Using Mobile Phones to Forecast Arterial Traffic through Statistical Learning

Ryan Herring* Aude Hofleitner† Saurabh Amin‡ Tania Abou Nasr§

Amin Abdel Khalek¶ Pieter Abbeel‖ Alexandre Bayen**

Submitted For Publication
89th Annual Meeting of the Transportation Research Board
August 1, 2009

Word Count:

Number of words:	6245
Number of figures:	7 (250 words each)
Number of tables:	0 (250 words each)
Total:	7995

*Corresponding Author, Department of Industrial Engineering and Operations Research, University of California, Berkeley, 2105 Bancroft Way, Suite 300, Berkeley, CA 94720, (510) 642-5667, ryanherring@berkeley.edu

†Department of Electrical Engineering and Computer Science, University of California, Berkeley, 2105 Bancroft Way, Suite 300, Berkeley, CA 94720, aude.hofleitner@polytechnique.edu, and Ecole Doctorale Ville, Transport et Territoire, Universit Paris-Est, Marne-La-Valle, France

‡Department of Civil and Environmental Engineering, Systems Engineering, University of California, Berkeley, 760 Davis Hall, Berkeley, CA 94720, amins@berkeley.edu

§California Center for Innovative Transportation, 2105 Bancroft Way, Suite 300, Berkeley, CA 94720, tania.abou-nasr@polytechnique.edu

¶California Center for Innovative Transportation, 2105 Bancroft Way, Suite 300, Berkeley, CA 94720, ana36@aub.edu.lb

‖Department of Electrical Engineering and Computer Science, University of California, Berkeley, 746 Sutardja Dai Hall #1758, Berkeley, CA 94720, pabbeel@cs.berkeley.edu

**Department of Civil and Environmental Engineering, Systems Engineering, University of California, Berkeley, 642 Sutardja Dai Hall, Berkeley, CA 94720, bayen@berkeley.edu

Abstract

This article introduces the new component of *Mobile Millennium* dedicated to arterial traffic. *Mobile Millennium* is a pilot system for collecting, processing and broadcasting real-time traffic conditions through the use of GPS equipped smartphones. Two algorithms that use data from GPS equipped smartphones to estimate arterial traffic conditions are presented, analyzed and compared. The algorithms are based on *Logistic Regression* and *Spatio-Temporal Auto Regressive Moving Average* (STARMA), respectively. Each algorithm contains a *learning* component, which produces estimates of spatio-temporal parameters for describing interactions between the states of arterial links in the network. Additionally, each algorithm contains an *inference* component, which gives the procedure for processing real-time data into short-term forecasts using these parameters. The algorithms are tested with simulation data obtained from Paramics software, and from a field test in New York. Both methods provide encouraging results in forecasting arterial traffic conditions using sparse GPS data.

1 Introduction

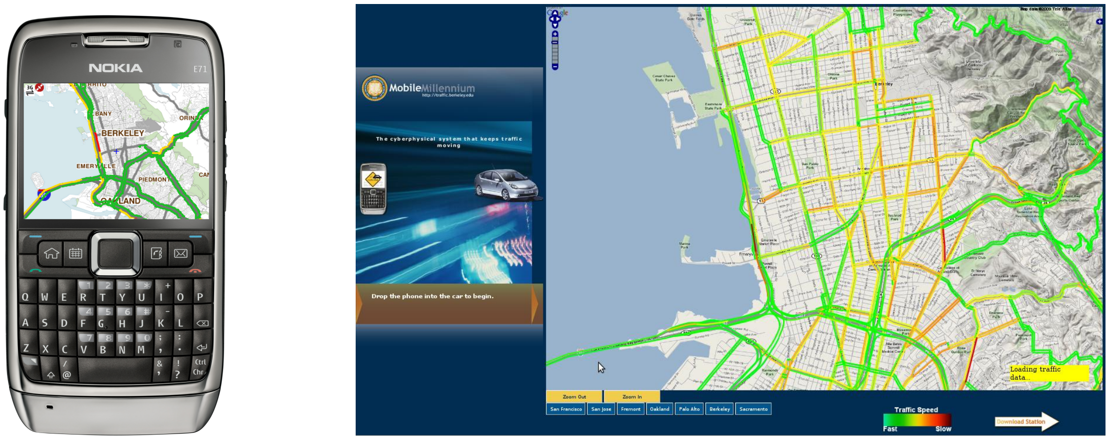
In the United States and numerous other parts of the world, traffic is an unavoidable part of economic activity. The 2007 Urban Mobility Report [3] states that traffic congestion causes 4.2 billion hours of extra travel in the United States every year, which accounts for 2.9 billion extra gallons of fuel, which cost taxpayers an additional $78 billion.

Numerous measures can be taken to address problems due to traffic congestion. An essential step is to create the ability to forecast traffic conditions with significant accuracy and reliability. Numerous challenges stand in the way of this type of effort. A significant portion of the transportation network has little or no dedicated infrastructure for collecting traffic data. Areas equipped with this infrastructure generally only cover highways and have high installation and maintenance costs in addition to providing data of variable reliability. An alternative to using dedicated communication infrastructure is to leverage an existing system such as the cellular phone network. The *Mobile Millennium* project [2] was conceived as a response to these challenges, to explore the capability of using cellular phones to provide traffic data.

The mobile internet is the underlying technology enabling the existence of the *Mobile Millennium* system. User-generated content (in the present case, smartphone measured traffic data) is sent to a central system, which provides information back to the cell phone owner for personal use. This "web 2.0" application framework is commonly referred to as "participatory sensing," which refers to the ad hoc process of voluntarily providing sensing data to a system. In general, there are a number of challenges to overcome with any nomadic sensing technology, including unknown location of upcoming measurements, sparsity of the data, and unpredictability of the frequency of data collection.

The *Mobile Millennium* system was officially launched on November 10, 2008 when the team released a software client for GPS enabled smartphones to the public, available for download (see figure 1(a)). Traffic conditions are broadcast back to drivers' mobile phones, enabling commuters to make more informed route and trip decisions. Additionally, traffic data can be analyzed in the *Mobile Millennium* live traffic visualizer, shown in figure 1(b), which is currently on display in the CITRIS [1] Tech Museum on the UC Berkeley campus. The deployment area of the pilot system is focused on commuters in Northern California, including the San Francisco Bay Area and Sacramento. The project is a follow up to the *Mobile Century* experiment, in which 165 UC Berkeley graduate students were hired to drive a 10-mile stretch of I880 in California for a day, demonstrating the feasibility of a real-time *highway* traffic estimation service using only GPS enabled devices [29].

This article focuses on estimating and forecasting *arterial* traffic conditions using only GPS enabled devices using two statistic models. Section 2 reviews previous traffic models and examines the need for a new (statistical) approach. Section 3 is a formal presentation of the problem with details of the data types and models used. Sections 4 and 5 present the details of the logistic regression and STARMA models, respectively. Results are presented on simulation and field experiment data in section 6. Further analysis and future directions are presented in the conclusion (section 7).



(a) Traffic client

(b) Web interface

Figure 1: ***Mobile Millennium* traffic information services.** (a) The *Mobile Millennium* traffic client. (b) The *Mobile Millennium* interface for highway and arterial traffic visualization.

2 Challenges and Motivation for the Statistical Learning Approach

The development of traffic theory has led to numerous modeling contributions since the pioneering work of Lighthill, Whitham and Richards, [15, 21], which relied on hydrodynamic theory. By nature, arterial traffic has very high variability, which make it challenging to use flow models for arterial networks. Studies have typically focused on modeling single intersections [22, 28, 4, 23] using dedicated traffic sensors. This modeling approach is difficult to adapt to a general traffic information system on a dense arterial network because it requires a high density of traffic sensors (which are prohibitively expensive at the scale of the arterial network). In particular, one of the major challenges of *Mobile Millennium* is that the system does not have access to flow counts. A statistical approach is suitable because sensing every vehicle is impractical and because this allows for the incorporation of other information types (such as human mobility patterns [10]). The present article article suggests to develop monitoring capabilities for arterial traffic in two directions: (1) using alternate data sources such as privacy aware cell phone information; (2) developing new arterial models based on statistical learning which overcome some major issues faced by analytical flow or queuing-based models. The motivations for these two items are described in the remainder of this section.

