Survey on traffic prediction in smart cities

Article *in* Pervasive and Mobile Computing · July 2018

DOI: 10.1016/j.pmcj.2018.07.004

2 authors:

Attila Mátyás Nagy
Budapest University of Technology and Economics
4 PUBLICATIONS **32** CITATIONS

SEE PROFILE

Vilmos Simon
Budapest University of Technology and Economics
54 PUBLICATIONS **174** CITATIONS

SEE PROFILE

Pervasive and Mobile Computing 50 (2018) 148–163

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Review

Survey on traffic prediction in smart cities

Attila M. Nagy [*], Vilmos Simon

Department of Networked Systems and Services Budapest University of Technology and Economics Budapest, Magyar Tudósok krt 2., Budapest, Hungary

ARTICLE INFO

Article history:
Available online 25 July 2018

Keywords:
Traffic flow prediction
Smart city
Prediction models
Intelligent transport

ABSTRACT

The rapid development in machine learning and in the emergence of new data sources makes it possible to examine and predict traffic conditions in smart cities more accurately than ever. This can help to optimize the design and management of transport services in a future automated city. In this paper, we provide a detailed presentation of the traffic prediction methods for such intelligent cities, also giving an overview of the existing data sources and prediction models.

Contents

[*] Corresponding author.
E-mail addresses: anagy@hit.bme.hu (A.M. Nagy), svilmos@hit.bme.hu (V. Simon).

https://doi.org/10.1016/j.pmcj.2018.07.004
1574-1192/© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Nowadays, smart city services are becoming more widespread than ever as cities are growing and becoming increasingly crowded as a result of urbanization and world population growth [1]. The term "smart city" in [2] refers to the use of information and communication technologies to sense, analyze and integrate key information from core systems in operating cities. At the same time, smart city services can make intelligent responses to different kinds of needs in terms of daily livelihood, environmental protection, and public safety, as well as the city's facilities and industrial and commercial activities. As smart city — related technologies develop rapidly (for example, the spread of IoT devices), more things become measurable. As a result, we have more and more usable data about the ecosystem of our cities.

Among the various notable goals of smart cities, construction of smart transportation systems and smart urban management systems are two of the key aims, which could significantly influence the lives of residents in future cities. Advanced Traffic Management Systems (ATMSs) and Intelligent Transportation Systems (ITSs) integrate information, communication, and other technologies and apply them in the field of transportation to build an integrated system of people, roads, and vehicles. These systems constitute a large, fully-functioning, real-time, accurate, and efficient transportation management framework [3].

In ATMSs and ITSs, it is a fundamental challenge to predict the next possible states of traffic with high precision, because this information helps to prevent unfortunate events like traffic jams or other anomalies on roads. The literature often refers to traffic as a flow, because it has similar properties to fluids. Thus when we speak about traffic flow prediction, we wish to predict the next state (it can be the volume, speed, density, or behavior) of the traffic flow based on historic and real-time data.

The rapid progress of urbanization has modernized many people's lives, but also brought remarkable challenges like traffic congestion [4] that can lead to increased energy/fuel consumption [5] and enormous emission of pollutants [6]. These phenomena have a great impact on the health and quality of life of city-dwellers. According to [7], laboratory studies indicate that transport-related air pollution may increase the risk of developing allergies and can exacerbate symptoms, particularly in susceptible subgroups; cohereas [8] showed that traffic jams increase the risk of heart attack.

Intelligent management systems (such as ATMS and ITS) can help overcome or significantly reduce the impact of such negative effects on city-dwellers. Forecasts can also support traffic control centers in managing the road network and allocating resources systematically, such as opening/closing lanes, dynamic parking pricing [9], or adaptive traffic lights [10] with a high level of automation [11].

When driving, knowledge of traffic forecasts for different routes is advantageous [12], and devices will be able to calculate more efficient routes and reduce travel time. Insight into vehicular flows within the smart city could make searching for parking spaces [13] much easier and faster. Moreover, it could be a useful source of information for emerging V2X-based traffic control systems [14], which could play an important role in the route planning of self-driving cars.

Another motivation for utilizing flow prediction is its marketing value. Marketers are highly interested in customer behavior, because understanding customer preferences helps them identify potential clients [15]. If it becomes possible to link information about the movement of particles in flows with useful marketing information, like gender or age, it could be very valuable for them. For example, if we have this kind of information, we could dynamically change the content of digital roadside billboards [16], or shop owners could open new stores in more frequented locations [17].

Similar to traffic flow prediction, pedestrian flow prediction is also an important challenge in smart cities. As the number of different types of mass events increases year by year, crowd incidents are becoming more and more common — for example stampedes or when participants in mass events crush one another in an overcrowded area [18]. To prevent such disasters, authorities need more reliable surveillance systems, in order to provide better monitoring of crowd movement in a real-time manner [18]. If the organizers and authorities can obtain valid information about crowd dynamics based on pedestrian flow prediction, they would be able to perform quick and targeted interventions. We want to highlight, that although the overwhelming majority of existing systems and scientific papers are focused on vehicle traffic flow prediction, the emergence of new data sources for smart cities could provide a renaissance of pedestrian flow prediction [19], which can significantly influence the health and life quality of city-dwellers as well.

The final goal would be to develop an integrated management system that would merge predictions for vehicular and other urban traffic flows such as pedestrian or bike. This new generation of smart city management systems could reveal high-level correlations between vehicular, pedestrian and bike flows within metropolises.

There are several survey papers which deal with traffic modeling and prediction. In [20], authors present a critical analysis of the mathematical methods for modeling vehicular traffic and crowd phenomena. These models describe the dynamics of traffic or pedestrian flows, typically based on physical flow models. These mathematical models are extremely useful for investigation based on simulation — for example, to examine the characteristics of an intersection. In our paper, we investigate other types of models, based on real-time traffic data from road infrastructure, where forecasts are made by using real-time information. Similar to our paper, some of the surveys [21–24] present traffic prediction method; however, we have tried to focus on today's frequently used and more novel models. Beside that, we also provide a detailed introduction and categorization of practical data sources and data models. There are also survey papers which explain how data fusion can be used in intelligent transportation systems [25,26].

The paper is organized as follows. In Section 2, we introduce and categorize the currently available data sources, with a special attention to publicly available, free data sets. In Section 3, the most frequently used models are collected, being

Table 1
List of different traffic sensor technologies and their capabilities [26,27].

Sensor technology	Vehicle classification	Multiple detection zone	Capital Cost ($K)
Inductive loop			3–8
Magnetic sensor			0.4–2
Video image processor	✓	✓	9–19
Microwave radar	✓	✓	9–13
Laser radar sensor	✓	✓	6–7.5
Active infrared	✓	✓	6–7.5
Passive infrared			0.7–1.2
Audio sensor	✓		3.7–8

classified by the special requirements of the flow prediction methods. In Section 4, we present the prominent flow prediction techniques thoroughly, by categorizing them as parametric and non-parametric models. Section 5 concludes the paper and points out the future directions of flow prediction research, together with our suggestion for a novel type of traffic flow prediction.

2. Data sources

In the first generation of ATMS and ITS systems, the sources of data utilized were different types of presence sensors in fixed positions, which were able to detect the presence of nearby vehicles. Initially, inductive loop detectors were the most popular, but nowadays a wide variety of sensors have become available [26].

In Table 1, the widely used fixed position sensor technologies are collected (an updated version of [26]). *Inductive loop* sensors are built into roads and detect the presence of a conductive metal object by measuring the change in the magnetic field. *Magnetic sensors* detect the presence of a ferrous metal object through the magnetic anomaly they cause in the Earth's magnetic field. *Video image processors* analyze the video image of roadway surveillance cameras and provide traffic flow data across several lanes. *Microwave radar sensors* transmit electromagnetic signals and receive echoes from objects of interest. *Infrared sensors* in active mode illuminate detection zones with low-power infrared energy transmitted by laser diodes and then use the reflected energy to detect vehicles. In passive mode, these sensors detect energy emitted by vehicles or energy emitted by the atmosphere and reflected by vehicles. *Laser radar sensors* are active sensors that transmit scanning infrared beams in the near infrared spectrum over one or more lanes. *Audio sensors* are passive sensors and use different audio signal processing techniques to calculate traffic density or volume.

Recently, the advent of GPS-equipped smartphones and vehicles has given rise to a new type of data source that could supplement presence-type sensors to gather more detailed information or get data about roads, which have not been covered with presence sensors yet. The real-time and historic traffic trajectories from GPS sources have enabled us to make better and improved predictions of traffic flow for ATMSs and ITSs. These trajectories can be collected also from the vehicles and from the pedestrians' mobile phones by utilizing mobile crowd-sensing techniques [28], providing us with valuable data about vehicle and pedestrian traffic trajectories.

In this section, we examine the two data sources, fixed position and moving sensors, and compare them by using the following criteria:

- Formal description of the data source
- Advantages and disadvantages
- Typically measured data types
- List of publicly available data sets.

2.1. Data from fixed position sensors

Traditional fixed position sensors are based on presence type detectors/sensors, which are deployed at a fixed position in space (indicated as p). Because of this property, these sensors always measure at a specific point of the road. (They might measure one or more lanes, depending on the capabilities of the sensor used) If we investigate a single direction of a road segment (for example, using an inductive loop detector), data from a traditional fixed position sensor can be described as an ordered sequence of measurements $\bar{\boldsymbol{m}}_{\boldsymbol{p}}$ in a given p position:

$$\bar{\boldsymbol{m}}_{\boldsymbol{p}} = \{m_{p,t}\} \qquad t = 1, 2, \ldots T \tag{1}$$

where $m_{p,t}$ is the value of the measurement at time t and position p.

