Research article

Time delay estimation of traffic congestion propagation due to accidents based on statistical causality

YongKyung Oh[1]**, JiIn Kwak**[2] **and Sungil Kim**[1,2,*]

[1] Department of Industrial Engineering, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea

[2] Artificial Intelligence Graduate School, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea

* **Correspondence:** Email: sungil.kim@unist.ac.kr; Tel: +82522173195.

Abstract: The accurate estimation of time delays is crucial in traffic congestion analysis, as this information can be used to address fundamental questions regarding the origin and propagation of traffic congestion. However, the exact measurement of time delays during congestion remains a challenge owing to the complex propagation process between roads and high uncertainty regarding future behavior. To overcome this challenge, we propose a novel time delay estimation method for the propagation of traffic congestion due to accidents using lag-specific transfer entropy (TE). The proposed method adopts Markov bootstrap techniques to quantify uncertainty in the time delay estimator. To the best of our knowledge, our proposed method is the first to estimate time delays based on causal relationships between adjacent roads. We validated the method's efficacy using simulated data, as well as real user trajectory data obtained from a major GPS navigation system in South Korea.

Keywords: statistical causality; transfer entropy; time delay estimation; traffic trajectory data; traffic incident analysis

1. Introduction

Traffic congestion represents a universal problem for urban life owing to the dramatic growth in vehicle use, expansion of the economy and infrastructure, and proliferation of delivery services, among other factors. Traffic congestion frequently spreads into adjacent roads [1], resulting in greater damage to the overall traffic network.

Consequently, the accurate estimation of time delays has become crucial in addressing fundamental questions regarding the origin points and propagation of traffic congestion. Figure 1 from [2] illustrates this study's objective by showing how the impact of a traffic accident propagates to incoming roads.

The black solid line represents the average speed on the road segment where the accident occurs, while the orange, green, and blue solid lines represent the average speeds on the three adjacent incoming roads. The average speed on the road segment was computed from GPS trajectory data provided by the NAVER Corporation using a map-matching process [3]. We can observe a time lag when the impact of an accident propagates to incoming roads. The time delay increases in the order of blue, green, and orange.
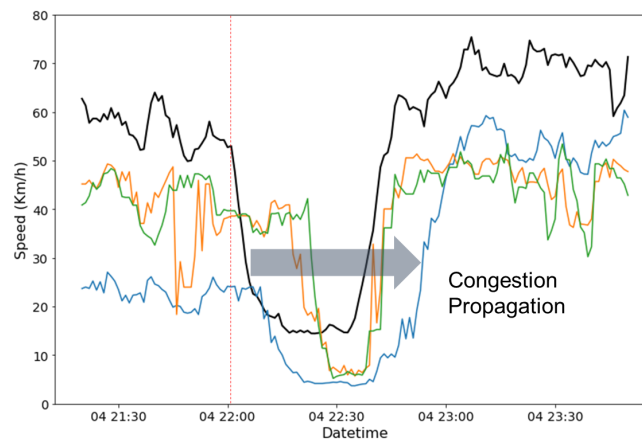

Figure 1. Motivating example: traffic congestion propagation [2].

However, certain aspects of traffic congestion propagation pose statistical challenges to the accuracy of time delay estimation. First, the average speed on an incoming road is affected by various geographic and topological characteristics, such as road length and width, as well as road network topology. The impact of a traffic accident is distributed among all incoming roads in a complex pattern according to these characteristics. Furthermore, the duration of congestion is dynamic. As shown in the figure, the road denoted in blue exhibits a longer congestion duration than the other incoming roads. That is, the time delay in traffic congestion propagation does not simply denote a temporal pattern shift, but involves complicated temporal dynamics. Finally, data uncertainty is inherent in average road speeds. Trajectory-based average road speeds may be highly volatile depending on the availability of user data over a specific period.

To overcome the challenges outlined above, we propose a novel time delay estimation method for traffic congestion propagation between roads using lag-specific transfer entropy (TE). Our main contributions are as follows:

- We provide a model-free approach to estimate congestion propagation delays using a lag-specific TE estimator in complex urban road systems.
- We quantify uncertainty in time delay estimation using bootstrap techniques. This uncertainty quantification is employed to evaluate the reliability of time delay estimates and serves as a basis for hyperparameter optimization.
- We show that decomposition and nonlinear normalization with a sliding window are effective time series preprocessing methods for revealing causal relationships between traffic speed data.
- We validate the proposed method through numerical simulations and real user trajectory data obtained from a major GPS navigation system in South Korea.

The remainder of this paper is organized as follows: Section 2 provides an overview of related studies. Section 3 presents crucial background information pertaining to the proposed method. Section 4 outlines our proposed time delay estimation method. Sections 5 and 6 validate the proposed method using simulated and real congestion propagation data, respectively. Finally, concluding remarks are presented in Section 7.

2. Related works

Various topics have been studied regarding the estimation of time delays due to traffic accidents, including travel time delay [4], incident duration [5, 6], real-time crash identification [7], and incident impact quantification [2, 8]. However, unlike the aforementioned studies, we focused on the propagation delay of traffic congestion caused by accidents in a road network.

2.1. Time delay estimation for congestion propagation

Although our proposed method is the first to estimate time delays for congestion propagation in road traffic networks, time delay estimation (TDE) is not a new problem. In digital signal processing, TDE refers to the task of ascertaining the differences in arrival times between signals received at the sensor array. The most widespread approach for TDE is cross-correlation [9, 10]. Supposing that signals are received from two sensors, the delay between the sensors can be estimated using the time lag that maximizes the cross-correlation between filtered versions of the received signal.

Limited attempts have been made to perform cross-correlation analyses with traffic speed data. Conventional TDE methods based on cross-correlation have been applied to a real road vehicle pass-by measurement to enable traffic monitoring using passive acoustic sensors [11]. Similarly, [12] conducted a cross-correlation analysis to prove the existence of a significant relationship between the current value of speed at a specific station, as well as past speed values at upstream and downstream stations in a freeway traffic network. We refer to this method as time-lagged cross-correlation (TLCC). This approach is not applicable in the presence of non-stationarity.

To quantify the TLCC level between two nonstationary time series at different scales, [13] proposed a time-lagged detrended cross-correlation analysis approach. This method, referred to as DCCA, divides an entire time series into overlapping boxes to handle nonstationarity [14].

The method proposed in the present study was compared with TLCC and DCCA for evaluation purposes.