2.1 Using Cell Phones as Traffic Probes

Experimental research on cell phone based traffic monitoring [5, 30, 24, 31] has investigated the ability to locate the position of users using trilateration- or triangulation-based methods. It has shown limited success for estimation of travel times due to the position measurement inaccuracy, particulary on short distances and dense networks [16, 12]. The complexity of traffic patterns in the arterial networks gives the use of GPS-based traffic information enormous growth potential.

2.2 Application of Machine Learning to Arterial Traffic

Machine learning techniques have been used to estimate and produce short term traffic predictions for both freeway [8, 27, 7, 6, 17] and arterial networks [26, 28, 14, 19, 25, 9]. These studies present encouraging results. One of the limitations of these approaches for our problem is that they present results for specific traffic variables (in particular for flow/density). The present article focuses on estimating and predicting congestion states and travel times. Congestion states, also referred to in the literature as *Level of Service* (LoS), represent traffic conditions on the road segment as experienced by the network user. They also represent the level of service offered by the network manager. They can be interpreted as a discrete representation of traffic states. Traffic states (for example travel time) have their own statistical distribution depending on traffic conditions.

In the remainder of this article, we present regression techniques corresponding to two different approaches:

- *Logistic Regression* model. An example of a neural network used as a clustering algorithm between discrete traffic congestion states.
- *Spatio-Temporal Auto-Regressive Moving Average* (STARMA) model. An example of a time series model in which the traffic variable studied (travel time) depends on the previous values of this variable.

To our knowledge, logistic regression has not been used in arterial traffic estimation or prediction. We compare its results to a more widely used model, the STARMA model. This comparison is of significant interest for the transportation community since it is between a discrete output (congestion states from the clustering of the logistic regression) and a continuous output (travel time estimations from the linear regression of the STARMA model). Furthermore, the system and results presented here are one of the first instantiations of real-time arterial monitoring using machine learning with streaming data collected from smartphone. *Mobile Millennium* is currently implemented and operational in all of Northern California [2].

3 Problem Formulation

This section formally presents the problem formulation, namely estimating LoS indicators which are the aggregate travel times and congestion states for an arterial road network. First, we introduce our sampling paradigm, the *Virtual Trip Line* in section 3.1. This leads to the problem of sensing on a graph (section 3.2) and the formal definitions of LoS indicators (section 3.3). The problem description of estimating the LoS indicators based on STARMA and logistic regression is presented in section 3.4.

3.1 Virtual Trip Line Sensing Infrastructure

A GPS-enabled smartphone is capable of recording its GPS location every few seconds. Over time, this vehicle trajectory information produces a rich history of the vehicle and the velocity field through which it evolves [11]. While this level of detail can be useful for traffic estimation, it can be privacy invasive, since the device is ultimately carried by a single user. Even if personally identifiable information from the data is replaced with a randomly chosen ID through a process known as pseudo-anonymization, it is still possible to re-identify individuals from trajectory data [13].

Virtual Trip Lines (VTLs) [12] are spatial triggers for phones to collect measurements and send updates. Each VTL consists of two GPS coordinates which make a virtual line drawn across a roadway of interest. Instead of time-based periodic sampling, VTLs trigger disclosure of speed and location updates by sampling in space, creating updates at predefined geographic locations on roadways of interest. Additionally, the travel time between pairs of VTLs can be extracted and this type of travel time data will be considered the primary data source used in this article.

3.2 Graph Model of the Road Network

Consider an arterial network with a total of N pairs of VTLs deployed. Each pair has a unique identification number $i \in \{1, \ldots, N\}$. The set of all VTL pairs is denoted by $\mathcal{V} = \{1, \ldots, N\}$. Each VTL pair has a segment of road in between with a possibility of one or more road features such as an intersection (with or without traffic lights), pedestrian walkways, stop/slow signs etc. The characteristics of these road features can be static (such as presence of a stop sign) or dynamic (such as phase of a signalized intersection) with respect to time. The travel time experienced by a vehicle traveling through a VTL pair depends on the characteristics of the road features as well as the demand-capacity restrictions imposed by the dynamics of traffic flow. We also assume that each VTL pair is associated with unidirectional traffic flow. For arterial links consisting for bidirectional traffic, we associate a VTL pair corresponding to each flow direction.

We say that the upstream (resp. downstream) VTL for the pair i is the VTL at which the traffic enters (resp. leaves) the corresponding stretch of road. For pair i, let the upstream and downstream VTLs be denoted by i_u and i_d respectively. Then the VTL sensor network can be represented as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of all VTL pair as defined earlier and $\mathcal{E}$ is the set of all edges. Two VTL pairs i and j form an edge directed from pair i to pair j, denoted e_{ij}, if i_d and j_u correspond to same VTL. Then i (resp. j) is called the upstream (resp. downstream) node of edge e_{ij}.

We define the set of first order neighbors for VTL pair j as

$$\mathcal{N}^1(j) = \{j\} \cup \{i \in \mathcal{V} : e_{ij} \in \mathcal{E}\} \cup \{k \in \mathcal{V} : e_{jk} \in \mathcal{E}\}$$

which is simply the set of all the upstream and downstream VTL pairs for the pair j (in which we include pair j itself).

We can extend the above definition to define n^{th} ($n \geq 1$) order neighbors as:

$$\left\{ \begin{array}{lcl} \mathcal{N}^0(j) & = & \{j\} \\ \mathcal{N}^n(j) & = & \mathcal{N}^{n-1}(j) \cup \left(\bigcup_{l \in \mathcal{N}^{n-1}(j)} \{i \in \mathcal{V} : e_{il} \in \mathcal{E}\} \cup \{k \in \mathcal{V} : e_{lk} \in \mathcal{E}\} \right) \end{array} \right. \tag{1}$$

3.3 Traffic Level of Service Indicators

We assume that for any VTL pair $i \in \mathcal{V}$, the travel time data is available at times $0 \leq t_1 \leq t_2 \leq \ldots$. As an alternative to travel time data, we can also compute the pace (travel time divided by the length of road for the VTL pair). We denote the data obtained at time t_1 for VTL pair i as $X_{t_1,i}$. Since the acceptable values of $X_{t_1,i}$ generally lie between a minimum and maximum value, we reject data that do not fall in this range. For VTL pair i, let us denote this range as $[\underline{X}_i, \overline{X}_i]$.

Since the data obtained is event-based, it cannot be directly used for training statistical models that needs regular sampling rates. We aggregate the travel time data in t second windows to obtain a time series of observations at times $k = 0, t, 2t, \ldots$. Here t is the aggregation interval. Henceforth, we will use k to denote the time interval $[(k-1)t, kt)$. The set of available observations during the time period k for any VTL pair i is denoted as $A_{k,i}$, that is,

$$A_{k,i} = \{X_{t_m,i} \,|\, (k-1)t \leq t_m < kt\}.$$

The *penetration rate* for VTL pair i during time k, denoted $p_{k,i}$, is the fraction of available observations out of the total number of vehicles $D_{k,i}$ traveling through pair i during time k:

$$p_{k,i} = \frac{A_{k,i}}{D_{k,i}}.$$

We define the spatial aggregation function for VTL pair i, $h_i(\cdot) : A_{k,i} \mapsto [\underline{X}_i, \overline{X}_i]$, as the function that aggregates the set of observations $A_{k,i}$ in to an *aggregate representative quantity*, denoted $Z_{k,i}$ with values in the range $[\underline{X}_i, \overline{X}_i]$. In the remainder of this article, $Z_{k,i}$ is an aggregated travel time (seconds). Thus, the aggregate travel time for VTL i during interval k is

$$Z_{k,i} = h_i(\{X_{t_m,i} \,|\, (k-1)t \leq t_m < kt\}).$$

We define the *mode* of a VTL pair as the categorical variable indicative of the extent of delay experienced in navigating through the VTL pair. For example, a binary mode classification can be *uncongested or congested.* Thus, the mode of a VTL pair can also interpreted as a *congestion state.* Let the mode of VTL pair i during time interval k be denoted as $Q_{k,i}$. In order to convert the total number of observations available at VTL i during time interval k into a mode of the VTL pair, we define a *congestion indicator function* $g_i(\cdot) : A_{k,i} \mapsto \{1, \ldots, M\}$ where M is the desired number of modes the VTL pairs should be classified to. Thus,

$$Q_{k,i} = g_i(\{X_{t_m,i} | (k-1)t \leq t_m < kt\}).$$

From a statistical modeling perspective, both the aggregate speed or travel time, $Z_{k,i}$, and the congestion state, $Q_{k,i}$, for $i \in \mathcal{V}$ and $k \in \{0, 1, \ldots\}$ can be considered as random processes generated by space-time varying traffic flow phenomena on the arterial network. Both $Q_{k,i}$ and $Z_{k,i}$ can be regarded as LoS indicators.

