

Analysing functional connectivity and causal dependence in road traffic networks with Granger causality

Md. Mahmud Hasan¹, Jiwon Kim¹

¹ School of Civil Engineering, The University of Queensland, Brisbane St Lucia, Australia

Email for correspondence: jiwon.kim@uq.edu.au

Abstract

Understanding the dependence structure of road segments is a key to building successful prediction models for traffic forecasting. In order to accurately predict the traffic state of a particular target link at a given time interval, the prediction model should incorporate traffic states of other links that are spatially and temporally correlated with the target link into the model structure. Given a potentially very large number of links in a network, however, identifying a subset of links whose traffic states are highly dependent is a challenging task. To tackle this problem, this paper proposes a statistical approach that uses Granger causality analysis to determine the causal dependence among time series data of link traffic flow measures and identify a group of links that are functionally connected to a given target link. The functional connectivity refers to the statistical dependence between two locations in terms of their observed traffic states, regardless of their structural connectivity, which refers to the static/physical connection or spatial adjacency determined by the underlying physical road network. As such, two distant links characterized by low structural connectivity can show high functional connectivity if traffic flow time-series from these two links show high statistical correlation or dependency. The Granger causality analysis has been widely applied to detect such a functional connectivity in various spatio-temporal systems. In this study, the Granger causality analysis is applied on the time-series of link measures (traffic volume or speed) collected from a road network in Brisbane, Australia in 2014 to discover the (causal) dependency structure of the links and understand dynamic changes in the dependence structure across different times of the day. The paper tests both bivariate and multivariate linear vector auto-regression (VAR) models to perform pair-wise and multivariate Granger causality analyses, respectively, and discusses the performance difference between these models. The study also discusses the impact of different choices of link measures (i.e., volume time-series vs. speed time-series) on the performance of identifying the causal structure and the capability of short term traffic prediction.

Key words: Traffic prediction, Granger causality, time-series analysis, causal dependence.

1. Introduction

The ability to accurately predict traffic states in a road network has long been considered central to urban traffic management and Intelligent Transport Systems (ITS). The increasing availability of rich sources of data and computing power has led to growing interests and demands for data-driven or machine learning models such as neural networks and Bayesian networks for large-scale network traffic prediction. One of the challenges in building such data-driven models is, however, to figure out input variables and model structure that are efficient in training with large

amounts of data while being effective in producing accurate prediction results. Including irrelevant variables in a model may decrease its prediction accuracy, not to mention increasing the computational complexity. In the context of short term traffic prediction, this means that we would want to identify a set of links whose traffic states are highly dependent so that we can consider only those relevant links when predicting a particular target link in the set.

Motivated by this need, i.e., identifying the spatial dependencies in a road network to improve the efficiency and effectiveness of short term traffic prediction models, this study proposes a statistical approach that uses Granger causality analysis to determine the causal dependence among time series data of link traffic flow measures. This Granger causality analysis method identifies directed functional or causal interactions of different variables in the time series data (Seth et al., 2015). Granger causality is based on two major principles (i) The cause happens prior to the effect and (ii) The cause makes unique changes in the effect (Granger 1969, 1980). A time series x is said to Granger cause another time series y , if regression for y which includes both past values of y and x is statistically significant than regression for y having only with past values of y (Arnold et al. 2007). Granger causality test has become an established method for analysing potential causal relationship (Li et al., 2015). Although this method is widely used in neuroscience (Dhamala et al., 2008; Barnett and Seth, 2014), economics and air transport studies (Fernandes and Pacheco, 2010; Vijver et al., 2014), it has not been widely explored in the area of road traffic research.

3. Test bed and data

For this study, approximately 11 km long section of a two-way Moggill road in Brisbane, Australia is selected as a test bed. The selected road section starts from Moggill road-Barkin road intersection and ends at Moggill road-Witton road intersection. The study site includes 32 road links in both inbound and outbound directions including minor approach roads. However, due to missing data and invalid flow and speed values, five links among these 32 links are excluded from the analysis. Figure 1 shows the roadways and links of the test bed.

Figure 1: Selected test bed (Google Earth view)

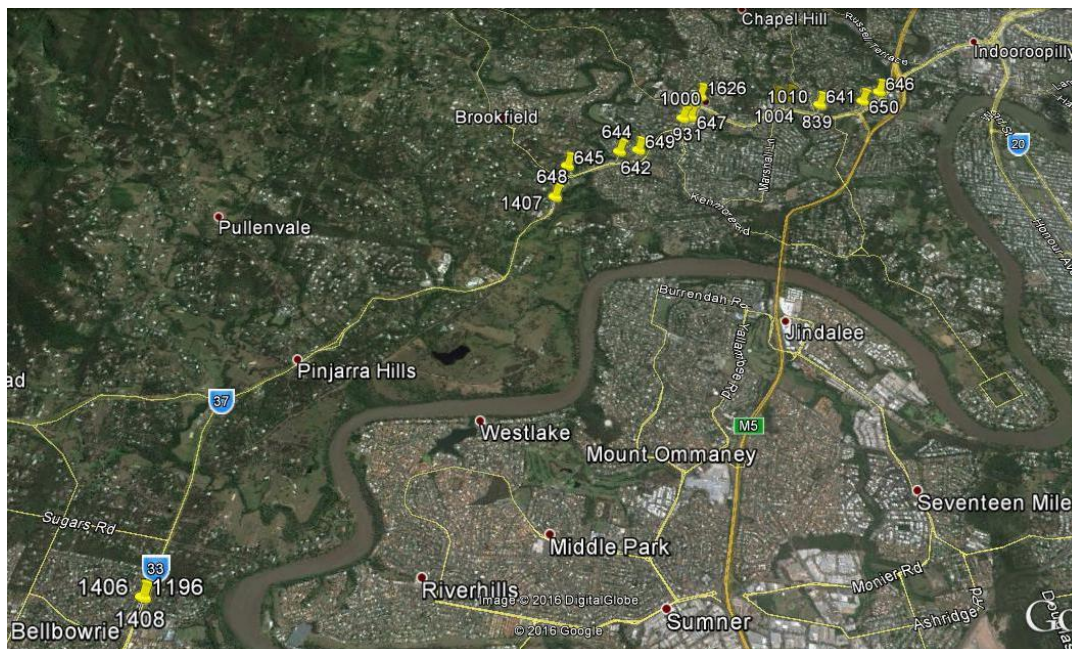


Table1 presents the basic information on the selected 27 links in the test bed which includes total length, number of lanes, design speed and direction of each link.

