes how an individual vehicle responds to the behavior of the vehicle(s) in front of it in order to keep a safe distance (Gazis, Herman, & Rthery 1961): $\ddot{x}_{n+1}(t+T) = (v\text{max} \cdot (\dot{x}_n(t) - \dot{x}_{n+1}(t))) \,/\, \left(k\text{max} \cdot (x_n(t) - x_{n+1}(t))^2\right)$, i.e., the acceleration of a vehicle is proportional to its speed difference with the leading vehicle and to the inverse square of its distance to the leading vehicle.

The traffic on different segments of a road network is related with each other by the **conservation of road traffic**: In an area and a time span, the net traffic volume flowing into this area plus the net traffic volume generated equals the increase of the traffic volume, $\int_{t_0}^{t_1} dt \left(\sum_{i \in \text{in ramps}} q(t, x_i) - \sum_{j \in \text{out ramps}} q(t, x_j) \right) = \int_V dx\, (k(t_1, x) - k(t_0, x) dx) + \int_{t_0}^{t_1} dt \int_V dx g(x, t)$, where *g(x,t)* is the number of new vehicles generated at time *t* and location *x*. In differential form, $\frac{\partial}{\partial x} q(x,t) - \frac{\partial}{\partial t} k(x,t) = g(x,t)$. Since our observables are the traffic speeds sampled by transport vehicles, we can rewrite the conservation equation in terms of traffic speed $\frac{dq}{dv}\frac{\partial v}{\partial x} - \frac{dk}{dv}\frac{\partial v}{\partial t} = g(x,t)$ and proceed to estimate the road parameters (e.g., $k\text{max}$ and $v\text{max}$ for fundamental diagram approximation), as well as the traffic variables (*q*, *v*, and *k*) at different times and locations in the road network. The earliest applications of the conservation equation to the study of traffic flow and more recent applications in traffic flow simulation can be found in (Lighthill & Whitham 1955; Richards 1956; Stephanopoulos & Michalopoulos 1979; Michalopoulos 1988).

More works on the road traffic theories are listed below. Gartner et al. (Gartner, Messer, & Rathi 2005) gave an up-to-date overview of traffic theories including the car-following models, traffic flow theory and traffic flow models at different types of road intersections. Highway Capacity Manual (hig 2000) contains concepts, guidelines, and computational procedures for computing the capacity and quality of service of various highway facilities. Kerner (Kerner 2004) proposed a three-phase theory of the road traffic and gave an interesting explanation of the "wide moving jam" phase. The traffic speed theories assign meanings to our terabytes vehicle tracking data. They also enable us to solve our traffic-related problems in model-driven parametric approaches and to understand the performance of data-driven non-parametric approaches.

Literature Review

We summarize below the previous work related to road traffic and having any of the keywords "GPS", "cellphone", "speed", "jam", "estimation" and "prediction".

Prashanth Mohan et al. (Mohan, Padmanabhan, & Ramjee 2008) reported their results on using the accelerometer, microphone, GSM radio and GPS in a smart phone to detect potholes, bumps, braking and honking. Their experiments are based on the data collected by several smart phones in several road trips for a total number of 27.5 hours travel time and 512 kilometers travel distance.

Smith at el. (Smith *et al.* 2003) gave a critical assessment of past studies of WLT (wireless location technology) based traffic monitoring and documented the evaluation of one of the most recent operational tests — the 2001 Virginia Department of Transportation (VDOT), Maryland State Highway Administration (MSHA), and US Wireless Corporation (USWC) effort in the Washington, D.C. region. Based on the results of the operational tests, they concluded that the early generation WLT-based system produced link speed estimates of moderate quality, but it showed some promise in future traffic monitoring applications. For the data sets used by Smith at el., the geographical precision is around 100 meters, the sample rate is 1 sample per several minutes, and the samples are collected by vehicles traveling along a specific road link under investigation.