3.4 Estimating Level of Service Indicators

If we had data from all the vehicles for all the VTL pairs over the entire time horizon of interest, the penetration rate $p_{k,i}$ would satisfy $p_{k,i} = 1$ for all $i \in \mathcal{V}$ and $k \in \{0, 1, \ldots\}$. We could then compute the entire probability distribution of $Z_{k,i}$ and $Q_{k,i}$. However, the challenge of arterial traffic state estimation and forecast is that the typical penetration rates are very low. Our focus in this article is to develop reliable estimation and forecasting methods for such situations.

We now describe the problem formulation for estimation or *nowcast*[1]. We typically only have data from a small percentage of the total number of vehicles ($p_{k,i} \sim 0.02 - 0.05$). Thus, the choice of the aggregation function $h_i(\cdot)$ (resp. the congestion indicator function $g_i(\cdot)$) becomes critical to obtain reliable estimates of $Z_{k,i}$ (resp. $Q_{k,i}$). For a given choice of $h_i(\cdot)$, the best estimate of the aggregate travel time or speed for VTL pair i during interval k is given by the conditional expectation of $Z_{k,i}$ given the aggregate travel times up-to (and excluding) the current time interval:

$$\hat{Z}_{k,i} = \mathbb{E}[h_i(A_{k,i})|h_j(A_{j,v}), j < k, v \in \mathcal{V}] = \mathbb{E}_{h_i}[Z_{k,i}|Z_{j,v}, j < k, v \in \mathcal{V}],$$

where notation $\mathbb{E}_{h_i}[\cdot]$ is used to indicate the dependence of the expectation on the aggregation function h_i. We now introduce the following *conditional independence assumption*: $Z_{k,i}$ is conditionally independent of all other data conditioned on the data from the past r time intervals for VTL pairs in the set $\mathcal{N}^s(i)$. Under this assumption, we can write

$$\hat{Z}_{k,i} \approx \mathbb{E}_{h_i}[Z_{k,i}|Z_{j,v}, k - r \leq j < k, v \in \mathcal{N}^s(i)] \tag{2}$$

Thus, $\hat{Z}_{k,i}$ only depends on data with r *temporal dependencies* in the past and s *spatial dependencies* from the neighbors. Similarly, for given choices of the aggregation function $h_i(\cdot)$ and the congestion indicator function $g_i(\cdot)$, we can write the conditional expectation of $Q_{k,i}$ given all the aggregate travel times up-to (and excluding) the current time interval as[2]

$$\begin{aligned}\hat{Q}_{k,i} &= \mathbb{E}[g_i(A_{k,i})|h_j(A_{j,v}), j < k, v \in \mathcal{V}] \\ &\approx \mathbb{E}_{h_i,g_i}[Q_{k,i}|Z_{j,v}, k - r \leq j < k, v \in \mathcal{N}^s(i)],\end{aligned} \tag{3}$$

In the statistics terminology, the quantities $Z_{k,i}$ and $Q_{k,i}$ in (2) and (3) are known as the *response variables*; the conditioned variables $Z_{j,v}$ and $Q_{j,v}$ are called the *dependent variables* or *covariates.* The present article compares the two estimators which we now introduce.

The first estimator is based on expressing (2) as a *linear regression problem.* For a temporal and spatial dependence of orders r and s respectively, we assume a linear dependence of response $Z_{k,i}$ on the covariates $Z_{j,v}$:

$$\hat{Z}_{k,i} = \beta_i^0 + \sum_{v \in \mathcal{N}^s(i)} \left(\sum_{j=k-r}^{k-1} \beta_i^{j,v} \, Z_{j,v} \right). \tag{4}$$

In order to make the notation concise, let $\mathbf{Z}_{k,i}^{r,s}$ be the $r \times \mathcal{N}^s(i)$ vector of covariates or dependent variables obtained by stacking the aggregate travel times $Z_{j,v}$ for $k - r \leq$

[1] Depending on the convention used, this can also be treated as one-step ahead forecast. In this article, we do not distinguish between one-step forecast and estimation.

[2] Alternatively, we can also condition $Q_{k,i}$ directly on congestion modes up-to (and excluding) the current time, that is, $\hat{Q}_{k,i} = \mathbb{E}_{g_i}[Q_{k,i}|Q_{j,v}, k - r \leq j < k, v \in \mathcal{N}^s(i)]$. However, we do not consider this type of estimator in this article.

$j < k$ and $v \in \mathcal{N}^s(i)$, β_i be the corresponding $r \times \mathcal{N}^s(i) + 1$ vector of parameters to be estimated. Then the equation (4) can be re-written as

$$\hat{Z}_{k,i} = \beta_i^\top \mathbf{Z}_{k,i}^{r,s},$$

where $^\top$ stands for the transpose of a vector. As described later in Section 5, instead of a simple regression model (4), we consider a STARMA model.

Our second estimator is based on expressing (3) as a logistic regression problem which assumes a linear dependence of the *logit* or the log-odds ratio of conditional expectation $\hat{Q}_{k,i}$ on the response variables. That is, for a temporal and spatial dependence of orders r and s respectively, we have

$$\log\left(\frac{\hat{Q}_{k,i}}{1 - \hat{Q}_{k,i}}\right) = \beta_i^\top \mathbf{Z}_{k,i}^{r,s}.$$

We can express this equation as

$$\hat{Q}_{k,i} = f_{\beta_i}(\mathbf{Z}_{k,i}^{r,s}) := \frac{1}{1 + \exp\left(-\beta_i^\top \mathbf{Z}_{k,i}^{r,s}\right)}, \qquad (5)$$

where the subscript β_i in $f_{\beta_i}(\cdot)$ encodes the dependence on the β_i.

We detail the implementation of a logistic regression estimator in Section 4 and a STARMA-based estimator in Section 5. However, two important points need to be mentioned. First, the above formulation can be modified to include the case of multiple steps forecast. For example, an $m-$step forecast at time k for VTL pair i can be written as

$$\hat{Z}_{k+m,i} = \mathbb{E}_{h_i}[Z_{k,i} | Z_{j,v}, j < k, v \in \mathcal{V}], \qquad (6)$$

where we consider data up to time k to predict traffic at time $k + m$.

Second, we note that for some VTL pairs and time intervals, we might not have any available data, that is, $A_{j,v} = \emptyset$ for some $j \in \{k - r, \ldots, k\}$ and $v \in \mathcal{N}^s(i)$. In this case, one has to employ a technique of *estimation with missing data*. We will briefly touch on the forecast problem for the STARMA model but will address the issue of missing data in later work.

