HRVOJE MARKOVIĆ, Ph.D.
E-mail: markovic@hrt.dis.titech.ac.jp
Hirota Lab., Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology
G3-49, 4259 Nagatsuta, Midori-ku,
Yokohama-city 226-8502, Japan
BOJANA DALBELO BAŠIĆ, Ph.D.
E-mail: bojana.dalbelo-basic@fer.hr
University of Zagreb,
Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb, Republic of Croatia
HRVOJE GOLD, Ph.D.
E-mail: hrvoje.gold@fpz.hr
University of Zagreb,
Faculty of Transport and Traffic Engineering
Vukelićeva 4, HR-10000 Zagreb, Republic of Croatia
FANGYAN DONG, Ph.D.
E-mail: tou@hrt.dis.titech.ac.jp
KAORU HIROTA, Ph.D.
E-mail: hirota@hrt.dis.titech.ac.jp
Hirota Lab., Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology
G3-49, 4259 Nagatsuta, Midori-ku,
Yokohama-city 226-8502, Japan

Science in Traffic and Transport
Original Scientific Paper
Accepted: June 1, 2009
Approved: Feb. 2, 2010

GPS DATA BASED NON-PARAMETRIC REGRESSION FOR PREDICTING TRAVEL TIMES IN URBAN TRAFFIC NETWORKS

ABSTRACT

A model for predicting travel times by mining spatio-temporal data acquired from vehicles equipped with Global Positioning System (GPS) receivers in urban traffic networks is presented. The proposed model, which uses k-nearest neighbour (kNN) non-parametric regression, is compared with models that use historical averages and the seasonal autoregressive integrated moving average (ARIMA) model.

The main contribution is provision of a methodology for mining GPS data that involves examining areas that cannot be covered with conventional fixed sensors. The work confirms that the method that predicts traffic conditions most accurately on motorways and highways (namely seasonal ARIMA) is not optimal for travel time prediction in the context of GPS data from urban travel networks. In all the examined cases, kNN approach yields a mean absolute percentage error that is twice as good as ARIMA, while in some cases it even yields a mean absolute percentage error that is an order of magnitude better.

The merit of the model is demonstrated using GPS data collected by vehicles travelling through the road network of the city of Zagreb. To evaluate the performance, the models mean absolute percentage error, mean error, and root mean square error are calculated. A non-parametric ranked Friedman ANOVA to test groups of three or more models, and the Wilcoxon matched pairs test to test significance between two models are used. The alpha levels are adjusted using the Bonferroni correction.

Today's commercial fastest-route guidance systems can readily incorporate the proposed model. Since the model yields travel times that are dependent on dynamic factors, these commercial systems can be made dynamic. Furthermore, the model can also be used to generate pre-trip information that will help users to save time.

KEYWORDS

travel time prediction, urban traffic, GPS data, k-nearest neighbour, seasonal ARIMA, non-parametric regression

1. INTRODUCTION

One of the main tasks in today's urban traffic control and route planning systems is to forecast various traffic conditions such as traffic flow, mean speed, and travel time. Travel time prediction has been recognised as one of the most valuable elements, especially for Advanced Traveller Information Systems (ATIS) and Advanced Traffic Management Systems (ATMS) in the context of intelligent transport systems. Since traffic conditions are significantly time-dependent, route guidance systems must be dynamic. For instance, the

routes that have higher speed limits may not be the optimal choice during certain times of day such as rush hour. Dynamic guidance systems try to find the fastest route by using algorithms that generate a travel time that changes according to the trip start time. The most commonly used algorithms are modifications of Dijkstra's shortest-path algorithm [1, 2].

There are many explanatory models that describe traffic conditions: DynaMIT [3], VISUM-online [4], Schreckenberg's cellular automata model [5], and Kerner's jam front propagation model [6]. In addition, there have been many attempts to estimate future conditions using data mining. Some are parametric linear and non-linear regression models [7-10], non-parametric regression models [11], ARIMA models [12], space-time ARIMA models [13-15], ATHENA models [16], Kalman filters [17], artificial neural networks [18-22], and support vector machines [23]. Emerging traffic data collection techniques make these extrapolation-based models easier to use. Older techniques, such as roadside sensors, cannot collect sufficient traffic data on spatially complex traffic networks due to coverage limitations. With the rise of GPS technology, vehicles travelling through road networks collect useful traffic data. Data mining can be used to predict future conditions. While significant work has been performed for motorways and highways, only limited work has been attempted for urban networks, where temporal dependence of travel time is more complex.

The main contributions considered are:

1. The travel times are predicted from GPS data. GPS data help applications cover spatially complex networks, which roadside sensors cannot do. This enables the exploration of large urban traffic networks.
2. A method for mining GPS data is developed. Since predicting travel times is not the primary motivation behind GPS technology, it is found that GPS data must be preprocessed before any travel time methods can be applied. The proposed preprocessing step involves map matching (that is, linking the GPS records with actual digital maps), temporal outlier detection (filtering records with unusually high travel times due to vehicles making stops), and reduction of travel time variability to ensure more accurate forecasts. The preprocessing step reduces travel time variability by using a non-equidistant aggregation interval approach.
3. The non-equidistant aggregation intervals approach is proposed as a novel way of handling the missing values. It enables the usage of GPS data even when coverage is low. This issue is critical when the data are collected from GPS transponders onboard delivery vehicles, as in the studied case.
4. Urban traffic networks are investigated. Urban traffic behaves differently from other types of traffic networks: specifically, travel times can show higher variability and chaotic behaviour [24]. GPS data enable the investigation of urban traffic networks, but both the GPS data and the nature of urban traffic networks (specifically, their volatility) introduce additional issues. The volatility of urban traffic, and appropriate confidence bounds can significantly impact real-time traffic forecasting [25].

   Three fundamentally different methods are used for travel time prediction: the historical averages method, the seasonal Autoregressive Integrated Moving Average (ARIMA) model, and the non-parametric *k*-nearest neighbour (*k*NN) model. The historical averages method is used as a baseline method and can be expected to produce the least accurate results. The seasonal ARIMA model is used because it is referred as the most accurate approach. The non-parametric *k*NN model is used because it is expected to be appropriate for urban traffic networks: that is, it is expected to be able to capture the chaotic behaviour associated with travel times.

   The most suitable method for analysing GPS data and urban traffic networks is identified by exploring the case study data. The forecast performance of the models is investigated by using the mean absolute percentage error, mean error, and root mean square error. Statistical significance between the models is determined by the mean rank for groups of more than two models by using the Friedman, and by Wilcoxon matched pairs test for the groups of two models. Additionally, the Bonferroni correction is performed.

All the analysed data came from 297 courier service vehicles travelling during a period of approximately 6 months (from October 2005 to April 2006) on the roads of the city of Zagreb, the capital of Croatia.

The original reasons to collect the data were to track a fleet of courier service vehicles and to construct and update a digital road map. The motivation in this paper was to examine the possibility of using the data for another application (i.e., to predict travel times). Subsequently, the predicted travel time can be used with fastest-route guidance systems, either en-route (during driving) or to confirm pre-trip information. Because of the original motivation for data collection, the sampling was defined spatially and not temporally. Specifically, sampling was not performed at constant time intervals, but rather using constant spatial intervals (100 metres).

Section 2 explains the data used for the analysis and describes the data preprocessing procedures. Section 3 gives theoretical foundations for the selected methods. Section 4 presents the experimental results.