Huisken (Huisken 2006) benchmarked the performance of different traffic jam prediction methods (using linear regression, ARMA, MLF/RBF/Elman/SOM neural networks respectively) using the traffic data (traffic speed and traffic flow rate) collected by inductive loops buried under the roads at two test sites. Each of the test sites is composed of two 10-kilometer-long highways connected with each other by a cloverleaf road intersection.

Lint (van Lint 2004) proposed to use (data-driven) neural network models to predict the travel time on a road link from the traffic information collected by inductive loops. He used simulated traffic information to benchmark his method.

Froehlich and Krumm (Froehlich & Krumm 2008) inspected the driving data (latitudes, longitudes, and timestamps) of 252 drivers in the *Seattle, WA* area — most of whom are Microsoft employees and affiliates — for an average duration of 15.1 days, with a sample rate of 6 seconds per sample when the vehicles are in operation, and collected in year 2005. They noted that 39.3% of the trips are repeated trips, and 40% of the repeated trips can be identified halfway.

Horvitz et al. (Horvitz *et al.* 2005) used a Bayesian network to predict traffic jams on the major highways in the *Seattle, WA* area. The prediction is based on the past and current traffic jams, accident reports, time of day, day of week, holiday and special event information, and weather information. The Bayesian network is trained from 15 months of (1) traffic jam information reported by WDOT (the Washington Department of Transportation) and (2) the aforementioned factors that are used to predict traffic jams. From what we know, the local drivers generally know the daily/weekly patterns, which account for about 80% of traffic speed variance, and they are more concerned about the irregular traffic conditions. The traffic jam and the traffic speed are both complex and hard-to-define phenomena (Kerner 2004). In this sense, the approach of Horvitz et al. is one by the machine learning researchers rather than one by the traffic engineers.

Data Preprocessing and Inspection

In this section, we will describe our inspection of the data set, as well as some patterns in it. A good statistical modeling of the city-wide road traffic should either exploit the patterns or implicitly reflect the patterns.

Inspection of the Data Quality

According to our inspections, most of the *(longitude, latitude)* pairs in the tracking data have sub-meter accuracy. Instantaneous vehicle speed can be estimated from consecutive *(longitude, latitude, time)* 3-tuples, since .00001 degree in both latitude and longitude corresponds to 1.1 meters in Costa Rica and since a vehicle moves straight statistically in a 10 second interval. The instantaneous vehicle speed estimated with the just described method differs from those by the vehicle mounted hardware by a RMS (root mean square) of 0.25 km/h. The vehicle heading data estimated by the vehicle mounted hardware agrees with the direction of the road segment that the vehicle is on by no more than 3° deviation for over 95% of the tracking records.

Mapping Geographical Coordinates to Roads

Since vehicles operate in roads, the road network provides both a structure for understanding the relationships of vehicles in different road segments and a way for describing the traffic information. Recall that the raw tracking data has only geographical coordinates and corresponding timestamps. Without mapping the geographical coordinates in the tracking data to roads, we may only be able to tell the drivers to avoid the horizontal strip that goes through the center line of *San Jose, San Jose, Costa Rica* and that is $1/4$ mile from north to south, With the map information considered, we might tell the drivers only to avoid the segment of Paseo Colon that connects the Pan-American Highway and that goes to the airport.