The type of measurement depends on the sensor capabilities. Some of them have only basic functionality, as they are only able to measure the traffic count (i.e., traffic volume), while others can also measure the speed or density of the flow. More advanced sensors are also capable of detecting the class of the vehicle, which provides a better insight into the characteristics of the traffic flow.

The biggest advantage of traditional fixed position sensors compared to moving sensors is that they are reliable data sources, capturing all vehicles passing by. On the other hand, a GPS sensor can only track one vehicle at a time. Therefore, aggregate statistics such as the number of vehicles or density of the flow can only be approximated to a certain precision depending on the number of available moving sensors in the area. The drawback of using data from traditional fixed position sensors is that we are not able to observe the exact paths of vehicles. Thus, it is hard to find relations between different road segments; we are only able to come up with rough estimates (e.g., spatial correlation analysis) based on the available sensor data. Unfortunately, the costs of the deployment and the maintenance of a big sensor network can be excessively high.

Valuable fixed position data can be harvested from Automated Fare Collection (AFC) systems. The main purpose of AFC systems is to make toll collection and management more convenient, but the gathered smart ticketing card data can be important input for the study of urban mobility patterns [29]. These systems are able to record the boarding location (entry point) and arriving location (exit point), so although the data is gathered in fixed positions, it is capable of revealing spatial correlations and providing additional information, since all transactions are paired with the card-holder's identity. More than ten papers [30–32] have been examined that deal with smart ticketing card data, but none of them used publicly available data sets.

The majority of data sets are not publicly accessible (due to legal limitations); however, there are some publicly available data sets which can be used for scientific purposes. The vast majority of scientific works [33–35] use the Caltrans Performance Measurement System (PeMS) data set [36]. This data set is provided by the California Department of Transportation. They collect the data in real-time from over 39,000 individual detectors like inductive loop sensors, magnetic sensors, or microwave radar sensors. These span the freeway system across all major metropolitan areas of the State of California. PeMS data are stored in the Archived Data User Service (ADUS), which provides over ten years of data for historic analysis. It integrates a wide variety of information from Caltrans and other local agency systems including traffic detectors, census traffic counts, incidents, vehicle classification, etc.

Another valuable data set is the Traffic Information Service (TRIS) Highway Englands [37]. This data set has provided the average journey time, speed, and traffic flow information for 15-minute periods since April 2015 on all motorways and A roads (known as the Strategic Road Network) managed by Highways England. It is worth remembering that the journey time and the speed is estimated based on the fusion of fixed sensors.

2.2. Data from moving sensors

Data from moving sensors are gathered from GPS sensors installed in smartphones or in vehicles (such as taxis or bicycles) utilizing mobile crowd-sensing techniques, providing us with valuable data about the vehicle and pedestrian traffic trajectories. Many crowd-sensing applications address tasks related to urban transportation systems, which include the tracking of public vehicles (buses, trams, subways and rentable bikes) or mapping bumps on the road in order to inform authorities quickly where to intervene [28].

Another promising source of moving sensor data is the Call Detail Record (CDR), stored by mobile network operators. A CDR record contains metadata (including position information) that describe a specific telecommunication transaction, but does not store the content of the transaction. CDR data covers large geographic areas, providing large sample sizes, and its positioning accuracy is appropriate for the analysis of commuter traffic. Therefore, CDR data can be utilized to determine the travel paths of mobile phone users [38]. These information can be exploited to investigate the mobility patterns of residents in smart cities, which are of great interest both from an economic and from a political perspective.

Contrary to fixed position sensors, these sensors are always in movement if the owner moves. In a formal way, data from GPS sensors can be interpreted as an ordered sequence of measurements where every GPS coordinate sample is associated with time data:

$$\bar{\boldsymbol{p}} = \{p_1 \rightarrow p_2 \rightarrow \ldots \rightarrow p_t \rightarrow \ldots\} \qquad t = 1, 2, \ldots T \tag{2}$$

where every p_t contains latitude and longitude coordinates and a t timestamp. In another sense, the data is a sequence of timestamped points, each of which contains the latitude and longitude. Because of this property, authors also refer to data from moving sensors as spatio-temporal data, where the *spatio* part is the GPS coordinate and the *temporal* part is the timestamp.

Data from moving sensors give us a new dimension of information. We can observe or identify exact paths and different motion patterns of vehicles and pedestrians, find connections between different road segments, etc. Furthermore, it makes possible to discover the real paths of pedestrians and bicyclists, since they can move more freely in space. Besides, it has significantly lower infrastructure costs than the fixed position solution, because the sensors are already installed in vehicles or in smartphones, so and we need only care about the data gathering and processing, not deployment or maintenance. Thus, roads which are not covered by fixed position sensors become observable as well.

Contrary to fixed sensors, aggregated data characteristics in the case of moving sensors can only be derived from multiple sensors moving along the same path. A huge drawback of spatio-temporal data is that if we do not get information from all the actors in traffic, we cannot be sure that our data is representative and describes all the possible trajectories. This can be assured only if a high proportion of vehicles or pedestrians provides data from their sensors. When using data from moving sensors, one should also deal with the uncertainty associated with GPS sensors [39]. This uncertainty makes it necessary to perform the map-matching task [40] to correct the deviation between the GPS points of vehicles and the road network after data preprocessing. In 2018, new smartphone GPS chips will be released [41]; and by utilizing L5 and L1 GPS signals together, they will be able to provide up to 1m of accuracy.

Similar to fixed position sensors, the majority of data sets used in the literature are not accessible (for instance, Google or Apple have been using spatio-temporal data in their services for a long time). There are some publicly available data sets, but less than for fixed position sensors. The T-Drive trajectory data set [42] contains one week of trajectories from 10,357 taxis. The total number of points in this data set is about 15 million, and the total distance of the trajectories reaches 9 million kilometers. In the Taxi Service Trajectory — Prediction Challenge (ECML PKDD 2015 Data Set) [43] they provide an accurate data set describing a complete year (from July 2013 to June 2014) of trajectories performed by the 442 taxis running in the city of Porto, in Portugal. Public CDR datasets can be found. For example, the D4D-Senegal Challenge dataset [44] contains CDR data about 9 million of Orange's customers in Senegal between January 1, 2013 and December 31, 2013. Another data set, which provides CDR data is a multi-source data set (combining telecommunication, weather, news, social networks, and electricity data) of urban life in the city of Milan and the Province of Trentino [45].

2.3. Environmental and seasonal information

The flows are typically influenced by the behavior and the cardinality of the entities of the flow; however, flows can also be significantly influenced by factors such as weather, seasonality (holidays, day of the week, seasons, school schedules), events, road construction, air quality or lighting conditions.

When there is heavy rainfall, pedestrians try to move under covered paths like underpasses and avoid open spaces, possibly leading to overcrowded areas. Car drivers always drive slower when there are slippery roads and poor visual conditions, leading to an increased risk of accidents, which in turn increases the probability and volume of traffic jams and congestions [46].

Seasonality also plays a major role. Traffic flows change periodically depending on the day of the week. Additionally, national holidays can cause a heavy loads on the road network and result in severe congestion. Seasons of the year also have a measurable effect on flows. For example roads and pedestrian paths nearby universities experience a significantly higher load in the period of termtime than during the examination period or summer break.

Utilizing these external data sources, like weather forecasts and seasonal impact, can lead to better and refined forecasts [47]. By ignoring them, huge inaccuracies could emerge, endangering the applicability of traffic flow prediction.

We believe that the previously introduced data sources should be used together in intelligent transportation systems that combine their advantages by applying data fusion techniques [25]. On important and high-capacity segments of smart city road networks, reliable traffic data is crucial for accurate traffic predictions and management, because these arterials significantly influence the whole road network. Currently, data from fixed position sensors can provide a fair traffic description on those arterials (if deployed in a massive number, which means high infrastructure cost). Nevertheless, data from moving sensors and external environmental monitoring stations should be used to provide complementary information, in order to fully understand the behavior of the different actors of the transportation system and to reliably measure the level of correlation between the different parts of the road network. Moreover, GPS trajectories provide suitable information for lesser important parts of the road network, where the deployment and maintenance of an expensive sensor network can imply unnecessary cost.

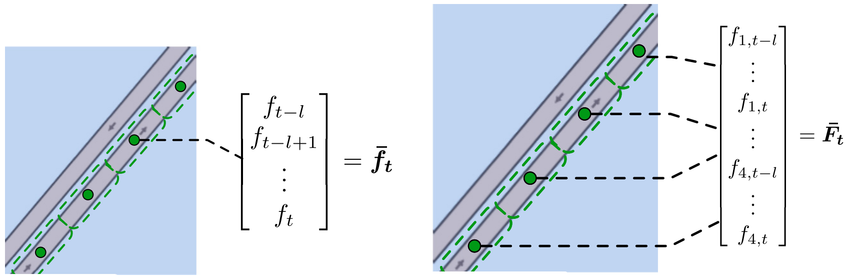
In the future, the spread of autonomous vehicles could bring about significant changes, because these vehicles can provide much more detailed information about itself and its driver. This new information would to be used to improve the services of future intelligent transportation systems. However, it also raises serious and unresolved privacy questions.