2. GPS DATA AND SPATIO-TEMPORAL PREPROCESSING

A global positioning system is a positional and navigational system that can be used to determine the location (and speed) of any GPS receiver. GPS data have already been used to estimate traffic congestion [26], to record information about traffic delays and to use this data for traffic monitoring and route planning applications [27].

In the study, GPS receivers produce a tabular log of record time, speed, latitude, longitude, course and GPS status. The record time is the time during which the record is generated. Generally, it is expressed in coordinated universal time UTC (i.e., as the number of seconds from 1.1.1970). The speed is the speed of the vehicle monitored by the GPS receiver in km/h. The latitude and longitude in the WGS84 geodetic system determine the location of the vehicle. The course is the angle at which the vehicle is travelling with reference to the North. Access to the GPS status, which indicates the data accuracy, is also available. A poor GPS status indicates records of questionable accuracy since they are generated from a small number of satellites or in the context of unsuitable satellite configurations. Each record also includes the identification number of the GPS receiver device. As each car has one receiver, this can also be interpreted as a vehicle identification number.

The devices in the vehicles were programmed to transmit information to the servers periodically. If the vehicle is moving, then the GPS device sends a position fix every 100 metres. If the vehicle is stationary, then the data are sent every five minutes. The initial amount of GPS data included 51,835,560 records.

The first step in data preprocessing is to eliminate records that have low GPS status. The second step is to do a map-match of the positions to link records with the appropriate road segments.

2.1 Map matching

Due to the limited accuracy of GPS and constraints on GPS signal reception in the urban environment (for example, multipath signal bouncing), GPS data are normally associated with a measurement error [28]. Both surrounding objects (such as buildings) and atmospheric conditions influence this error.

Collected GPS data feature these measurement errors. Certain data points are off-road even though all vehicles travelled on the roads at all times. To accurately determine vehicle location, these data must be preprocessed to match the trajectory of the vehicle movement to the link that the vehicle travelled through. This technique is known as map matching [28].

The used digital road network is represented in the database by vectors. Onboard systems use information about road networks to map current vehicle positions onto appropriate road segments. These systems represent the vehicle trajectory as a sequence of historical positions. For real-time applications, the task of map matching can be quite time-consuming. Accordingly, in a trade-off between speed and accuracy, entire trajectories are not used, but instead only the most recent positions are used. In addition, if the onboard system is navigational as well as positional, the destination can be known in advance. This information can be used to ensure effective mapping.

Many map matching algorithms are currently in use (for more information, see [28]). The authors did not develop a map-matching algorithm, since one was already available with the data. In the experiments, a map-matching algorithm that was developed by the Mireo Company [29] was used. Originally it was used during the creation of digital maps from GPS data, yielding maps with ±5 metre accuracy in 95% of cases [29].

The most important step in preprocessing stage is to identify the outliers. Outliers are observations that are numerically distant from the rest of the samples. In a sense, map matching can be said to be a process of identifying spatial outliers and correcting their values. After map matching has been done, temporal outliers must be removed. These outliers are sample travel time values that should not be used in the process of modelling.

2.2 Temporal outlier detection

Data used for the analysis were acquired by courier service vehicles that make frequent stops. For that reason, some of the sample travel times have disproportionately higher values than other samples obtained for the same link.

The values that do not follow the characteristic distribution of the data are referred to as outliers. Outliers are not necessarily error values: they can indicate unusual behaviour within the underlying process and highlight anomalies. Identifying outliers is one of the main challenges associated with data mining. In a modelling process, outliers can negatively impact the accuracy of the final model. Specifically, in a regression analysis, where the sums of the squares of the distances are minimised to form a model, outliers can significantly influence the regression line. Because of this, outliers must be detected before developing a model for travel time prediction.

One of the most widely used methods to detect outliers is a Box Plot technique, which has already been applied in travel time prediction [30-31]. Bajwa et al. used the technique on highway data (Tokyo Metropoli-

tan Expressway) to reduce variability and achieve higher estimation accuracy. They used the 25th percentile as the lower quartile and the 75th percentile as the upper quartile, and the interquartile range to model lower and upper boundaries. They used 1.5 times the interquartile range to define lower and upper boundaries (that is, they used inner fences). For the experiments described in this paper, to make sure that the modelling stage receives more data, outer fences (that is, three times the interquartile range) are used. The boundaries are defined as:

$$\textit{lower boundary} = \\ = \textit{lower quartile} - 3\,(\textit{upper quartile} - \textit{lower quartile}),$$

and

$$\textit{upper boundary} = \\ = \textit{upper quartile} + 3\,(\textit{upper quartile} - \textit{lower quartile}),$$

where

lower quartile is the 25th percentile and *upper quartile* is the 75th percentile.

Every sample time below the lower boundary and above the upper boundary is marked as an outlier and is excluded from our modelling of travel time.

2.3 Reduction of travel time variability

There are two types of temporal travel time variability: short- and long-term variability. Short-term variability of vehicle travel time is the result of traffic signal phases in urban networks. Long-term variability is the result of evolving traffic patterns during the day (i.e., congestion). While preserving long-term variability is crucial, a reduction of short-term travel time variability plays a key role in our ability to accurately estimate travel time.

Torday and Dumont [32] have used microsimulations and the floating car data technique to show how to reduce short-term variability in urban networks using appropriate sub-link definitions. However, this approach is not used, because the amount of data does not allow it. Using simulations, they have also shown that minimising the aggregation interval reduces variability [32]. They suggested that the aggregation period should be a multiple of the duration of the traffic lights signal switch periods.

As a result of the small sample size, it is not possible to model all the effects that may be present in analysed travel time data. Therefore, the main motivation is to bind the long term variability caused by congestion. Luckily, if there are two intervals with the same duration but at different times of day, the one compiled during congestion will feature more samples. The other may not even have a single sample (e.g., on Sundays or during the night, when traffic moves more fluidly).

This reality inspired the use of a non-equidistant aggregation intervals approach. Experiments with different durations and placements of time intervals are performed, and finally, the following settings are selected. There are several intuitive reasons for such day-part divisions. Intuitively, during the night (i.e., from 20:00 to 06:00) there is no congestion and all the samples can be aggregated into a single value; from 06:00 to 10:00 when congestion is severe, aggregation intervals are set to 15 minutes; between 10:00 and 15:00, to take into account medium term variability during the day, aggregation intervals of 1 hour are used; from 15:00 to 18:00 aggregation periods are again 15 minutes; and finally from 18:00 to 20:00 aggregation is performed in 1-hour intervals. Since there are very few data for Saturdays and Sundays (the courier service makes few weekend deliveries), a 24-hour aggregation interval is used. In cases when there was an aggregation interval that contained zero samples, the travel time duration was set to equal the median travel time of the corresponding link.

Since the investigated regression methods require time series with a fixed time step, all the aggregation intervals are divided into equally sized segments, i.e., the size of the shortest time interval (15 minutes). For instance, if the duration of a certain aggregation interval is one hour, it is broken into four 15-minute intervals for which the time series value is the same as for the original aggregation interval.

3. TRAVEL TIME PREDICTION METHODS

Although there are a number of methods to predict traffic conditions, most of the work has been performed on data collected using roadside sensors [11, 30, 31]. Additionally, most such research is concerned with motorway traffic and not with urban traffic conditions. Given the nature of urban networks and the reality that all of the analysed travel time data were collected by vehicles equipped with GPS receivers, three methods to predict travel times are selected: the historical averages method, the seasonal ARIMA and the non-parametric *k*NN model.