We subsequently mapped the *(longitude, latitude)* sequences in the tracking data set to the corresponding *(road segment, offset)* sequences with the Viterbi algorithm for hidden Markov models. One reason for this preprocessing step is that many interesting phenomena of road traffic happen at the road level. Another reason is that the tracking records at the *(longitude, latitude)* level are also very sparse, even after they are aggregated into 10 meters by 10 meters cells. Since two consecutive samples in the data set are around 10 seconds apart in time, there exists a strong relationship between the road(s) or road link(s) corresponding to the two samples. As a result, we can expect a better mapping by making sense of a (consecutive) sequence of vehicle positions than by simply choosing the road segments that have the shortest distances to the individual vehicle positions. The hidden Markov model that describes the behavior of an operating vehicle is constructed in the following way. The latent state S_t of a vehicle (at sample time *t*) is the road link that the vehicle is on at time *t*. Corresponding to latent state S_t is the observed *(longitude, latitude)* pair of the vehicle at time *t*. The probability of observing *(longitude, latitude)* when the vehicle is in state S_t is $\mathbb{P}((\text{longitude, latitude})|S_t)$. The state transition $S_t \to S_{t+1}$ of the vehicle from time *t* to time $t+1$ is either that the vehicle keeps on the same road link ($S_t = S_{t+1}$) or that the vehicle passed through several road intersections and arrives at different road link ($S_t \neq S_{t+1}$). The conditional probability $\mathbb{P}(S_{t+1}|S_t)$ defines how likely that a vehicle on road link S_t at time *t* is on road link S_{t+1} at time $t+1$, and it is 0 when the transition from S_t to S_{t+1} is unlikely. Given a sequence of observations $\{(\text{latitude}_i, \text{longitude}_i)\}_{i=1}^N$, the Viterbi path is then the best estimation of the corresponding road link sequence that maximizes the likelihood function $\prod_{i=2}^N \mathbb{P}(S_i|S_{i-1}) \cdot \mathbb{P}((\text{latitude}_i, \text{longitude}_i)|S_i)$.

We inspected a randomly selected set of 50 traces involving different vehicles operating on different days. By our inspection, 100% of the corresponding *(road segment, offset)* sequences make sense. The error rate of the Viterbi algorithm should be close to our empirical estimation.

Histogram of Road Usages

The road usage in the tracking data obeys the 80-20 rule (power law). In other words, the vehicles are on a small number of road segments most of the time, and they are widely dispersed throughout a large number of road segments for the rest of the time. This road usage pattern coincides with the pattern noticed by Lammer (Lammer, Gehlsen, & Helbing 2006): In the data set, there is one major highway — Pan-American Highway (Route 1, 1a, 1b, 2, 2a) — which is 657 km in length, and has 1,959,478 tracking records on it. There are another 34 numbered highways, with total length 1187 km, and corresponding to 1,863,884 tracking records. There are 391 named/numbered arterial roads, with total length 4739 km, and corresponding to 2,530,482 tracking records. The collector roads and residential streets are unnamed. Their total length is 33002 km and they correspond to 2,609,084 tracking records.

The 80-20 rule of the road usage has some indications on the road traffic speed estimation and travel time prediction. A vast amount of local roads are infrequently used, they are less likely to have large variations in traffic conditions including traffic jams, and they are sparsely sampled by the delivery vehicles. Their traffic conditions are concerned about by a few the drivers. A few roads, such as some highways and arterial roads, are densely used, they are more likely to have large variation in traffic conditions, and they are densely sample by the delivery vehicles. Their traffic conditions are concerned by many of the drivers.

Periodic Patterns of Road Traffic

Different road segments and different directions along the same road segment have different temporal patterns on traffic speed. These patterns on a road segment can be understood and modeled in terms of how the multitude of vehicles pass through it in both directions: Where are they from, where do they go, and when do the trips happen. Since the road segments are connected into a network, the patterns are related to each other through the network connection. On a major highway that connects one administrative district to another district, the traffic speed normally drops significantly to an identical level during peak hours (around 9:00 am ˜ 10:00 am and/or around 4:00 pm ˜ 5:00 pm during weekdays). The peak hours are relatively longer on Monday mornings and Friday afternoons. The road traffic speed is likely to remain high for the rest time of a week. This pattern is induced by people commuting between where they live and where they work. Since this behavior is stable in the tracking data set, we can believe that commuters accounts for a large fraction of travelers along the major highway. Many other highways, arterial roads, and collector roads close to the major highways in the road network show similar patterns. In contrast, the traffic speed on a local street is more determined by factors such as the POIs on and around the street, and whether the street connects a major highway. On many local streets, speed slow-downs during evenings and weekends can be observed from the tracking data. Many of the local streets are very infrequently sampled by the transportation vehicles. The weekly traffic speed pattern of a section of a major highway and the periodogram of the same section are plotted in Figure 1.