2.2. Traffic causality analysis

Various traffic causal analysis methods have been developed to identify causal relationships among congested roads and detect congestion propagation patterns. The authors of [15] forecasted future traffic flow by ranking input variables to identify a subset of the Bayesian network as the set of cause nodes using the Pearson correlation coefficient. A two-step mining architecture has been proposed to capture the origin points of road traffic congestion [16]. TE was used to reveal the delay propagation network among multiple airports with a time series of airport delays [17]. However, the aforementioned approaches primarily focus on revealing causal relationships and do not provide information regarding estimated time delays during congestion propagation.

3. Preliminary

3.1. Bootstrap for Markov chains

Suppose that $\{X_t\}_{t\geq 1}$ is a stationary Markov chain with a finite state space $S = \{s_1, \dots, s_n\}$, where $n \in \mathbb{N}$. Let $\mathbf{P} = (p_{ij}) \in \mathbb{R}^{n \times n}$ be the transition probability matrix of the chain and the stationary distribution by $\pi = (\pi_1, \dots, \pi_n)$. Thus, for any $1 \leq i, j \leq n$, $p_{ij} = P(X_{t+1} = s_j | X_t = s_i)$ and $\pi_i = P(X_t = s_i)$. Given a time series $\{X_1, \dots, X_L\}$ of size L from a stationary Markov chain, π_i and p_{ij} can be estimated as

$$\hat{\pi}_i = \frac{1}{L} \sum_{t=1}^{L} \mathbb{1}(X_t = s_i), \quad \hat{p}_{ij} = \frac{1}{\hat{\pi}_i L} \sum_{t=1}^{L} \mathbb{1}(X_t = s_i, X_{t+1} = s_j). \tag{3.1}$$

The bootstrap observations $\{X_1^*, \dots, X_L^*\}$ can now be generated using the estimated transition matrix and marginal distribution in Eq (3.1) [18].

1) Generate a random variable X_1^* from the discrete distribution on $\{1, \dots, n\}$ that assigns mass $\hat{\pi}_i$ to $s_i, 1 \leq i \leq n$.

2) Generate a random variable X_{t+1}^* from the discrete distribution on $\{1, \dots, n\}$ that assigns mass $\hat{p}_{ij}$ to $j, 1 \leq j \leq n$, where s_i is the value of X_t^*.

3) Repeat Step (2) until a simulated time series $\{X_1^*, \dots, X_L^*\}$ has been obtained.

3.2. Lag-specific transfer entropy

TE is a measurement of directed information flow [19] based on the concept of Shannon entropy [20] in the field of information theory. For a discrete random variable I with probability distribution $p(i)$, the Shannon entropy represents the average number of bits required to optimally encode independent draws, calculated as follows:

$$H(I) = -\sum_i p(i) \log_2 p(i). \tag{3.2}$$

Equation (3.2) can be easily extended to the concept of conditional entropy using two discrete random variables I and J:

$$H(I|J) = -\sum \sum p(i, j) \log_2 p(i|j) \tag{3.3}$$

This equation can be used to measure information flow between two discrete random variables.

Consider two discrete random variables, I and J, with marginal probability distributions $p(i)$ and $p(j)$, and joint probability $p(i, j)$. Suppose both variables represent stationary Markov processes of orders k and l, respectively. For the order k Markov process I, Eq (3.2) can be extended to

$$H^{(k)}(I) = -\sum_i p(i_t, i_{t-1}^{(k)}) \log p(i_t | i_{t-1}^{(k)}),$$

where $i_{t-1}^{(k)} = (i_{t-1}, \dots, i_{t-k})$. Analogously, the information flow from J to I is measured by quantifying the deviation from the generalized Markov property $p(i_t | i_{t-1}^{(k)}) = p(i_t | i_{t-1}^{(k)}, j_{t-u}^{(l)})$ for an arbitrary source–target lag u, as follows:

$$T_{J \to I}^{(k,l)}(t, u) = \sum p(i_t, i_{t-1}^{(k)}, j_{t-u}^{(l)}) \log \frac{p(i_t | i_{t-1}^{(k)}, j_{t-u}^{(l)})}{p(i_t | i_{t-1}^{(k)})}. \tag{3.4}$$

Equation (3.4) preserves the computational interpretation of TE as an information transfer, which is the only relevant option in keeping with Wiener's principle of causality [21]. The transfer entropy is known to be biased in small samples [22]. To correct any bias, [22] proposed the effective transfer entropy (ETE):

$$ETE_{J \to I}^{(k,l)}(t, u) = T_{J \to I}^{(k,l)}(t, u) - T_{J_{\text{shuffled}} \to I}^{(k,l)}(t, u). \tag{3.5}$$

where $T_{J_{\text{shuffled}} \to I}^{(k,l)}$ indicates the transfer entropy using a shuffled version of time series J. The shuffling process randomly draws values from the original time series and realigns them to generate a new time series. Thus, shuffling eliminates any time series dependencies of J, as well as statistical dependencies between J and I. Note that $T_{J_{\text{shuffled}} \to I}^{(k,l)}$ converges to zero as the sample size increases and any nonzero value of $T_{J_{\text{shuffled}} \to I}^{(k,l)}(t, u)$ is a result of the small sample effect. To ensure estimation consistency, shuffling is repeated, and the average of the shuffled transfer entropy estimates across all iterations serves as an estimator for the small sample bias, which is subsequently subtracted from the Shannon or Rényi transfer entropy estimate to correct any bias.

4. Methodology

4.1. Problem definition

Suppose that congestion information is transferred from a source road to a destination road with a time delay of u. The objective of time delay estimation is to estimate u given previously observed traffic speed data on the two roads denoted as $\{X_t\}_{t=1}^L$ and $\{Y_t\}_{t=1}^L$, respectively. Therefore, the time delay estimation task entails formulating a function $f(\cdot)$ that computes the source–target lag u,
$\left[\{X_t\}_{t=1}^L, \{Y_t\}_{t=1}^L\right] \xrightarrow{f(\cdot)} u.$

The proposed time delay estimation algorithm comprises three steps. First, bootstrapping for each time series is performed using preprocessing methods. Next, the transfer entropy computation provides the estimated time delay lag u^*. Finally, the distribution of time delay lag is estimated to determine the existence of a statistical causal relationship.

4.2. Preprocessing and bootstrapping

Consider a time series of congested traffic speed data, $\{X_t\}_{t=1}^L$, which has the properties of scale dependence, nonlinearity, and nonstationarity. To effectively identify the causal relationship within such a time series, appropriate preprocessing methods are essential.