4 Logistic Regression

We now discuss the estimator based on the logistic model (5) to estimate the congestion state $Q_{k,i}$ for a VTL pair i and time interval k. Suppose that $Q_{k,i}$ is binary-valued, that is $Q_{k,i} = \{0, 1\}$ and $M = 2$. When $Q_{k,i} = 1$ (resp. $Q_{k,i} = 0$), we say that the VTL pair i during interval k is in the *congested mode* (resp. *uncongested mode*). Then the estimator $\hat{Q}_{k,i}$ gives the conditional probability of the $Q_{k,i}$ given the dependent variables:

$$\begin{aligned}\hat{Q}_{k,i} = \mathbb{E}_{h_i,g_i}[Q_{k,i} | \mathbf{Z}_{k,i}^{r,s}] &= 1 \cdot \mathbb{P}_{h_i,g_i}[Q_{k,i} = 1 | \mathbf{Z}_{k,i}^{r,s}] + 0 \cdot \mathbb{P}_{h_i,g_i}[Q_{k,i} = 0 | \mathbf{Z}_{k,i}^{r,s}] \\ &= \mathbb{P}_{h_i,g_i}[Q_{k,i} = 1 | \mathbf{Z}_{k,i}^{r,s}]\end{aligned}$$

Now using (5), we can write the conditional probability of $Q_{k,i}$ given the aggregate travel time for r temporal and s spatial dependencies as

$$\mathbb{P}_{h_i,g_i}[Q_{k,i}|\mathbf{Z}^{r,s}_{k,i};\beta_i] = [f_{\beta_i}(\mathbf{Z}^{r,s}_{k,i})]^{Q_{k,i}}[1 - f_{\beta_i}(\mathbf{Z}^{r,s}_{k,i})]^{1-Q_{k,i}}$$

We now assume that for a VTL pair i, the response process $\{Q_{k,i}\}$ and the covariate process $\{\mathbf{Z}^{r,s}_{k,i}\}$ is available for a number of time intervals $k = 0, \dots, K$. Introducing the conditional independence assumption that the response variable $Q_{k,i}$ is independent of all other data given $\mathbf{Z}^{r,s}_{k,i}$. Then the joint conditional probability of $\{Q_{k,i}\}$ given $\{\mathbf{Z}^{r,s}_{k,i}\}$ (also known as the conditional likelihood) can be expressed as

$$\mathbb{P}_{h_i,g_i}[\{Q_{k,i}\}^K_{k=0}|\{\mathbf{Z}^{r,s}_{k,i}\}^K_{k=0};\beta_i] = \prod_{k=0}^{K}[f_{\beta_i}(\mathbf{Z}^{r,s}_{k,i})]^{Q_{k,i}}[1 - f_{\beta_i}(\mathbf{Z}^{r,s}_{k,i})]^{1-Q_{k,i}}$$

For a given training data $\{Q_{k,i}\}^K_{k=0}$ and $\{\mathbf{Z}^{r,s}_{k,i}\}^K_{k=0}$, the *best* estimate of parameter β_i is obtained by maximizing the logarithm of the conditional likelihood which we state explicitly as follows:

$$\mathcal{L}(\beta_i; \{Q_{k,i}\}^K_{k=0}, \{\mathbf{Z}^{r,s}_{k,i}\}^K_{k=0}) = \sum_{k=0}^{K}\left(Q_{k,i}\cdot\beta_i^\top\mathbf{Z}^{r,s}_{k,i} - \log\left[1 + \exp\left(\beta_i^\top\mathbf{Z}^{r,s}_{k,i}\right)\right]\right)$$

The optimal estimate so obtained and denoted β_i^*, is called the *maximum likelihood estimate* (MLE). A number of standard iterative methods, all similar to Newton-Raphson method, can be used to obtain the MLE β_i^*. Examples of such method include Fisher scoring method, iterative reweighted least squares etc. Due to space limitations, we omit the details of the algorithm and refer the reader to [18].

Once the parameters are learned, *validation* can be done on a similar data set as the one used to obtain β_i^*. Validation is done to assess the ability of the learned model to correctly estimate the traffic status (congestion state in this case) on previously unseen data.

5 STARMA

We now discuss the STARMA model which is a more efficient estimator than the simple linear regression model (4). The number of parameters to be estimated for (4), given by $r \times |\mathcal{N}^s(i)| + 1$ ($|A|$ is the cardinality of A), can increase significantly as the spatial dependency s increases. In order to explain the model, we first present the *spatio-temporal autoregressive* (STAR) model and subsequently generalize to a full STARMA model.

Following (1), the set of n order neighbors ($0 \le n \le s$) for a VTL pair i can be expressed as follows

$$\mathcal{N}^s(i) = \bigcup_{n=0}^{s}\mathcal{N}^n(i)\backslash\mathcal{N}^{n-1}(i).$$

Here we adopt the convention that $\mathcal{N}^0(i)\backslash\mathcal{N}^{-1}(i) = \{i\}$. Now, for the linear regression model (4), for any temporal order j, ($k - r \le j < k$) and spatial order n, ($0 \le n \le s$), we introduce the assumption that

$$\text{For all } v \in \mathcal{N}^n(i)\backslash\mathcal{N}^{n-1}(i), \quad \beta_i^{j,v} \equiv \beta_i^{j,n}, \tag{7}$$

and the definition of *n-th order, spatially-weighted travel time* as

$$\varphi_i^{(n)}(Z_j) = \frac{\sum_{l\in\mathcal{N}^n(i)\backslash\mathcal{N}^{n-1}(i)} w_{i,l}^{(n)} Z_{j,l}}{\sum_{l\in\mathcal{N}^n(i)\backslash\mathcal{N}^{n-1}(i)} w_{i,l}^{(n)}}, \tag{8}$$

where $Z_j = (Z_{j,1}, \ldots, Z_{j,N})$ is the vector of aggregate travel times for all the N VTL pairs during time interval j and $w_{i,l}^{(n)}$ are the pre-defined *spatial weights of order* n for $Z_{j,l}$.

Under the assumption (7) and the definition (8), we can now write the STAR model of *autoregressive* (AR) temporal order r and spatial order s as

$$Z_{k,i} = \sum_{j=k-r}^{k-1} \sum_{n=0}^{s} \beta_i^{j,n} \varphi_i^{(n)}(Z_j) + \epsilon_{k,i} \tag{9}$$

where $\epsilon_{k,i}$ is the normally distributed error term with variance σ^2 with the properties that $\mathbb{E}[\epsilon_{k,i}] = 0$ for all k and $i \in \mathcal{V}$; and for all $i, j \in \mathcal{V}$

$$\mathbb{E}[\epsilon_{k,i}\epsilon_{k+s,j}] = \begin{cases} \sigma^2 & \text{if } s = 0 \\ 0 & \text{otherwise.} \end{cases}$$

The number of parameters to be estimated for the STAR model (9), including σ^2, is $r(s+1)+1$ which is (typically) much smaller than $r \times \mathcal{N}^s(i) + 1$ for (4). The STAR model can now be generalized to STARMA model of autoregressive temporal order r and spatial order s, and *moving average* (MA) temporal order p and spatial order q as[3]

$$Z_{k,i} = \sum_{j=k-r}^{k-1} \sum_{n=0}^{s} \beta_i^{j,n} \varphi_i^{(n)}(Z_j) - \sum_{j=k-p}^{k-1} \sum_{n=0}^{q} \alpha_i^{j,n} \varphi_i^{(n)}(\epsilon_j) + \epsilon_{k,i}, \tag{10}$$

where $\epsilon_j = (\epsilon_{j,1}, \ldots, \epsilon_{j,N})^\top$.

Here $\alpha_i^{j,n}$ are the moving average parameters. The total number of parameters (including σ^2) to be estimated for the STARMA model (10), denoted as STARMA(r, s, p, q) are $r(s+1) + p(q+1) + 1$.

Following [20], we adopt the assumption in this article that that STARMA parameters are same for VTL pairs, that is, $\alpha_1^{j,n} = \ldots = \alpha_N^{j,n} \equiv \alpha_{j,n}$ and $\beta_1^{j,n} = \ldots = \beta_N^{j,n} \equiv \beta_{j,n}$. Then model (5) can be vectorized for all VTL pairs $i \in \mathcal{V}$ as

$$Z_k = \sum_{j=k-r}^{k-1} \sum_{n=0}^{s} \beta^{j,n} \Phi^{(n)}(Z_j) - \sum_{j=k-p}^{k-1} \sum_{n=0}^{q} \alpha^{j,n} \Phi^{(n)}(\epsilon_j) + \epsilon_k. \tag{11}$$

where $\Phi^{(n)}(\cdot) = (\varphi_1^{(n)}(\cdot), \ldots, \varphi_N^{(n)}(\cdot))^\top$ and $\epsilon_k = (\epsilon_{k,1}, \ldots, \epsilon_{k,N})^\top$.