Road traffic speed is a tricky concept. It has many incompatible definitions and is widely argued in traffic theory works. Concerning our vehicle tracking data set, we point out that the root mean square of the speed difference of any two tracking records about a vehicle around 10 seconds apart is 9.35 km/h. On the other hand, we also noted that the road traffic speed is very predictable given the instantaneous speeds of a moving vehicle. Thus traffic speed is a reasonable concept and can be estimated from vehicle speeds, but traffic speed is by no means equivalent to vehicle speed. This observation from the tracking data coincides with the common sense about road traffic speed. A vehicle driver normally observes that other vehicles around his vehicle travel in the same speed on the road. He normally does not change lanes frequently on a road in order to be faster, because he believes (at least partially) that passing over other vehicles does not make him much faster.

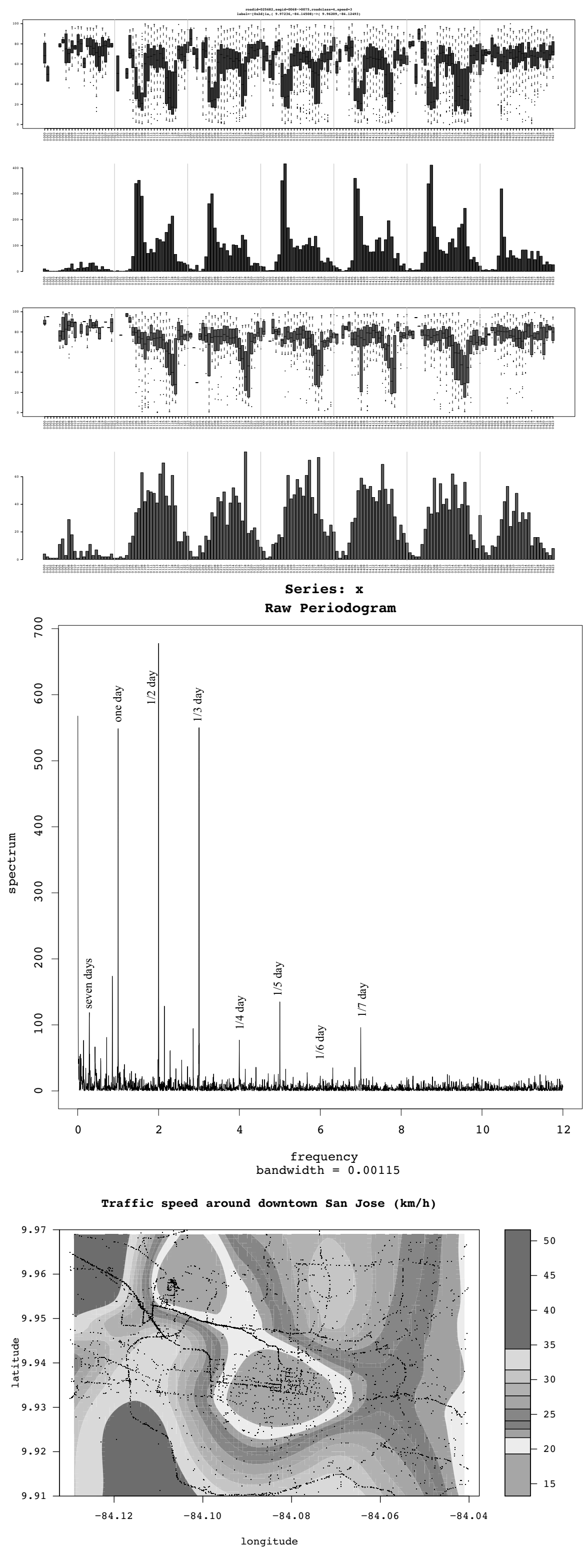

Figure 1: Top: weekly traffic speed pattern on a section of Pan-American Highway. Middle: Periodogram of the road traffic speed on a section of Pan-American Highway. Bottom: A closer distance to the downtown corresponds to a lower traffic volume and a lower speed.

