Transportation Engineering 2 (2020) 100013

Contents lists available at ScienceDirect

Transportation Engineering

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

Smart transportation planning: Data, models, and algorithms

Zahra Karami[a], Rasha Kashef[b,*]

[a] Computer Science Department, Ryerson University, Toronto, Canada
[b] Electrical, Computer, and Biomedical Engineering Department, Ryerson University, Toronto, Canada

ARTICLE INFO

Keywords:
Smart transportation
Machine learning
Time-series
Prediction Models

ABSTRACT

By developing cities and increasing population, smart transportation becomes an essential component of modern societies. Extensive research activities using machine learning techniques and several industrial needs have paved the way for the emerging field of smart transportation. This paper presents data, methods, and models that are essential for intelligent planning of transportation. In particular, the current data sources for gathering information to control or forecast traffic are described, connected Vehicles (CVs) that bring smart and green transportation to modern life is also discussed. Clustering Analysis as an effective unsupervised machine learning method in trip distribution and generation and traffic zone division is discussed in the paper. Various machine learning techniques and models that use time series prediction are introduced in this paper including ARIMA, Kalman filtering, Holt winters'Exponential smoothing, Random walk, KNN Algorithm, and Deep Learning. Finally, a discussion on the main advantages and drawbacks of these models, as well as the business adoption of the forecasting models are presented.

1. Introduction

In modern civilization, the increasing population faces mobility, sustainability challenges, and transportation play a crucial part in that it connects people, spreads different cultures, and ultimately promotes evolution. As innovations started to burst in the last two centuries, the transportation system became more sophisticated. The complex transportation system brought both advantages and disadvantages: while transportation tools have been more affordable, traffic congestions appeared much more often as a consequence. Therefore, systematic and efficient transportation planning and learning are essential to address and handle the complex transportation system. Spatiotemporal forecasting is an example of such a complex learning task. This task is mainly challenging due to the complexity of spatial dependencies and temporal dependencies. The main target of traffic forecasting is to predict the future flow using previous traffic speeds as well considering the other parameters such as road conditions. Reoccur incidents such as collisions, or bad weather can generate models out of the pattern. While the traditional techniques for the transportation data collection process were through surveys and census, gathering corresponding data was time-consuming and costly. Thus, people are now utilizing and incorporating data from different sources into the analysis, primarily non-traditional clustering data, including GPS and GIS data. Meanwhile, mobilized data through personal devices are also becoming more crucial with the increasing popularity of cell phones and smart public transportation cards. After getting the data, the trip generation has always been used as the first step to build up the transportation forecast model. Time series analysis, which predicts values based on time, is now the most acknowledged and practiced prediction methodology.

Future research in smart transportation planning requires a road map to answer the following question: which models are used for intelligent forecasting models, which data sources are available, and what is the available business adoption to the forecasting models. This paper discusses several analytical techniques and models in addressing transportation planning issues and the metrics in measuring the quality and accuracy of these techniques. In this paper, we will cover state-of-art time series models, including the autoregressive moving average (ARIMA) model, KNN, KD, Holt-Winters' exponential smoothing model, random walk model, and deep learning. Connected Vehicles (CV) is the next generation of smart transportation systems; this technology brings intelligent and green transportation to modern life. In this paper, connectivity, safety, and traffic management are introduced to pave the way for discussion on connected vehicles and smart transportation planning.

The organization of the paper is as follows: Section 2 discusses the data sources in transportation analysis. Connectivity, safety, and management in connected vehicles are presented in Section 3.

Clustering Analysis and its adoption in transportation planning, including trip distribution and planning, are discussed in Section 4. We then introduce the machine learning techniques applied in traffic flow and time series prediction in Section 5. Section 6 introduced some Deep

* Corresponding author.
E-mail addresses: zahra.karami@ryerson.ca (Z. Karami), rkashef@ryerson.ca (R. Kashef).

https://doi.org/10.1016/j.treng.2020.100013
Received 11 June 2020; Received in revised form 9 July 2020; Accepted 10 July 2020

Learning methods used to further enhance the performance of time series prediction models discussed in Section 5. The accuracy of time series prediction models is presented in Section 7. Business opportunities for traffic forecasting are discussed in Section 8.

2. Data sources in smart transportation analysis

In smart transportation planning, various data sources are used in developing intelligent systems; some of these data sources are characterized by different configurations, properties, and sizes, as shown next.

2.1. GPS and GIS data

GPS and GIS technologies are considered new sources of collecting transportation data, especially travel data. GPS data could provide real-time spatial and temporal information. It shows the travel behavior, including distance, travel speed, trip time, and other information in digital formats at the same time, which could reduce the burden of reporting information. There are two categories of GPS data, the real-time tracking GPS and the logging GPS. The real-time tracking GPS devices record the running vehicle's coordinates per second. There are a few drawbacks to using GPS data. Firstly, statistics are not representing all the time because of the unnecessary selection criteria. Uncleared criteria would affect the accuracy of the result. Also, this method requires that all surveyed subjects must have GPS devices. Therefore, the collected data may not be complete and representative. Moreover, the cost of collecting data is quite high because of the complex collection process and the extra costs of hiring data collectors [1].

2.2. Traffic flow data source

Sensors and detectors are installed on sites along the highways to collect vehicle volume data. These data include characteristics such as traffic flow (denoted by volume/hour), lane occupancy, and average speed of vehicles. The collected data is then analyzed using different methods and models to derive data-driven solutions [2].