[3] More generally, the AR spatial order s (resp. the MA spatial order q) can vary with the temporal order r (resp. p). However, we do not consider this generalization in this article.

For given training data $\{Z_k\}$, $(k = 0, \ldots, K-1)$, the best estimate of the parameters $A := [\alpha^{j,n}]_{p\times(q+1)}$, $B := [\beta^{j,n}]_{r\times(s+1)}$ and σ^2 is given by maximizing the conditional likelihood expressed as

$$\mathbb{P}(\{Z_k\}_{k=0}^{K-1}; A, B, \sigma^2) = (2\pi)^{-\frac{KN}{2}} \left|\sigma^2 \mathbf{I}_{KN\times KN}\right|^{-\frac{1}{2}} \exp\left(-\frac{S(A,B)}{2\sigma^2}\right) \tag{12}$$

where $I_{KN\times KN}$ is the identity matrix, $S(A,B) := (\epsilon_0, \ldots, \epsilon_{K-1})^\top (\epsilon_0, \ldots, \epsilon_{K-1})$ and according to (11), we have

$$\epsilon_k = Z_k - \sum_{j=k-r}^{k-1} \sum_{n=0}^{s} \beta^{j,n} \Phi^{(n)}(Z_j) + \sum_{j=k-p}^{k-1} \sum_{n=0}^{q} \alpha^{j,n} \Phi^{(n)}(\epsilon_j).$$

The *maximum likelihood estimate* parameters, denoted A^*, B^*, are obtained by maximizing the logarithm of the conditional likelihood (12), and the corresponding σ^* is estimated by

$$\sigma^* = \sqrt{\frac{S(A^*, B^*)}{KN}}.$$

For further details, we refer the reader to [20].

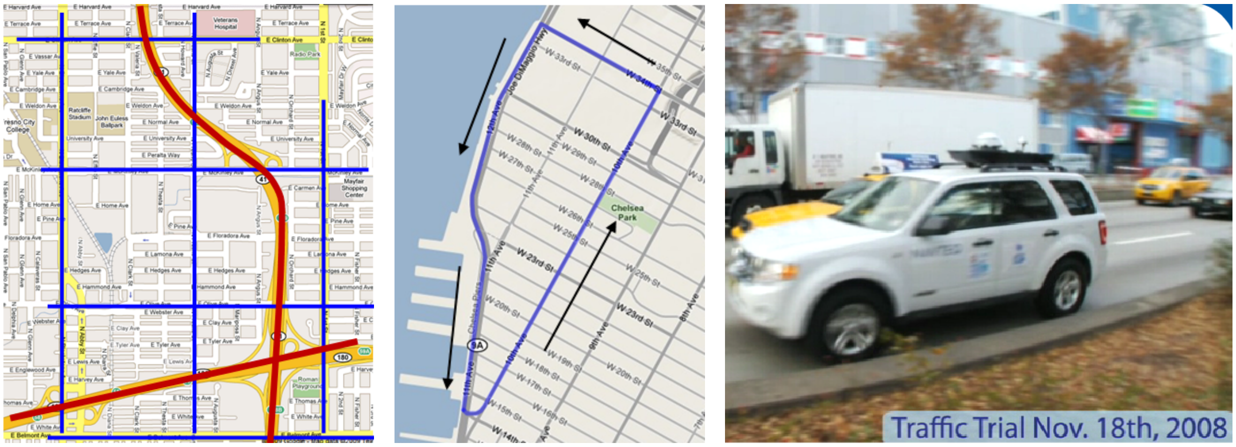
6 Results

This section presents the results from logistic regression based classification and STARMA-based continuous linear regression. Each algorithm is implemented and tested on simulation and field experiment data, described in section 6.1. The framework for quantifying accuracy is described in section 6.2. Results are then presented for one-step forecast (section 6.3), followed by multi-step forecast for the STARMA model (section 6.5). Additionally, a study of the effect of the *penetration rate* on the forecast accuracy (section 6.4) is presented.

6.1 Simulation and Field Experiment Data

There are two data sets used in this article. The first set was generated from Paramics micro-simulation software. The road network modeled consists of 1,961 nodes, 4,426 links, 210 zones and is based on the SR41 corridor in Fresno, CA. We specifically analyzed a sub-network that includes 9 arterial roads, 20 signals and 15 stop signs. Paramics simulates every car in the network. From this simulation, we extract the position of every vehicle at one-second time intervals. This provides detailed information about speed and travel time through the network. The sub-network studied in this article includes 380 different links, each one of which is characterized with a specific length, a number of lanes, a direction, a speed limit and signal information. 99 VTLs were placed on different links, which corresponds to 156 different pairs of VTLs, in order to capture travel times along links and through intersections.

The second data set was obtained as part of the official *Mobile Millennium* launch demonstration in New York City at the *ITS World Congress*. Twenty drivers, each carrying a GPS equipped cell phone, drove for 3 hours (9:00am to 12:00pm) around a



(a) Paramics Map (b) New York Map (c) Test Vehicle

Figure 2: **Experiment Design.** (a) Map of the Paramics network in Fresno, CA. (b) Experiment route for New York City field test used to collect the data (arrows represent the direction of traffic of probe vehicles). (c) Test vehicle used for the New York test.

2.4 mile loop of Manhattan (see figure 2). This number of drivers constituted approximately 2% of the total vehicle flow through the road of interest. The experiment was repeated 3 times in order to use two of the experiments as training data for the models and the other to validate the model results. The operational capabilities of the system were demonstrated at the *ITS World Congress* [2] on November 18, 2008, when live arterial traffic was displayed for conference attendees.

6.2 Validation Framework

In order to compute the accuracy of the model, one needs to define the "ground truth" state of traffic. In this article, travel times are aggregated into a single value per time interval (5 minutes for Paramics, 15 minutes for New York). This single value per time interval is considered the true state for the interval. Determining ground truth for the logistic regression method requires classifying each time interval as congested or uncongested. The STARMA method uses the average travel time during each interval as the ground truth value. Both of these methods correspond to choosing appropriate $h_i(\cdot)$ and $g_i(\cdot)$ functions as described in section 3.3.

The aggregation function $h_i(\cdot)$ should capture the pattern of change in pace over different intervals to provide an aggregate quantity that is sufficiently representative of the congestion state, thus providing better accuracy in training the model and obtaining the logistic regression parameters. Based on extensive testing and simulation, it is observed that aggregating the travel times based on the entire data available in an interval fails to capture the congestion state due to the high variance of travel times when a link is congested. The probes most affected by congestion should thus have more weight in the aggregation process. A simple yet fairly effective data-driven aggregation method is as follows: given the set of observations for VTL pair i and interval k, $A_{k,i}$

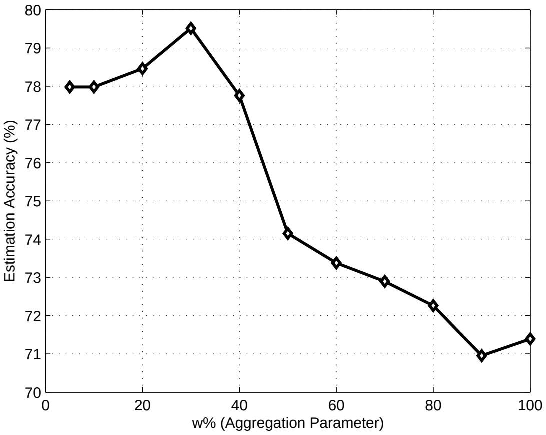


Figure 3: Average estimation accuracy vs. aggregation parameter w.

is sorted such that $t_{m_1} < t_{m_2} \implies X_{t_{m_1},i} > X_{t_{m_2},i}$, then take

$$Z_{k,i} = h_i(\{X_{t_m,i} \,|\, (k-1)t \le t_m < kt\}) := \frac{1}{w.M_{k,i}} \sum_{m=1}^{\lfloor w.M_{k,i} \rfloor} X_{t_m,i},$$

where $M_{k,i}$ is the number of observations in $A_{k,i}$ and $0 < w \le 1$ is the fraction of observations used for aggregation. The symbol $\lfloor a \rfloor$ denotes the floor value of a. In words, the aggregate pace is the mean of the $100 \times w\%$ observations with highest pace or equivalently the worst observations. The simulation results for different values of w are shown in figure 3. From an application-driven point of view, we select the w that maximizes estimation accuracy, in the present case $w = 0.3$. At this value, the travel time envelope of the time series of observations is best captured.