The historical averages method is used as a baseline method. It is used because of the issues with sample size. It is challenging to acquire large sample sizes for every aggregation period, for each past date and for every link in a large-scale urban network.

Since the literature states that the seasonal ARIMA model outperforms neural networks and *k*NN non-parametric regression [11, 12], it is included among the implemented models. On the other hand, it is questionable whether the seasonal ARIMA model would give satisfactory results, given that it models processes that are non-deterministic with linear state transitions. Disbro and Frame [24] showed that traffic flow behaves chaotically, especially in cases frequently found in urban traffic networks (i.e., during congestion

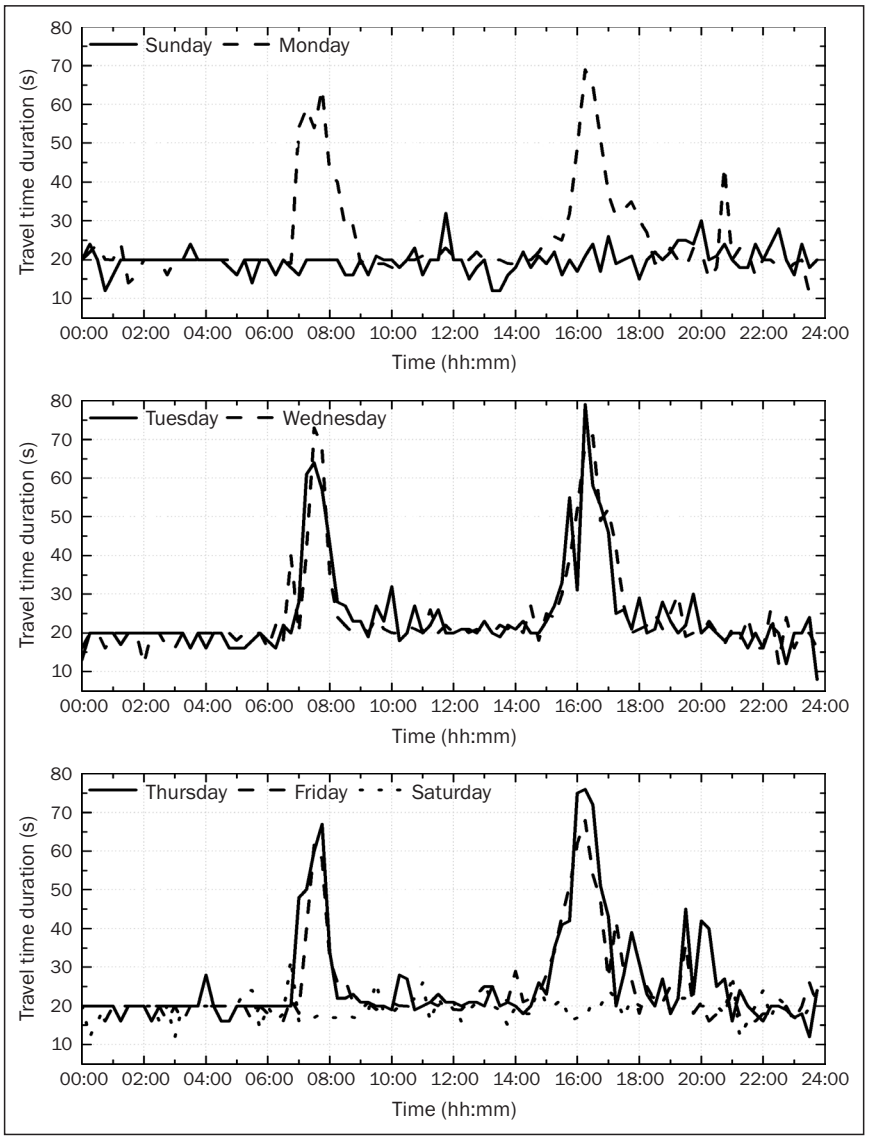

Figure 1 - Weekday profiles of an exemplar link

periods). Given that chaotic systems are described by processes that are deterministic and feature non-linear state transitions, it motivated the use of the non-parametric *k*NN model.

3.1 Historical averages method

The historical averages method is a very simple model in which every weekday is one case that can be described by a series of aggregated values. *Figure 1* shows a graph in which every weekday is described by an appropriate time series.

Time series are stored as records in the database. Each record has a weekday attribute defining the day of the week, a timestamp defining 15-minute periods during the day, and a duration giving the average travel time for a given link. The user supplies the values for the required link identifier, weekday and time of day. The returned value denotes the predicted link travel time.

3.2 Seasonal ARIMA model

The seasonal ARIMA model was proposed by Box and Jenkins [33, 34] for analysing time series $\{X_t\}$. In order to define a seasonal ARIMA process formally, the backshift operator B with order of differencing j is used to transform the time series: $B^j X_t = X_{t-j}$. The seasonal differencing with seasonal period s is given as:

$$X_t - X_{t-s} = X_t - B^s X_t = (1 - B^s) X_t. \quad (1)$$

The seasonal differencing with seasonal period s and order of seasonal differencing D is written as $(1 - B^s)^D X_t$. In terms of the backshift operator, the non-seasonal differencing is defined in a similar manner as $X_t - X_{t-1} = (1 - B) X_t$, and the non-seasonal differencing of order d as $X_t - X_{t-d} = (1 - B)^d X_t$.

The time series $\{X_t\}$ is called a seasonal ARIMA (p, d, q)(P, D, Q)$_s$ process if for any non-negative integers d and D, the differenced series $Y_t = (1-B)^d(1-B^s)^D X_t$ is a stationary autoregressive moving average (ARMA) process satisfying:

$$\varphi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)e_t. \quad (2)$$

The backshift operator B is given by Eq. (1). The functions φ, Φ, θ, and Θ are polynomials defined as

$$\varphi(z) = 1 - \sum_{i=1}^{p}\varphi_i z^i,\ \Phi(z) = 1 - \sum_{i=1}^{P}\Phi_i z^i,$$

$$\theta(z) = 1 - \sum_{i=1}^{q}\theta_i z^i, \text{ and } \Theta(z) = 1 - \sum_{i=1}^{Q}\Theta_i z^i,$$

respectively. The coefficients φ_i, Φ_i, θ_i, and Θ_i are unknowns and have to be found. The time series $\{e_t\}$ corresponds to errors, known as white noise that can be found from standard ARIMA(p,q) model

$$X_t = e_t + \sum_{i=1}^{p}\varphi_i X_{t-i} + \sum_{i=1}^{q}\theta_i e_{t-i}$$

having unknown coefficients φ_i and θ_i. Errors e_t are identically and normally distributed with mean zero, variance σ^2 and $\text{cov}(e_t, e_{t-k}) = 0, \forall k \neq 0$, in other words, $\{e_t\} \sim \text{N}(0, \sigma^2)$. Additionally, p is the non-seasonal autoregressive order, q is the non-seasonal moving average order, P is the seasonal autoregressive order and Q is the seasonal moving average order. Similarly, d is the order of non-seasonal differencing and D is the order of seasonal differencing.

Background on the theoretical seasonal ARIMA process and its usage in forecasting traffic conditions is also given by Williams and Hoel [35] and by Smith et al. [11].