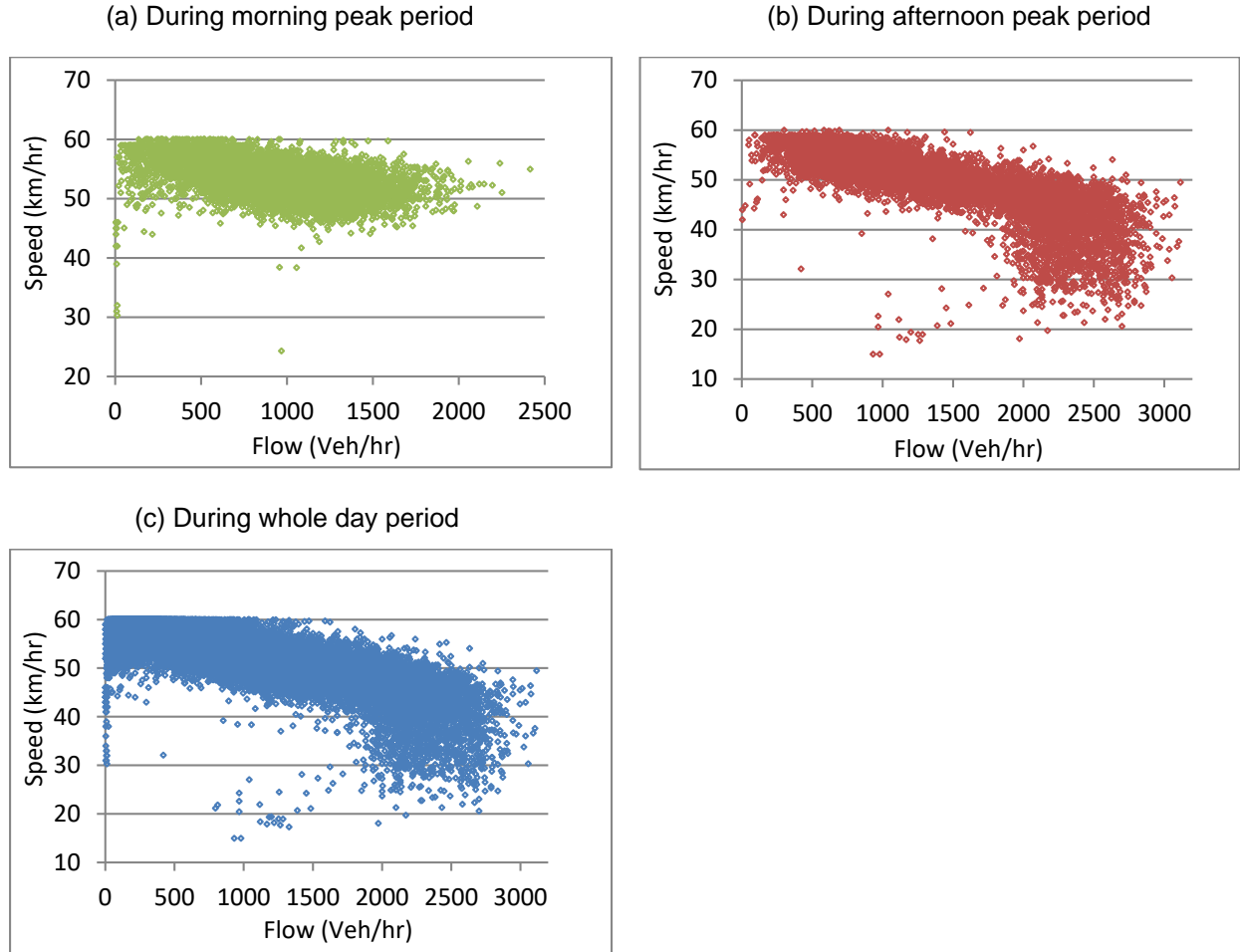
Table 1: Details of each link of the test bed

Link ID	Length (m)	No. of Lanes	Design speed (km/hr)	Type of link	Direction
466	182	1	50	Minor road	Eastbound
641	357	2	60	Major arterial	Eastbound
642	212	1	60	Major arterial	Eastbound
643	149	3	60	Major arterial	Eastbound
644	620	1	70	Major arterial	Eastbound
645	468	1	70	Major arterial	Northbound
646	267	1	60	Major arterial	Northeast bound
647	213	1	60	Major arterial	Southbound
649	189	1	60	Major arterial	Southwest bound
650	267	1	60	Major arterial	Southwest bound
651	357	3	60	Major arterial	Westbound
652	149	1	60	Major arterial	Westbound
653	212	1	70	Major arterial	Westbound
654	620	1	70	Major arterial	Westbound
789	288	1	50	Minor road	Westbound
839	183	1	50	Minor road	Southbound
931	343	1	50	Minor road	Northbound
989	1356	1	60	Minor road	Northbound
999	517	3	60	Major arterial	Eastbound
1004	1050	1	60	Major arterial	Northeast bound
1006	220	3	60	Major arterial	Southwest bound
1010	517	1	60	Major arterial	Westbound
1173	172	1	50	Minor road	Westbound
1196	607	1	50	Minor road	Westbound
1406	965	1	60	Major arterial	Northbound
1407	1340	1	70	Major arterial	Northeast bound
1408	685	1	60	Major arterial	Southbound

In this study, two traffic parameters are considered which are traffic flow and vehicle speed. Traffic data are obtained from Queensland Department of Transport and Main Road (DTMR) through Public Traffic Data System (PTDS). Traffic data were collected from loop detectors in 2014 which include traffic flow and speed in every 15 minutes and details of all links of roads in South East Queensland. The traffic data are divided into four parts based on the time of day i.e. morning peak period (6am-11am), daytime off-peak period (11am- 4pm), afternoon peak period (4pm-10pm), and night time off-peak period (10pm-6am). In this study, we consider morning

peak traffic, afternoon peak traffic and whole day traffic. For the analysis, six traffic parameter cases are selected for Granger causality test as follows: (1) traffic flow at morning peak period, (2) traffic flow at afternoon peak period, (3) traffic flow for the whole day period, (4) speed at morning peak period, (5) speed at afternoon peak period, and (6) speed for the whole day period. Figure 2 presents a typical speed-flow diagram of a link based on the obtained data.

Figure 2: Example of speed-flow diagram of a link in the test bed at morning peak period, afternoon peak period and whole day period.



4. Model description

4.1. Model formulation

In this study, Granger causality method of analysis is employed to find out the spatial relationship among the road links. Granger causality test can be performed by the following regression model. Let \bar{x}_{t-1} be the lagged variable of \bar{x}_t , \bar{y}_{t-1} be the lagged variable of \bar{y}_t and \mathbf{A}, \mathbf{B} represent vectors of coefficients.

$$y_t = a_0 + \mathbf{A} \cdot \bar{y}_{t-1} + \mathbf{B} \cdot \bar{x}_{t-1} \quad (1)$$

$$y_t = a_0 + A \cdot \bar{y}_{t-1} \quad (2)$$

Then applying F test to obtain a p value for whether or not the first equation results in a better regression model than second equation with statistical significance. If the p value rejects the null hypothesis that x does not Granger cause y, then it is said that x Granger causes y.

Granger causality among the variables can be tested by a pairwise or Vector Auto Regression (VAR) method (Kamarianakis and Prastacos, 2003). The pairwise Granger causality test cannot fulfil the aim of the study since conditional probabilities among different variables are needed for actual spatial relationship. Thus, a Granger causality test based on VAR is selected for this study.

In a VAR model, each of the variables is considered as dependent variable once and the rest are taken as independent variables. This model can be developed by bivariate or multivariate time series in which every variable is explained by its own lagged and current values and past values of other variables (Zivot and Wang, 2006).

Let, $Y_t = (y_{1t}, y_{2t}, \dots, y_{nt})$ is a $(n \times 1)$ vector of time series variables. The basic p-lag vector VAR model can be written as

$$Y_t = c + \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \dots + \Pi_p Y_{t-p} + \varepsilon_t \quad (3)$$

where Π_i are $(n \times n)$ coefficient matrices, c is the intercept, $t = 1, \dots, T$ and ε_t is an $(n \times 1)$ unobservable zero mean white noise vector process (serially uncorrelated or independent) with time invariant covariance matrix, Σ .