To this end, we first decompose the time series into a trend and its residual, as follows:

$$\forall t, X_t = \mathcal{T}_t + R_t = \frac{1}{m} \sum_{j=0}^{m-1} X_{t-j} + R_t, \tag{4.1}$$

where $\mathcal{T}_t$ and R_t are the trend and residual components, respectively. The trend component is a moving average of order m, representing the mean forefront value at time t. The purpose of the trend component is to smooth the time series for estimating the underlying trend. After extracting the underlying trend from $\{X_t\}_{t=1}^L$, the residual time series $\{R_t\}_{t=1}^L$ is assumed to be a stationary Markov process. The assumption of Markovian property in traffic speed is not a novel notion. Many prior traffic speed prediction and modeling studies have been conducted under this assumption [23–27]. Based on $\{R_t\}_{t=1}^L$,

we can generate the bootstrap residuals $\{R_t^{*(b)}\}_{t=1}^L$ as explained in Section 3.1. Subsequently, we can easily obtain a bootstrap time series $\{X_t^{*(b)}\}_{t=1}^L$ by $\mathcal{T}_t + R_t^{*(b)}$ for $t = 1, \ldots, L$.

Nonlinear normalization with a sliding window is then applied to the obtained time series to address the scale-dependency, nonlinearity, and nonstationarity of traffic speed data. To ensure the data are scale-independent and close to linear [29], the standard normal cumulative distribution function Φ is applied. A sliding window technique has similarly been employed to handle a nonstationary time series in [28]. Let $\mathbf{X}_{t,w} = \left\{X_k^{*(b)}\right\}_{k=t-w+1}^t$ be the forefront sequence of $X_t^{*(b)}$ with a sliding window size of w, and $F_{25,t}$, $F_{50,t}$ and $F_{75,t}$ be the 25, 50 and 75th percentiles of $\mathbf{X}_{t,w}$, respectively. Note that these percentiles depend on the location of the sliding window. Then, a normalized time series $\left\{\tilde{X}_t^{*(b)}\right\}_{t=1}^L$ can be obtained by

$$\tilde{X}_t^{*(b)} = \Phi\left(0.5 \times \frac{X_t^{*(b)} - F_{50,t}}{F_{75,t} - F_{25,t}}\right). \tag{4.2}$$

To verify the effectiveness of the nonlinear normalization method expressed in Eq (4.2), we compared its performance with that of existing normalization methods [28, 29], including the min-max method $\left(\tilde{X}_t^{*(b)} = \frac{X_t^{*(b)}}{\max \mathbf{X}_{t,w}}\right)$ and the z-score method $\left(\tilde{X}_t^{*(b)} = \frac{X_t^{*(b)} - \mu(\mathbf{X}_{t,w})}{\sigma(\mathbf{X}_{t,w})}\right)$ with and without a sliding window technique (see Section 5).

4.3. Time delay estimation

Using Eq (3.5), the time lag in a causal relationship $J \to I$ can be estimated by solving the following optimization problem:

$$\hat{u} = \underset{u \in \mathbb{N}}{\operatorname{argmax}}\ ETE_{J \to I}^{(k,l)}(t, u). \tag{4.3}$$

In this study, we assume $k = \ell = 1$.

To compute the lag-specific ETE in Eq (4.3), we discretize continuous data using symbolic encoding. This discretization can be performed by partitioning the data into a finite number of bins. We denote the bounds specified for the n bins by $q_1, q_2, \ldots, q_{n-1}$, where $q_1 < q_2 < \cdots < q_{n-1}$. For the normalized time series in Eq (4.2), we obtain the encoded time series $\{J_t^{*(b)}\}_{t=1}^L$ by the following equation:

$$J_t^{*(b)} = \begin{cases} 1 & \text{for } \tilde{X}_t^{*(b)} \le q_1 \\ 2 & \text{for } q_1 < \tilde{X}_t^{*(b)} < q_2 \\ \vdots & \vdots \\ n & \text{for } \tilde{X}_t^{*(b)} \ge q_{n-1}. \end{cases} \tag{4.4}$$

The choice of bins depends on the distribution of data. In the case where tail observations are of particular interest, binning is typically based on empirical quantiles, such that the left and right tail observations are allocated into separate bins. In this study, we implemented symbolic encoding with $n = 3$ based on 5% and 95% empirical quantiles, thereby emphasizing speed extremes caused by dynamic speed changes and traffic accidents.

Consequently, we obtain $\{J_t^{*(b)}\}_{t=1}^L$ from $\{\tilde{X}_t^{*(b)}\}_{t=1}^L$, and $\{I_t^{*(b)}\}_{t=1}^L$ from $\{\tilde{Y}_t^{*(b)}\}_{t=1}^L$, respectively, for $b = 1, \ldots, B$. Given $\{J_t^{*(b)}\}_{t=1}^L$ and $\{I_t^{*(b)}\}_{t=1}^L$, $b = 1, \ldots, B$, we obtain bootstrap observations of the time lag, $u^{*(1)}, \ldots, u^{*(B)}$ using Eq. (4.3).

4.4. Uncertainty quantification of time delay estimates

Suppose those bootstrap observations of the time lag follow a distribution $\mathcal{G}$,

$$u^{*(1)}, \ldots, u^{*(B)} \sim \mathcal{G},$$

which is unknown in practice. Let μ and σ^2 denote the mean and variance of $\mathcal{G}$, respectively, which can be estimated by

$$\hat{\mu}_B = \frac{1}{B} \sum_{b=1}^{B} u^{*(b)}, \quad \hat{\sigma}_B^2 = \frac{1}{B} \sum_{b=1}^{B} \left(u^{*(b)}\right)^2 - \hat{\mu}_B^2.$$

Proposition 1 implies that 1) the bootstrap estimate $\hat{\mu}_B$ is an unbiased estimate of μ, and 2) $\frac{1}{B}\hat{\sigma}_B^2$ quantifies the uncertainty of $\hat{\mu}_B$. That is, we can evaluate the uncertainty of the bootstrap estimate $\hat{\mu}_B$ using $\frac{1}{B}\hat{\sigma}_B^2$, which is practically useful because μ is unknown. This approach can be applied to hyperparameter tuning. In this study, we used a grid search to determine a set of hyperparameters (length of time series (L) and sliding window size (w)) that minimizes $\frac{1}{B}\hat{\sigma}_B^2$.