2.3. Smart card

Smart card data are used to analyze personal travel patterns using specific transportation tools. The benefit of using smart card data is that data can show the travel start time, end time, and travel direction. Based on the frequency of different destinations, the management team can predict the traffic flow and create a proper stuff schedule. However, since smart card data can only show the traffic data under certain transportation, the flexibility of data is limited [3].

2.4. Mobile phone

Since trips on transit are getting complex, data from in-vehicle segments could not reflect all the segments within a trip. The usage of mobile phone data can provide an accurate picture of the user's location. Smartphones provide increasingly feasible to collect individual-level location data over a long period and with the low respondent burden, forming personal travel records. Mobile phone data provide a high-resolution image of the travel path, including in-vehicle or out-vehicle segments trips [9]. However, there are some drawbacks associated with the data. The volume of personal location data would be enormous, and data cleaning would be a tough job. Moreover, data collection of mobile phone data might be inconsistent if the devices are without power. Charging equipment would be required under the possession of data collection.

2.5. Call detail record (CDRs)

Call detail records (CDRs) are the result of the rapid advancement in mobile technology, are automatically collected by mobile phone carriers for billing. CRDs contain information as timestamped and longitude-latitude coordinates of anonymized customers. Therefore, the information gathered from CRDs is collected more regularly and cost-effectively than traditional travel surveys, commonly performed once or twice per decade. CDRs can effectively capture individual' trip routes and are compatible with the analysis of transportation models [5].

3. Connected vehicles

Connected Vehicles (CVs) are known as the next generation of smart transportation systems. This Technology will bring intelligent and green transportation to modern life and pave the way for various applications such as road safety or service-based applications. Moreover, connected vehicles are emerging as the Internet of Things (IoT) in transportation known as the Internet of Vehicles (IoV). Connected vehicle technology increases the efficiency and reliability of autonomous vehicles by improving efficiency and drivers' comfort while increasing mobility and safety [15].

3.1. Connectivity in connected vehicles

Connected vehicles are wireless connectivity-enabled vehicles that can communicate with their internal and external environments [15]. This connectivity is called vehicles-to-x (V2X), which supports the interaction of vehicle, and it is provided on different levels, i.e., vehicle-to-sensor on-board (V2S), vehicle-to-vehicle (V2V), vehicle-to-road infrastructure (V2R), and vehicle-to-Internet (V2I). These levels of connectivity will discuss further in this part. Wireless Technology is the most knowns infrastructure in CVs besides the other alternatives such as Bluetooth, Ultra-Wideband, etc. The wireless is used as a solution to provide vehicle-to-x connectivity.

3.1.1. Intra-vehicle connectivity

Nowadays, modern cars are facilitated by various types of sensors for different purposes. The intra-vehicles wireless sensors are still challenging due to some facts and characteristics. Sensors need to be connected to Electrical control units (ECU) that require a cable connection that leads space in cars and, in some cases, may add considerable weight to cars. Sensors are stationary, so they need maintenance of aftermarket to update or repair. There are also some concerns related to data security and reliability. These sensors provide data transmissions requiring low-latency and high reliability to satisfy the needs and requirements of the real-time control system.

3.1.2. Inter- vehicle connectivity

The Inter-Vehicle communication (V2V) enables a productive connection between vehicles that can have a key role in increasing road safety for passengers without the assistance of any built infrastructure by using a vehicular ad hoc network (VANET). However, the VANET faces some challenges in the urban area is surrounding obstacles like buildings that can lead to a broken connection and connection lost. Another VANET problem is the data flow disconnection due to the limited range of V2V communication.

3.1.3. V2I and V2R connectivity

With recent advances in technology, the internet-connectivity becomes a significant property of modern vehicles. Internet-connected cars (V2I) enables vehicles to use a various range of online applications and services. V2R is an effective solution to overcome many issues, including road safety and traffic congestions. These aims are achieved by connecting the vehicle and ITSs infrastructures, such as a street sign, traffic lights, and road sensors.

3.2. Safety in connected vehicles

The ability that vehicles can communicate with other vehicles (V2V) and infrastructure (V2I) provides the opportunity to enhance road safety. Connected-vehicles technology can influence all aspects of driving decision-making by enhancing reliable decisions during the operational time. Autonomous cars are the primary beneficiary of this technology. Humans have higher reaction time and uncertainty compared to robots, while robots can consider more variables into account to make a decision that results in more stable behaviors [17]. The V2V and V2I communications provide information about related to road and vehicle. Driver behavior can be influenced based on the information received, and this ability allows to enhance the performance of driver assistance technologies. Using the on-board sensors, the systems can adjust their acceleration and space based on the leader or following vehicle behavior. The V2I communication provides informative details of changes in the speed limit, work zone condition, weather condition, roadway condition, geometry. The entire information received by V2V and V2I helps drivers' decisions with having an optimal and safe lane-selection, route choice, and speed.