For example, a multivariate VAR (with 2 lag order) model equation has the following form

$$Y_{1t} = c_1 + \Pi_{11}^1 y_{1,t-1} + \Pi_{12}^1 y_{2,t-1} + \dots + \Pi_{1n}^1 y_{n,t-1} + \Pi_{11}^2 y_{1,t-2} + \Pi_{12}^2 y_{2,t-2} + \dots + \Pi_{1n}^2 y_{n,t-2} + \varepsilon_{1t} \quad (4)$$

$$Y_{2t} = c_2 + \Pi_{21}^1 y_{1,t-1} + \Pi_{22}^1 y_{2,t-1} + \dots + \Pi_{2n}^1 y_{n,t-1} + \Pi_{21}^2 y_{1,t-2} + \Pi_{22}^2 y_{2,t-2} + \dots + \Pi_{2n}^2 y_{n,t-2} + \varepsilon_{2t} \quad (5)$$

$$Y_{nt} = c_n + \Pi_{n1}^1 y_{1,t-1} + \Pi_{n2}^1 y_{2,t-1} + \dots + \Pi_{nn}^1 y_{n,t-1} + \Pi_{n1}^2 y_{1,t-2} + \Pi_{n2}^2 y_{2,t-2} + \dots + \Pi_{nn}^2 y_{n,t-2} + \varepsilon_{nt} \quad (6)$$

After the formation of VAR model, the Granger causality test can be performed by hypothesis testing of F test with zero restriction. In this test, a given time series is considered as Granger cause of another time series if at least one value in the coefficient vector is found non-zero by statistical significance test (Bahodori and Liu, 2012). To check whether a variable such as θ_2 is granger cause of $\theta_{n,t}$ in VAR Equation 6, the null hypothesis is $H_0: \Pi_{n2}^1 = \Pi_{n2}^2 = 0$ and alternative hypothesis is $H_1: \Pi_{n2}^1 \neq 0$ and/or $H_1: \Pi_{n2}^2 \neq 0$.

The restricted model (model with zero restriction) is:

$$\theta_{n,t} = c_n + \Pi_{n1}^1 \theta_{1,t-1} + \dots + \Pi_{nn}^1 \theta_{n,t-1} + \Pi_{n1}^2 \theta_{1,t-2} + \dots + \Pi_{nn}^2 \theta_{n,t-2} + \varepsilon_{n,t} \quad (7)$$

The unrestricted model is:

$$\theta_{n,t} = c_n + \Pi_{n1}^1 \theta_{1,t-1} + \Pi_{n2}^1 \theta_{2,t-1} + \dots + \Pi_{nn}^1 \theta_{n,t-1} + \Pi_{n1}^2 \theta_{1,t-2} + \Pi_{n2}^2 \theta_{2,t-2} + \dots + \Pi_{nn}^2 \theta_{n,t-2} + \varepsilon_{n,t} \quad (8)$$

The test statistic can be written as

$$F_0 = \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{n - (k + 1)}} \quad (9)$$

where SSR_r is the sum of the squared residuals of the restricted model and SSR_{ur} is the sum of the squared residuals of the unrestricted model, n is the number of observations, k is the number of independent variables in the unrestricted model and q is the number of restrictions or the number of coefficients being jointly tested (Blackwell, 2008).

The calculated F_0 value is then compared with the critical value of F with significance value (0.05 or 0.01). If the calculated value is higher than the critical value, then it rejects the null hypothesis and, therefore, it can be said that the time series Granger causes the target time series.

4.2. Model development and estimation

In this study, six traffic scenarios based on different traffic parameters and time periods are considered for model development. These scenarios are: traffic flow at morning peak period, afternoon peak period and whole day period and speed at morning peak period, afternoon peak period and whole day period. Therefore, six different VAR models are developed in this paper to find out how causal relationship changes with time of day. These VAR models are developed by using software package Gretl.

Traffic parameters of all links on the roadway are considered as endogenous variables in multivariate VAR analysis. One of the most important conditions of VAR or Granger causality model is that time series of each variable should be stationary. To meet this criterion, every variable needs to be evaluated by unit root test such as Augmented Dickey Fuller (ADF) test. The following equations are estimated for each of the time series:

$$\Delta y_t = bD_t + d_0y_{t-1} + d_1\Delta y_{t-1} + d_2\Delta y_{t-2} + \dots + d_p\Delta y_{t-p} + u_t \quad (10)$$

where y_t may contain a unit root, D_t is a vector of deterministic variables (a constant and linear trend; sometimes dummy variables for a break in the intercept or the slope of the trend), Δ is the first difference operator, u_t is a disturbance term, t is the time, p denotes the number of lags used and α_t is the error term, b and d 's are parameters (Weber, 2001). The null hypothesis ($H_0: \theta = 0$: The series Δy_t is non-stationary and it needs to be differenced to make it stationary) can be rejected if θ is statistically significant with negative sign. Lag order of this test is selected by BIC as in VAR model. The test results show that alternative hypothesis ($H_1: \theta < 0$) is accepted for all variables with 95% confidence level. It means that all data are stationary and every variable is within unit root of 1. Also, same result is found by Engle-Granger co-integration test (Engle and Granger, 1987) which shows that residuals of ADF regression are not auto-correlated or not co-integrated. So, the data used for Granger causality satisfy the condition of stationary. Figure 3 and Figure 4 show that all of the time series are within in the circle of unit root of 1.

Figure 3: VAR Inverse root in relation to the unit circle plot of flow at morning peak period, afternoon peak period and whole day period (From left to right)

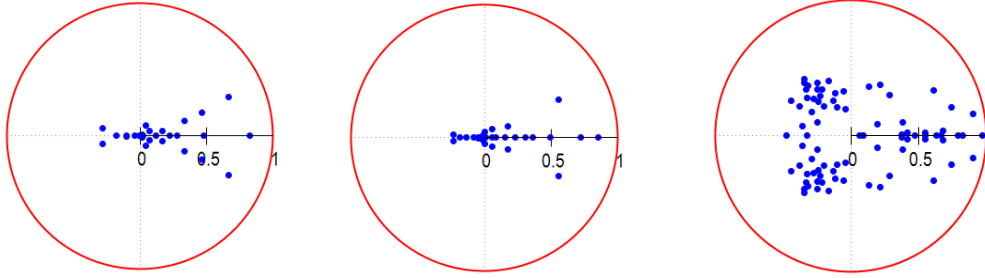
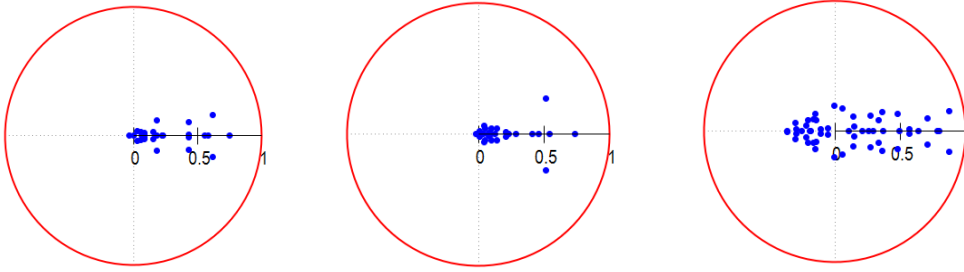


Figure 4: VAR Inverse root in relation to the unit circle plot of speed at morning peak period, afternoon peak period and whole day period (From left to right)