3. Data models

When we create a new traffic flow prediction data model, we have to identify the relevant aspects and features of the traffic flow and determine the right granularity of our data model. Besides that, we need to make various design decisions when creating our model — for example, whether we can utilize only one or more data sources or use basic or fused data. Still, there are some basic properties of traffic flow which have to be considered during the data model's design such as its spatio-temporal property, which means that it has both spatial and temporal qualities.

In the case of the spatial property, it should be defined what level of spatial details we are interested in. Sometimes we want to examine the flow of an exact point in space [33,35], which can be interpreted as a microscopic view of the flow; and sometimes we want to know correlations between larger areas [39,48–50], a macroscopic way to study the flows.

We also have to deal with the timeliness of our model. In the literature, different time intervals are used, which mainly depend on the desired prediction horizon of the selected prediction model. Generally, narrow intervals – for example 10 s – are meaningless for traffic prediction [48]. We have found that the most common time intervals are in minute scale (5–10 min) [33,39,49], but there are also many papers claiming that longer time intervals would be more effective, like quarter or half an hour [34,35,48,49]. The minute scale can be useful when our primary goal is to avoid traffic jams or to compute journey times between shorter distances. On the other hand, longer time intervals can identify long-term tendencies. Thus they are useful for city planning or global traffic control. Another parameter, which has to be considered carefully together with the time interval, is the resolution of the used data source. For example, with low sampling frequencies we are not capable of making accurate predictions for time intervals of fifteen or thirty minutes.



(a) Univariate vector-based model (b) Multivariate vector-based model

Fig. 1. Example of different vector-based models.

The rest of this section presents and gives a detailed analysis of the widely adopted data models in the literature. We have categorized the data models found in the literature as scalar-, vector- and matrix-based data models. Scalar-based data models use simple scalar values to describe the actual state of traffic flow, while vector and matrix models use vectors and matrices, respectively.

3.1. Scalar-Based models

The simplest data model for traffic flow prediction utilizes data from fixed position sensors without any preprocessing. As stated in Eq. (1), every $m_{p,t}$ measurement contains a scalar value and the timestamp of the measurement at position p. Data from moving sensors can also be used in this model, but the measured data of GPS traces need to be preprocessed, extracting the values for only the examined p positions. The goal of the traffic prediction is to determine the $m_{p,t+1}$ value based on $\bar{\boldsymbol{m}}_{\boldsymbol{p}}$ where p denotes the position. Ordered sequences of scalar values can be modeled by time series, which is a well-known data model.

Time series can be characterized by the following features. *Trend* is a gradual increase or decrease of the series' values over time, which can be global, local, or both. It can also be linear or non-linear. A *seasonal cycle* is a repetitive, predictable pattern in the series' values, while a *non-seasonal cycle* is a repetitive, possibly unpredictable pattern. Shifts in the time series' values that cannot be explained by the model are referred to as *residuals*.

In many cases, the Autoregressive Integrated Moving Average (ARIMA) [51,52] (or an extended variant like SARIMA [53], KARIMA [54], etc.) is used for prediction of time series. Apart from these, some other prediction models like the Kalman filter [55], Bayesian networks [56], or Neural networks [57] are also utilized in the related works.

The biggest drawback of using a scalar based model is that it hides the spatial correlations, and it describes only the temporal ones. Sometimes we are only interested in the flow at a specific point of the road. For this, the scalar-based model is the right approach. Still, in most cases we also want to understand the flow's behavior over the whole trajectory, which is not possible with this model.

Scalar-based models also have scaling problems, because the calculation of predictions for every sensor in a large sensor network is a computationally intensive task.

3.2. Vector-Based models

Instead of using one scalar value to characterize the flow at a point in space and time, vector-based models define a vector that describes the actual state of the flow (also known as the state vector). As input data sources for vector-based models, one can use fixed position (if one can assume that there are enough installed sensors on roads) or moving position sensors, or a mixture of them. As for the spatial factor, we can distinguish univariate and multivariate versions. An univariate vector model observes one sensor, while the multivariate versions observe more sensors.

In related works, authors most frequently use vector-based models with the K-Nearest Neighbors (KNN) prediction model (or some variant of it), because it is analogous to a data model, which is used by KNN.

3.2.1. Univariate vector models

A univariate vector model can be observed in Fig. 1a for a given sensor of the transportation network, where the state vector of the current flow at time t can be defined as:

$$\bar{\boldsymbol{f}}_{\boldsymbol{t}} = \{f_{t-l}, f_{t-l+1}, \ldots, f_t\} \tag{3}$$

where l (lag) is the number of time intervals handled by the model. f_{t-l} denotes the measured value at time $t - l$. The goal is to predict the next f_{t+1} value.

In most cases, authors use disjoint time intervals [58], but some works utilize overlapping intervals. To mention an example for overlapping time intervals, we let $l = 2$ and suppose we are using 5-minute time intervals. Then f_t denotes the actual interval, f_{t-1} denotes the previous 5-min interval and f_{t-2} denotes the previous 10-min interval. Unfortunately, similar to scalar-based models, the univariate vector model is not able to describe the spatial correlation, either.

3.2.2. Multivariate vector models

In a multivariate vector model (Fig. 1b), we observe more sensors of a transportation network. Thus, the state vector of the current flow at time t can be defined as:

$$\bar{\boldsymbol{F_t}} = \{f_{1,t-l}, \ldots, f_{1,t}, f_{2,t-l} \ldots, f_{j,t}\} \tag{4}$$

or

$$\bar{\boldsymbol{F_t}} = \{f_{1,t-l}, \ldots, f_{j,t-l}, f_{1,t-l+1}, \ldots, f_{j,t}\} \tag{5}$$

where l (lag) is the number of time intervals handled by the model, and j is the number of sensors present in the state vector. The goal of traffic prediction here is to determine the value of $f_{x,t+1}$, where x is an arbitrary sensor and $t + 1$ is the next time interval. As you can see, the ordering of the values of multivariate vector based models can vary. Some authors order the values by time [34] like in Eq. (4), and some use space ordering [39] like in Eq. (5).

The multivariate vector model uses disjoint time intervals. It can also identify spatial correlations on road segments as the elementary units of space [39,49,50]. To reveal the spatial correlations, a possible method is to examine the target road segment along with its upstream and downstream road segments. A more general approach is to define a radius r, within which the sensors are examined together. For a complex road network, the graph structure has to be considered, so r needs to be replaced by the hop distance.

3.3. Matrix-Based models

In the case of a matrix-based model, matrices are defined that describe the actual state of the flow. Fixed position (if we can assume that there is enough installed sensors on roads), moving position sensors or a mixture of them can be utilized as data sources. Unlike scalar- or vector-based models, matrix-based models always identify both spatial and temporal correlations. These models can be categorized as a macroscopic model, meaning that instead of examining one point in space in a detailed way, they are suitable for identifying correlations between bigger areas.

Beside the KNN prediction model [59] (or its extended variants), the most popular prediction models used for matrix-based models in the related works are the Convolutional Neural Network (CNN) [60], the custom made neural networks [50], and the Bayesian networks [61]. Based on the properties of the matrix-based models, they can be classified into two major groups: the time–space matrix models and the region matrix models.

3.3.1. Time–space matrix models

The time–space matrix model is an improved version of the multivariate vector based models. The matrix serves as a time–space image (from here comes the motivation to utilize convolutional neural network), where the x-axis of the matrix represents the time dimension and the y-axis the space dimension (typically a road segment). The cells contain measured values of the flow at a specific time in a point of space [48]. In a formal way, a time–space matrix can be defined as in Eq. (6):

$$\boldsymbol{M} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,L} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ m_{P,1} & m_{P,2} & \cdots & m_{P,L} \end{bmatrix} \tag{6}$$

where L is the length of time intervals, P is the number of measurement sensors, and $m_{i,j}$ is the measurement of the ith position at time j (see Fig. 2). Since the density of measurement sensors is mostly constant throughout a particular road segment, the number of measurement sensors is proportional to the length of the road segment.

One strategy can be to choose the set of measurement positions within a pre-defined r radius (similar to the multivariate vector model) by considering the road network as a graph.

3.3.2. Region matrix models

Region matrix models are a relatively new approach for modeling the time–space correlation in the case of flow prediction. Regions are arbitrarily shaped areas of the city, and the goal is to find the relevant relations between these areas from a flow prediction perspective. It is important to highlight that only moving sensors can be utilized as data sources. Data from fixed position sensors do not contain spatial correlation information. Contrary to GPS trajectories, data from fixed position sensors are not able to reveal relations between regions.