3.3. Smart traffic mangament in connected vehicles

The connected vehicles (CVs) are very promising to alleviate traffic congestion via smart traffic management. CVs technology provides real-time data about traffic conditions that lead to better traffic management by improving data quality [16]. The two types of car-following stability have been identified related to CVs: local stability and string stability. Local stability refers to the vehicle's response to its leader's acceleration decisions. String stability is defined for a group of vehicles and investigates the behavior of the entire group in response to its leader. The main focus of traffic flow management studies is to investigate the string stability of traffic flow. Accordingly, a model representing the connected-vehicles environments is Automated Highway Systems (AHS), where fully autonomous vehicles are operated on a set of designated lanes to have automation and connectivity. An AHS's performance is a function of vehicle movement strategies (control laws) and decisions of the Traffic Management Center (TMC). AHS investigates the root of traffic congestion and proposes a series of actions to eliminate it, and on some levels, it can prevent traffic congestion in advance. Besides the term traffic management, it can guarantee a collision-free system.

4. Clustering analysis in smart transportation

Clustering is an unsupervised machine learning technique, has proven its efficiency is developing intelligent transportation systems. Clustering has been applied in various categories in transportation planning as trip generation, traffic zone division, and trip distribution, as discussed below.

4.1. Trip generation

The first step of the traditional transportation forecast model is the trip generation [8]. In this step, the goal is to estimate the number of trips produced or originated in each traffic analysis zone. However, with the new dynamics of compiling different data sources within one analysis, the traditional approach could no longer accommodate the problem. Alternatively, clustering analysis is used to assess the origin-destination trips and the traffic zone division. With advancements in mobile technology and the rising popularity of mobile phone usage, more spatial-temporal information becomes available. Call detail records (CDR) are produced with the use of a mobile phone. Each record contains the anonymous user ID, timestamp, and geographic information of the phone user. The data is collected through the phone carrier from its users and can be acquired in real-time. However, information needs to be extracted from the CDR data to satisfy further transportation research needs [4]. It uses clustering analysis to convert CDR data into clustered locations then make inferences about the origins and destinations.

4.2. Traffic zone division

Traffic zone division is mainly based on big data from mobile phone base stations [11]. Traffic zone division acts as an essential input to accurately calibrate the travel demand, forecasting model. Traffic zoning simplifies complex urban traffic network and serves as a fundamental of traffic planning. Traditional division methods rely heavily on the social, economic characteristics of the area and natural and administrative boundaries. Since the slow update, and difficulty in quantifying these characteristics, the traditional method cannot reflect the characteristic timely and consistently. However, by clustering the CDRs collected, the traffic zones can be found through a data-driven approach in real-time [10]. This increases the reliability and accuracy of further analysis that uses the traffic zone division information. The traffic zone inference can be drawn from combining geographic data and timestamps stored in the CDRs records. This data-driven method alleviates bias from assigning traffic zone subjectively from social-economic information and based the assignment on transportation data, making the traffic zone division more relevant in the later stage of traffic analysis.

4.3. Trip distribution

Understanding the origin and destination (O-D) trip helps determine the magnitude of f total daily travel in a given transportation system. To estimate average daily origin-destination trips and solve travel flow problems, an agglomerative clustering algorithm is used to determine the clustered locations or destinations that stand for any place where the objects spend some time [6]. The agglomerative clustering is the bottom-up approach of the hierarchical clustering; each observation starts in its cluster. Then two groups of observations that have the smallest distance are merged. The combining process is repeated until there is only one cluster left in the end. The advantage of the agglomerative clustering algorithm is that it allows the classification of clusters in the spatial scale.

4.4. Similarity measures

Similarity and dissimilarity measures are core factors of machine learning techniques. They measure the strength or the divergence of the data-points' relationship. These measures impact the learning process and results. For transportation planning methods that employ both supervised and unsupervised analysis, some of the most common similarity measures including Euclidean distance, Manhattan Distance, Minkowski Distance, Cosine Coefficient, and Jaccard Similarity [4,6,8,10].

5. Traffic flow prediction: time series forecasting models

The traffic flow on most expressways exhibits the characteristic weekly pattern. Peak hours during weekdays often present in the mornings and evenings. Weekends peak hours usually occur around noon. The roads are congested during rush hours and less congested during non-peak hours. This pattern can be captured by time series analysis. Time series regression is a regression on time, and it captures factors such as trends, seasonal variations, cycles, and irregular components. Based on learning outcomes from the historical traffic condition data analysis, time series model predicts the traffic conditions for a future time. Traffic forecasting studies usually falling into two main categories: a data-driven approach and a knowledge-driven approach. In transportation and operational research, knowledge-driven methods usually apply queuing theory and simulate user behaviors in traffic in time series. Some of the popular data-driven methods will be introduced next.

5.1. ARIMA model

Auto-Regressive Integrated Moving Average (ARIMA) is a class of models that can predict based on historical values. It is known as one of the most precise methods for predicting traffic flow compared to the other available methods. An ARIMA model consists of three parts: autoregressive (AR) part, differencing (I) part, and moving average (MA) part. AR Part in ARIMA is a linear regression model that uses its own lagged values as a predictor, and it works well when the lags are not correlated. Hence, in the ARIMA, the time series should be stationary. As a result, the most common attitude to removing non-stationarity from any time-series data is to differentiate it. The I (for "integrated") indicates that the data values have been substituted with the difference between current values and the previous values. Sometimes, depending on the complexity of the series, the differencing may have processed more than once. ARIMA models are defined by ARIMA(p,d,q) where parameters p, d, and q are non-negative integers, p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model. If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'. Seasonal ARIMA models are usually denoted ARIMA(p,d,q)(P,D,Q)m, where m refers to the number of periods in each season, and the uppercase P,D,Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model. An ARIMA(p,d,q) model is given by:

$$\begin{aligned} p &: y_t = \sum_{i=1}^{p} \alpha_i y_{t-1} + Z_t \\ d &: y'_t = y_t - y_{t-1} \\ q &: y_t = Z_t + \sum_{j=1}^{q} \beta_j Z_{t-1} \end{aligned} \quad (1)$$