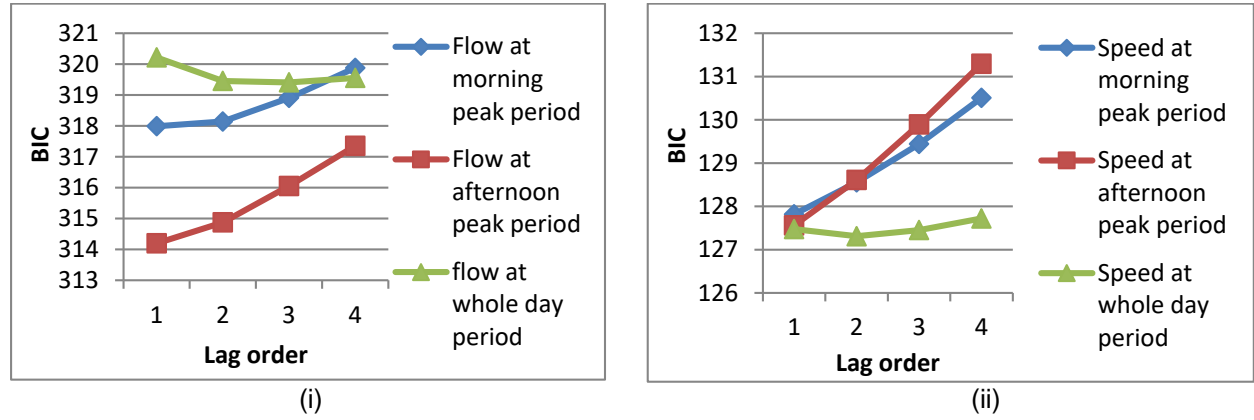


The next step is to select lag order for Granger causality test. This study used 15 minute interval traffic data. So the data are considered to have four lags within an hour. Therefore, maximum lag order is selected as 4. Then actual lag order is calculated by BIC (Schwarz Bayesian Information Criteria) score. The formulation of this measure is

$$\text{BIC} = -2l(\hat{\theta}) + k \log n \quad (11)$$

where $l(\hat{\theta})$ represents the maximum log-likelihood as a function of the vector of parameter estimates ($\hat{\theta}$) and k is the number of independently adjusted parameters within the model. BIC is negatively related to the likelihood and positively related to the number of the parameter. So $k \log n$ means that the penalty for adding extra parameters grows with sample size. This ensures, asymptotically, one will not select a larger model over a correctly specified parsimonious model. Among different lag orders, the one with the lowest BIC score is considered as actual lag order for modelling (Cottrell and Lucchetti, 2016). Figure 5 depicts BIC score and lag order selection for each of the six cases.

Figure 5: BIC score and lag order for (i) flow based model and (ii) speed based model



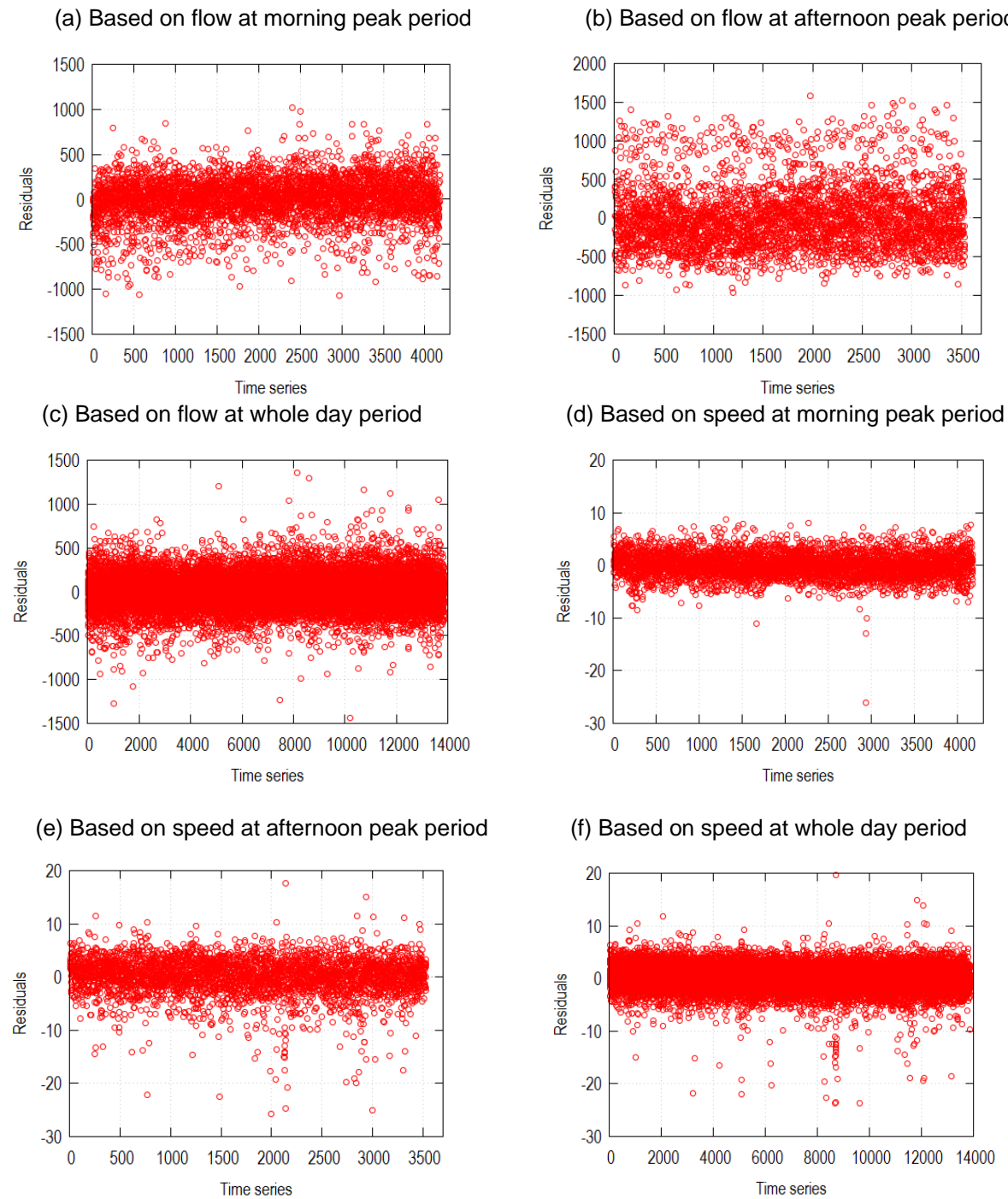
Durbin-Watson test are also evaluated in order to find out whether the residuals of the model variables are auto correlated or not. The test statistics

$$d = \sum_{t=2}^n (z_t - z_{t-1})^2 / \sum_{t=1}^n z_t^2 \quad (12)$$

where $z_t \dots z_n$ are the residuals, n is number of observations (Durbin and Watson, 1971). In all cases, Durbin Watson test statistics (d) values are in between 2 to 2.1 which are higher than upper critical value of Durbin Watson statistics ($d_u = 1.78$ at level of significance $\alpha = 0.05$). Therefore, it can be said that model residuals are not auto-correlated meaning that errors in past have no effect on errors at present.

Residual vs. Time series plots (Figure 6) for six time periods of analysis depict that errors are independent or they are not in the form of positive or negative correlation. In fact, no trend can be identified from Figure 6. Speed residuals plot has lesser scatter points compared to flow residuals plot. In the residual illustrations, a number of outliers are found in the afternoon peak periods for both flow and speed. This is due to the fact that link 651 has been considered here as an example of target link which is in the outbound direction. Traffic flow rate are much higher during afternoon peak period and therefore, traffic congestion may occur often.

Figure 6: Residuals vs. Time series for flow and speed based analysis of morning peak period, afternoon peak period and whole day period



4.3. Model validation

As mentioned in the introduction, the main motivation of this study is to build a data-driven traffic prediction model in a more systematic manner. Based on the assumption that the prediction accuracy of the traffic state of a particular target link is improved by including only those links that are statistically dependent to the target link into the model, this study proposes a Granger causality-based method that aims to automatically identify a set of links that are statistically dependent each other. In order to evaluate the effectiveness of the proposed method, we build a simple traffic prediction model using a Bayesian Network (BN) to test the prediction accuracy of the BN model under different combinations of input variables. A Bayesian Network (BN) is a probabilistic graphical model that represents probabilistic relationships among a set of variables via a directed acyclic graph (DAG). A BN consists of a set of nodes and a set of edges, where nodes represent random variables and edges connecting pairs of nodes represent direct dependencies between variables. Recently, a BN has been increasingly used in the studies for traffic state or congestion prediction (Pascale and Nicoli, 2011; Kim and Wang, 2016).