3.3 *k*NN non-parametric regression

In the previous chapter, a linear parametric model (ARIMA) was introduced. There, the main idea was to form a model that could satisfactorily approximate the entire instance space. On the other hand, in instance-based learning, as represented by the *k*NN model, only a local approximation of the target function that applies in the neighbourhood of the new forecast instance needs to be constructed [36]. For that reason, there are no restrictions on the data being modelled (specifically, there is no requirement for stationarity, unlike in the ARIMA model). The model consists of past (historical) values that are stored and subsequently used to determine the values for new instances.

An arbitrary instance x is represented by attributes (or features) denoted as a_i, i = 1, 2, ..., n, and its feature vector or state space is $[a_1, a_2, a_3, \ldots, a_n]$. The observed instance can then be viewed as a point in an n-dimensional space represented by the values of the attributes $a_i(x)$, i = 1, 2, ..., n

$$x := (a_1(x), a_2(x), a_3(x), \ldots, a_n(x)). \quad (3)$$

Each training sample has a known target function value $f(y_i)$ and can be written as:

$$(y_i, f(y_i)),\ i = 1,2,\ldots,N, \quad (4)$$

The k-nearest neighbour forecasting can then be defined as follows. Given the test point x and N training samples $(y_i, f(y_i))$, $i = 1,2,\ldots,N$, find the k nearest training inputs $y_1, y_2, \ldots, y_k$ to x with respect to the given distance function $d(x, y_i)$, $i = 1,2,\ldots,N$.

If the forecasting is performed for a regression problem then the forecast value is

$$\hat{f}(x) = h(g_1(f(y_1), d(x,y_1)), g_2(f(y_2), d(x,y_2)), \ldots, \\ , g_k(f(y_k), d(x,y_k))) \quad (5)$$

If $g_i(f(y_i), d(x,y_i)) = f(y_i)$, $i = 1,2,\ldots,k$ and h is a simple average function then the regression problem forecasting is given by Eq. (6).

$$\hat{f}(x) = \frac{1}{k}\sum_{i=1}^{k} f(y_i) \quad (6)$$

Another frequently used simple model is appropriate if $g_i(f(y_i), d(x,y_i)) = f(y_i)/d(x,y_i)$, $i = 1,2,\ldots,k$ and h is defined as a quotient of the sum of $g_i(f(y_i), d(x,y_i))$, $i = 1,2,\ldots,k$ and the sum of the inverses of the distances $d(x,y_i)$, $i = 1,2,\ldots,k$. In this case, the regression problem forecasting is given by Eq. (7).

$$\hat{f}(x) = \sum_{i=1}^{k}\frac{f(y_i)}{d(x,y_i)} \Bigg/ \sum_{i=1}^{k}\frac{1}{d(x,y_i)} \quad (7)$$

The distance between the test point and the training sample y_i, $i \in [1, N]$ is determined by the standard Euclidean distance

$$d(x, y_i) = \sqrt{\sum_{r=1}^{n} \left| a_r(x) - a_r(y_i) \right|^2},$$

where n is the dimensionality of the state vector, and $a_r(x)$ and $a_r(y_i)$ are features of the test point and the training sample, respectively.

Distance metrics can also be weighted in such a way that some features contribute more or less to the overall distance. There is an infinite number of distance metrics and the standard Euclidean distance is chosen as the measure of the distance between the instances for the purpose of forecasting travel time.

Different state vectors can be used for the *k*NN regression. More precisely, there is an infinite number of possible state vectors. The most reasonable features to be used are present and time-lagged values of the time series $x(t) = [V(t), V(t-1), V(t-2), \ldots, V(t-d)]$ where d is the selected lag. However, in forecasting traffic flow, Smith et al. [11] have shown that using past average values yields more accurate forecasts. They used a hybrid model $x(t) = [V(t), V(t-1), V(t-2), V_{\text{hist}}(t), V_{\text{hist}}(t+1)]$. If their traffic flow is considered in the context of travel time, then $V(t)$ is the travel time at the present interval and $V_{\text{hist}}(t)$ and $V_{\text{hist}}(t+1)$ are the historical average travel times for the weekday and the time of day associated with time t. There is a sound justification for the use of past average values. The attractor of the

chaotic system is the value to which the system settles when time approaches infinity. This occurs as the kNN approach tries to rebuild the attractor of the process that generates the time series [37] and the average of past values puts each instance on the cyclic pattern of the attractor. Various state spaces are investigated, and Section 4 shows the results. The mean absolute percentage error (MAPE - for the formal definition see Section 4) is used to determine which state space should be used and to determine the state space that produces the lowest MAPE for forecasting purposes.

The required number of neighbours k must be determined experimentally. This is done by determining the MAPE for models with different numbers of neighbours and selecting the one with the lowest value for forecasts. A small number of neighbours could have too much variance and could result in loss of generality, while too large number of neighbours could introduce too much bias into the forecast [38].

In the context of regression analyses, there is an infinite number of possible forecast estimations. The most common ones are straight averages (Eq. (8)) and averages that are weighted by the inverse of the distance (Eq. (9)). Other forecasts include heuristics to assure more accurate estimates. While forecasting traffic flow, Smith et al. [11] obtained the best kNN forecasts with the hybrid approach, which adjusts by both $V_{\text{hist}}(t)$ and $V_{\text{hist}}(t+1)$, and weights by the inverse of distances (Eq. (10)). Again, to find the most accurate forecast estimation, MAPE is used.

$$\hat{V}(t+1) = \frac{1}{k}\sum_{i=1}^{k} V_i(t+1) \tag{8}$$

$$\hat{V}(t+1) = \sum_{i=1}^{k} \frac{V_i(t+1)}{d_i} \Bigg/ \sum_{i=1}^{k} \frac{1}{d_i} \tag{9}$$

$$\hat{V}(t+1) = \sum_{i=1}^{k} \frac{V_i(t+1)\frac{1}{2}\left(\frac{V_c(t)}{V_i(t)} + \frac{V_{hist,c}(t+1)}{V_{hist,i}(t+1)}\right)}{d_i} \Bigg/ \sum_{i=1}^{k} \frac{1}{d_i} \tag{10}$$

In Eqs. (8) – (10) k is the number of nearest neighbours, $\hat{V}(t+1)$ is the forecast time series value at time $t+1$ (corresponding to the forecast value introduced in Eq. (5)), $V_i(t+1)$ is the time series value of the i-th nearest neighbour at time $t+1$ (corresponding to the known function value $f(y_i)$ introduced in Eq. (4)), d_i is the Euclidean distance between the instance being forecast and the i-th nearest neighbour, $V_c(t)$ is the current time series value, $V_{hist,c}(t+1)$ is the historic average value for the estimated time series value, aggregated by weekday and time of the day with respect to $t+1$, and $V_{hist,i}(t+1)$ is the historic average value for the i-th nearest neighbour time series value aggregated by weekday and time of the day with respect to $t+1$.

4. CASE STUDY: URBAN NETWORK OF THE CITY OF ZAGREB

A set of data collected from 1 October 2005 to 21 April 2006 has been studied. The data are divided into two groups. The first group contains data from 1 October 2005 to 7 April 2006, and is used to develop the model. The other group contains data from 7 April 2006 to 21 April 2006, and is used to evaluate the model. It should be noted that the validation group contains only two weeks' worth of data. Two weeks are chosen because this time period corresponds to two seasonal lags in the obtained ARIMA model. Non-equidistant time intervals are used to average the travel time. The final time series resolution is 15 minutes. To develop the model and to test its performance, 20 random links out of the 100 links with the greatest number of matched records are selected. *Table 1* gives descriptive statistics for the links used to build the models. Furthermore, Section 4.6 presents results for four additional links used to illustrate the evaluation process for the model.