Region matrix models provide the highest level of macroscopic view. In the case of these models, determining the right size or the shape of the regions remains an open problem. Fortunately, there are some known methods to tackle this

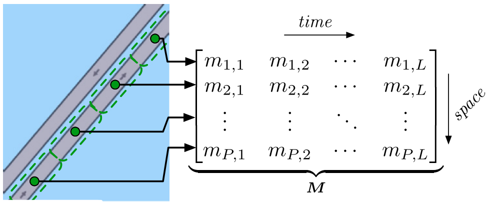
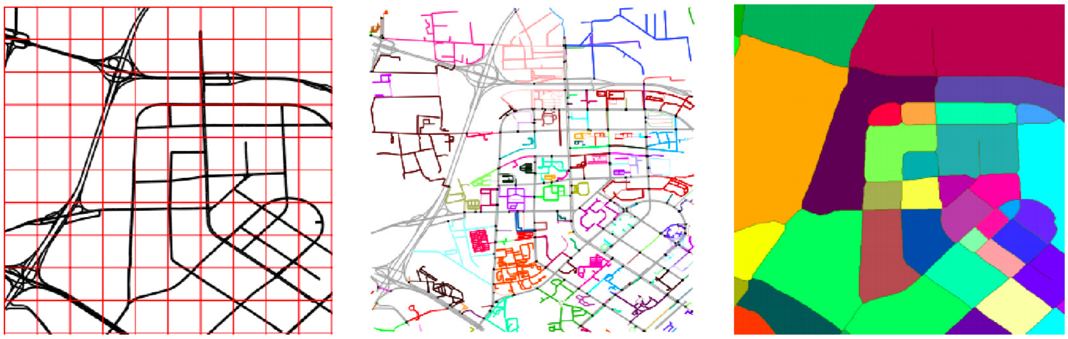

time
space
$\begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,L} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ m_{P,1} & m_{P,2} & \cdots & m_{P,L} \end{bmatrix}$
M

Fig. 2. Illustration of the time–space matrix model.



(a) Grid-based map segmentation (b) Hierarchical map segmentation (c) Morphology-based map segmentation

Fig. 3. Map segmentation methods [64].

challenge. Grid-based segmentation (see Fig. 3a) simply divides the map using a grid. It can be useful, for instance, to predict drivers' destinations by mapping their historic trips [62], predict in/out flows between grid cells [50], or to suggest the most profitable grid cells for taxi drivers [63]. The drawback of grid-based segmentation is that it does not take into account the logical cohesion between city areas, such as the road network or districts that form functional entities within the city.

In hierarchical segmentation (see Fig. 3b), authors use the road hierarchy to generate the regions [65]. Specifically, the road networks are first divided into areas by high level roads, then the partition process is performed recursively for each area. A drawback of this method is that only the road network is handled and segmented into regions (not covering the areas between the roads, which can also contain paths for pedestrians or bicycles). To solve this problem, one can use morphology-based map segmentation (see Fig. 3c), where the hierarchical segmentation has been extended with morphological operators such as intersection, union, inclusion, or complement. Thus an urban road network can be efficiently segmented into regions [64].

The majority of papers only focus on examining the in/out flows between regions [50]. The formal mathematical definition [50] is introduced for grid-based segmentation. Let $\mathbb{P}$ be a collection of trajectories at the tth time interval. For a grid (i, j) that lies at the ith row and the jth column, the inflow and outflow of the crowds at the time interval t are defined respectively as:

$$x_t^{in,i,j} = \sum_{Tr\in\mathbb{P}} |\{k > 1 | g_{k-1} \notin (i,j) \wedge g_k \in (i,j)\}| \tag{7}$$

$$x_t^{out,i,j} = \sum_{Tr\in\mathbb{P}} |\{k \geq 1 | g_k \in (i,j) \wedge g_{k+1} \notin (i,j)\}| \tag{8}$$

where $Tr : g_1 \to g_2 \to \dots \to g_{|Tr|}$ is a trajectory in $\mathbb{P}$, and g_k is the geospatial coordinate; $g_k \in (i, j)$ means the point g_k lies within grid (i, j), and vice versa; while $|\cdot|$ denotes the cardinality of a set. At the tth time interval, inflow and outflow in all $I \times J$ regions can be denoted as a tensor $\boldsymbol{X_t} \in R^{2\times I\times J}$, where $(\boldsymbol{X}_t)_{0,i,j} = x_t^{in,i,j}$, $(\boldsymbol{X}_t)_{1,i,j} = x_t^{out,i,j}$. Note that it can be extended to arbitrarily shaped convex regions.

4. Prediction models

In the literature, there are numerous prediction models utilized for traffic flow prediction; but before we go into detail on those, predictability should be defined, since it is a fundamental traffic flow property that influences the selection of the prediction model.

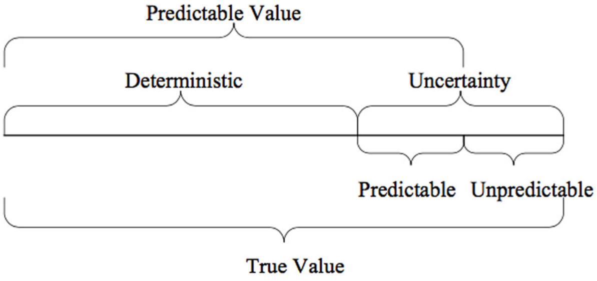


Fig. 4. Predictable and unpredictable components of a variable [67].

According to [66], traffic flow predictability denotes the possibility for prediction to satisfy some precision requirement over a desired prediction time horizon. In other words, what is the maximum amount of time can be predicted by a model with bounded error? The true value of a prediction is composed of a predictable component and an error [67], which includes both prediction error and unpredictability of uncertainty. The predictable value is derived from the deterministic part and the predictable part of uncertainty (depicted on Fig. 4).

Predictability of traffic flow depends on whether the model is able to predict the uncertain portion of traffic flow with the required precision. Uncertainty is influenced by many factors, like weather, days of week, events, road construction, lighting conditions, etc. Incorporating external environmental factors and fused data [25] into the model is crucial to decrease the error of prediction and increase the predictable part of uncertainty. The predictability of traffic flows is directly related to the prediction time horizon. Intuitively, prediction accuracy decreases as the prediction horizon increases [68].

A traffic flow can either be stationary or non-stationary [69], and this property has to be considered when the suitable model is chosen. For example, a simple Autoregressive (AR) model, which depends on the time series under investigation being stationary, is inadequate if the time series does not have this property. Therefore, it is impossible to obtain meaningful predictions with inappropriate models.

Authors of [67] have found that the spatial–temporal relationship between the points of the road network can improve traffic predictability, as the evolution of traffic conditions can be expressed as a form of interaction among traffic flows on upstream and downstream road links [70]. They have used the cross-correlation function (CFF) to measure this relationship.

Traffic is predictable in the sense that it does not vary significantly during weekdays and during most months of the year [71]. In [72], similar results were obtained as well, since they established a relatively high daily predictability of traffic conditions (despite the absence of any a priori knowledge of drivers' origins and destinations) and quite different travel patterns between weekdays and weekends.

Taking into account the previous considerations, the predictability of traffic flows, and consequently the precision of the predictions, can be significantly improved.

In the following subsections, we present prominent empirical traffic prediction approaches often used in the literature. (See Table 2.) We highlight the advantages and disadvantages of the models, which can be categorized into three main categories: naive, parametric, and non-parametric models. A thorough analytical comparison of the existing models is not easy, because the performance of the given prediction model depends heavily on the exact scenario, the available data sources, and the parameters of the environment. Therefore, some of the models would outperform the others in a given scenario; but in a different scenario, their performance would deteriorate. Hence, the appropriate model should be chosen by examining beforehand the given scenario and understanding the effect of the various parameters.

4.1. Naive models

Naive models are the simplest approach for traffic flow prediction. These methods use only basic statistical assumptions like statistical average, and they are not able to utilize any external properties of the traffic to reduce the uncertainty of the prediction. Naive models are widely used in real-world applications because of their simplicity and relatively accurate results with small prediction horizons. Compared to naive models, both parametric and non-parametric models can achieve higher accuracy.

The Instantaneous Travel Times (ITT) approach uses the actual measured value as the prediction of the next measurement. This method is very fast, because no calculation is needed. Yet, the precision of the prediction is bad in most cases [73], especially for long prediction horizons.

Historical Average (HA) makes predictions based only on the arithmetic average of the past values [74]. It is more accurate than ITT, and it can outperform some more complex prediction models on longer horizons, but it cannot handle sharp changes in traffic flow.

We have found that, in real world navigation services [75], one of the most popular approaches is the combination of ITT and HA [69], where the real-time traffic data is combined with historic averages for each road segment.

Table 2
A comparative list of traffic prediction models.

Model name	Model type	Data models	Integrates environmental data	Model has spatial property	Handles nonlinearity	Handles nonstationarity
Instantaneous Travel Times	Naive	Scalar				
Historical Average	Naive	Scalar				
ARIMA	Parametric	Scalar				✓[a]
SARIMA	Parametric	Scalar				✓[a]
STARIMA	Parametric	Scalar		✓		✓[a]
KARIMA	Parametric	Scalar				✓[a]
ARIMAX	Parametric	Scalar	✓	✓		✓[a]
VARMA	Parametric	Vector		✓		
Kalman Filter	Parametric	Scalar				✓
Bayesian Networks	Non-Parametric	Scalar, Vector, Matrix		✓	✓	✓
K-Nearest Neighbors	Non-Parametric	Vector	✓	✓	✓	✓
Feed Forward Neural Network	Non-Parametric	Scalar	✓	✓	✓	✓
Time Delayed Neural Network	Non-Parametric	Scalar			✓	✓
Recurrent Neural Network	Non-Parametric	Scalar			✓	✓
Long–short Term Recurrent Neural Network	Non-Parametric	Scalar			✓	✓
Gated Recurrent Unit Neural Network	Non-Parametric	Scalar			✓	✓
Convolutional Neural Network	Non-Parametric	Scalar, Vector, Matrix		✓	✓	✓
Combination of CNN and FFNN	Non-Parametric	Scalar, Vector, Matrix	✓	✓	✓	✓
Combination of CNN and LSTM	Non-Parametric	Scalar, Vector, Matrix	✓	✓	✓	✓

[a] A differentiation step is necessary to handle non-stationary data.