Proposition 1. *Let $u^{*(1)}, \ldots, u^{*(B)}$ be a bootstrap sample, and $E(u^{*(b)}) = \mu$, $Var(u^{*(b)}) = \sigma^2$. Then, the sample mean $\hat{\mu}_B = \frac{1}{B} \sum_{b=1}^{B} u^{*(b)}$ approximately follows $\mathcal{N}\left(\mu, \frac{1}{B}\hat{\sigma}_B^2\right)$, where $\hat{\sigma}_B^2$ is the sample variance of the bootstrap sample.*

Proof. As $\hat{\sigma}_B^2 \to \sigma^2$ in probability, $\frac{\sqrt{B}(\hat{\mu}_B - \mu)}{\hat{\sigma}_B} = \frac{\sigma}{\hat{\sigma}_B} \frac{\sqrt{B}(\hat{\mu}_B - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$ by the central limit theorem and Slutsky's theorem [30]. Thus, $\hat{\mu}_B$ approximately follows a normal distribution, $\hat{\mu}_B \sim \mathcal{N}\left(\mu, \frac{1}{B}\hat{\sigma}_B^2\right)$. $\qquad\square$

To determine whether the bootstrap estimate $\hat{\mu}_B$ is reliable, we compare $\hat{\sigma}_B^2$ with a predetermined threshold σ_T^2. That is, if $\hat{\sigma}_B^2 > \sigma_T^2$, we can conclude that $\hat{\mu}_B$ is not reliable, and congestion is therefore not propagated on the corresponding road segment.

To determine an appropriate value of σ_T^2, we employ the concept of the tolerance interval (TI). The TI is a statistical interval in which a specified proportion γ of a population will fall with a certain level of confidence $(1 - \alpha)$. By definition, a $(\gamma, 1 - \alpha)$-TI of $\hat{\mu}_B$,

$$TI = \left[\mu - k_{\gamma,\alpha} \sqrt{\frac{1}{B}\hat{\sigma}_B^2}, \mu + k_{\gamma,\alpha} \sqrt{\frac{1}{B}\hat{\sigma}_B^2}\right],$$

satisfies

$$P\left[P(\hat{\mu}_B \in TI) \geq \gamma\right] = 1 - \alpha,$$

where $k_{\gamma,\alpha}$ is the tolerance factor [31]. Then, σ_T^2 can be determined by $k_{\gamma,\alpha} \sqrt{\frac{1}{B}\sigma_T^2} = 1$ (min) to yield a ± 1 minute TI. With $B = 100$, $\gamma = 0.9$, and $\alpha = 0.01$, the present study uses $\sigma_T^2 = \frac{100}{k_{0.9,0.01}^2} = 5.05^2$.

Consider the propagation of traffic congestion caused by accidents. Here, we define the propagation path by a sequence of incoming roads in the direction opposite to the traffic flow, where kth element in the propagation path is denoted as Hop(k-1). Let $\hat{\mu}_{B,k-1}$ and $\hat{\sigma}_{B,k-1}^2$ denote the bootstrap estimate and sample variance at Hop(k-1), respectively. Hop0 corresponds to the road where the accident occurred. We consider Hop(k-1) to be statistically *significant* if $\hat{\mu}_{B,k-1}$ and $\hat{\sigma}_{B,k-1}^2$ satisfy the following conditions: 1) $\hat{\sigma}_{B,k-1}^2 < \sigma_T^2$ and 2) $\hat{\mu}_{B,k-2} < \hat{\mu}_{B,k-1}$. The second condition states that the time delay between Hop0 and Hop(k-1) must exceed that between Hop0 and Hop(k-2).

5. Simulation studies

The proposed method was validated using simulated data. Two-time series – $\{X_t\}_{t=1}^{120}$ and $\{Y_t\}_{t=1}^{120}$ – were generated by

$$X_t = \begin{cases} 100 + \epsilon_{x,t} & \text{for } t < 10 \\ 0.95X_{t-1} + \epsilon_{x,t} & \text{for } 10 \leq t < 95 \text{ ,} \\ 1.10X_{t-1} + \epsilon_{x,t} & \text{for } t \leq 120 \end{cases}$$

$$Y_t = \begin{cases} 70 + \epsilon_{y,t} & \text{for } t < 10 \\ 0.5X_{t-u_0} + 20 + \epsilon_{y,t} & \text{for } t \geq 10 \end{cases} \text{ ,}$$

where $\epsilon_{x,t} \sim \mathcal{N}(0, 2)$ and $\epsilon_{y,t} \sim \mathcal{N}(0, 2)$. A predetermined source–target lag (u_0) exists such that a significant information flow from X to Y is formed, but not vice versa. Figure 2(a) depicts two time series with $u_0 = 10$, where the black solid and red dashed lines represent $\{X_t\}_{t=1}^{120}$ and $\{Y_t\}_{t=1}^{120}$, respectively. This simulation represents a typical congestion propagation scenario between two adjacent roads R_X and R_Y, assuming that there was a traffic accident on R_X at $t = 10$ and that congestion was resolved at $t = 95$. With the time shift $u_0 = 10$, the congestion on R_X propagates to R_Y.

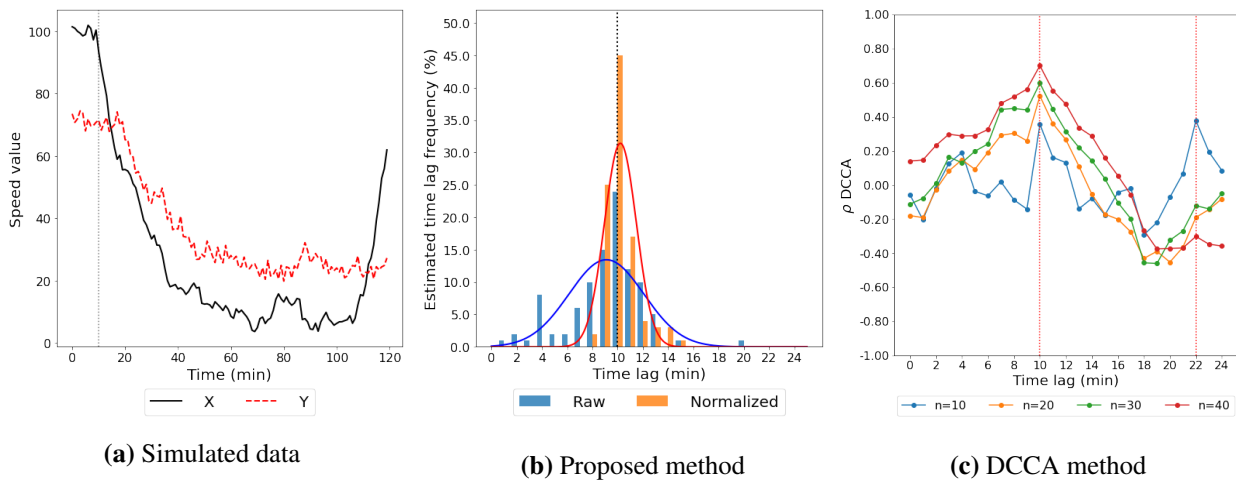**(a)** Simulated data   **(b)** Proposed method   **(c)** DCCA method

Figure 2. Results of simulation study.