4.1 Forecast performance measures

The measures used for the model's forecast performance are: mean absolute percentage error—MAPE

$$\left(MAPE = 1/n\sum_{i=1}^{n}\left|(A_i - F_i)/A_i\right|\right),$$

mean error—ME

$$\left(ME = 1/n\sum_{i=1}^{n}(A_i - F_i)\right),$$

and root mean squared error—RMSE

$$\left(RMSE = \sqrt{1/n\sum_{i=1}^{n}(A_i - F_i)^2}\right),$$

where n is the number of samples, A_i is the known (observed) value of the i-th sample, and F_i is the forecast value of the i-th sample. MAPE is used to estimate the size of the forecasting error, ME is used to determine whether the forecasts are biased, and RMSE is used to determine whether the error distribution features outliers.

Table 1 - Descriptive statistics for the links used to build the models.

	Mean	Minimum	Maximum	Standard deviation
Length of the link (m)	526.10	192	1152	278.877
Median travel time (s)	32.75	12	60	15.033
Upper-outlier boundary for the travel time (s)	93.60	36	220	56.351
Number of matched records	21852.65	13596	44467	8003.862

Although MAPE gives guidance as to which method might be better, it does not offer any statistical confidence. For that, non-parametric Friedman ranked ANOVA [34] tests whether there is a significant difference in absolute percentage errors between the methods. For every forecast point and for every method, the absolute percentage error is calculated. The H_0 hypothesis is that the medians of the errors for all the methods are equal. If the α -value is small enough (for all cases <0.05), then there is evidence that the H_0 hypothesis can be rejected. Similarly, to test the difference between two methods, the Wilcoxon matched pairs test is performed on the absolute percentage errors. Additionally, Bonferroni correction adjusts the alpha values.

4.2 Historical averages results

Table 2 gives the results from the historical averages model. For all 20 random links, the mean MAPE equals 0.1738, which is relatively good, but the maximum MAPE of 0.3409 suggests that for certain links, this method performs unsatisfactorily. Maximum values of ME (9.0364 s) and RMSE (26.3311 s), show that for some links this method is both biased and sensitive to extreme values. Overall, the historical averages model results show that there are some effects that cannot be modelled as time-of-day and day-of-the-week dependencies.

Table 2 - MAPE, ME and RMSE calculated by historical averages.

	Mean	Minimum	Maximum	Standard deviation
MAPE	0.1738	0.0727	0.3409	0.0763
ME (s)	0.2494	-2.5352	9.0364	2.4463
RMSE (s)	7.9026	2.2388	26.3311	6.2410

4.3 Seasonal ARIMA results

Using Box and Jenkins procedure [33, 34], travel time is determined to be an $\text{ARIMA}(1,0,1)(0,1,1)_{672}$ process that matches the results obtained by Smith et al. [11] for traffic flow forecasting. A seasonal lag of 672 corresponds to one week, because one week encompasses 672 15-minute intervals. From the investigations of all 20 random links, travel time is described by the same $\text{ARIMA}(1,0,1)(0,1,1)_{672}$ model. Table 3 lists the results.

Table 3 - MAPE, ME and RMSE obtained for the XXX_FORMULA_XXX model.

	Mean	Minimum	Maximum	Standard deviation
MAPE	0.1632	0.0096	0.3362	0.0882
ME (s)	0.1315	-3.3806	7.6772	2.9892
RMSE (s)	7.0932	2.3725	19.6569	5.1591

The minimum MAPE of 0.0096 suggests that some links can be modelled quite accurately by seasonal ARIMA. The maximum MAPE of 0.3362, however, suggests that for some links, seasonal ARIMA may not be the most suitable model. The mean value (0.1315 s) and standard deviation (2.9892 s) of ME suggest that forecasts for 20 random links are, in general, not strongly biased.

4.4 *k*NN non-parametric regression results

Experiments with a range of lagged values in state spaces and with different numbers of neighbours are performed. The simulations include lag values from 0 to 10 and from 1 to 30 nearest neighbours. Additionally, straight averages, weightings by the inverse of distance, and a hybrid state space are also used. In total, for all 20 random links and all possible *k*NN parameters, 11×30×3×20 = 19800 executions are performed each for two weeks of 15 minute data (i.e.,

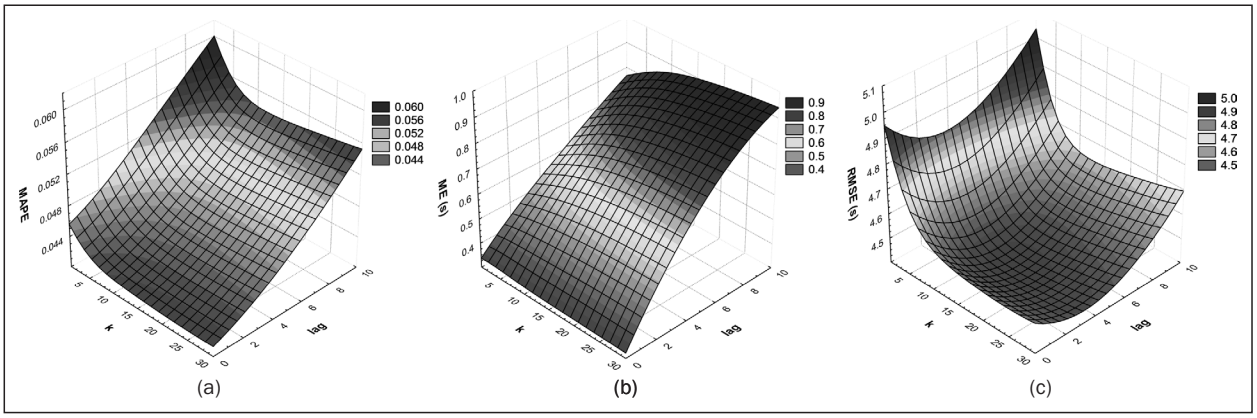

Figure 2 - Dependence of MAPE (a), ME (b), and RMSE (c) on the number of lagged values (lag) and the number of nearest neighbours (k) for kNN in the context of a hybrid state space. The 3D surface plot is generated with the use of a distance-weighted least square fit.

1344 forecasting points). It is found that *k*NN with a hybrid state space yields smaller MAPE values.

Figure 2 shows the dependence of MAPE, ME and RMSE on the number of lagged values in the state space and the number of neighbours when a hybrid state space is used. It can be seen that, generally, a high number of lagged values results in higher MAPE and ME in a manner that is independent of the number of neighbours, while generally, a low number of neighbours results in higher RMSE values.

The *k*NN model with the smallest MAPE values is proposed as the preferred model. Specifically, *k*NN with a hybrid state space, with one lagged value (lag=1), and 26 neighbours (k=26) is proposed. Table 4 gives the results for all 20 random links, obtained using the proposed *k*NN model. The mean (0.0218), maximum (0.0423), and minimum (0.0018) values of MAPE show that the proposed *k*NN performs very well for all 20 random links. Moreover, low values of ME and RMSE indicate that *k*NN forecasts are neither biased nor sensitive to extreme values.

Table 4 - MAPE, ME and RMSE for our proposed kNN model.