Methodologies and Results

In this section, two approaches are given for road traffic speed estimation and prediction. The support vector method can be used to approximate a function $f(x)$ in terms of a linear combination of a finite number of basis functions $\sum_{i=1}^{n} \alpha_i \langle \phi(x), \phi(x_i) \rangle y_i$ where the bilinear operator $\langle \bullet, \bullet \rangle$ is a dot product, and the parameters α_i are trained by from the training set $\{(x_i, y_i), i = 1, \cdots, n\}$ by the structural risk minimization (SRM) principle (Vapnik 1998; Cristianini & Shawe-Taylor 2000). In the dynamic Bayesian network approach, the traffic in the road network is described in the following way. The traffic speed in a road link has a Gaussian distribution conditioned on the traffic flow rate $v_x(t) \sim \mathcal{N}(v_x(q_x(t)), \sigma_x(t)^2$. When road link 1 and road link 2 merges into road link 3, the traffic flow rates of the three road links satisfy $q_3(t) = q_1(t) + q_2(t)$. With this description and the EM (estimation maximization) algorithm, we can "learn" the relationship between traffic speed and traffic flow rate, estimate the (latent) flow rates, and predict future speed and flow rate on different road links. The dynamic Bayesian network takes the form of the latent structure influence model (Dong & Pentland 2007).

We used a data driven non-parametric approach (support vector regression with Laplacian kernel) to estimate road traffic speeds. Short term traffic speed prediction and the travel time prediction for vehicles on a road segment is based on the tracking records before the time of the predictions. To assess the performance of our non-parametric approach to predict the travel time through a road segment, we randomly separated out the tracking data for 25% of the delivery vehicles for validation. The average relative error for highway travel time prediction is 14%, and more than 95% predictions have less than 20% relative errors. In comparison, the travel time prediction based on road average speeds is 22%, and 30% of the predictions have more than 20% relative errors.

The support vector method improves the travel time prediction by finding the "abnormal" points — the points (t_i, x_i) whose speeds deviate from the expected ones (which can be roughly treated as the prior derived from historical data) by a value $\Delta v_{t_i, x_i}$ greater than the given threshold ϵ — and expressing the speed correction $\Delta v(t, x) = \sum_{i=1}^{N} \alpha_i \cdot \Delta v_{x_i, t_i} \cdot \exp(-\sigma \sqrt{(x - x_i)^2 + (t - t_i)^2})$ (which is added to the prior speed to give the estimated instantaneous traffic speed) as a linear combination of the $\Delta v_{t_i, x_i}$'s corresponding to those abnormal points. Since The Laplacian kernel $K((x, t), (x_i, t_i)) = \exp(-\sigma \sqrt{(x - x_i)^2 + (t - t_i)^2})$ decreases exponentially fast when the point of speed-estimation (t, x) goes away from the sample point (t_i, x_i), the influence of $\Delta v_{t_i, x_i}$ quickly becomes negligible with increasing distance between (t, x) and (t_i, x_i). We set ϵ to be 10%, and choose *C* and σ by 3-fold cross-validation.

Figure 2 illustrates the results of traffic speed estimation and travel time prediction on road 25608 (the segment of Pan-American highway from Juan Santamara International Airport to downtown San Jose) and road 34682 (the segment of Pan-American highway from downtown San Jose to Juan Santamara International Airport) on 02/29/2008. This figure is representative of the results for other roads and on other days according to our experiments. In the four plots, the axes going horizontally and vertically are respectively hour-of-day (from 00:00:00 to 24:00:00) and location-of-road (from 0 kilometer to 12 kilometers). Plot one and plot three from top to bottom visualize the results of estimating the traffic speeds on road 25608 and 34682 respectively. In the two plots, the colored points correspond to the tracking records, and different colors represent different vehicles. The solid lines represent the *estimated* tracks of *imaginary* vehicles that go from location 0 to location 12-kilometers of the roads, starting from different times of the day. Plot two and plot four visualize the results of predicting travel times from known tracking records on road 25608 and 34682 respectively. In the two plots, the dotted lines correspond to the tracking records used for traffic speed estimation. The points that are not blue represent the tracks to be estimated, and different colors represent different vehicles. The blue points represent the predicted tracks for those vehicles. As evidenced by the four plots, better estimation of traffic speed and prediction of travel time are possible because the tracking records provide evidence of instantaneous traffic speeds.