In this study, nodes in a BN represent a set of links and edges represent dependency relations between links. Each node is a discrete random variable that represents traffic measure θ , which can be either flow or speed, of the corresponding link and takes one of four discrete states {very low, low, high, very high} in terms of the value of θ . The discretization is performed for each link based on the percentage fraction of the respective maximum value. The discretised four states are defined as follows: very low (<0.25 of the maximum value), low (>0.25 to <0.5 of the maximum value), high ($>=0.5$ to <0.75 of the maximum value) and very high ($>=0.75$ of the maximum value). The full description for the discretization and state definition is presented somewhere else (Kim and Wang, 2016) and interested readers are referred to that paper.

Five different scenarios are tested in terms of the variable selection for a BN model. Given a target link,

- Scenario 1: includes those links identified by the Granger causality analysis into the model.
- Scenario 2: includes only the nearest upstream and downstream links of the target link into the model.
- Scenario 3: includes both Granger causal links as well as the immediate upstream and downstream links to the model.
- Scenario 4: includes all links in the same direction of target link into the model.
- Scenario 5: includes all links in both directions into the model.

Then each of these scenarios is tested for two traffic measures i.e. flow and speed, separately at three different periods of time such as morning peak period, afternoon peak period and whole day period. As such, a total of 30 BN models were specified.

The specification of each BN model structure (i.e., edge connection) is based on the actual network connectivity of road links. In these Bayesian networks, upstream and downstream links of target link are placed at upstream and downstream of target node. Since traffic parameters of upstream link have effects on traffic parameters of downstream link, the edge or arrow between two nodes is directed from upstream to downstream. The links of minor road in the intersections are connected to their nearest downstream link of major road and the arrow is connected from minor road's node towards major road's node. Bayesian network model developed by Granger causal links, Granger links with upstream-downstream links and all links of the roads include

links on both ways of the road. In these BN structures, the nearest node in opposite direction of target node is connected with target node by directing an arrow from opposite directional node to the target node. All other nodes in the opposite direction follow traffic movement direction from upstream to downstream. However, other two scenarios based BN structures include only the links in the same direction of target link.

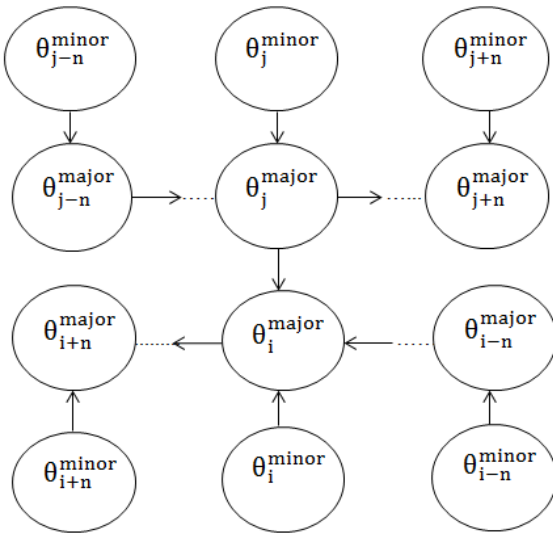
This paper uses software package Netica to build BN models. Figure 7 provides graphical illustrations of the tested BN models, where traffic parameters (flow and speed) are indicated as θ , target link and nearest opposite directional link of target link are identified as subscripts i and j respectively. Also, major or minor road link is mentioned as a superscript major or minor. The notations used in Figure 7 are summarized in Table 2.

Table 2: Summary of notations for BN variables

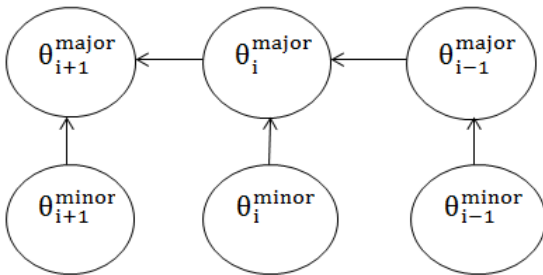
Notation	Description
θ_i^{major}	traffic parameter at target link
$\theta_{i-1}^{\text{major}}$	traffic parameter at the nearest upstream of target link
$\theta_{i+1}^{\text{major}}$	traffic parameter at the nearest downstream of target link
$\theta_{i-n}^{\text{major}}$	traffic parameter at the n^{th} distant upstream link of target link
$\theta_{i+n}^{\text{major}}$	traffic parameter at the n^{th} distant downstream link of target link
θ_j^{major}	traffic parameter at the nearest opposite directional link of target link
$\theta_{j-1}^{\text{major}}$	traffic parameter at the nearest upstream of opposite directional link j
$\theta_{j+1}^{\text{major}}$	traffic parameter at the nearest downstream of opposite directional link j
$\theta_{j-n}^{\text{major}}$	traffic parameter at the n^{th} distant upstream link of opposite directional link j
$\theta_{j+n}^{\text{major}}$	traffic parameter at the n^{th} distant downstream link of opposite directional link j
θ_i^{minor}	minor road approach link connecting to target link
$\theta_{i-n}^{\text{minor}}$	minor road link connecting to the n^{th} distant upstream of target link
$\theta_{i+n}^{\text{minor}}$	minor road link connecting to the n^{th} distant downstream of target link
θ_j^{minor}	minor road link connecting to link j
$\theta_{j-n}^{\text{minor}}$	minor road link connecting to the n^{th} distant upstream link of link j
$\theta_{j+n}^{\text{minor}}$	minor road link connecting to n^{th} distant downstream link of link j link

Figure 7: Bayesian network models for testing five variable selection scenarios

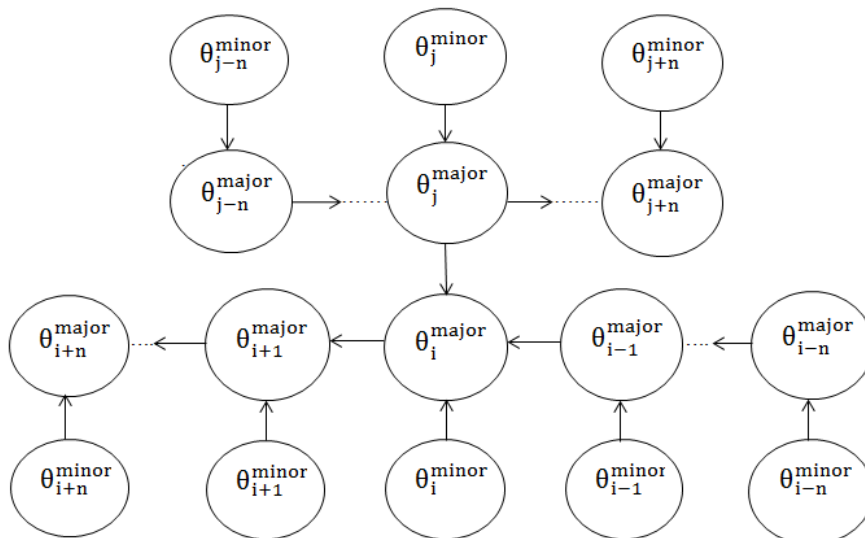
(a) Scenario 1: Granger causal links based Bayesian network structure:



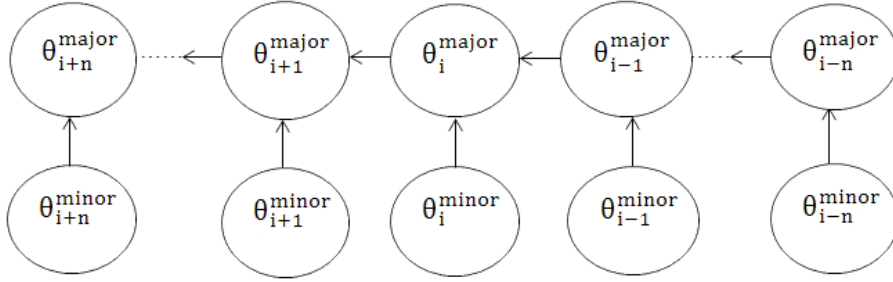
(b) Scenario 2: Nearest upstream-downstream links based Bayesian network structure:



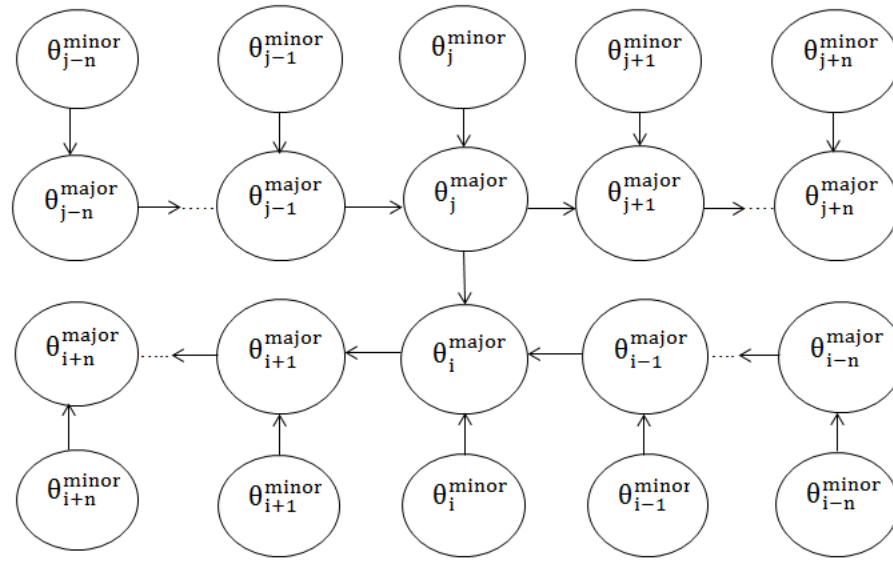
(c) Scenario 3: Granger causal links with nearest upstream downstream links based Bayesian network structure:



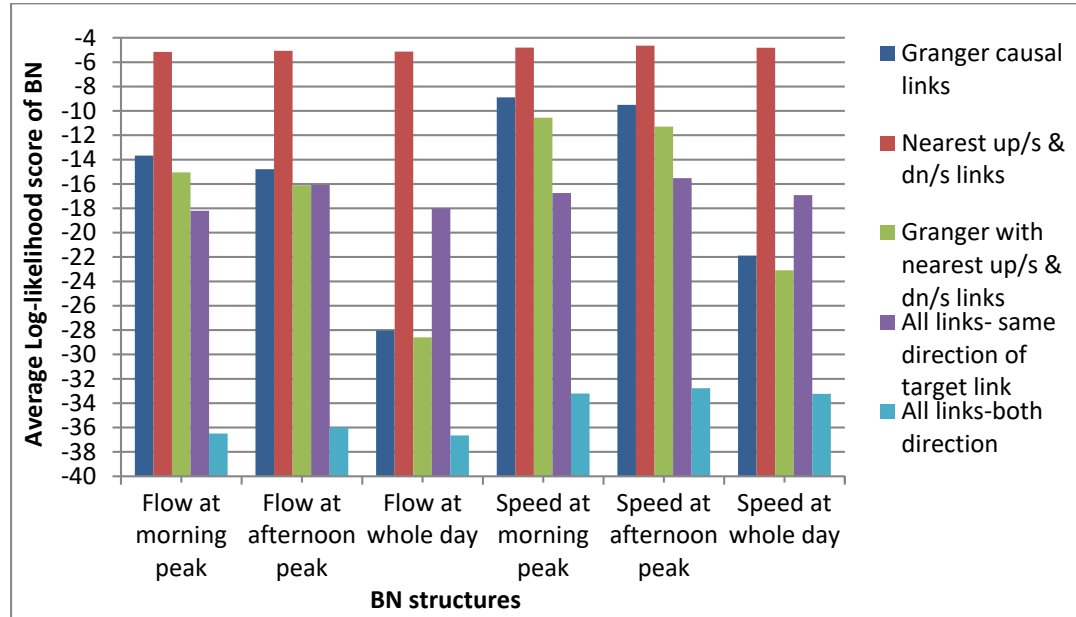
(d) Scenario 4: All links in the same direction of target link based Bayesian network structure:



(e) Scenario 5: All links in both directions based Bayesian network structure:



This study uses log-likelihood as the scoring function for BN structures. Figure 8 presents the average log-likelihood score of proposed BN structures where higher the value, better the model. It is found that nearest upstream-downstream based BN models have the highest log-likelihood score in all cases. Granger causal links based BN structures have the second highest log-likelihood values in the morning peak and afternoon peak periods. However, for whole day period, all links in the same direction of target link based BN models have higher log-likelihood values than that of Granger causal links based BN models. In fact, log-likelihood value depends on number of nodes and connection arrows. Therefore, the nearest upstream-downstream links based BN structures which have only 3 or 4 nodes produce the highest log-likelihood scores. Also, during whole day period Granger causal links based BN structures include more number of nodes than morning peak and afternoon peak. Therefore, all links in the same direction of target link based BN has the second highest log-likelihood values during whole day period. Overall, it can be said that the proposed BN structures describe the data quite closely because of having significant higher score in log-likelihood.

Figure 8: Average Log-likelihood score of proposed BN model structures.

To validate the proposed BN models, data are separated into training dataset and testing dataset. In this study, 80% of total data is taken as a training set and 20% of total data as a testing dataset. Bayesian model is developed by the training dataset and testing dataset is used for validation. In the validation process, each model predicts the traffic parameter state of the target node by testing data of all other nodes. The error rate of prediction is found by confusion matrix (Table 3) which represents the number of cases those are predicted inaccurately. The prediction accuracy of the model is calculated as 100 minus the error rate of the prediction.

Table 3: Confusion matrix for developed BN structures

Actual traffic parameter condition	Prediction of traffic parameter condition			
	Very low	Low	High	Very high
Very low	$TRUE_{very\ low}$	$FALSE_{low}$	$FALSE_{high}$	$FALSE_{very\ high}$
Low	$FALSE_{very\ low}$	$TRUE_{low}$	$FALSE_{high}$	$FALSE_{very\ high}$
High	$FALSE_{very\ low}$	$FALSE_{low}$	$TRUE_{high}$	$FALSE_{very\ high}$
Very high	$FALSE_{very\ low}$	$FALSE_{low}$	$FALSE_{high}$	$TRUE_{very\ high}$

$$\text{Error percentage of prediction} = (\text{Total false prediction} / \text{Total number of cases}) * 100\% \quad (13)$$

$$\text{Prediction accuracy percentage} = 100 - \text{Error percentage of prediction} \quad (14)$$

5. Results and discussion

In this study, all the traffic links on the roadway including minor road approaches are considered for evaluation of spatial relationship. Granger causality test is conducted on each of the six cases i.e. traffic flow and speed at morning peak period, afternoon peak period and whole day

period. This study considers 99% confidence level while testing Granger causality in order to get more accurate causal relationship. The results of the Grange causality show that spatial connectivity depends on traffic parameters as well as time of the day. Each traffic parameter case is found to have different Granger causal links and even within same traffic parameter, Granger causal relationship changes over time of the day. In fact, three time periods have different number of observations. For example, the number of Granger causal links is higher during whole day period because of having much higher number of observations compared to other two time periods. Also, the two traffic parameters have different ranges of values i.e. traffic flows are found to have higher ranges (from 4 to 4480 veh/hr) whereas speed values have smaller ranges (from 7 to 70 km/hr). This is one of the key factors of providing different Granger causal links by these two traffic parameters. Figure 9 and Figure 10 illustrate that Granger causality relationship depends on time of the day as well as traffic parameters. As an example, outbound link 651 in Moggill road is considered here as a target link and all other links are taken as predictor links.