The training phase of logistic regression requires as input a congestion threshold along with the aggregate travel times $Z_{k,i}$. Since the congestion threshold should be chosen to be consistent with the choice of aggregate travel times to provide meaningful classification, we define the congestion threshold, T_i as the mean of the $100 \times w\%$ observations in D_i with highest travel time where D_i is the set of available observations in all intervals and w is essentially be the same value chosen for aggregation ($w = 0.3$ in this section). This corresponds to choosing

$$Q_{k,i} = g_i(\{X_{t_m,i} | (k-1)t \le t_m < kt\}) = I(h_i(\{X_{t_m,i} \,|\, (k-1)t \le t_m < kt\}) > T_i),$$

where $I(\cdot)$ is the indicator function. The STARMA model does not use a g_i function because it forecasts a continuous quantity.

The logistic regression algorithm produces a probability of congestion for each VTL pair studied. If this probability is greater than .5, then the forecasted state is congested. The accuracy of the logistic regression forecasts is defined as the percentage of correctly forecasted states over all intervals and VTL pairs studied. For the STARMA model, the accuracy is defined as the percentage error between the forecasted travel time value and the actual travel time value as defined by the h_i function described earlier.

6.3 Short-Term Forecast

Both regression methods are designed to do one-step (short-term) forecasts. For each data set (as described in section 6.1), the performance of each model was evaluated by dividing the data set into a training set and a validation set. For the Paramics simulation data, the training set consisted of three simulation runs and the validation set consisted of a separate, fourth simulation run. For the New York experiment data, two days of data were used for training and the other day for validation. Through a-priori experimentation, the temporal dependency for the logistic regression model was set to $r = 1$ for the logistic regression, $r = 2$ for the STARMA model. The spatial dependency is varied for comparison in the result figures described in the following paragraph.

The Paramics simulations give information about every vehicle. For testing the methods, only a subset of the data is used for training and inference, corresponding to the penetration rate. This was incorporated into the following analysis by requiring each regression method to produce estimates for the validation data set using only a small percentage of the available travel times. Figure 4 displays the one-step forecast results of the logistic regression and STARMA methods on the Paramics validation set respectively, using a penetration rate of 5%. Similarly, figure 5 displays the one-step forecast results on the New York validation set.

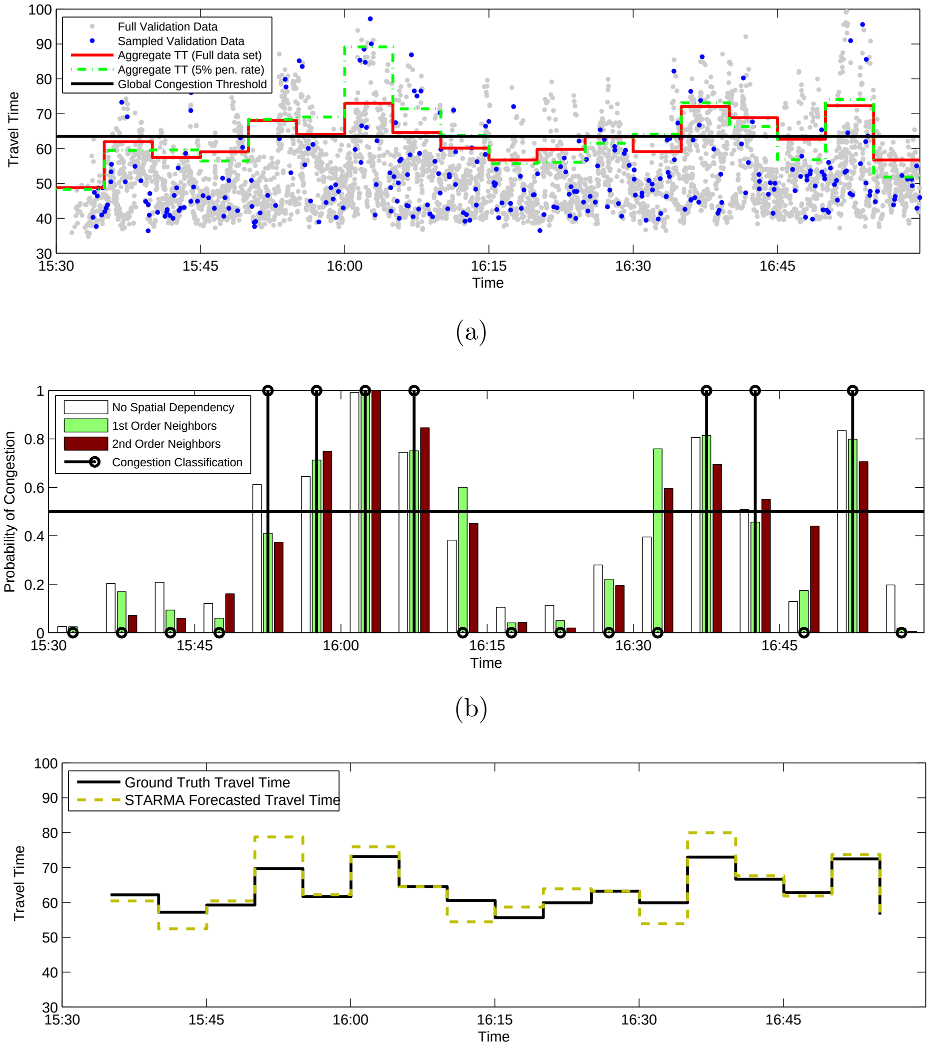
6.4 Penetration Rate Study

The value of 5% for the penetration rate used in section 6.3 was chosen based on the prospects for future adoption of GPS equipped cell phones running traffic information software (such as that provided by *Mobile Millennium*). Therefore, a study of the effect of the penetration rate on results is of interest to quantify the influence of technology adoption on estimation and forecast accuracy. Figure 6 shows the one-step forecast accuracy for the logistic regression and STARMA methods as a function of the penetration rate. From these figures, one can infer that 2% penetration rate can give reasonably good results, while 5% and higher give very accurate results. We also note that using spatial neighbors of order 1 (direct neighbors) generally provides better results. One can interpret this as indicating that second order neighbors lead to an overfit model while no neighbors lead to an underfit model.

6.5 Multi-Step Forecast

The STARMA model is capable of producing forecasts of any number of steps by using the output of the model as input for the next time interval. It is not straightforward to do the same for the logistic regression model since it has an output that is fundamentally different from the input it requires. Therefore, the discrete output of the logistic regression model must be transformed back to a continuous value in order to do forecast in the same way. This avenue is not considered in this article and is left as further research.