The proposed method was applied to simulated data with $m = 2$ and $w = 20$, as described in Section 4. Figure 2(b) compares the distributions of bootstrap observations obtained from two-time series without normalization, to those obtained from two time series with nonlinear normalization. The red and blue lines in the figure denote the values of $\left(\hat{\mu}_B, \hat{\sigma}_B^2\right)$ in a distribution form, that is, $\left(9.54, 3.94^2\right)$ and $\left(10.30, 1.35^2\right)$, respectively. Here, a normal distribution is used for visualization purposes. We confirmed that nonlinear normalization with a sliding window improved the accuracy of the time delay estimation, as $1.35^2 < 3.94^2$.

Table 1. Simulation results with comparison methods.

	TLCC	DCCA(10)	DCCA(20)	DCCA(30)	DCCA(40)
$\hat{u}$	11.00	22.00	10.00	10.00	10.00

For comparison purposes, the TLCC and DCCA methods were also applied to the simulated data. The DCCA method requires a hyperparameter n, which indicates the size of the overlapping box [14]. Here, we used multiple values of n (10, 20, 30, 40) to obtain the results of time delay estimation, as shown in Figure 2(c) for the DCCA method. Furthermore, Table 1 summarizes the results of conventional TDE methods. These results show that both the TLCC and DCCA methods generally yield reasonable time delay estimates for the simulated data. In particular, it is recommended to set n to be greater than 20.

Table 2. Simulation results comparison ($u_0 = 10$) with $B = 100$.

| Decomposition | Normalization | Metrics | Window length | | | | | Average |
			10	20	30	40	120 (all)	
false	none	$\hat{\mu}_B$	–	–	–	–	11.23	11.23
		$\hat{\sigma}_B^2$	–	–	–	–	7.03	7.03
		MAE	–	–	–	–	6.13	6.13
	min-max	$\hat{\mu}_B$	12.93	12.64	13.14	12.71	14.89	13.26
		$\hat{\sigma}_B^2$	7.21	7.49	7.27	7.39	6.72	7.21
		MAE	6.73	6.94	6.92	6.82	7.20	6.92
	z-score	$\hat{\mu}_B$	13.06	13.51	12.97	13.29	14.84	13.53
		$\hat{\sigma}_B^2$	6.98	6.97	7.00	7.15	6.73	6.97
		MAE	6.49	6.67	6.52	6.75	7.23	6.73
	nonlinear	$\hat{\mu}_B$	13.29	12.86	12.28	13.17	14.27	13.17
		$\hat{\sigma}_B^2$	7.12	7.24	7.13	7.07	6.89	7.09
		MAE	6.77	6.66	6.36	6.63	7.03	6.69
true	none	$\hat{\mu}_B$	–	–	–	–	9.54	9.54
		$\hat{\sigma}_B^2$	–	–	–	–	3.94	3.94
		MAE	–	–	–	–	2.45	2.45
	min-max	$\hat{\mu}_B$	14.12	10.57	8.88	10.28	16.97	12.16
		$\hat{\sigma}_B^2$	4.24	2.81	2.51	3.53	4.94	3.61
		MAE	4.75	1.88	1.95	2.26	7.38	3.64
	z-score	$\hat{\mu}_B$	10.09	10.28	10.63	11.97	16.83	11.96
		$\hat{\sigma}_B^2$	2.87	3.58	3.90	4.19	5.90	4.09
		MAE	1.58	2.26	2.55	2.75	7.96	3.42
	nonlinear	$\hat{\mu}_B$	10.78	10.30	10.78	10.99	13.38	11.25
		$\hat{\sigma}_B^2$	2.98	1.35	1.49	2.04	6.72	2.91
		MAE	1.53	0.94	1.25	1.72	6.04	2.30

To investigate the proposed method's performance, we conducted simulation experiments under various settings of (1) decomposition, (2) normalization and (3) length of the sliding window. Performance was evaluated using $\hat{\mu}_B$, $\hat{\sigma}_B^2$ and MAE, where $MAE = \frac{1}{B} \sum_{i=1}^{B} |\hat{u}_i - u_0|$. Values closer to u_0 indicate a more accurate estimate $\hat{\mu}_B$. Likewise, smaller values of $\hat{\sigma}_B^2$ and MAE indicate better results. As summarized in Table 2, the decomposition technique improved the overall performance, and

nonlinear normalization with $w = 20$ generally performed better than all other normalization methods in terms of $\hat{\sigma}_B$ and MAE. This implies that nonlinear normalization with $w = 20$ produced the most precise and accurate estimates among the tested schemes.

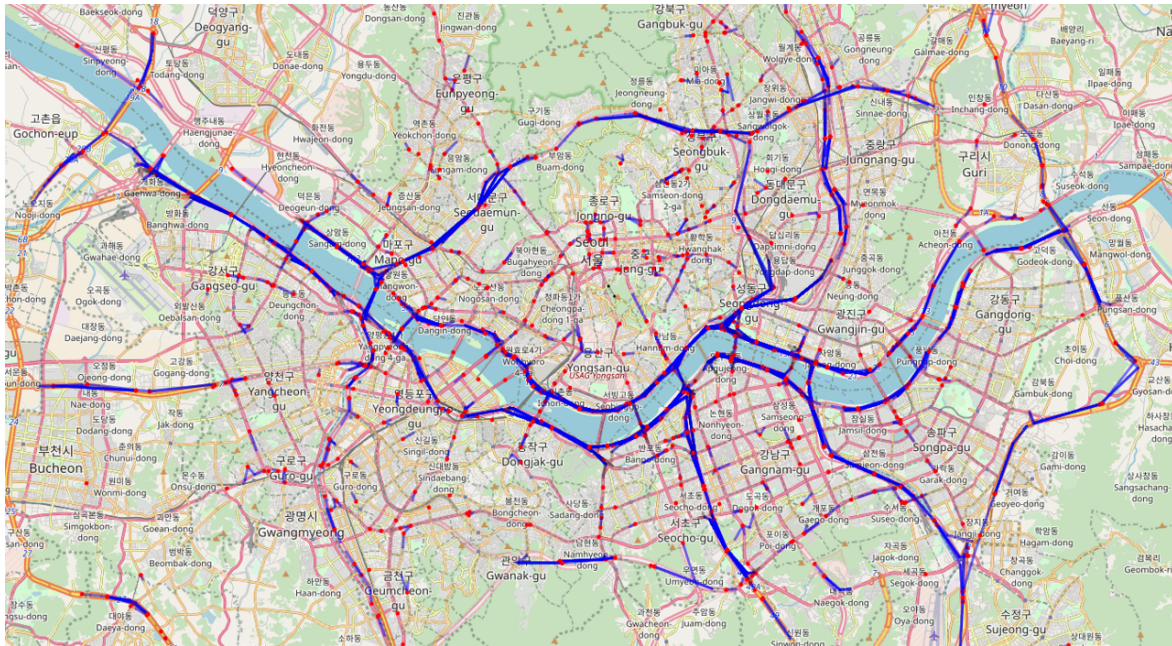## 6. Real data example: Accident-driven traffic congestion propagation

Figure 3. Locations of traffic accidents and their significant propagation paths in the city of Seoul from September 2020 to February 2021 (red dot: location of the accident, blue line: significant propagation path).