Figure 9: Granger causal links of target link at different time period based on traffic flow

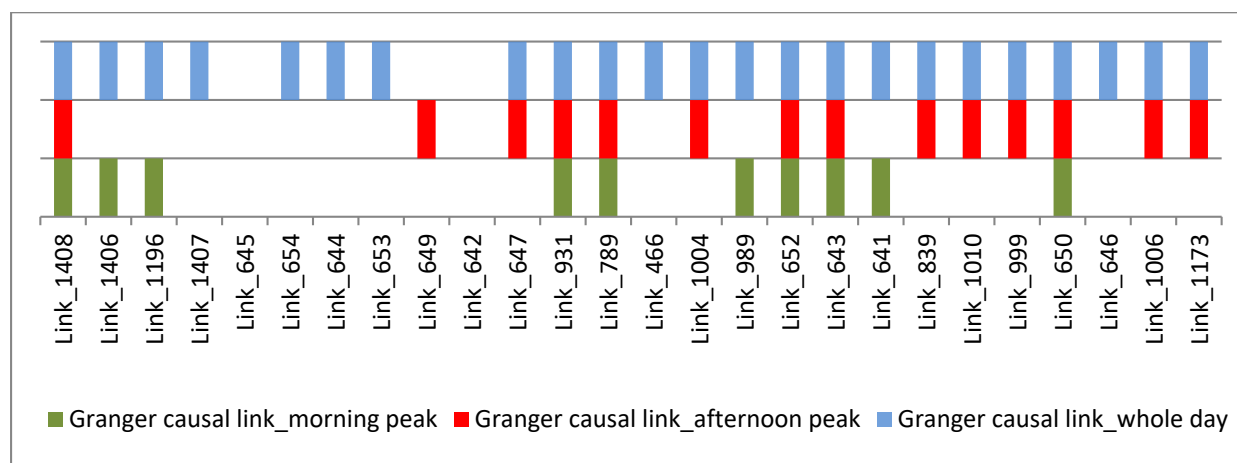


Figure 10: Granger causal links of target link at different time period based on speed

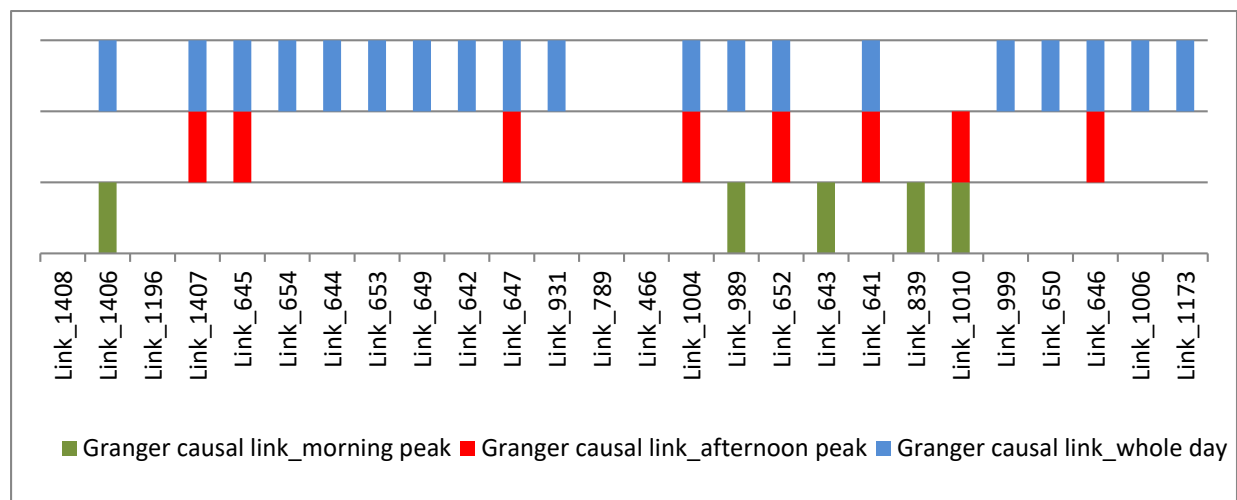


Figure 11 and Figure 12 illustrate that the spatial relationship among target link and upstream-downstream links changes with traffic parameters as well as time periods. In these figures, red circle represents target link (Link 651 is taken as an example) and green circles indicate Granger causal links.

Figure 11: Granger causal links based on flow at (a) morning peak period, (b) afternoon peak period, and (c) whole day period

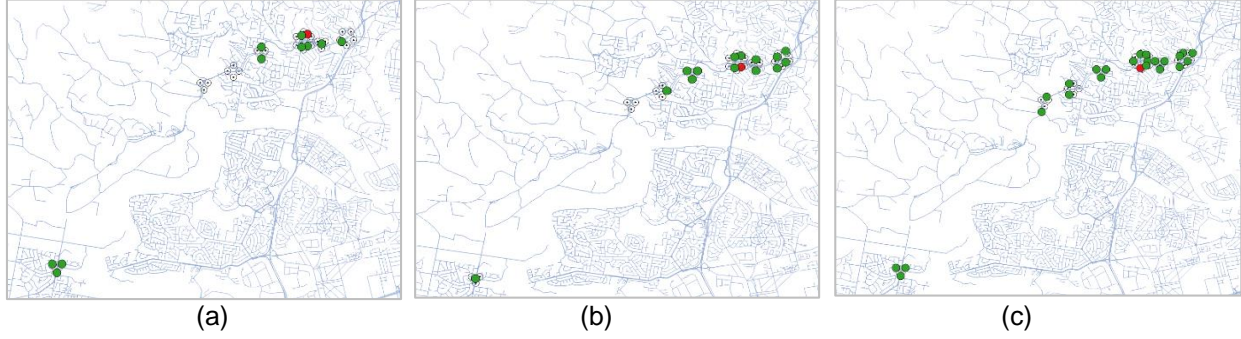
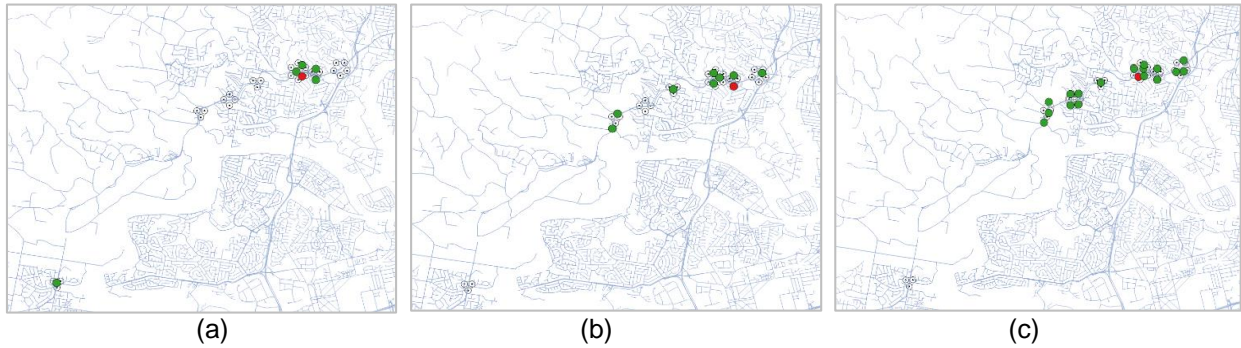


Figure 12: Granger causal links based on speed at (a) morning peak period, (b) afternoon peak period, and (c) whole day period



To evaluate prediction accuracy of BN models, each of all road links (in total twenty links) in the main road is selected as a target link. However, the minor road approach links are not taken as target links because data are not available for their nearest upstream links. Prediction accuracy percentage of each of these twenty links is calculated using Eq.14 and the average value is obtained for different measures (flow vs. speed), time-of-day conditions, and variable selection scenarios. The average prediction accuracy for each case is presented in Table 4 and Figure 13.