Where y_t is time-series observation, y_{t-1} is the previous observation, α_p is the coefficient of the auto regressive process, β_q is the coefficient of the moving average process and Z_t is a white noise sequence. Finally, The ARIMA prediction equation using the backshift operator is defined as:

$$\emptyset(B)(1-B)^d y_t = \theta(B)\, Z_t \quad (2)$$

Where $\emptyset$ is the polynomial degree of p and θ is the polynomial degree of q. Since the traffic flow reveals a strong seasonal pattern due to peak and off-peak traffic conditions, which is usually repeating during a period, the seasonal ARIMA (SARIMA) model is a suitable choice to model traffic flow behavior. The trend is filtered through the differencing part [12]. Then, the model eliminates random shock factors by generating the moving averages. Suppose the one observation of data involves the number of vehicles passed by a detector in a 5-min interval, the moving average is generated by calculating the average of four 5-min intervals. This is based on the assumption that future moving average is equal to historical moving averages. If the average number of vehicles observed in a 5-min interval for the past five days during peak hours is 100, then it is believed that the number of vehicles that will pass by the detector within a 5-min interval during peak hours is also 100.

5.2. Kalman filtering

The Kalman filter algorithm is one of the methods that can be used to predict real-time traffic flow in urban areas. This method can be implemented by both fix-sensors and connected vehicles (CV). Since CVs do not need any infrastructure or installation, using the data that is driven from connected vehicles is much low-cost than fixed sensors. Besides, CV technology is more precise. However, the idea of connected vehicles may take a long time to become available [13]. Despite batch estimation techniques that need historical measured data in each step to predict the next step, the Kalman filter is a recursive estimator. This means that it only needs to store the last state to update the prediction. For this reason, it is mentioned as a light method to predict the traffic. To use the Kalman filter to estimate the traffic flow, one must specify the following matrices:

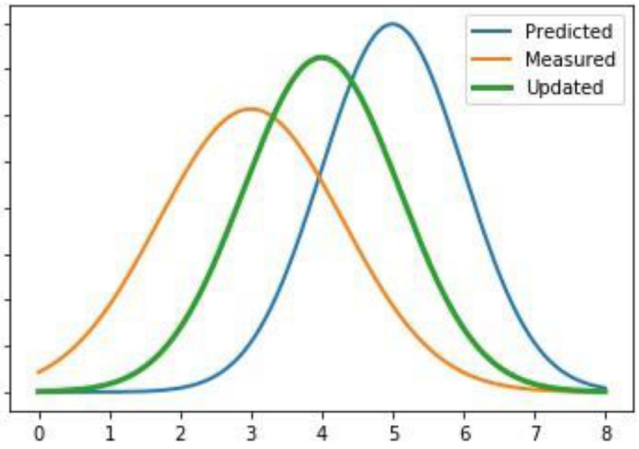

Fig. 1. The Kalman forecasting two-step process.

- F_t, the state-transition model, which maps previous states into the next state x_{t+1} .
- H_t, the observation model, which maps the true state space into the observed space
- Q_t, the covariance of the process noise;
- R_t, the covariance of the observation noise;

So, the prediction model can be written as:

$$x_{t+1} = F_t x_t + w_t \quad (3)$$

Where $x_t$is the state variable in step t. Also, the model includes a noise vector w_t which is assumed to be drawn from a zero-mean multivariate normal distribution, $\mathcal{N}$, with covariance Q_t: $w_t \sim \mathcal{N}(0, Q_t)$. At time t, an observation z_t of the true state x_t is made according to:

$$z_t = H_t x_t + v_t \quad (4)$$

Where v_t is the observation noise, which is assumed to be zero-mean Gaussian white noise with covariance R_t : $v_t \sim \mathcal{N}(0, R_t)$. The Kalman filter is mentioned as a consistent and suitable method for high volatile traffic flows, for the reason that in each step of work, the model updates repeatedly and presents the real-time traffic flow. The Kalman filter can be applied to multi-input and multi-output. It works in a two-step process: "Predict" and "Update". In the predict step, the Kalman filter produces estimates of the current state variables, and their uncertainties. Once the outcome of the next measurement is observed, these estimates are updated using a weighted average, with more weight being given to estimates with higher certainty. This process is shown in Fig. 1. This predicted state estimate is also known as the a priori state estimate. In the update step, the predicted state is combined with current observation information to improve the state estimate. If an observation is unavailable for some reason, the update may be. Likewise, if multiple independent observations are available simultaneously, multiple update steps may be performed (typically with different observation matrices H_t).

5.3. Holt winters' exponential smoothing

Holt-Winters' trend and seasonal smoothing technique is a generalized version of exponential smoothing, and this model deals with variations in trends and seasonality factors over time. Seasonality is defined as the tendency of time-series data to present behavior that replicates itself in each period. L is the season length in periods. The observations can illustrate a linear climbing trend, which is called exponential growth or a damped trend. In the case of traffic flow at weekends, the seasonality is additive. The exponential smoothing is a procedure for repeatedly

Table 1
Weights for an observation for a value of $y = 0.2$ at each time instance of K.