	Mean	Minimum	Maximum	Standard deviation
MAPE	0.0218	0.0018	0.0423	0.0119
ME (s)	0.5990	0.0301	2.6035	0.7240
RMSE (s)	3.0388	0.8030	11.8426	3.1187

4.5 Forecasting performance of the models

For all 20 random links, the results obtained across all models are compared. For all the links, with respect to ranked Friedman ANOVA, the null hypothesis (that the medians of the errors for all the models are equal) is rejected. Additionally, the Wilcoxon matched pairs test is performed. This result is used to determine inter-group differences in means. The Bonferroni correction of α-value is performed and this results in an α-value of 0.016666667. For 5 links out of 20, the Wilcoxon matched pairs test null hypothesis at both the original and the Bonferroni-corrected α significance level cannot be rejected. For all five of these links, the null hypothesis for historical averages and the seasonal ARIMA model cannot be rejected. For historical averages and the proposed *k*NN model, as well as for the seasonal ARIMA and proposed *k*NN models for all 20 random links, the hypothesis at the Bonferroni corrected α significance level can be rejected.

Figure 3 (a) shows the MAPE for 20 random links, as well as the mean for all the links obtained with historical averages, seasonal ARIMA and the proposed *k*NN model. It can be seen that the proposed *k*NN yields a lower MAPE for all the links. The maximum MAPE is 0.04229. In most cases, ARIMA yields lower values than historical averages do, and this approach reaches a maximum value of 0.3362 while the historical average model reaches a maximum value of 0.3409 for MAPE. *Figure 3 (b)* gives the calculated mean Friedman rank for 20 random links, and the mean rank for performance on all links, with respect to historical averages, the seasonal ARIMA and the proposed *k*NN model. For the links where the Wilcoxon matched pairs test null hypothesis cannot be rejected, the obtained α-values are shown. For all the examined cases, the proposed *k*NN yields the lowest mean rank. Additionally, when compared to the other two methods using the Wilcoxon matched pairs test, the null hypothesis can be rejected.

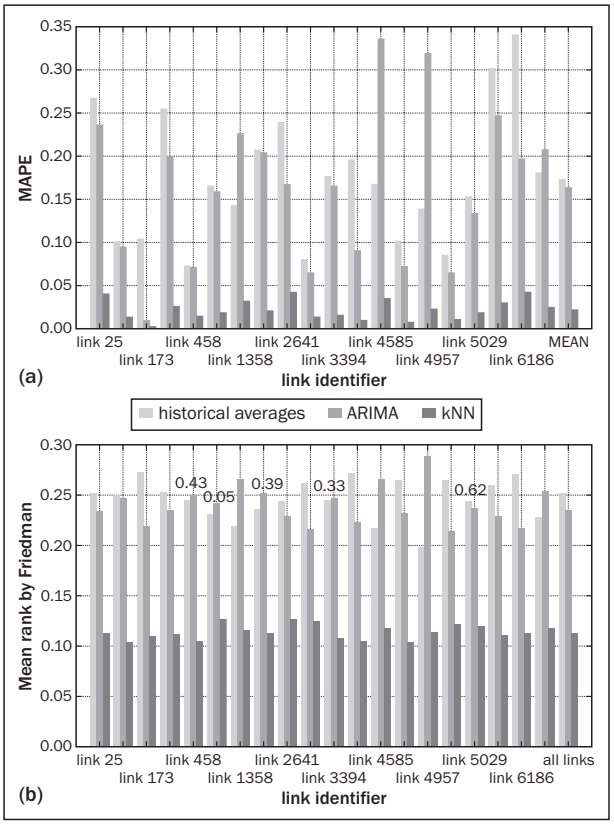

Figure 3 - Comparison of MAPE values (a) and mean Friedman rank (b) as generated using historical averages, seasonal ARIMA, and the proposed kNN model

Figure 4 shows the ME and the RMSE for examined links obtained with historical averages, seasonal ARIMA and the proposed *k*NN model. It can be seen that the proposed *k*NN in some cases yields a higher ME than both historical averages and the seasonal ARIMA model. However, the maximum ME values for both historical averages and the seasonal ARIMA model are higher than the maximum ME for the proposed *k*NN. Overall, for the proposed *k*NN, the ME is positive for all the examined cases, but its absolute value is never more than three seconds. In addition, in all of these cases, the proposed *k*NN model yields lower RMSE values.

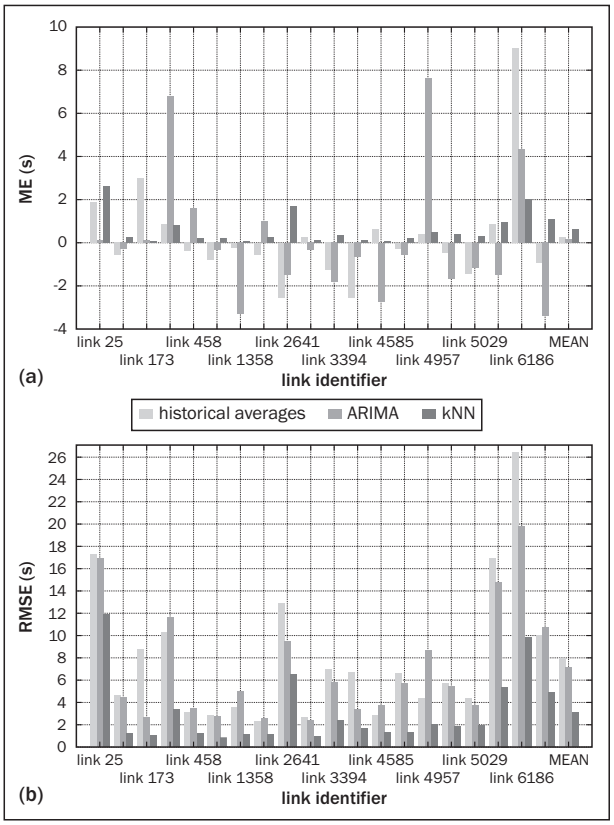

Figure 4 - Comparison of ME (a) and RMSE (b) as generated using historical averages, seasonal ARIMA, and the proposed kNN model

4.6 Evaluation of the model on selected cases

To evaluate the proposed model, four selected links are used. They are shown in *Figure 5*. These links are chosen because they are elements of roads in different parts of the city, each with different characteristics. Link 4562 (a) is a section of a bridge. It has only one input and one output connecting link. Link 2619 (b) has more than one dominant output link, so it represents the opposite situation. Other links, link 2775 (c) and 947 (d), are somewhere between those two extreme cases. Link 4562 is the only link, out of the selected 4, that is one of the aforementioned 20 random links.

Table 5 lists the properties of the selected links, while *Table 6* shows the results. The results are presented for historical averages, seasonal ARIMA, best-performing *k*NN and the proposed *k*NN model. Again, to find the best-performing *k*NN model for a given link, from 0 to 10 lagged values, from 1 to 30 neighbours, and the straight average, weighted by inverse of distance, and a hybrid state space are used. The main purpose of this experiment is to determine how similarly the proposed *k*NN performs to the optimal *k*NN for a given link. *Table 6* gives the mean rank in the context of Friedman, MAPE, ME and RMSE values. In all cases, the null hypothesis with respect to both the ranked Friedman ANOVA and the Wilcoxon matched pairs test at the Bonferroni corrected α significance level is rejected.