In this section, the results of multi-step forecast for the STARMA model are presented. Figure 7 shows the forecast results for the New York data set. The best results for the first step forecast are obtained for an autoregressive temporal order of 1, a spatial order of 2, a moving average temporal order of 1 and a spatial of 1. The two

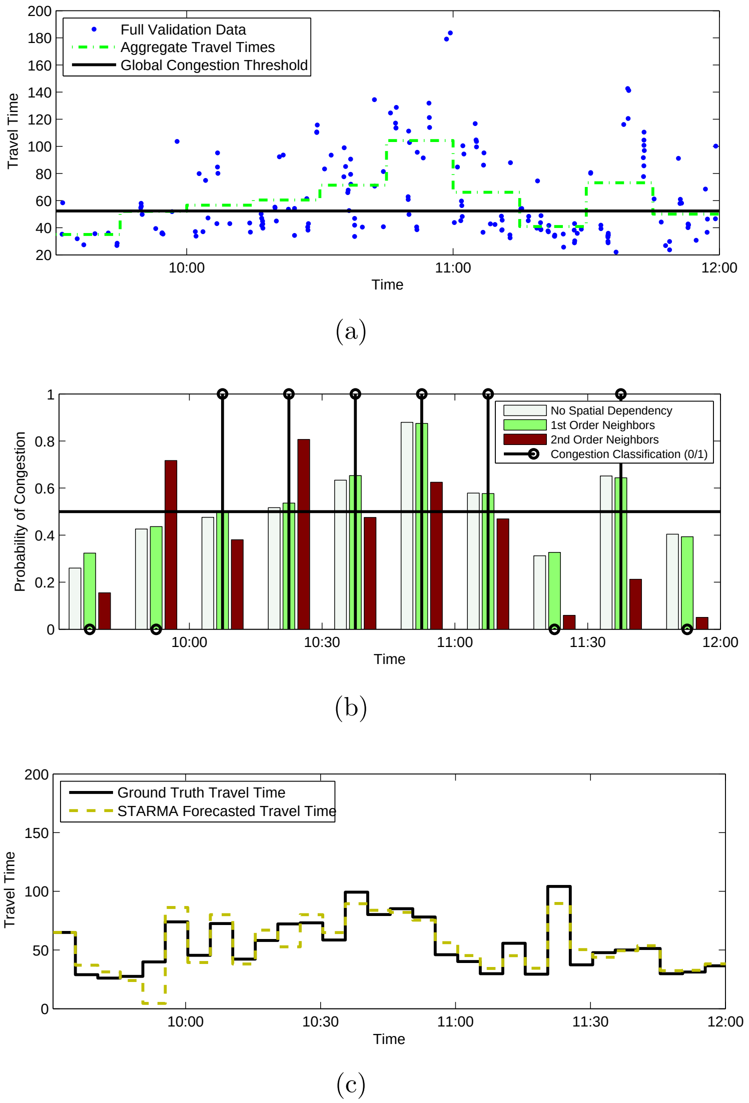


(a)

(b)

(c)

Figure 4: **One-step forecast validation results on a given VTL pair of the Paramics simulation network (penetration rate: 5%)**. (a) Travel time data of the VTL pair and its aggregate value on 5 minutes time intervals. Both the data and the aggregate value are shown for the whole data set and for a 5 % penetration rate. (b) One-step forecast of the congestion state produced by the logistic regression algorithm. The bars represent the probability of congestion estimated by the models for different levels of spatial dependency. The real state of congestion is represented with circles. (c) One-step forecast of travel time produced by the STARMA algorithm.

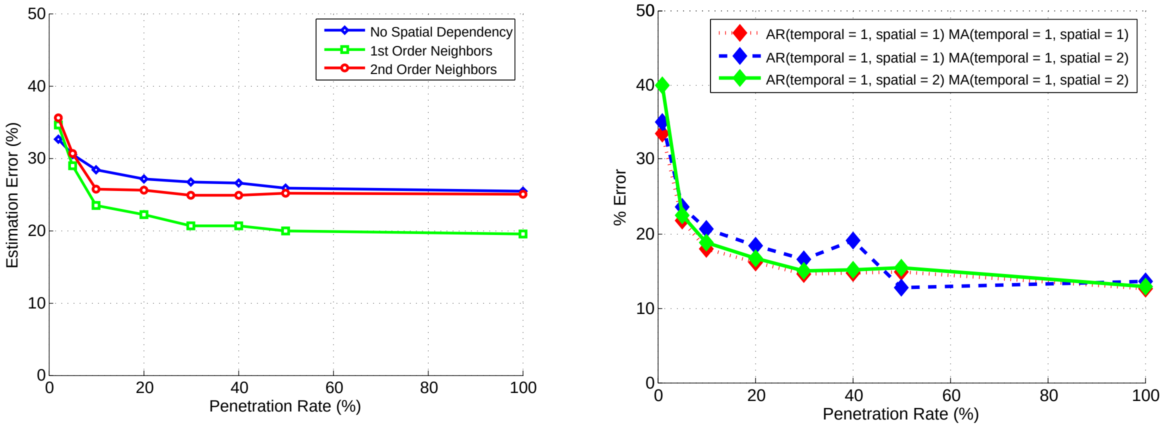


(a)

(b)

(c)

Figure 5: **One-step forecast validation results for logistic regression on one VTL pair of the New York network.** (a) Travel time data of the VTL pair and its aggregate value on 15 minutes time intervals. (b) One-step forecast of the congestion state produced by the logistic regression algorithm. The bars represent the probability of congestion estimated by the models for different levels of spatial dependency. The ground truth state of congestion is represented with circles. (c) One-step forecast of travel time produced by the STARMA algorithm.



(a)

(b)

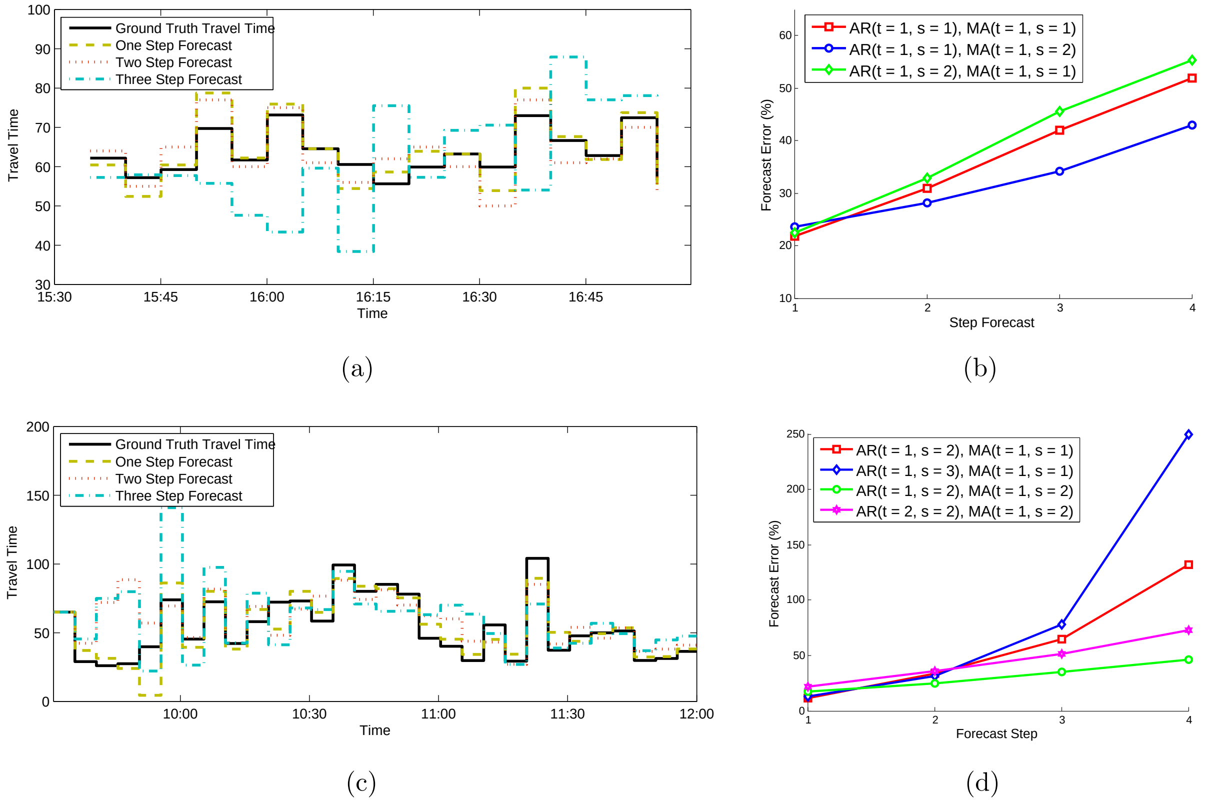
Figure 6: **Average one-step forecast error vs. penetration rate for all VTL pairs in the Paramics dataset.** (a) Logistic Regresion Forecast Classification Error. (b) STARMA Travel Time Forecast Error.

plots for which the moving average temporal and spatial orders are both equal to 1 show the best result for the first step forecast, but the error becomes quickly significant when the forecast step increases. On the other hand, the two other plots for which the moving average orders are one temporally and two spatially show a worse result for the first step forecast but considerably better results for more than one step. The choice of the parameters is therefore a very important step and should take into consideration the performance of the forecasting for more than one step ahead. Analysis of a larger data set is necessary to come to a statistically significant conclusion about the best way to chose the spatio-temporal parameters for the STARMA model.