Observations	Weight	Value
Y_k	0.2	0.20
Y_{k-1}	0.2	0.16
Y_{k-2}	0.2	0.13
Y_{k-3}	0.2	0.10
Y_{k-4}	0.2	0.082

updating a prediction in the light of more recent experience. Exponential smoothing assigns exponentially decreasing weights as the observation gets older. In other words, most current observations are given relatively more weight in prediction than the older observations. For example, Table 1 shows the weights for observation for a value of $y = 0.2$ at each time instance. The weights decrease exponentially, and the recent observations have a more significant impact on the forecast. Generally, in metropolis cities, weekday traffic flow patterns differ from the weekend traffic flow patterns, and The Holt-Winters exponential smoothing works well when the data has both trend and seasonality. Hence, this model gives highly competitive forecasts and match considerably well with the observed traffic flow data during peak hours. Time-series methods like ARIMA develop a model where the prediction is a weighted linear sum of recent past observations. Holt-Winters' Exponential Smoothing methods are similar in that a prediction is a weighted sum of past observations, but to forecast new data assigns weights to previous data. The new data is predicted based on old information. The weight of older traffic flow data decreases exponentially because the older data has less impact on predicting new data. The most current observation of traffic data receives the maximum weight..

In Traffic flow prediction, Triple Exponential Smoothing is used for data that shows trend and seasonality [14]. To forecast the additive model, we need to calculate and update three indicators of level, trend, and seasonality as defined below:

I: smoothing level index, L_k

$$R_k = \propto * \frac{y_k}{S_{k-L}} + (1 - \propto) * (R_{k-1} + b_{k-1}) \tag{5}$$

II: smoothing trend index, b_k

$$b_k = \beta * (S_k - S_{k-1}) + (1 - \beta) * b_{k-1} \tag{6}$$

III: smoothing the seasonal Index, S_k

$$S_k = \gamma * \frac{y_k}{R_k} + (1 - \gamma) * S_{k-L} \tag{7}$$

α, β, γ are the level, trend and seasonal smoothing parameters respectively. The Seasonality Index (SI) of a period indicates how much this period typically deviates from the annual average. By using these three indicators, the multiple-step ahead can be predicted as:

$$Y_{k+T} = (R_{k-1} + T * b_{k-1}) * S_{k+T-L} \tag{8}$$

5.4. Random walk model

One of the simplest and most used naïve models in time series modeling is the random walk model. This model assumes that each period of variable takes a random step away from the previous value, and the steps are independently and identically distributed in size. This means the model uses the most recent observations for forecasting the next level. Hence, the new data equal to the old observation plus a random error (with zero mean and constant variance).

$$y_t = y_{t-1} + \varepsilon_t \tag{9}$$

Where $\varepsilon_t \sim N(0, \sigma^2)$ that generated from the normal distribution and y_t is observed data in the instance of time t. The Random walk model is applied to non-stationary traffic flow data. One of the limitations of the random walk model is that it does not depend on historical data and it is good for a one-step forward prediction. In this case, the random walk model can be used to forecast the next 24 hours' traffic flow condition [12].

5.5. KNN algorithm and KD tree

K-nearest neighbor (KNN) Algorithm as one of the machine learning methods has been implemented to tackle urban traffic issues. The KNN algorithms hold a collection of training instances. Each instance consists of a series of features and is associated with a target that is the most similar, near the target to an instance. The similarity is calculated based on distance metrics such as Euclidean distance that is indicated as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{10}$$

The k in the KNN algorithm points to k nearest points to the new instance. The target is indicated by finding the closest object to the new instance. For time series prediction, each target represents a collection of values in time series [21]. For example, the collection includes daily time-series observations for several months, and the goal is the prediction of future days. Time series can contain a repetitive pattern. The goal of the KNN algorithm is finding the most similar patterns in that past and hoping that the current observation shows similar values. Another area of using KNN in smart transportation is the best route to the destination. In this term, euclidean distance is used to find a similar area that categorizes the information about the time needed and the radius of the area that is affected by congestion or car accidents in the traffic network. The KNN algorithm is based on coordinates and midpoints of roads in different cities to select the best road with less traffic for drivers. Dispersing heavy traffic to light traffic areas is the optimization process to solve the traffic flow problem. The advantage of this algorithm is that the accuracy rate is relatively high. However, one of the drawbacks of the KNN algorithm is that the calculation requires a long time when k is too large; Another drawback is that KNN is sensitive to irrelevant features and the scale of data. Because of those drawbacks, KNN performance well when in short-term traffic prediction, but it is not consistent with using in real-time analysis for the long-term. To monitor real-time traffic change for the long-term and get up-to-date information quickly. KD Tree is introduced to build the data index that expects to search space more effectively [7]. K denotes the number of dimensions, and the KD tree is the multi-dimensional data structure that divides the space into different sections. Then the algorithm calculates the nearest point between the given location instead of calculating all points each turn. In Table 2, a summarization of the advantages, the disadvantages, and applications for ARIMA, Kalman Filtering, Holt Winter's, Random Walk, and KNN is presented.