4.2. *Parametric models*

Parametric models are based on a finite θ set of known parameters about the modeled population (in our case the traffic flow). For instance, if a process is modeled with the Poisson distribution, the value of a single (λ) parameter should be determined. Given the parameters, future predictions (x) are independent of the observed data, $\mathcal{D}$:

$$P(x|\theta, \mathcal{D}) = P(x|\theta) \tag{9}$$

Therefore, θ captures everything there is to know about the data, which indicates that only parameters of the model have to be determined using the input data. Knowledge of these parameters (depending on the parametric model used) can be utilized in these models, which could aid in the understanding different behaviors of traffic flow. An important advantage of parametric models is that they use significantly less training data than non-parametric models, because the complexity of the parametric model is bounded. Nevertheless, this could be also a disadvantage, because even if the amount of data is unbounded, these models are not able to utilize additional information which can be present in a large data set.

In the following subsections, we will present three major parametric models that are used for traffic flow prediction: traffic simulation models, time series models, and Kalman filters.

4.2.1. *Traffic simulation models*

Traffic simulation models are mathematical models which help plan and design transportation systems. Instead of using historic and real-time traffic data, these models simulate the traffic; since in the design phase of a road network, there is no historic data available. It is important to highlight that future traffic levels can be estimated with traffic simulation models to validate the suitability of the design of a transportation system; however, these models are not capable of predicting the next state of traffic based on historic and real-time data.

The basic elements of traffic simulation models were established by Beckmann, McGuire, and Winsten [76]. In these models, traffic is simulated by an Origin-Destination Origin–Destination matrix (OD), which describes vehicle movement in a certain area. The OD matrix features a cell representing the number of trips from origin (row) to destination (column). Traffic simulation models can be categorized according to their scope as microscopic, macroscopic, or mesoscopic.

In the *microscopic* view, every individual actor on the roads and their interactions are modeled in a multi-agent system [77] where every agent keeps a record of its trip including basic information or behavior (such as lane changing behavior or gap acceptance behavior, when conflicts could arise with other vehicles). A typical microscopic simulation approach is the Cellular Automata (CA), where roads are divided into cells which can be empty or occupied by a vehicle, and the time is discretized to steps of δt. CA has the ability to reproduce a wide range of different traffic phenomena, and due to the simplicity of the model, it is numerically very efficient. This model has been combined with OD predictions to come up with network-wide traffic predictions [78].

In the *macroscopic* view, only global variables of a road network are considered –such as the density, speed or traffic count –where these variables are determined for each road segment of the road network [79]. To allocate traffic on the simulated road network, there are two methods; static and dynamic assignment.

One of the differences between the static and dynamic approaches is that the dynamic assignment models normally require time-dependent OD matrices to model traffic flows over time, while static assignment models only need at least one static OD matrix of overall trip demand of actors on the road network [80]. The changes in traffic count and density in time can be modeled with macroscopic models, complemented with formulas stemming mainly from hydrodynamics according to [81].

Mesoscopic models are combinations of macroscopic and microscopic models [82]. First, traffic is allocated to different road segments using macroscopic models, after which individual cars are moved through the network based on the calculated microscopic traffic variables. The big advantage of the *mesoscopic* approach is that a wider variety of phenomena can be modeled, such as the operation of installed traffic signals, freeway merges, weaving sections, or high-occupancy vehicle lanes.

4.2.2. Time series models

The next parametric models are time series models utilizing historic and real-time traffic data to predict the next traffic state, while using scalar-based data models (i.e. time series) to model the traffic flow as we mentioned in Section 3.1.

The basic time series models assume that the value of the series at time t depends linearly only on its previous values with added random noise [51]. Autoregressive (AR) and Moving Average (MA) components are used to model the time series, forming an Autoregressive Moving Average (ARMA) model. AR forecasts the variable of interest using a linear combination of past values of the variable, while MA uses past forecast errors in a regression-like model. Because most time series have non-stationary behavior in practice, the ARMA model can be generalized to handle non-stationarity by applying differentiation (computing the differences between consecutive observations). This extension of the ARMA model is called Autoregressive Integrated Moving Average (ARIMA). Drawbacks of using time series models are that they cannot deal with non-linear processes, and it is hard to integrate environmental data sources in to them.

The first studies concentrated on the usability of ARIMA models on the scalar-based data model for traffic flow prediction, applying the well-known Box–Jenkins method for model identification [83]. The Box–Jenkins model identification method selects the appropriate model (whether to use and autoregressive and/or a moving average component) based on checking stationarity, identifying seasonality, and estimating the order and parameters of ARIMA models.

[51] also focused on the application of using time series models to study the arterial travel time prediction problem for urban roadways. Their study indicated the potential and effectiveness of using ARIMA models for predicting travel times. In [52], authors used subset ARIMA, which is represented by only a few non-zero coefficients instead of the original coefficient vectors determined by the Box–Jenkins method. For example, if we use an AR(4) model, we use the last 4 past values of the time series to predict the next value; whereas AR(1 4) model uses only the last and fourth last values of time series for prediction. They observed that the subset ARIMA model gave more stable and accurate results than other time-series models, especially a full ARIMA model.

To deal with the seasonal property of traffic flow, a new ARIMA-based approach has been proposed. The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is one of the popular univariate time series models in the field of short-term traffic flow forecasting [56]. The parameters of the SARIMA model are commonly estimated using classical (maximum likelihood estimate and/or least square estimate) methods. [53] presented the theoretical basis for modeling univariate traffic condition data streams as Seasonal Autoregressive Integrated Moving Average processes. Also in [84], SARIMA basic exponential smoothing models were developed and tested on data sets.

In [54], the Kohonen Autoregressive Integrated Moving Average (KARIMA) method, a hybrid method for short-term traffic forecasting, was introduced. This technique uses as an initial classifier a Kohonen self-organizing map, in which the road network is mapped over in a hexagonal pattern. Each traffic class, which comes as the result of an unsupervised learning using historic traffic data, has an individually tuned ARIMA model.

To exploit the spatio-temporal relations between road segments, [74] and [85] proposed the Space-Time Space–Time Autoregressive Integrated Moving Average (STARIMA), a multivariate extension of standard univariate ARIMA, which expresses each observation at time t and location p as a weighted linear combination of previous observations lagged both in space and time. Also in [74], a Vector Autoregressive Moving Average (VARMA) model was proposed for traffic flow prediction that describes a set of time series by using $N \times N$ autoregressive and moving average parameter matrices to represent all auto-correlations and cross-correlations within and among the N time series under study.

Another multivariate model is the Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX), which includes independent predictor variables such as another time series. This model is also referred to as the vector ARIMA or dynamic regression model. The ARIMAX model is similar to a multivariate regression model, but allows one to take advantage of auto-correlations that may be present in residuals of the regression to improve the accuracy of the forecast. According to [86], application of ARIMAX provided improved forecast performance over univariate forecast models.

In other papers, a hybrid model is adopted by researchers to model and forecast the scalar-based data model. [87] proposed an ARIMA model with Generalized Autoregressive Conditional Heteroscedasticity (ARIMA-GARCH) for traffic flow prediction. However, [87] concluded that for ordinary traffic flow prediction, the standard ARIMA model is more accurate than the hybrid model.

4.2.3. Kalman filters

The Kalman Filter is an efficient recursive filter that estimates the internal state of a linear dynamic system from a series of noisy measurements. This is a linear model, and authors mainly use it with scalar-based data models, but it may be applied to model short-term stationary or non-stationary traffic flow prediction.

[88] uses a Kalman Filter to fuse data from fixed position sensors and moving sensors for the estimation of traffic density. Subsequently, the estimated data is utilized for predicting the density of future time intervals using a time series regression model. The models were estimated and validated using both real and simulated data. Both estimation and prediction models performed well, despite the challenges arising from heterogeneous traffic flow conditions.

[55] utilized a Kalman Filter to implement a SARIMA + Generalized Autoregressive Conditional Heteroscedasticity (GARCH) structure. However, a conventional Kalman filter cannot update its process variances in real time. Therefore, this paper investigated and tested an adaptive Kalman Filter approach using real world traffic flow data. Similar to the previous one, an online adaptive model was proposed, taking into account historic off-line data [89]. The algorithm was extended to a more general and flexible state-space model, and the predictions were computed recursively with a Kalman Filter.

In [90], authors proposed a short-term highway traffic prediction method based on a structural state space model. The true state of traffic was decomposed to regular traffic pattern, structural deviation, and random fluctuation components. The proposed model is incorporated into a Kalman Filter-based algorithmic framework, together with an adaptive scheme for determining the variances of random errors.

4.3. Non-Parametric models

Non-parametric models assume that the data distribution cannot be defined in terms of a finite set of parameters, but they can often be defined by assuming an infinite dimensional θ. Thus, usually more data is required than for parametric models. Non-parametric models are more flexible than parametric models, because the amount of information that θ can capture about the data can grow as the amount of data grows. The advantage of these models is that they are able to handle non-linear, dynamic processes, and they can exploit spatial–temporal relationships as well. Some of them are also capable of integrating environmental data sources, which can increase the accuracy of predictions in extreme cases. The drawback of non-parametric models is the model training or prediction itself, can be a computationally intense task is comparison to parametric models, since huge amounts of data have to be processed.