Two types of datasets were considered from various sources: A traffic dataset provided by the NAVER corporation navigation team, and an accident dataset provided by the Korean National Police Agency. The traffic dataset encompasses trajectory-based speed and traffic road networks of the major metropolitan area of Seoul, where nearly half of the country's population resides. The speed data are described by GPS trajectories. A GPS trajectory consists of a series of points with latitude, longitude, and timestamp information generated during travel. To align a sequence of observed user positions within the road network, we used a map-matching process [3]. Each accident record includes the reported time, information source, category, incident description, and point of origin described by both geographical coordinates and road segment ID, as described in Table 3.

The proposed method was validated on 3197 real traffic accidents that occurred between September 2020 and February 2021 in Seoul. Figure 3 presents the accident locations, as denoted by red stars, where the blue lines indicate significant propagation paths. Let $\hat{\mu}_{B,k-1}$ and $\hat{\sigma}^2_{B,k-1}$ denote the bootstrap estimate and sample variance at Hop(k-1), respectively, for $k = 2, 3, 4$. In this study, we investigated up to $k = 4$. $k = 1$ was excluded because Hop0 refers to the road where the accident occurred.

Table 4 summarizes the time delay estimation results for all 3197 traffic accidents. To ensure consistency within results, the hyperparameter $(w, L) = (60, 180)$ was used for all accidents based on the

Table 3. An example of a real accident record.

Event ID	3786580
Created datetime	2021-01-30 19:00
Information source	Korean National Police Agency
Category	Accident
Description	Traffic accident on the first lane from Guro IC on Nambu Beltway to Anyang Bridge
Longitude	126.87657
Latitude	37.48874

Table 4. Summary of time delay estimation results for 3197 traffic accidents.

	Number of roads	Significant roads	Significance ratio	Average time delay (min)
Hop1	5,036	3,483	69.16%	8.95
Hop2	6,856	4,479	65.33%	11.10
Hop3	9,721	6,139	63.15%	11.97

grid search. There are 5036 roads at Hop1 associated with accidents, approximately 69.16% of which were revealed to be significant, with an average time delay of 8.95 minutes. For Hop2 and Hop3, 65.33% and 63.15% of the roads were revealed to be significant with average time delays of 11.10 and 11.97 minutes, respectively.

We selected two representative cases among the traffic data to detail how the proposed method identifies causal relationships and estimates time lag. Case 1 represents a simple road network with few propagation paths, whereas Case 2 represents a complex road network with many propagation paths. For comparison purposes, the TLCC and DCCA methods were also applied with equivalent settings to those used in the simulation study.

6.1. Case 1: Simple traffic network

The accident occurred on September 8, 2020, at 06:44 AM. The blue star in Figure 4 denotes the exact location of the accident. Case 1 has one propagation path $[A, B, C, D]$. The black, red, blue, and green line segments in the figures indicate Hop0, Hop1, Hop2, and Hop3, respectively. Time delays were estimated using average speed data recorded at one-minute intervals from the previous hour to the subsequent two hours based on the time when the accident was reported.

Figure 5(a),(b) show the results of time delay estimation for the propagation path $[A, B, C, D]$. It is apparent from the significantly smaller values of $\hat{\sigma}_B^2$ that the time series with nonlinear normalization produced a more consistent estimate of time delays that increased with each hop. We, therefore, conclude that the congestion effect of the accident propagated along the path to Hop1, Hop2 and Hop3 at 3.60, 7.30, and 19.97 min after the accident, respectively.

Unlike the proposed method, both TLLC and DCCA failed to identify the congestion propagation effect of the accident. Furthermore, the DCCA method produced inconsistent time delay estimates for varying values of n. Note that both the TLLC and DCCA methods do not provide uncertainty quantification of the time delay estimates.
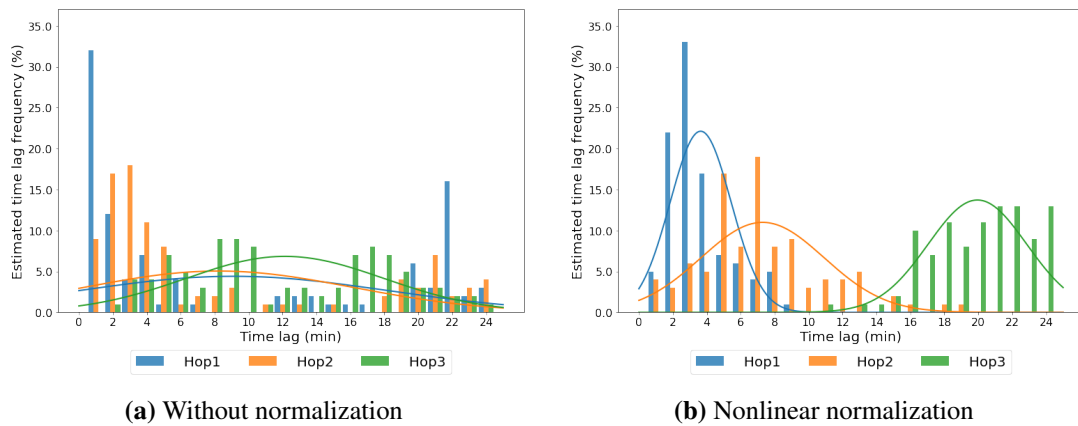
Figure 4. Traffic accident for Case 1.

(a) Without normalization

(b) Nonlinear normalization

Figure 5. Time delay estimation for Case 1.

Table 5. Results of time delay estimation for Case 1.