6. Deep learning time-series prediction methods

Time-series traffic flow prediction methods have their own's difficulties. Time-Series often contain temporal dependencies that lead to low stability, high data requirements, or poor adaptability. In light of development in deep learning, some hybrid methods have improved the time series predictions and alleviate their problems, some of these methods are defined next. Hybrid deep modeling has been applied in various applications including traffic flow prediction [26] and short-term forecasting [30].

6.1. LSTM and ARIMA

This model takes the features of improved long short-term memory neural network (LSTM)), which is derived from the Recurrent Neural

Table 2
Compariosn between time-series prediction methods.

Method	Advantages	Disadvantages	Applications
ARIMA	It is more precise than the other methods	It needs huge historical data	Traffic Flow [18], Road traffic prediction [26], urban roadway travel time prediction [27]
Kalman filtering	It is a recursive estimator and does not need historical measured data	It is suitable for short-term prediction. The accuracy is not so satisfying	Traffic flow prediction [28], Traffic Stream Density estimation [29], short-term traffic flow prediction [13].
Holt Winters' Exponential Smoothing	It works well when the data has both trend and seasonality	Finding the seasonality period among the data is challenging	passenger flow predicton [30], network traffic modeling [31], daily traffic prediction [32]
Random Walk Model	It depends only on the current observation and not on the previous values	It is suitable for short-term prediction	urban traffic simulation and optimizing [33], Intelligent Transportation modeling [34], mobile ad-hoc networks [35]
KNN	It has a simple implementation It is suitable when there is little or no prior knowledge about the distribution of the data.	Poor performance on large number of data, It is very sensitive to irrelevant or redundant features	Intelligent transportation [22], Traffic Volume Forecasting [23], travel time prediction [24], traffic state prediction [25]

Networks (RNN) model, and combines it to ARIMA to increase the accuracy of ARIMA model. In LSTM architecture, additional to standard feedforward connections in RNN, units have feedback connections. A common LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell. This model of LSTM provides an ability to learn longterm dependency information. To address the over-fitting problem of the natural network, Hinton proposed a solution named dropout [19] that is used to improve the accuracy of the network. Self-Adaptive Probabilities LSTM Neural Network (SDLSTM) poses the neural network unit by a certain probability during the training process and then retain the unit in the next training epoch, and repeats this process. In this way, the network structure changes in each training process that helps with reducing the over-fitting in the training process. According to the empirical result, the SDLSTM method shows higher accuracy in comparison to LSTM in the time of heavy traffic [18]. In terms of traffic flow predictions, the predictions can be improved by combining SDLSTM and ARIMA with non-equal interval; this means using intervals with a different unit of time for each method. Then, combining the result of the singular module achieves an accurate prediction of traffic flow data. The experimental results demonstrate that the method based on the SDLSTM - ARIMA model has higher accuracy than the similar method using only ARIMA [18].

6.2. STL and CNN

The main goal of this hybrid technique [20] is merging the data from two sources to complement each other, especially in the case of combine knowledge and data-driven systems in a single framework. Complementing Deep Neural networks (DNN) with expert knowledge can reduce the dependency on the data [20]. In this technique, first, two sources of data are fed separately to the STL decomposition module, which decomposes the input signal into its constituent parts. An STL stands for Seasonal and Trend decomposition using Loess that is a statistical method of decomposing a Time Series data into three components containing (i) seasonality, (ii) trend, and (iii) residual. The second component, the trend, is separated from the rest signals (seasonality and residual). Hence, the output of STL decomposition contains two signals, trend and the rest of the signal. These signals are then given to their respective CNN (Convolutional neural network) estimators as inputs. In this section, the convolutional neural network (CNN) was chosen as a Deep Learning model because it is generally easier to optimize [20]. Finally, the overall output of the model is the sum of the output of the two CNNs, which is shown in Fig. 2.

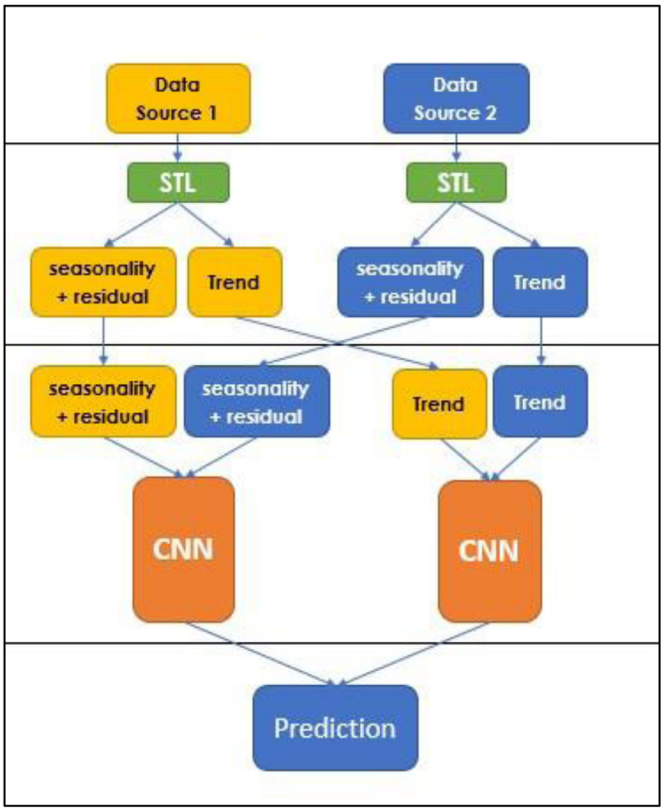

Fig. 2. Overview of DeepEX architecture.