4.3.1. Bayesian networks

A Bayesian Network is a directed graphic model for representing conditional dependences between a set of random variables [91]. As a non-parametric model, it is able to handle non-linear and non-stationary processes. It can be extended to work with spatio-temporal data, and it has the ability to cope with incomplete data. However, the authors of [91] mentioned that this approach might not be optimal for incidents and accidents, because collecting large amounts of data during incidents and accidents is usually difficult. Thus, other techniques might be helpful, such as abnormality detection methods, etc.

Most of the Bayesian Network-based work uses time–space matrix-based data models and Gaussian Mixture Models (GMMs) with Competitive Expectation Maximization (CEM) to calculate joint distributions. GMM is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters, while CEM is a well-founded iterative statistical algorithm which calculates the unknown parameters of Gaussian distributions using maximum likelihood estimation.

[61] investigated a spatio-temporal Bayesian Network predictor. This approach incorporated all the spatial and temporal information available in a transportation network to carry out traffic flow forecasting. In a transportation network, there are usually many road segments related to or providing information about the traffic flow of the road segment under investigation. However, using all the related segments as input variables (nodes) would involve much irrelevance and redundancy, as well as being prohibitive computationally. To solve this problem, authors of [61] adopted the Pearson Correlation Coefficient to rank the input variables (traffic flows) for prediction, and the best-first strategy was employed to select only a subset as nodes of a Bayesian network.

[56] used a scalar-based data model (i.e. time series), and instead of using classical inference, the Bayesian Method is employed to estimate the parameters of a SARIMA model. The Markov Chain Monte Carlo Method was used to calculate the posterior probability distributions of the Bayesian model, where “posterior” means after taking into account the relevant evidence related to the particular case under examination.

In [92], authors proposed an adaptive Bayesian Network, which accomplishes online classification of the traffic flow into categories such as free flow or traffic congestion. Since the degree of spatio-temporal dependence between the road segments depends on the determined traffic category, the network topology of the Bayesian Network can be optimized using mutual information learning, which measures the mutual dependence between the two nodes in the Bayesian Network. In [93], another Bayesian Network-based model was proposed, where authors used the Multiregression Dynamic Model (MDM) to improve forecasts. MDM is designed to preserve certain conditional independent structures over time across a multivariate time series.

4.3.2. K-Nearest neighbors models

K-Nearest Neighbors (KNN) is a simple regression algorithm that uses the actual and the closest *k* (based on a distance measure) previously measured traffic flow states to predict the next states of the traffic flow using vector- or matrix-based data models. It should be highlighted that the closest *k* flow states can come from different road segments.

Similar to Bayesian Networks, it can deal with non-linear and non-stationary properties of traffic flows, and it can handle noisy data and be computationally effective if the training data are vast. A serious drawback of this model is that the optimal value of parameter *k* (number of nearest neighbors) needs to be determined. In case of distance-based learning, it is not clear which distance function gives the best result. Computation cost can be high if the algorithm has to compute distances between actual and all previous traffic states for every prediction, and it is difficult to integrate environmental variables. For determining the value of *k*, authors of [39,94] proposed different procedures. For the distance computation problem, one can use indexing (e.g., K-D Tree) or better neighbor search strategies [95] to reduce computation costs.

Most of the articles [58,94] use univariate vector models as a data model. However, some of the articles [34,39] use multivariate vector models or time–space matrix models [59] in order to describe spatio-temporal relationships. [59] proposed an improved KNN model to enhance forecasting accuracy based on spatio-temporal correlations, where the nearest neighbors are selected according to the Gaussian Weighted Euclidean Distance. [39] presented different MapReduce-based approaches on a Hadoop platform, in which they exploit the spatial and temporal correlations between nearby road segments to improve the performance.

4.3.3. Neural networks

Nowadays, Neural Networks (NN) are the most widely used prediction models for traffic flow prediction, because they are able to model non-linear, stationary/non-stationary behavior and are highly extensible. This means that the spatio-temporal property can be also considered, and the environmental data sources can be integrated easily. All these properties decrease the unpredictable portion of uncertainty, which could significantly decrease the prediction error in extreme cases as well.

For a long time, NNs have been used for times series prediction where data is modeled by scalar models. Authors began to examine traffic flow as a time series using the standard Feed Forward Neural Networks (FFNNs) with a backpropagation algorithm [57].

Backpropagation is an algorithm for supervised learning of artificial neural networks using gradient descent. Given an artificial neural network and an error function, the method calculates the gradient of the error function with respect to the neural network's weights. These NNs can perform better than simple parametric models. However, they cannot exploit the spatio-temporal property of traffic flows.

While traditional NNs build on the assumption that their input data are independent of each other, Time Delayed Neural Networks (TDNNs) and Recurrent Neural Networks (RNNs) are capable of finding temporal relationships, because they allow information to persist [96]. TDNNs augment the input scalar model with delayed copies, while RNNs retain an internal state (memory) by using a directed cycle in neurons. Unfortunately RNNs have problem with long-term dependencies [97]. Therefore, other authors use Long-Short Long–Short Term Recurrent Neural Networks (LSTMs) [98], which are an extension of RNNs and are capable of learning long- or short-term dependencies. [33] uses Gated Recurrent Units (GRUs) for traffic flow prediction, which have the same capabilities as LSTMs, but they have a simpler architecture, because they have fewer parameters, since they lack an output gate. According to the authors of [33], LSTM and GRU have better performance than ARIMA, and GRUs perform a little better and usually converge faster than LSTMs.

To consider spatial correlations, authors have started to use Convolutional Neural Networks (CNNs), which have the ability to find the spatial correlations over a map, made possible by the usage of convolution layers. To exploit this ability, different approaches [48,50,60] use time–space matrix or grid-based region matrix data models.

The most popular solution is a combination of CNN and LSTM. In [98], the spatial dependencies of network-wide traffic could be captured by CNNs, and the temporal dynamics could be learned by LSTMs. Besides those, there some other custom models [50,99,100] that also utilize global environmental data sources, such as weather or traffic condition reports. They can capture spatial and temporal properties of traffic flow, and they are able to handle non-linear behavior as well. Thus, it seems that these models are the most appropriate approaches at this moment.

5. Conclusions and outlook

In our survey, we have studied different traffic flow prediction models in depth, motivated by their possible contribution to ATMS and ITS systems to forecast potential traffic conditions, thereby solving traffic management problems in smart cities.

In Section 2, we examined currently available data sources used for traffic flow prediction. In our opinion, the enumerated data sources should be used together, because every data source has its own advantage. That way, one can achieve the best result by fusing them in an appropriate model.

Among fixed position sensors, the sensors able to scan more lanes at the same time (e.g., video image processors or laser radar sensors) could be more cost effective than other fixed position solutions. These can be used to implement crowd surveillance tasks in cities as well, so they could be the eyes of future smart cities due to their versatility and flexibility.

With moving sensors, we can identify exact paths, speeds, and moving patterns of vehicles and pedestrians, which can reveal direct connections between adjacent road segments. Moving sensors have minimal infrastructure cost compared to fixed position sensors, and they are important data sources in areas that are not covered by fixed position sensors.

In addition, traffic flow is influenced by many factors like weather, the day of the week, random events, road construction, lighting conditions, etc. Consequently, integration of external environmental factors is also crucial to decrease the error of prediction.

After examining data sources we analyzed many different data models often used in the literature. We found that the right data model strongly depends on the focus of the study. Sometimes we want to examine the flow of an exact point in space, which can be interpreted as a microscopic view of the flow. In other cases, we want to determine correlations between bigger areas, which is a macroscopic way to study the flows. In general, it seems that the most promising data models take advantage of the spatio-temporal property of traffic flows, such as time–space matrix models or region-based models. Nevertheless, there is a need for a data model that also works when the particles of the flow do not move in the same direction as vehicles. This kind of model could be very useful for the investigation of real-time pedestrian flows.

Before prominent empirical traffic prediction approaches were introduced, the predictability of traffic flows was examined as it is a fundamental knowledge for selecting the appropriate prediction model. Traffic flows are non-linear, mostly non-stationary processes influenced by many factors such as weather, the day of the week, unexpected events, roads construction, and lighting conditions. It also has significant spatio-temporal properties. Enumerated articles usually made comparisons to other methods to predict traffic flows; however, the outcome of these comparisons depended highly on the data sources used, the settings of the model's parameters, the traffic scenario, etc. We also wished to highlight the advantages and disadvantages of the models.

We have found that the most prominent prediction models are the non-parametric ones, because they are able to handle non-linear, stationary or non-stationary, dynamic processes, and they can exploit the spatio-temporal relationship of traffic flows as well. Some of them are also capable of incorporating environmental data sources, which can increase the accuracy of the predictions in most cases. From the investigated non-parametric models, different neural networks are currently the most popular (such as CNN, LSTM, or their combination) because of the properties introduced in Section 4.3.3.