For all four selected links, the proposed *k*NN performs better than both historical averages and the seasonal ARIMA model with respect to the mean rank according to Friedman, MAPE, and RMSE. For link 2619, the proposed *k*NN results in an ME higher than the one for ARIMA, but the acquired MAPE is almost seven times lower. In addition, a substantially lower RMSE can be observed.

When the proposed *k*NN is evaluated against the best-performing *k*NN, it can be seen that the differences for all four links with respect to mean rank, MAPE, ME, and RMSE are relatively small. The greatest difference is for link 2775, where the proposed *k*NN gives a 3.4 % greater MAPE than the best-performing *k*NN. However, it is still 4 % lower than the MAPE associated with the seasonal ARIMA model.

5. CONCLUSION

One of the main tasks of this paper is to define a model that can predict spatio-temporally dependent travel times for urban road networks from GPS data used for automatic digital road map creation. The majority of the work presented in the literature on travel time prediction has been performed on data collected using roadside sensors and other techniques. Additionally, the data collected to date have focused on motorways. In this paper, a model based on GPS data

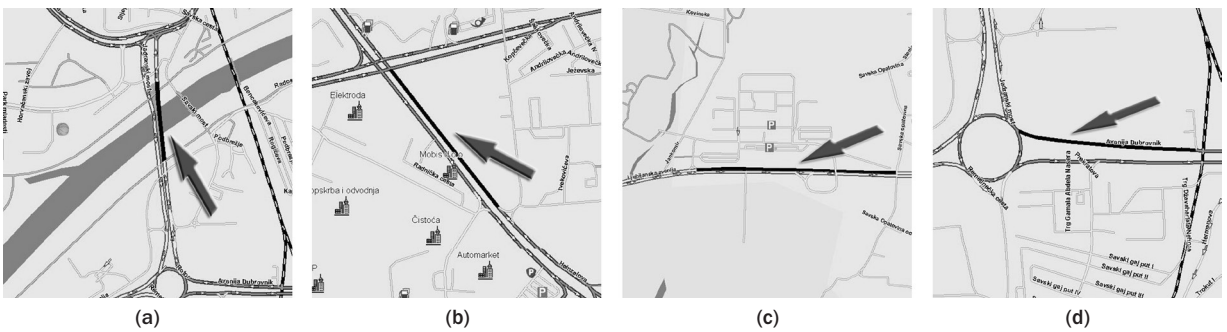

Figure 5 - Location of link 4562 (a), link 2619 (b), link 2775 (c), and link 947 (d)

Table 5 - Properties of the links used to evaluate the models

Link identifier	Name of the corresponding street	Length of the link (m)	Number of matched records	Median travel time (s)	Upper-outlier boundary for the travel time (s)
4562	Jadranski most (bridge)	339	17331	20	64
2619	Heinzelova Street	429	15619	96	364
2775	Ljubljana Avenue	1101	55366	60	132
947	Dubrovnik Avenue	396	22678	32	120

Table 6 - Mean Friedman rank and MAPE, ME and RMSE for selected links as given by historical averages, seasonal ARIMA, the proposed kNN and the best performing kNN

Method	Mean rank	MAPE	ME (s)	RMSE (s)
Link 4562				
Historical averages	3.6830	0.1962	-2.5343	6.5885
ARIMA (1, 0, 1)(0, 1, 1)$_{672}$	3.2106	0.0907	-0.6663	3.3118
Best performing *k*NN (hybrid state space, lag=1, k=8)	1.5164	0.0054	0.1176	1.0743
Proposed *k*NN (hybrid state space, lag=1, k=26)	1.5900	0.0092	0.0919	1.6198
Link 2619				
Historical averages	3.5599	0.3941	7.2405	42.4838
ARIMA (1, 0, 1)(0, 1, 1)$_{672}$	3.2299	0.3322	4.2783	38.0210
Best performing *k*NN (hybrid state space, lag=0, k=30)	1.4780	0.0433	6.3643	29.2083
Proposed *k*NN (hybrid state space, lag=1, k=26)	1.7321	0.0497	7.4477	30.1610
Link 2775				
Historical averages	3.6990	0.1623	9.1623	14.0586
ARIMA (1, 0, 1)(0, 1, 1)$_{672}$	2.9100	0.0919	4.9754	11.8567
Best performing *k*NN (weighted by inverse of distance with non-hybrid state space, lag=1, k=2)	1.3947	0.0176	0.7316	4.7470
Proposed *k*NN (hybrid state space, lag=1, k=26)	1.9963	0.0520	3.7704	7.7468
Link 947				
Historical averages	2.9851	0.2236	-1.1105	11.8424
ARIMA (1, 0, 1)(0, 1, 1)$_{672}$	3.4368	0.1648	2.3306	11.5170
Best performing *k*NN (hybrid state space, lag=0, k=24)	1.7094	0.0335	1.8596	8.0602
Proposed *k*NN (hybrid state space, lag=1, k=26)	1.8687	0.0423	2.1502	8.4637

collected for urban road networks is presented. In this framework, methods for preparing GPS data for modelling, map matching, outlier detection and reducing travel time variability are demonstrated. The non-equidistant aggregation intervals approach is implemented to handle insufficient GPS data coverage.

Three different travel time prediction methods are investigated and implemented. The most basic method, the historical averages method is used only for reference, and, as expected, it produced very poor results. The seasonal ARIMA model and the *k*NN models are the other two methods that are explored. Seasonal ARIMA is used because the available literature commonly presents it as the most suitable approach for the prediction of traffic conditions on motorways and dual carriageways. Since there are some effects that are typical for urban networks, seasonal ARIMA was expected not be the most suitable method for this type of data. Surprisingly, in all the examined cases, *k*NN proved to be the most accurate method.

All the analysed links are part of the urban traffic network of Zagreb. The proposed kNN model is determined by analysing 20 random links out of the 100 links that featured the greatest data coverage. Then the proposed model is evaluated on four selected links. Specific links are chosen to illustrate different construction and congestion issues. The analysed data are collected for a period greater than six months. The proposed model can also be applied to predict travel times in other cities. The size of the city is not an issue: for large urban environments, a grid computer could be used to ensure fast performance. The only limit of the model is the coverage of the GPS data. This is not, however, an issue for most developed urban environments, where it is often necessary to use a GPS.

For the links presented in this paper, the forecasting mean absolute percentage error for the baseline

method (historical averages) ranges from 7.27% to 39.41%, for the seasonal ARIMA model from 0.96% to 33.62%, and for the proposed *k*NN from 0.18% to 5.20%. Additionally, the mean error and root mean square error for forecasts show that the historical averages model gives the least accurate forecasts, the proposed *k*NN model gives the most accurate forecasts, and seasonal ARIMA gives forecasts with intermediate accuracy.

The experiments provide justification for the use of the *k*NN method in travel time prediction. To the best of the authors' knowledge, no other published research has shown that the *k*NN approach can perform better than seasonal ARIMA. There are two reasons for this. Firstly, in this study, GPS data are used for travel time prediction, and secondly, the data are for urban traffic networks. Since seasonal ARIMA and *k*NN non-parametric regression are usually used to model different systems (non-deterministic with linear state transitions as opposed to deterministic with non-linear state transitions), this contribution may suggest that traffic in urban networks behaves chaotically.

Because of the lack of coverage and the way in which the GPS data are sampled in the study, the authors were unable to apply certain very interesting methods. One such method is space-time ARIMA. Future work should attempt to determine whether STARIMA would be the most appropriate method, since it can model the influences that neighbouring links exert on each other. Such broader perspectives, enabled by additional GPS data, may also include examining the performance of the proposed model when an entire route map is analysed.