7. Accuracy and validations measures

Smart transportation modeling systems are relying on information retrieved from machine learning methods to reach important decisions. The most known measures for evaluation for supervised and unsupervised prediction methods are MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), Precision, Recall, and F-measure. It is always best to present the results of multiple goodness of fit measures when evaluating models as opposed to those using only single evaluation measures. The accuracy of the time-series model can be tested with root mean square error of prediction (RMSE), mean absolute percentage er-

ror (MAE), and standard errors (SE).

$$\text{RMSE} = \frac{\sqrt{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}}{n} \tag{11}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{\hat{y}_i}\right| \tag{12}$$

$$\text{SE} = \frac{\sigma}{\sqrt{n}},\ \sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - \mu_y\right)^2}{n-1}} \tag{13}$$

Where $\hat{y}_i$ is the predicted value and y_i is the actual value and μ_y is the mean value. Among these statistics, MAPE statistic is a well suited measurement tool for traffic data because the traffic flow measurement performs variabilities by order of magnitude between the daily peak hours and trough hours. MAPE statistic is useful and illustrative when nominal levels of the processed present [12]. The traffic flow data was observed at different nominal levels. For example, some of the detectors installed on highways observed an aggregate number of vehicles that have passed by in 5 min. The detectors observed the average speed of the vehicles. The time variable is the only independent variable the time series model takes into consideration. It does not capture factors such as weather conditions, road pavement conditions, lighting along the expressways, pedestrians, etc. A regression model includes more independent variables that can be implemented the capture the effect mentioned above.

8. Business opportunities: traffic forecast applications

Overall, there are abundant business opportunities with the development of advanced transportation data analytics Companies capture opportunities to forecast real-time traffic situations, which could be provided directly or be utilized as marketing insights after being sorted and analyzed with the techniques discussed earlier. One example of came up with is to deliver the traffic forecast results to individual users. Just like the weather forecast, companies can develop an app called Traffic Forecast. Google map, for instance, currently shows the real-time traffic flows only. If a consumer wants to know the time required to get home from work in the next few hours so that the consumer could plan when to leave, Google map will not have the ability to do so. Google can then implement such a model to show accurate predicted time to travel and the fastest route to take to satisfy consumer demand. Google can also add a function called Traffic Forecast, just like a weather forecast application, that shows the traffic conditions for the next 24 h in a city. The analytical traffic planning can also be built within current business programs: for example, Uber drivers can follow the instruction of which route is the fastest. Also, time series analysis would be beneficial to the transportation department of governments, as it provides a traffic forecast with high precisions. In this way, it is easier to plan when to direct traffic once some congestions or events would cause traffic stagnation. The analysis could either be written into the current traffic management systems or set up separately with no significant burdens. The implementation of this time series analysis would significantly reduce congestion since it provides forecasts on future traffic status and reduces the probabilities of encountering accidents at the same time.

9. Conclusion

It is now more accessible for the government and individuals to foresee the status of urban transportation status through advances in transportation planning development. Still, the analytical process and data extractions have been challenges with relatively higher precision demand. The combination of mobile and smart card data has connected each individual's activities more closely. Which provides the foundation of further data analytics; the use of clustering techniques enabled higher precision due to the nature of its formation, which automates the information and grouped spontaneously. It is suggested to improve mobile data accuracy, such as adding an extra amount of monitoring spots as it is one of the most crucial parts of the process. However, current technologies need a long time to become available to a high degree. Based on a comprehensive review of the previous studies, developing traffic prediction methods is still an open research area. Therefore this paper has provided a road map for future research in smart transportation planning, focusing on various data sources, forecasting models and their properties and configurations, and future business opportunities for transportation analytical models.

Declaration of Competing Interest

None.