To our knowledge, there is no widespread system available that is able to handle other types of traffic, such as pedestrian or bike. In our opinion, there is a need to develop an integrated management system able to handle vehicular, pedestrian, and bike traffic flows simultaneously. This new generation of smart city management systems could reveal correlations between vehicular, pedestrian, and bike flows of future cities, which could potentially aid in city planning and resource management tasks.

References

[1] D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel, Nature (2006) 462–465.

[2] H. Qin, H. Li, X. Zhao, Development status of domestic and foreign smart city, Glob. Presence 9 (2010) 50–52.

[3] S.-h. An, B.-H. Lee, D.-R. Shin, A survey of intelligent transportation systems, in: Computational Intelligence, Communication Systems and Networks, CICSyN, 2011 Third International Conference on, IEEE, 2011, pp. 332–337.

[4] O. Al-Kadi, O. Al-Kadi, R. Al-Sayyed, et al., Road scene analysis for determination of road traffic density, Front. Comput. Sci. 8 (4) (2014) 619–628.

[5] A.T. Chin, Containing air pollution and traffic congestion: transport policy and the environment in Singapore, Atmos. Environ. 30 (5) (1996) 787–801.

[6] M. Rosenlund, F. Forastiere, M. Stafoggia, D. Porta, M. Perucci, A. Ranzi, F. Nussio, C.A. Perucci, Comparison of regression models with land-use and emissions data to predict the spatial distribution of traffic-related air pollution in Rome, J. Exposure Sci. Environ. Epidemiol. 18 (2) (2008) 192–199.

[7] M. Krzyżanowski, B. Kuna-Dibbert, J. Schneider, Health Effects of Transport-Related Air Pollution, WHO Regional Office Europe, 2005.

[8] A. Peters, S. Von Klot, M. Heier, I. Trentinaglia, Hörmann, Exposure to traffic and the onset of myocardial infarction, New England J. Med. (2004) 1721–1730.

[9] Z.S. Qian, R. Rajagopal, Optimal dynamic parking pricing for morning commute considering expected cruising time, Transp. Res. C (2014) 468–490.

[10] J. de Gier, T.M. Garoni, O. Rojas, Traffic flow on realistic road networks with adaptive traffic lights, J. Stat. Mech. Theory Exp. 2011 (04) (2011) P04008.

[11] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, C. Chen, Data-driven intelligent transportation systems: A survey, IEEE Trans. Intell. Transp. Syst. (2011) 1624–1639.

[12] J. Yousaf, J. Li, L. Chen, J. Tang, X. Dai, Generalized multipath planning model for ride-sharing systems, Front. Comput. Sci. 8 (1) (2014) 100–118.

[13] INRIX, Inrix parking solutions, 2017. URL http://www2.inrix.com/parking-solutions. (Accessed 21 November 2017).

[14] N. Varga, L. Bokor, A. Takács, J. Kovács, L. Virág, An architecture proposal for V2X communication-centric traffic light controller systems, in: ITS Telecommunications, ITST, 2017 15th International Conference on, IEEE, 2017, pp. 1–7.

[15] B.-W. Chen, W. Ji, Intelligent marketing in smart cities: crowdsourced data for geo-conquesting, IT Professional 18 (4) (2016) 18–24.

[16] O. Capio, Traffic-based media selection, US Patent 8, 447, 421, May 21 2013. URL https://www.google.com/patents/US8447421.

[17] INRIX, Inrix retail, 2017. URL http://inrix.com/industries/retail/. (Accessed 21 November 2017).

[18] A.M. Nagy, V. Simon, Integrated mass surveillance system for large scale events, in: Smart Cities Conference, ISC2, 2016 IEEE International, IEEE, 2016, pp. 1–6.

[19] A. Vemula, N. Patil, V. Paharia, A. Bansal, M. Chaudhary, N. Aggarwal, D. Bansal, K. Ramakrishnan, B. Raman, Improving public transportation through crowd-sourcing, in: Communication Systems and Networks, COMSNETS, 2015 7th International Conference on, IEEE, 2015, pp. 1–6.

[20] N. Bellomo, C. Dogbe, On the modeling of traffic and crowds: A survey of models, speculations, and perspectives, SIAM Rev. 53 (3) (2011) 409–463.

[21] E. Bolshinsky, R. Friedman, Traffic Flow Forecast Survey, Tech. Rep., CSD, Technion, 2012.

[22] A. Ermagun, D. Levinson, Spatiotemporal traffic forecasting: Review and proposed directions, Transp. Rev. (2018) 1–29.

[23] J. van Hinsbergen, F. Sanders, Short term traffic prediction models, 2007.

[24] E.I. Vlahogianni, J.C. Golias, M.G. Karlaftis, Short-term traffic forecasting: Overview of objectives and methods, Transp. Rev. 24 (5) (2004) 533–557.

[25] N.-E. El Faouzi, H. Leung, A. Kurian, Data fusion in intelligent transportation systems: Progress and challenges–A survey, Inf. Fusion 12 (1) (2011) 4–10.
[26] L.A. Klein, M.K. Mills, D.R. Gibson, Traffic Detector Handbook: -Volume II, Tech. Rep., 2006.
[27] U.S. department of transportation, intelligent transportation systems. https://www.itscosts.its.dot.gov/its/benecost.nsf/CostHome. (Accessed 05 April 2018).
[28] A. Petkovics, V. Simon, I. Godor, B. Böröcz, Crowdsensing solutions in smart cities towards a networked society, Endorsed Trans. Internet Things 15 (2015).
[29] T. Li, D. Sun, P. Jing, K. Yang, Smart card data mining of public transport destination: A literature review, Information 9 (1) (2018) 18.
[30] K. Mohamed, E. Côme, L. Oukhellou, M. Verleysen, Clustering smart card data for urban mobility analysis, IEEE Trans. Intell. Transp. Syst. 18 (3) (2017) 712–728.
[31] C. Zhong, E. Manley, S.M. Arisona, M. Batty, G. Schmitt, Measuring variability of mobility patterns from multiday smart-card data, J. Comput. Sci. 9 (2015) 125–130.
[32] C. Zhong, M. Batty, E. Manley, J. Wang, Z. Wang, F. Chen, G. Schmitt, Variability in regularity: Mining temporal mobility patterns in london, Singapore and Beijing using smart-card data, PLoS One 11 (2) (2016) e0149222.
[33] R. Fu, Z. Zhang, L. Li, Using LSTM and GRU neural network methods for traffic flow prediction, in: Chinese Association of Automation, YAC, IEEE, 2016, pp. 324–328.
[34] P. Dell'Acqua, F. Bellotti, R. Berta, A. De Gloria, Time-aware multivariate nearest neighbor regression methods for traffic flow prediction, IEEE Trans. Intell. Transp. Syst. 16 (6) (2015) 3393–3402.
[35] Y. Lv, Y. Duan, W. Kang, Z. Li, F.-Y. Wang, Traffic flow prediction with big data: a deep learning approach, IEEE Trans. Intell. Transp. Syst. 16 (2) (2015) 865–873.
[36] PeMS data set. http://pems.dot.ca.gov/. (Accessed 03 October 2017).
[37] Highways England network journey time, traffic speed and traffic flow data. http://tris.highwaysengland.co.uk/. (Accessed 3 October 2017).
[38] H. Dong, M. Wu, X. Ding, L. Chu, L. Jia, Y. Qin, X. Zhou, Traffic zone division based on big data from mobile phone base stations, Transp. Res. C 58 (2015) 278–291.
[39] D. Xia, H. Li, B. Wang, Y. Li, Z. Zhang, A map reduce-based nearest neighbor approach for big-data-driven traffic flow prediction, IEEE Access 4 (2016) 2920–2934.
[40] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, Y. Huang, Map-matching for low-sampling-rate GPS trajectories, in: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2009, pp. 352–361.
[41] S.K. Moore, Superaccurate GPS chips coming to smartphones in 2018, 2017. URL https://spectrum.ieee.org/tech-talk/semiconductors/design/superaccurate-gps-chips-coming-to-smartphones-in-2018. (Accessed 18 November 2017).
[42] J. Yuan, Y. Zheng, X. Xie, G. Sun, Driving with knowledge from the physical world, in: Proceedings of the 17th ACM SIGKDD, ACM, 2011, pp. 316–324.
[43] Taxi service trajectory - prediction challenge, ECML PKDD 2015 dataset. http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html. (Accessed 07 October 2017).
[44] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, V.D. Blondel, D4D-Senegal: the second mobile phone data for development challenge, 2014. ArXiv preprint arXiv:1407.4885.
[45] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, B. Lepri, A multi-source dataset of urban life in the city of Milan and the Province of Trentino, Sci. Data 2 (2015) 150055.
[46] T. Maze, M. Agarwai, G. Burchett, Whether weather matters to traffic demand, traffic safety, and traffic operations and flow, Trans. Res. Rec. J. Trans. Res. B (2006) 170–176.
[47] S. Dunne, B. Ghosh, Weather adaptive traffic prediction using neurowavelet models, IEEE Trans. Intell. Transp. Syst. 14 (1) (2013) 370–379.
[48] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, Y. Wang, Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction, Sensors 17 (4) (2017) 818.
[49] A. Abadi, T. Rajabioun, P.A. Ioannou, Traffic flow prediction for road transportation networks with limited traffic data, IEEE Trans. Intell. Transp. Syst. 16 (2) (2015) 653–662.
[50] J. Zhang, Y. Zheng, D. Qi, R. Li, X. Yi, DNN-based prediction model for spatio-temporal data, in: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2016, p. 92.
[51] D. Billings, J.-S. Yang, Application of the ARIMA models to urban roadway travel time prediction-a case study, in: Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on, vol. 3, IEEE, 2006, pp. 2529–2534.
[52] S. Lee, D. Fambro, Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting, Transp. Res. Rec. J. Transp. Res. Board (1678) (1999) 179–188.
[53] B.M. Williams, L.A. Hoel, Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results, J. Transp. Eng. 129 (6) (2003) 664–672.
[54] M. Van Der Voort, M. Dougherty, S. Watson, Combining Kohonen maps with ARIMA time series models to forecast traffic flow, Transp. Res. C 4 (5) (1996) 307–318.
[55] J. Guo, W. Huang, B.M. Williams, Adaptive kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification, Transp. Res. C 43 (2014) 50–64.
[56] B. Ghosh, B. Basu, M. O'Mahony, Bayesian time-series model for short-term traffic flow forecasting, J. Transp. Eng. 133 (3) (2007) 180–189.
[57] G. Huisken, E.C. van Berkum, A comparative analysis of short-range travel time prediction methods, in: Transportation Research Board Annual Meeting, CD-Rom, 2003.
[58] S. Li, Z. Shen, F.-Y. Wang, A weighted pattern recognition algorithm for short-term traffic flow forecasting, in: Networking, Sensing and Control, ICNSC, 2012 9th IEEE International Conference on, IEEE, 2012, pp. 1–6.
[59] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, J. Sun, A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting, Transp. Res. C 62 (2016) 21–34.
[60] Y. Li, R. Yu, C. Shahabi, Y. Liu, Graph convolutional recurrent neural network: Data-driven traffic forecasting, 2017. ArXiv preprint arXiv:1707.01926.
[61] S. Sun, C. Zhang, Y. Zhang, Traffic flow forecasting using a spatio-temporal bayesian network predictor, in: International Conference on Artificial Neural Networks, Springer, 2005, pp. 273–278.
[62] J. Krumm, E. Horvitz, Predestination: Where do you want to go today? Computer 40 (4) (2007).
[63] J.W. Powell, Y. Huang, F. Bastani, M. Ji, Towards reducing taxicab cruising time using spatio-temporal profitability maps, in: SSTD, Springer, 2011, pp. 242–260.
[64] N.J. Yuan, Y. Zheng, X. Xie, Segmentation Of Urban Areas Using Road Networks, MSR-TR-2012–65, Tech. Rep., 2012.
[65] H. Gonzalez, J. Han, X. Li, M. Myslinska, J.P. Sondag, Adaptive fastest path computation on a road network: a traffic mining approach, in: Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB Endowment, 2007, pp. 794–805.
[66] A. Sang, S.-q. Li, A predictability analysis of network traffic, Comput. Netw. 39 (4) (2002) 329–345.
[67] Y. Yue, A.G. Yeh, Y. Zhuang, Prediction time horizon and effectiveness of real-time data on short-term traffic predictability, in: Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE, IEEE, 2007, pp. 962–967.