7 Conclusion

This article presented two statistical learning algorithms for estimating and forecasting arterial traffic conditions on a network. A first implementation inside the *Mobile Millennium* system demonstrates both algorithms' ability to successfully forecast arterial travel times when sufficient training data is available. In summary, this work has achieved the following:

1. It established the validity of a new data collection paradigm on arterial roadways, namely the inter-VTL travel time data collection method for travel time estimation and forecast at low penetration rates.
2. It created data aggregation methods for capturing trends in arterial travel times (functions $h_i(\cdot)$ and $g_i(\cdot)$).
3. It applied logistic regression and STARMA methods for learning spatio-temporal parameters used for estimating arterial link travel times.
4. It validated both models using a training/validation partition of the data, including a Paramics simulation data set and the results from three field tests in New York City.



(a)

(b)

(c)

(d)

Figure 7: **Forecast accuracy.** (a) and (c): Forecast error for a VTL pair in the Paramics and New York networks, respectively. (b) and (d): Average forecast error as a function of the number of forecast steps into the future, Paramics and New York networks, respectively. One step is 5 minutes. In (b) and (d), t represents the temporal dependency and s represents the spatial dependency.

5. It analyzed the effect of penetration rate on forecast accuracy.
6. It analyzed the forecast horizon and its effect on the forecast accuracy.
7. It implemented real-time algorithms inside the *Mobile Millennium* system, public displaying arterial traffic information.

This work is foundational to future research using statistical learning techniques to forecast travel times in urban networks. Extensions to other statistical learning methods beyond logistic regression and STARMA are needed to assess which technique is most appropriate and efficient to the case of arterial travel times. Additionally, there are a number of analyses that could extend the current work. In particular, the following questions are open future research topics:

- Given the segment by segment travel time forecasts, how can accurate forecasts of route travel times be determined?
- The current work requires a full training set on which to operate and needs an aggregated data value (h_i function). How can the data requirements be relaxed while maintaining high accuracy? The goal here is to fill in “missing” data using knowledge of typical traffic patterns.
- How can the results from the specific examples here be generalized to all roads by using common features such as speed limit, number of lanes, number of signals, number of stop signs, etc.? The goal here is to be able to estimate spatio-temporal model parameters in locations where no validation data yet exists.

Those questions are part of the ongoing work in *Mobile Millennium.* Current efforts are focused on giving more accurate forecasts of arterial travel times on a network-wide scale.

Acknowledgements

The authors are grateful to Jeff Ban and Peng Hao for the production of the simulation data used in the present article with Paramics software. They thank Steve Andrews for the logistics of the field tests in New York City used to collect data used in the present article. They also wish to thank Quinn Jacobson, Toch Iwuchukwu and Ken Tracton for their work in setting up the operational system to collect data in New York City.

References

[1] CITRIS, Center for Information Technology Research in the Interest of Society. http://www.citris-uc.org/.

[2] The *Mobile Millennium* Project. http://traffic.berkeley.edu.

[3] TTI, Texas Transportation Institute: Urban Mobility Information: 2007 Annual Urban Mobility Report. http://mobility.tamu.edu/ums/.

[4] X. Ban, R. Herring, P. Hao, and A. Bayen. Delay pattern estimation for signalized intersections using sampled travel times. In *Proceedings of the 88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.

[5] H. Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C*, 15(6):380–391, December 2007.

[6] B.S. Chen, S.C. Peng, and K.C. Wang. Traffic modeling, prediction, and congestion control for high-speed networks: a fuzzy AR approach. *IEEE Transactions on Fuzzy Systems*, 8(5), May 2000.

[7] H. Chen, S. Grant-Muller, L. Mussone, and F. Montgomery. A study of hybrid neural network approaches and the effects of missing data on traffic forecasting. *Neural Computing & Applications*, 10(3), April 2001.

[8] G. A. Davis and N. L. Nihan. Nonparametric regression and Short-Term freeway traffic forecasting. *Journal of Transportation Engineering*, 117(2):178–188, March 1991.

[9] N. Geroliminis and A. Skabardonis. Prediction of arrival profiles and queue lengths along signalized arterials by using a Markov decision process. *Transportation Research Record*, 1934(1):116–124, May 2006.

[10] M. Gonzalez, C. Hidalgo, and A. Barabasi. Understanding individual human mobility patterns. *Nature*, (453):779–782, June 2008.

[11] J.C. Herrera, D. Work, J. Ban, R. Herring, Q. Jacobson, and A. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century experiment. *Submitted to Transportation Research Part C*, December 2008.

[12] B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J. C. Herrera, and A. Bayen. Virtual trip lines for distributed privacy-preserving traffic monitoring. In *The Sixth Annual International conference on Mobile Systems, Applications and Services (MobiSys 2008)*, Breckenridge, U.S.A., June 2008.

[13] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, 5(4):38–46, March 2006.

[14] M. Kamarianakis and P. Prastacos. Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches. *Transportation Research Record*, 1857:74–84, May 2004.

[15] M. J. Lighthill and G. B. Whitham. On kinematic waves. II. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 229(1178):317–345, May 1955.

[16] H. Liu, A. Danczyk, R. Brewer, and R. Starr. Evaluation of cell phone traffic data in minnesota. *Transportation Research Record*, 2086:1–7, December 2008.

[17] I. Nagy, M. Karny, P. Nedoma, and S. Varacova. Bayesian estimation of traffic lane state. *International Journal of Adaptive Control and Signal Processing*, 17(1):51–65, November 2003.

[18] A. Ng. Lecture notes. CS 229: Machine learning. *Stanford University*, 2003.

[19] T. Park and S. Lee. A Bayesian approach for estimating link travel time on urban arterial road network. In *Computational Science and Its Applications ICCSA 2004*, pages 1017–1025. Perugia, Italy, May 2004.

[20] P.E. Pfeifer and S.J. Deutsch. A three-stage iterative procedure for space-time modeling. *Technometrics*, 22(1):35–47, February 1980.

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

[22] A. Skabardonis and N. Geroliminis. Real-time estimation of travel times on signalized arterials. In *Proceedings of the 16th International Symposium on Transportation and Traffic Theory*, University of Maryland, College Park, MD, July 2005.

[23] A. Skabardonis and N. Geroliminis. Real-Time monitoring and control on signalized arterials. *Journal of Intelligent Transportation Systems*, 12(2):6474, March 2008.

[24] B. L. Smith and Smart Travel Laboratory. *Cellphone probes as an ATMS tool.* Center for ITS Implementation Research, University of Virginia, VA, June 2003.

[25] A. Stathopoulos and M. Karlaftis. A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C*, 11(2):121–135, April 2003.

[26] S. Sun, C. Zhang, and G. Yu. A Bayesian network approach to traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 7(1), March 2006.

[27] J.W.C. Van Lint, S. P. Hoogendoorn, and H.J. Van Zuylen. Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C*, 13(5-6):347–369, August 2005.

[28] E. I. Vlahogianni, N. Geroliminis, and A. Skabardonis. On traffic flow regimes and transitions in signalized arterials. In *Proceedings of the 86th TRB Annual Meeting, January, Washington, D.C.*, January 2007.

[29] D. Work, O.P. Tossavainen, S. Blandin, A. Bayen, T. Iwuchukwu, and K. Tracton. An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. In *Proceedings of the 47th IEEE Conference on Decision and Control*, pages 5062–5068, Cancun, Mexico, December 2008.

[30] JL. Ygnace. Travel time estimation on the san francisco bay area network using cellular phones as probes. *California PATH Program, Institute of Transportation Studies, University of California at Berkeley*, September 2000.

[31] Y. Yim and R. Cayford. Investigation of vehicles as probes using global positioning system and cellular phone tracking: field operational test. *California PATH Program, Institute of Transportation Studies, University of California at Berkeley*, February 2001.