References

[1] A. Huang, D. Levinson, Axis of Travel: Modeling Non-Work Destination Choice With GPS Data, Transportation Research Part C: Emerging Technologies, 2015.
[2] Lau, P.L.B. & Marakkalage, S. & Zhou, Y. & UL Hassan, N. & Yuen, C. & Zhang, M. & Tan, U.-.X.. (2019). A survey of data fusion in smart city applications.
[3] Schmöcker, J.-.D. & Kurauchi, F.. (2017). Public transport planning with smart card data.
[4] L. Alexander, S. Jiang, M. Murga, M. C.González, Origin–destination trips by purpose and time of day inferred from mobile phone data, Transp. Res. Part C 58 (2015) 240–250.
[5] Vanhoof, M. & Reis, F. & Smoreda, Z. & Ploetz, T.. (2018). Detecting home locations from CDR data: introducing spatial uncertainty to the state-of-the-art.
[6] Qi Shi, M. Abdel-Aty, Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways, Transp. Res. Part C (2015) 58, doi:10.1016/j.trc.2015.02.022.
[7] J.L. Bentley, J.H. Friedman, Data structures for range searching, ACM Comput Surv 11 (4) (1979) 397–409.
[8] K. Button, D.A. Hensher, Handbook of Transport Systems and Traffic Control, 2nd, Pergamon, Amsterdam, 2007 Chapter 3.
[9] A. Carrel, P.S. Lau, R.G. Mishalani, R. Sengupta, J.L. Walker, Quantifying transit travel experiences from the users' perspective with high-resolution smartphone and vehicle location data: methodologies, validation, and example analyses, Transp. Res. Part C 58 (2015) 224–239.
[10] Duan, Y. (2014). Traffic Flow Prediction with Big Data: A Deep Learning Approach. Retrieved September 09, 2014, from http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6894591
[11] 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. Part C 58 (2015) 278–291, doi:10.1016/j.trc.2015.06.007.
[12] Ghosh, B. (2004). Time-series modeling for forecasting vehicular traffic flow in Dublin. Retrieved November/December 2004.
[13] A. Emami, M. Sarvi, S.A. Bagloee (2019), Using Kalman filter algorithm for short-term traffic flow prediction in a connected vehicle environment.
[14] P.S. Kalekar (2004), Time series forecasting using holt-winters exponential smoothing.
[15] N. Lu, N. Cheng, N. Zhang, X. Shen, and JW. Mark, (2014), Connected vehicles: solutions and challenges.
[16] A. Talebpour, HS. Mahmassani (2016), Influence of connected and autonomous vehicles on traffic flow stability and throughput.
[17] M. Treiber, A. Kesting, D. Helbing, Influence of reaction times and anticipation on stability of vehicular traffic flow, Transport. Res. Rec. 1999 (1) (2007) 23–29.
[18] B. Liu, X. Tang, J. Cheng, P. Shi (2018). Traffic flow combination forecasting method based on improved LSTM and ARIMA.
[19] N. Srivastava, G. Hinton, A. Krizhevsky, et al., 'Dropout: a simple way to prevent neural networks from overfitting, J. Machine Learn. Res. 15 (1) (2014) 1929–1958.
[20] M.A. Chattha, S.A. Siddiqui, M. Munir, I. Malik, L. van Elst, A. Dengel, and S. Ahmed (2019), DeepEX: bridging the gap between knowledge and data driven techniques for time series forecasting.
[21] F. Martínez, MP. Frías, F. Charte, AJ. Rivera, Time series forecasting with KNN in R: the tsfknn package, R J. (2019).
[22] G. Zhang, F. Li, Application of the KNN algorithm based on KD tree in intelligent transportation system, in: 2014 IEEE 5th International Conference on Software Engineering and Service Science, Beijing, 2014, pp. 832–835, doi:10.1109/ICSESS.2014.6933695.
[23] Z. Wang, S. Ji, B. Yu, "Short-term traffic volume forecasting with asymmetric loss based on enhanced KNN method, Math. Prob. Eng. 2019 (2019) 4589437–4589448.
[24] S. Tak, et al., Real-time travel time prediction using multi-level k-nearest neighbor algorithm and data fusion method, Comput. Civil Build. Eng. 2014 (2014) 1861–1868.
[25] S. Oh, Y-Ji Byon, H. Yeo, "Improvement of search strategy with k-nearest neighbors approach for traffic state prediction, IEEE Trans. Intell. Transp. Syst. 17.4 (2015) 1146–1156.
[26] B. Alsolami, R. Mehmood, A. Albeshri, "Hybrid statistical and machine learning methods for road traffic prediction: a review and tutorial, in: Smart Infrastructure and Applications, Springer, Cham, 2020, pp. 115–133.

[27] D. Billings, J.-S. Yang, "Application of the ARIMA models to urban roadway travel time prediction-a case study, 2006 IEEE International Conference on Systems, Man and Cybernetics. Vol. 3, 2006.
[28] S.V. Kumar, Traffic flow prediction using Kalman filtering technique, Procedia Eng. 187 (2017) 582–587.
[29] MA. Aljamal, HM. Abdelghaffar, HA. Rakha, "Developing a Neural–Kalman Filtering Approach for Estimating Traffic Stream Density Using Probe Vehicle Data, Sensors 19.19 (2019) 4325.
[30] Q. Lai, et al., A hybrid short-term forecasting model of passenger flow on high-speed rail considering the impact of train service frequency, Math. Problems Eng. 2017 (2017) 1828102–1828111.
[31] R. Jašek, A. Szmit, M. Szmit, "Usage of modern exponential-smoothing models in network traffic modelling, in: Nostradamus 2013: Prediction, Modeling and Analysis of Complex Systems, Springer, Heidelberg, 2013, pp. 435–444.
[32] A. Pompigna, F. Rupi, "Comparing practice-ready forecast models for weekly and monthly fluctuations of average daily traffic and enhancing accuracy by weighting methods, J. Traffic Transp. Eng. 5.4 (2018) 239–253.
[33] Yu Cheng, T. Zhang, J. Wang, "Multi-agent system model for urban traffic simulation and optimizing based on random walk, in: Advances in Neural Network Research and Applications, Springer, Berlin, Heidelberg, 2010, pp. 703–711.
[34] K.-.C. Chu, Modeling and Probing Strategy for Intelligent Transportation System Utilizing Lagrangian Traffic Data, Diss (2016).
[35] H. Babaei, M. Fathy, M. Romoozi, "Modeling and optimizing random walk content discovery protocol over mobile ad-hoc networks, Perform. Evaluat. 74 (2014) 18–29.