[68] R.B. Noland, J.W. Polak, Travel time variability: a review of theoretical and empirical issues, Transp. Rev. 22 (1) (2002) 39–54.

[69] B.L. Smith, B.M. Williams, R. Oswald, Comparison of parametric and nonparametric models for traffic flow forecasting, Transp. Res. C 10 (4) (2002) 303–321.

[70] J. Whittaker, S. Garside, K. Lindveld, Tracking and predicting a network traffic process, Int. J. Forecast. 13 (1) (1997) 51–61.

[71] A. Stathopoulos, M. Karlaftis, Temporal and spatial variations of real-time traffic data in urban areas, Transp. Res. Rec. J. Transp. Res. Board (1768) (2001) 135–140.

[72] J. Wang, Y. Mao, J. Li, Z. Xiong, W.-X. Wang, Predictability of road traffic and congestion in urban areas, PLoS One 10 (4) (2015) e0121825.

[73] H. van Lint, M. Schreuder M.S.c, Travel time prediction for variable message sign panels: Results and lessons learned from large-scale evaluation study in the Netherlands, in: Transportation Research Board 85th Annual Meeting, no. 06–2045, 2006.

[74] Y. Kamarianakis, P. Prastacos, Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches, Transp. Res. Rec. J. Transp. Res. Board (1857) (2003) 74–84.

[75] Waze, Routing server - waze. URL https://wiki.waze.com/wiki/Routing_server#Routing_requests. (Accessed 13 November 2017).

[76] M. Beckmann, C.B. McGuire, C.B. Winsten, Studies in the Economics of Transportation, Tech. Rep., 1956.

[77] C. Xiong, Z. Zhu, X. He, X. Chen, S. Zhu, S. Mahapatra, G.-L. Chang, L. Zhang, Developing a 24-hour large-scale microscopic traffic simulation model for the before-and-after study of a new tolled freeway in the Washington, DC, J. Transp. Eng. 141 (6) (2015) 05015001.

[78] M. Miska, Real time traffic management by microscopic online simulation, T2007/1, January, 2007.

[79] D. Ngoduy, R. Wilson, Multianticipative nonlocal macroscopic traffic model, Comput.-Aided Civ. Infrastruct. Eng. 29 (4) (2014) 248–263.

[80] C. Kemper, Dynamic traffic flow model–a new approach with static data, in: Proceedings of the 5th European Congress and Exhibition on Intelligent Transport Systems, ITS, 2005, pp. 1–13.

[81] H. Liu, H.J. Van Zuylen, H. Van Lint, Y. Chen, K. Zhang, Prediction of urban travel times with intersection delays, in: Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE, IEEE, 2005, pp. 402–407.

[82] Y. Chiu, L. Zhou, H. Song, Development and calibration of the anisotropic mesoscopic simulation model for uninterrupted flow facilities, Transp. Res. B 44 (1) (2010) 152–174.

[83] N.L. Nihan, K.O. Holmesland, Use of the box and jenkins time series technique in traffic forecasting, Transportation 9 (2) (1980) 125–143.

[84] B. Williams, P. Durvasula, D. Brown, Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models, Transp. Res. Rec. J. Transp. Res. Board (1644) (1998) 132–141.

[85] W. Min, L. Wynter, Real-time road traffic prediction with spatio-temporal correlations, Transp. Res. C 19 (4) (2011) 606–616.

[86] B. Williams, Multivariate vehicular traffic flow prediction: evaluation of arimax modeling, Transp. Res. Rec. J. Transp. Res. Board (1776) (2001) 194–200.

[87] C. Chen, J. Hu, Q. Meng, Y. Zhang, Short-time traffic flow prediction with ARIMA-GARCH model, in: Intelligent Vehicles Symposium, IV, 2011 IEEE, IEEE, 2011, pp. 607–612.

[88] A. Anand, G. Ramadurai, L. Vanajakshi, Data fusion-based traffic density estimation and prediction, J. Intell. Transp. Syst. 18 (4) (2014) 367–378.

[89] F. Yang, Z. Yin, H. Liu, B. Ran, Online recursive algorithm for short-term traffic prediction, Transp. Res. Rec. J. Transp. Res. Board (1879) (2004) 1–8.

[90] C.-C. Lu, X. Zhou, Short-term highway traffic state prediction using structural state space models, J. Intell. Transp. Syst. 18 (3) (2014) 309–322.

[91] S. Sun, C. Zhang, G. Yu, A Bayesian network approach to traffic flow forecasting, IEEE Trans. Intell. Transp. Syst. 7 (1) (2006) 124–132.

[92] A. Pascale, M. Nicoli, Adaptive bayesian network for traffic flow prediction, in: Statistical Signal Processing Workshop, SSP, 2011 IEEE, IEEE, 2011, pp. 177–180.

[93] C.M. Queen, C.J. Albers, Intervention and causality: forecasting traffic flows using a dynamic bayesian network, J. Amer. Statist. Assoc. 104 (486) (2009) 669–681.

[94] B. Sun, W. Cheng, P. Goswami, G. Bai, Flow-aware wpt k-nearest neighbours regression for short-term traffic prediction, in: Computers and Communications, ISCC, 2017, IEEE, 2017, pp. 48–53.

[95] S. Oh, Y.-J. Byon, H. Yeo, Improvement of search strategy with k-nearest neighbors approach for traffic state prediction, IEEE Trans. ITSs 17 (4) (2016) 1146–1156.

[96] X. Zeng, Y. Zhang, Development of recurrent neural network considering temporal-spatial input dynamics for freeway travel time modeling, Comput.-Aided Civ. Infrastruct. Eng. 28 (5) (2013) 359–371.

[97] Y. Bengio, P. Frasconi, P. Simard, The problem of learning long-term dependencies in recurrent networks, in: Neural Networks, IEEE International Conference on, 1993, pp. 1183–1188.

[98] H. Yu, Z. Wu, S. Wang, Y. Wang, X. Ma, Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks, 2017. ArXiv preprint arXiv:1705.02699.

[99] N.G. Polson, V.O. Sokolov, Deep learning for short-term traffic flow prediction, Transp. Res. C 79 (2017) 1–17.

[100] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, Y. Liu, Deep learning: A generic approach for extreme condition traffic forecasting, in: Proceedings of the 2017 SIAM International Conference on Data Mining, SIAM, 2017, pp. 777–785.