Traffic speed estimation based on the above non-parametric approach works by considering nearby vehicle speed abnormalities. Predicting traffic speed abnormality (e.g., traffic jam) in advance is a harder problem since we need to "learn" a **functional** rather than a function, $V : v(i, x, t') \cdot 1_{t' < t} \rightarrow v_{j,y}(t + \Delta t)$. Specifically, we are given the traffic speed at different locations *x* of different road segments *i* up to time *t*, (which is a function of *x*, *i* and *t'¡t* respectively), and we are required to estimate the speed at location *y* of road *j* at a future time $t + \Delta t$. This functional is non-linear since $V(\alpha v_1 + \beta v_2) \neq \alpha V(v_1) + \beta V(v_2)$ generally. While the functional can be estimated with support vector methods in a model-driven parametric approach, and we can use our knowledge on the traffic flow theory to assess the performance of any (blind) data-driven non-parametric approach, we will give below a dynamic Bayesian network approach, since the latter is much simpler.

We can represent the road network by a directed graph $\{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{V}$ is the set of road links and $\mathcal{E}$ is the set of transitions from one road link to another road link. For each road link $i \in \mathcal{V}$, its traffic flow rate is $\min(\sum_{(j,i) \in \mathcal{E}} q_j(t), q_i^{\max}(t)) + g_i(t)$ where $q_i^{\text{max}}(t)$ is the maximum flow rate of road link *i*, and $g_i(t)$ is the traffic flow generated at sample time t. Both $q_i^{\text{max}}(t)$ and $g_i(t)$ is an (unknown) periodic function. The traffic speed at road link i has Gaussian distribution around a non-linear function $v_i(t) = \mathcal{N}(v_i(q_i(t)), \sigma_i^2(t))$, where *v* and *q* satisfies the fundamental diagram approximation. Our goal is to learn the parameters g_i, $k_i^{\max}$ and $v_i^{\max}$, and to estimate $q_i(t)$.

With the EM algorithm similar to (Dong & Pentland 2007), we can predict traffic jams half an hour to two hours on highways. We note that the further ahead in time we predict, the more information and computation we need, and the traffic conditions of the less number of places we can predict with required precision based on limited amount of information. Since the further we look into the future, the more uncertainty is going to add into our computation.