LITERATURE

[1] **E.W. Dijkstra**: "*A note on two problems in connexion with graphs*", Numerische Mathematik. 1, 1959, pp. 269–271

[2] **T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein**: "*Introduction to algorithms*", 2nd Ed., MIT Press and McGraw-Hill, 2001, pp. 595–601

[3] **M. Ben-Akiva, M. Bierlaire, H.N. Koutsopoulos, R. Mishalani**: "*Real time simulation of traffic demand-supply interactions within DynaMIT*", Applied optimization, Vol. 63, 2002, pp. 19-34

[4] **M. Fellendorf, K. Nokel, N. Handke**: "*VISUM-online - traffic management for the EXPO 2000 based on traffic model*", Traffic Technology International, 2000

[5] **K. Nagel, M. Schreckenberg**: "*A cellular automaton model for freeway traffic*", Journal de Physique I, 1992, pp. 2221-2229

[6] **B.S. Kerner, H. Rehborn, M. Aleksic**: "*Forecasting of Traffic Congestion*", Traffic and Granular Flow '99, Springer, Heidelberg, 2000

[7] **J. Rice, E. van Zwet**: "*A simple and effective method for predicting travel times on freeways*", IEEE trans. on intelligent transportation systems, Vol. 5, 2004, pp. 200-207

[8] **X. Zhang, J.A. Rice**: "*Short-term travel time prediction using a time-varying coefficient linear model*", Berkeley: Tech. Rep. Dept. Statistics, Univ. California, 2001

[9] **H. Sun, H.X. Liu, H. Xiao, B. Ran**: "*Short term traffic forecast using the local linear regression model*", TRB Paper no 03-3580, Transportation research board, 2003

[10] **M.P. D'Angelo, H.M. Al-Deek, M.C. Wang**: "*Travel time prediction for freeway corridors*", Transportation Research Record, No. 1676, 1999, pp. 184-191

[11] **B.L. Smith, B.M. Williams, R.K. Oswald**: "*Comparison of parametric and nonparametric models for traffic flow forecasting*", Transportation Research Part C: Emerging Technologies Vol. 10, 2002, pp. 303-321

[12] **M. Van Der Voort, M. Dougherty, S. Watson**: "*Combining Kohonen maps with ARIMA time series models to forecast traffic flow*", Transportation Research Part C: Emerging Technologies, Vol. 4, No. 5, 1996, pp. 307-318

[13] **Y. Kamarianakis, P. Prastacos**: "*Space-time modeling of traffic flow*", Methods of spatial analysis – spatial time series analysis, ERSA Proceedings, Dortmund, 2002

[14] **Y. Kamarianakis, P. Prastacos**: "*Forecasting traffic flow conditions in an urban network*: Comparison of multivariate and univariate approaches", Transportation Research Record, Vol. 1857, 2003, pp. 74-84

[15] **H. Liu, D.E. Brown**: "*A new point process transition density model for space-time event prediction*", IEEE transactions on systems, man and cybernetics, Part C, Vol. 34, No. 3, 2004, pp. 310-324

[16] **M. Danech-Pojouh, M. Aron**: "*ATHENA: a method for short-term inter-urban motorway traffic forecasting*", Recherche Transports Sécurité, English version of the report, Vol. 6, 1991, pp. 11-16

[17] **M. Chen, S. Chien**: "*Dynamic freeway travel time prediction using probe vehicle data: Link-based vs. path-based*", Transportation Research Board, Vol. 1768, 2001, pp. 157-161

[18] **D. Park, L.R. Rilett**: "*Forecasting freeway link travel times with a multilayer feedforward neural network*", Computer aided civil and infrastructure engineering, Vol. 14, No. 6, 1999, pp. 357-368

[19] **H. Dia**: "*An object-oriented neural network approach to short-term traffic forecasting*", European Journal of Operational Research, Vol. 131, No. 2, 2001, pp. 253-261

[20] **H. Yin, S.C. Wong, J. Xu, C.K. Wong**: "*Urban traffic flow prediction using a fuzzy-neural approach*", Transportation research Part C: Emerging technologies, Vol. 10, No. 2, 2002, pp. 85-98

[21] **S. Ishak, P. Kotha, C. Alecsandru**: "*Optimization of dynamic neural networks performance for short-term traffic prediction*", Transportation Research Record, ISSU 1836, 2003, pp. 45-56

[22] **R.K. Oswald, W.T. Scherer, B.L. Smith**: *Traffic flow forecasting using approximate nearest neighbour nonparametric regression*, Research Report No. UVACTS-15-13-7, Center for transportation studies at the University of Virginia, 2001

[23] **C.H. Wu, J.M. Ho, D.T. Lee**: "*Travel-time prediction with support vector regression*", IEEE transactions on intel-

ligent transportation systems, Vol. 5, No. 4, 2004, pp. 276-281

[24] **J.E. Disbro, M. Frame**: "*Traffic flow theory and chaotic behaviour*", Transportation research record, Vol. 1225, 1989, pp. 109-115

[25] **T. Tsekeris, A. Stathopoulos**: "*Real-Time Traffic Volatility Forecasting in Urban Arterial Networks*", Transportation Research Record, No. 1964, 2006, pp. 146-156

[26] **F. Batool, S.A. Khan**: "*Traffic estimation and real time prediction using ad hoc networks*", Proc. of the IEEE symposium on emerging technologies, 2005

[27] **B. Hull, V. Bychkovsky, Y. Zang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, S. Madden**: "*CarTel: a distributed mobile sensor computing system*", Proc. of the 4th international conference on embedded networked sensor systems, ACM Press, 2006, pp.125-138

[28] **S. Brakatsoulas, D. Pfoser, R. Salas, C. Wenk**: "*On map-matching vehicle tracking data*", Proc. of the 31st very large databases conf., 2005, pp. 853-864

[29] Mireo Company, 2008. www.mireo.hr, accessed on 2008/10/29

[30] **S. Bajwa, E. Chung, M. Kuwahara**: "*Sensitivity analysis of short-term travel time prediction model's parameters*", Proc. of the 10th ITS world congress, 2003

[31] **S.I. Bajwa, E. Chung, M. Kuwahara**: "*Performance evaluation of an adaptive travel time prediction model*", Proc. of the 2005 IEEE intelligent transportation systems, 2005

[32] **A. Torday, A.G. Dumont**: "*Parameters influencing probe vehicle based travel time estimation accuracy*", Proc. of. the 4th Swiss transport research conference, 2004

[33] **G.E.P. Box, G.M. Jenkins**: "*Time series analysis, forecasting and control*", Revised Edition, Holden-Day, Oakland, California, 1976

[34] StatSoft, Inc., 2005. STATISTICA (data analysis software system), version 7.1. www.statsoft.com, accessed Jan. 2007.

[35] **B.M. Williams, L.A. Hoel**: "*Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results*", Journal of transportation engineering, Vol. 129, part 6, 2003, pp. 664-672

[36] **R.O. Duda, P.E. Hart, D.G. Stork**: "*Pattern recognition*", John Wiley and Sons, New York, 2001

[37] **F.J. Mulhern, R.J. Caprara**: "*A nearest neighbour model for forecasting market response*", International journal of forecasting, Vol. 10, 1994, pp. 191–207

[38] **Schaal, S.**: "*Nonparametric regression for learning*", Proceedings of the conference on prerational intelligence, Germany, 1994