	Hop1		Hop2		Hop3	
	$\hat{\mu}_B$	$\hat{\sigma}_B^2$	$\hat{\mu}_B$	$\hat{\sigma}_B^2$	$\hat{\mu}_B$	$\hat{\sigma}_B^2$
TLCC	7.00	–	7.00	–	0.00	–
DCCA(10)	0.00	–	19.00	–	0.00	–
DCCA(20)	0.00	–	24.00	–	17.00	–
DCCA(30)	0.00	–	10.00	–	16.00	–
DCCA(40)	0.00	–	2.00	–	8.00	–
without normalization	9.62	83.70	8.07	61.31	12.08	34.83
nonlinear normalization	3.60	2.88	7.30	13.83	19.97	8.93

6.2. Case 2: Complex traffic network

The accident occurred on September 4, 2020, at 10:16 PM, and affected five propagation paths: $[A, B, C, D]$, $[A, E, F, G]$, $[A, H, I, J]$, $[A, H, K, M]$ and $[A, H, K, L]$. Figure 6 depicts the exact location of the accident, along with the five propagation paths. For each path, the previous hour and the subsequent two hours were considered for time delay estimation. Figure 7 and Table 6 present the results of time delay estimation. In Path 1, no specific causal relationship could be found, as shown in Figure 7(a). This finding is supported by the corresponding high values of $\hat{\sigma}_B^2$ ($> \sigma_T^2 = 5.05^2$) in Table 6. Similarly, we can conclude that the congestion effect of the accident on road A propagated along the second hop of Paths 2 and 3, and the third hop of Paths 4 and 5. Moreover, the values of $\hat{\sigma}_B^2$ in Table 6 indicate that the congestion effect propagated along Path 3 up to Hop2, as depicted in Figure 7(b), and along Paths 4 and 5 up to Hop3, as depicted in Figure 7(c). For Paths 4 and 5, the time delay estimates are (8.23, 15.65, 22.06) and (8.22, 15.55, 20.75), respectively.

Figure 6. Traffic accident for Case 2.

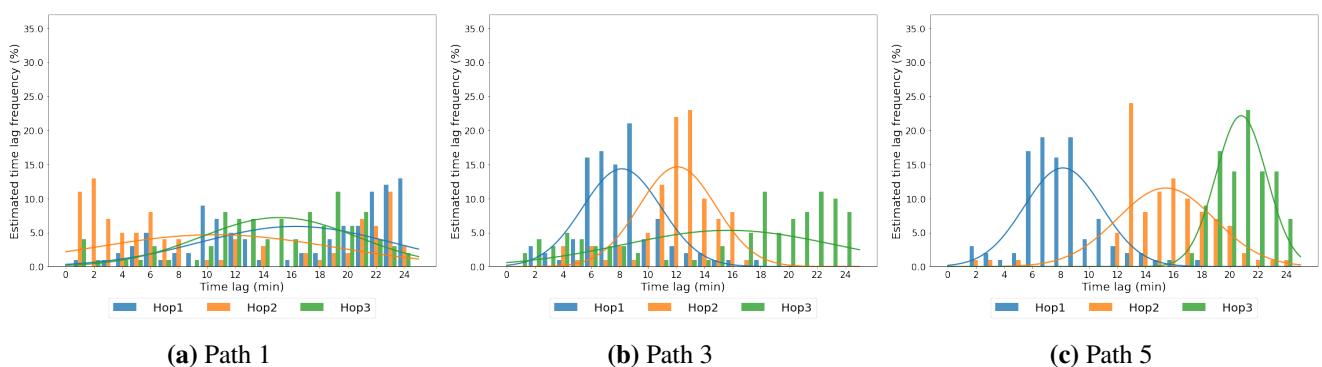**(a)** Path 1   **(b)** Path 3   **(c)** Path 5

Figure 7. Time delay estimation for Case 2.

As in Case 1, both the TLCC and DCCA methods failed to estimate consistent time delays. The DCCA method with 30 overlapping boxes ($n = 30$) obtained comparable results with the proposed method only for Path 3, as seen in Table 6. From the results of both cases, it can be concluded that

the proposed method identifies causal relationships and estimates the time lag more accurately than the conventional TDE methods.

Table 6. Results of time delay estimation for Case 2.

Propagation Path	Methods	Hop1		Hop2		Hop3	
		$\hat{\mu}_B$	$\hat{\sigma}_B^2$	$\hat{\mu}_B$	$\hat{\sigma}_B^2$	$\hat{\mu}_B$	$\hat{\sigma}_B^2$
Path 1 $[A, B, C, D]$	TLCC	24.00	–	24.00	–	23.00	–
	DCCA(30)	8.00	–	6.00	–	6.00	–
	Proposed	16.03	47.09	11.59	73.56	15.17	30.46
Path 2 $[A, E, F, G]$	TLCC	22.00	–	12.00	–	4.00	–
	DCCA(30)	2.00	–	5.00	–	5.00	–
	Proposed	4.15	3.72	10.49	6.97	8.64	16.32
Path 3 $[A, H, I, J]$	TLCC	24.00	–	24.00	–	1.00	–
	DCCA(30)	8.00	–	12.00	–	14.00	–
	Proposed	8.21	7.97	12.13	7.03	15.72	57.16
Path 4 $[A, H, K, M]$	TLCC	24.00	–	24.00	–	24.00	–
	DCCA(30)	8.00	–	2.00	–	19.00	–
	Proposed	8.23	7.62	15.65	9.01	22.06	3.24
Path 5 $[A, H, K, L]$	TLCC	24.00	–	24.00	–	24.00	–
	DCCA(30)	8.00	–	2.00	–	2.00	–
	Proposed	8.22	7.67	15.55	11.01	20.75	3.17

7. Conclusions

Traffic congestion spreads its effects to the inflow roads, creating a causal relationship between the accident site and adjacent roads. To identify the said relationship, we developed a new method for estimating differences in congestion time. The proposed method utilizes a lag-specific TE estimator with decomposition and normalization techniques. Furthermore, we conducted extensive performance comparisons under varying experimental settings and found that the proposed decomposition and non-linear normalization methods yield substantial performance improvements. We also confirmed that the proposed method produces more stable and robust results than the conventional TDE methods. Thus, the proposed time delay estimation method helps quantitatively understand the propagation of traffic congestion.

Moreover, the bootstrap technique and its density estimation of statistical functionals enable the uncertainty quantification of time delay estimates. This uncertainty quantification allows us to evaluate the reliability of time delay estimates and serves as a basis for optimal hyperparameter tuning. Specifically, $\hat{\sigma}_B^2$ serves as a key indicator of a causal relationship between two-time series. We developed a rigorous and practical guidance for decision-making based on the tolerance interval.

In this study, we only considered the method to obtain accurate time delay estimates using historical traffic data. Eventually, the proposed method will be used to predict time delays in GPS navigation systems, thereby providing users with more accurate estimated arrival times using real-time traffic data.

Using our proposed method as a foundation, real-time delay prediction methods can be developed in future work.