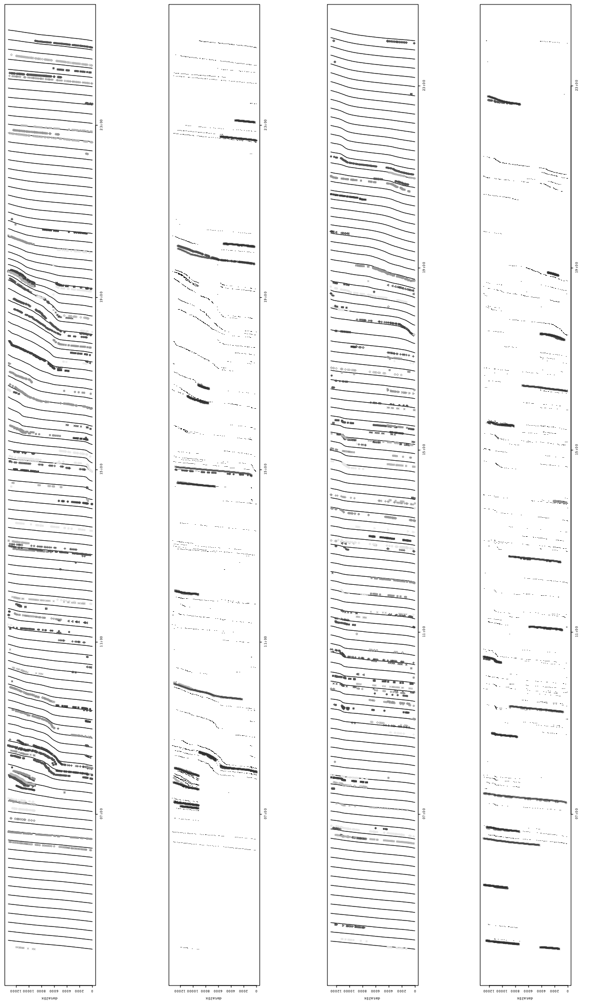

Figure 2: Traffic speed estimation and travel time prediction on road 25608 (the segment of Pan-American highway from Juan Santamara International Airport to downtown San Jose) and road 34682 (the segment of Pan-American Highway from downtown San Jose to Juan Santamara International Airport) on 02/29/2008.

Conclusions

The paper studies an important type of human group behavior — the city-wide road traffic. The study is based on a multi-terabyte data set containing latent information. We assign a meaning to the data set by mapping its *(longitude, latitude)* sequences to the corresponding *(road segment, offset)* sequences , and by making use of the known results from traffic engineering. The combination of domain knowledge and machine learning methods both generate better understanding of the data set and shed light on its latent information.

References

Cristianini, N., and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Dong, W., and Pentland, A. 2007. Modeling influence between experts. *Artificial Intelligence for Human Computing* 170–189.

Froehlich, J., and Krumm, J. 2008. Route prediction from trip observations. *Society of Automotive Engineers (SAE) 2008 World Congress*.

Gartner, N. H.; Messer, C. J.; and Rathi, A. K., eds. 2005. *Revised Monograph on Traffic Flow Theory*. Federal Highway Administration.

Gazis, D. C.; Herman, R.; and Rthery, R. W. 1961. Nonlinear follow the leader models of traffic flow. *Operations Research*.

Greenshields, B. D. 1935. A study of traffic capacity. In *Highway Research Board Proceedings*, volume 14, 448–477.

2000. *Highway Capacity Manual 2000*. Washington, DC: Transportation Research Board.

Horvitz, E.; Apacible, J.; Sarin, R.; and Liao, L. 2005. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *UAI*, 275–283.

Huisken, G. 2006. *Inter-Urban Short-Term Traffic Congestion Prediction*. Ph.D. Dissertation, University of Twente.

Kerner, B. S. 2004. *The Physics of Traffic: Empirical Freeway Pattern Features, Engineering Applications, and Theory (Understanding Complex Systems)*. Springer.

Lammer, S.; Gehlsen, B.; and Helbing, D. 2006. Scaling laws in the spatial structure of urban road networks. *Physica A: Statistical Mechanics and its Applications* 363(1):89–95.

Lighthill, M. J., and Whitham, G. B. 1955. On kinematic waves ii: A theory of traffic flow on long crowded roads. In *Proceedings of the Royal Society*, number 1178 in A229, 137–145.

Michalopoulos, P. G. 1988. Analysis of traffic flow at complex congested arterials. *Transportation Research Record* 1194(77-86).

Mohan, P.; Padmanabhan, V. N.; and Ramjee, R. 2008. Traffic sense: Rich monitoring of road and traffic conditions using mobile smartphones. Technical Report MSR-TR-2008-59, Microsoft Research India, Bangalore.

Richards, P. I. 1956. Shock waves on the highway. *Operations Research* 4(1):42–51.

Smith, B. L.; Zhang, H.; Fontaine, M.; and Green, M. 2003. Cellphone proves as an atms tool. Technical Report Smart Travel Lab Report No. STL-2003-01, University of Virginia.

Stephanopoulos, G., and Michalopoulos, P. G. 1979. Modelling and analysis of traffic queue dynamics at signalized intersectons. *Transportation Research* 13A:295–307.

van Lint, J. W. C. 2004. *Reliable Travel Time Prediction for Freeways*. Ph.D. Dissertation, Technische Universiteit Delft.

Vapnik, V. 1998. *Statistical Learning Theory*. John Wiley & Sons.