Acknowledgments

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. H. Nguyen, W. Liu, F. Chen, Discovering congestion propagation patterns in spatio-temporal traffic data, *IEEE Trans. Big Data*, **3** (2016), 169–180. https://doi.org/10.1109/TBDATA.2016.2587669

2. J. Y. Lee, J. I. Kwak, Y. K. Oh, S. Kim, Quantifying incident impacts and identifying influential features in urban traffic networks, *Transportmetrica B Transp. Dyn.*, **2022** (2022), 1–22. https://doi.org/10.1080/21680566.2022.2063205

3. P. Newson, J. Krumm, Hidden markov map matching through noise and sparseness, in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, (2009), 336–343. https://doi.org/10.1145/1653771.1653818

4. F. G. Habtemichael, M. Cetin, K. A. Anuar, Incident-induced delays on freeways: Quantification method by grouping similar traffic patterns, *Transp. Res. Record*, **2484** (2015), 60–69. https://doi.org/10.3141/2484-07

5. A. Garib, A. E. Radwan, H. Al-Deek , Estimating magnitude and duration of incident delays, *J. Transp. Eng.*, **123** (1997), 459–466. https://doi.org/10.1061/(ASCE)0733-947X(1997)123:6(459)

6. D. Nam, F. Mannering, An exploratory hazard-based analysis of highway incident duration, *Transp. Res. Part A Policy Pract.*, **34** (2000), 85–102. https://doi.org/10.1016/S0965-8564(98)00065-2

7. M. Zhu, H. F. Yang, C. Liu, Z. Pu, Y. Wang, Real-time crash identification using connected electric vehicle operation data, *Accid. Anal. Prev.*, **173** (2022), 106708. https://doi.org/10.1016/j.aap.2022.106708

8. D. Cao, J. Wu, X. Dong, H. Sun, X. Qu, Z. Yang, Quantification of the impact of traffic incidents on speed reduction: A causal inference based approach, *Accid. Anal. Prev.*, **157** (2021), 10616. https://doi.org/10.1016/j.aap.2021.106163

9. C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, *IEEE Trans. Acoust. Speech Signal Process.*, **24** (1976), 320–327. https://doi.org/10.1109/TASSP.1976.1162830

10. M. Souden, J. Benesty, S. Affes, Broadband source localization from an eigenanalysis perspective, *IEEE Trans. Audio Speech Lang. Process.*, **18** (2009), 1575–1587. https://doi.org/10.1109/TASL.2009.2038556

11. P. Marmaroli, X. Falourd, H. Lissek, A comparative study of time delay estimation techniques for road vehicle tracking, in *Acoustics 2012*, 2012.

12. S. R. Chandra, H. Al-Deek, Cross-correlation analysis and multivariate prediction of spatial time series of freeway traffic speeds, *Transp. Res. Record*, **2061** (2008), 64–76. https://doi.org/10.3141/2061-08

13. C. Shen, Analysis of detrended time-lagged cross-correlation between two nonstationary time series, *Phys. Lett. A*, **379** (2015), 680–687. https://doi.org/10.1016/j.physleta.2014.12.036

14. R. T. Vassoler, G. F. Zebende, Dcca cross-correlation coefficient apply in time series of air temperature and air relative humidity, *Phys. A Stat. Mech. Its Appl.*, **391** (2012), 2438–2443. https://doi.org/10.1016/j.physa.2011.12.015

15. S. Sun, C. Zhang, Y. Zhang, Traffic flow forecasting using a spatio-temporal bayesian network predictor, in *International Conference on Artificial Neural Networks*, (2005), 273–278. https://doi.org/10.1007/11550907_43

16. S. Chawla, Y. Zheng, J. Hu, Inferring the root cause in road traffic anomalies, in *2012 IEEE 12th International Conference on Data Mining*, (2012), 141–150. https://doi.org/10.1109/ICDM.2012.104

17. Y. Xiao, Y. Zhao, G. Wu, Y. Jing, Study on delay propagation relations among airports based on transfer entropy, *IEEE Access*, **8** (2020), 97103–97113. https://doi.org/10.1109/ACCESS.2020.2996301

18. W. Härdle, J. Horowitz, J. P. Kreiss, Bootstrap methods for time series, in *Handbook of Statistics*, **30** (2012), 3–26. https://doi.org/10.1111/j.1751-5823.2003.tb00485.x

19. T. Schreiber, Measuring information transfer, *Phys. Rev. Lett.*, **85** (2000), 461. https://doi.org/10.1103/PhysRevLett.85.461

20. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Technical J.*, **27** (1948), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

21. M. Wibral, N. Pampu, V. Priesemann, F. Siebenhühner, H. Seiwert, M. Lindner, et al., Measuring information-transfer delays, *PloS One*, **8** (2013), e55809. https://doi.org/10.1371/journal.pone.0055809

22. R. Marschinski, H. Kantz, Analysing the information flow between financial time series, *Eur. Phys. J. B Condens. Matter Complex Syst.*, **30** (2002), 275–281. https://doi.org/10.1140/epjb/e2002-00379-2

23. W. C. Hong, P. F. Pai, S. L. Yang, R. Theng, Highway traffic forecasting by support vector regression model with tabu search algorithms, in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, (2006), 1617–1621. https://doi.org/10.1109/IJCNN.2006.246627

24. S. R. Chandra, H. Al-Deek, Predictions of freeway traffic speeds and volumes using vector autoregressive models, *J. Intell. Transp. Syst.*, **13** (2009), 53–72. https://doi.org/10.1080/15472450902858368

25. E. I. Vlahogianni, M. G. Karlaftis, J. C. Golias, Short-term traffic forecasting: Where we are and where we're going, *Transp. Res. Part C Emerging Technol.*, **43** (2014), 3–19. https://doi.org/10.1016/j.trc.2014.01.005

26. D. Pavlyuk, Short-term traffic forecasting using multivariate autoregressive models, *Procedia Eng.*, **178** (2017), 57–66. https://doi.org/10.1016/j.proeng.2017.01.062

27. Z. Song, Y. Guo, Y. Wu, J. Ma, Short-term traffic speed prediction under different data collection time intervals using a sarima-sdgm hybrid prediction model, *PloS One*, **14** (2019), e0218626. https://doi.org/10.1371/journal.pone.0218626

28. E. Ogasawara, L. C. Martinez, D. De Oliveira, G. Zimbrão, G. L. Pappa, M. Mattoso, Adaptive normalization: A novel data normalization approach for non-stationary time series, in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, (2010), 1–8. https://doi.org/10.1109/IJCNN.2010.5596746

29. J. Wang, S. Su, Y. Li, J. Chen, D. Shi, Desaturated probability integral transform for normalizing power system measurements in data-driven manipulation detection, in *2019 IEEE Power & Energy Society General Meeting (PESGM)*, (2019), 1–5. https://doi.org/10.1109/PESGM40551.2019.8973800

30. G. Casella, R. L. Berger, *Statistical Inference*, Cengage Learning, 2021.

31. V. Witkovsky, On the exact two-sided tolerance intervals for univariate normal distribution and linear regression, *Austrian J. Stat.*, **43** (2014), 279–292.