First, it is observed that prediction accuracy is significantly higher when using speed as model variable than when using flow. This may pertain to the shape of fundamental diagram (FD). In FD, speed decreases or increases monotonically with density, while flow follows a parabolic path. As such, flow can be more complex to predict, compared to speed, as the same level of flow can be observed in both uncongested and congested phases making the classification task more complex. This explanation is also supported by the fact that the prediction accuracy varies more widely with time-of-day selection in the flow case than in the speed case. For instance, in Figure 13, the prediction accuracy of flow is higher for the afternoon peak than the morning or whole-day periods, which can be because the portions of FD during the afternoon peak

contained a path that is more monotonic than the portion covering the morning peak or the entire curve associated with the whole day. However, a further investigation needs to be carried out to draw conclusions about the impact of traffic parameter and time-of-day choice on prediction accuracy.

Next, we compare different variable selection scenarios within each case of traffic parameter and time-of-day choices. Overall, the difference in prediction accuracy appears to be small in terms of the absolute magnitude of prediction accuracy percentage. However, when we take a closer look at its relative difference in Figure 13(b), it shows patterns that provide insights into the performance of different scenarios and potential benefit of the proposed method. An overall pattern shows that scenarios 1, 3, and 5 produce higher prediction accuracy than scenarios 2 and 4. More specifically, we can make the following observations:

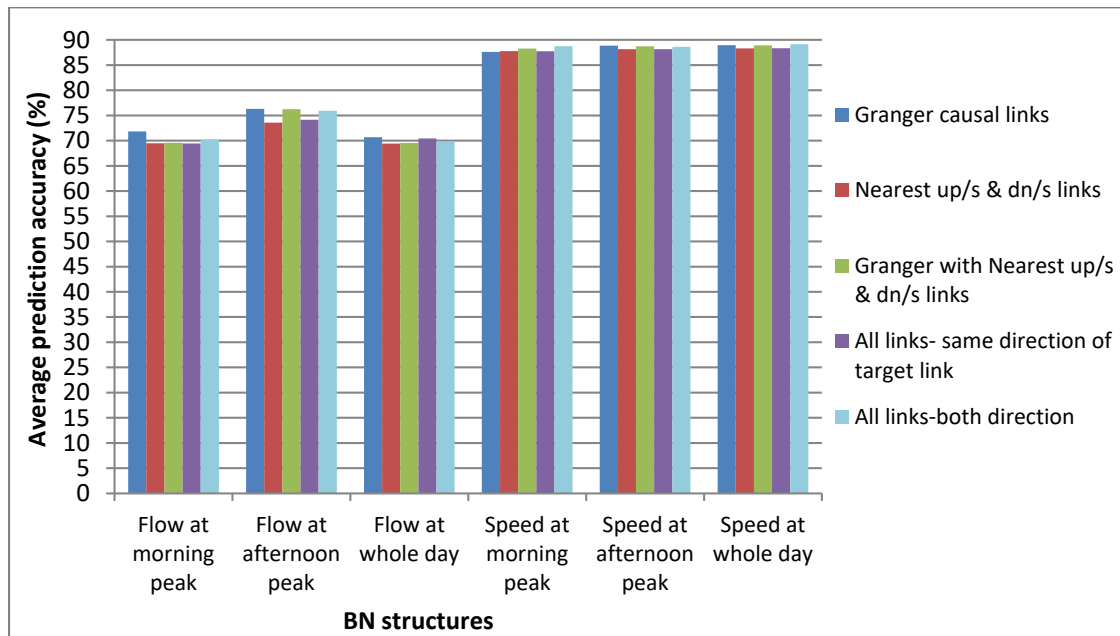
- **Scenario 2 vs. Scenario 3:** The models that only include the nearest upstream and downstream links (scenario 2) could be further improved by adding Granger-causal links (scenario 3).
- **Scenario 1 vs. Scenario 3:** The models that only include Granger-causal links (scenario 1) perform equally well or better (in flow cases) than the models that contain both Granger-causal and the nearest up/downstream links (scenario 3).
- **Scenario 1 vs. Scenario 5:** The models that only include Granger-causal links (scenario 1) perform equally well or better than the models that contain all links in both directions (scenario 5).
- **Scenario 4 vs. Scenarios 5 and 2:** If we remove the links in the opposite direction (scenario 4) from the models with all links (scenario 5), the performance of those models decreases and they (scenario 4) perform equally poorly as those with the nearest up/downstream links only (scenario 2).

Although it is difficult to generalize the observed patterns beyond the tested data sample, the proposed method (scenario 1) seems to offer an appropriate trade-off between *model accuracy* and *model simplicity*; that is, models with Granger-causal links (scenario 1) are more accurate than the simplest model (scenario 2) and simpler than the most comprehensive model (scenario 5).

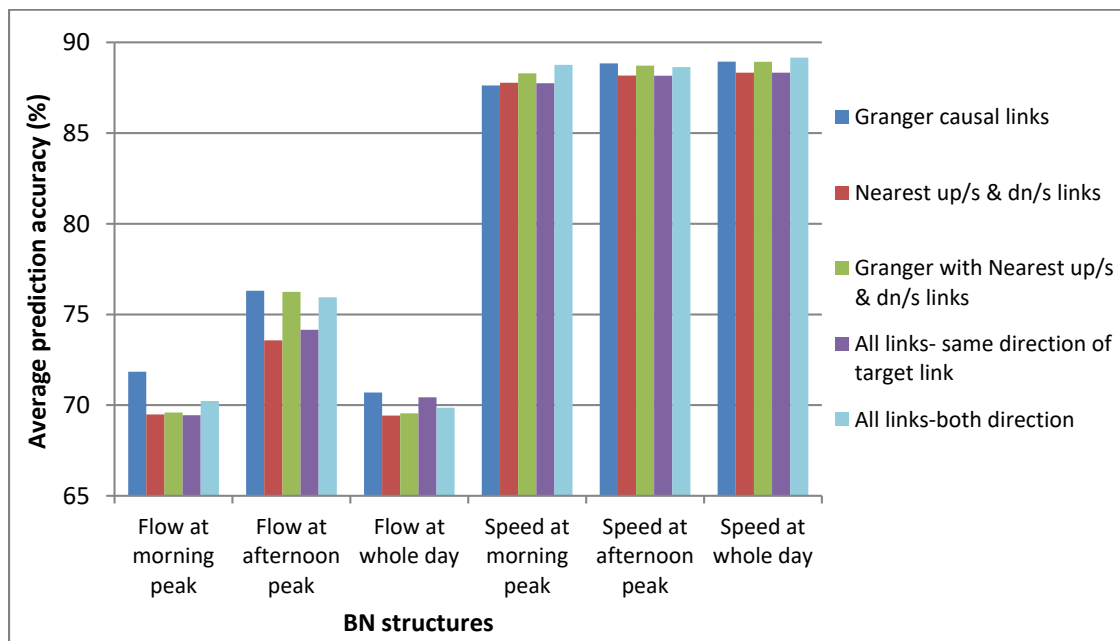
Table 4: Average prediction accuracy percentages of BN structures in different time and variable selection scenarios

Variable selection scenario \ Time period	Flow at morning peak period	Flow at afternoon peak period	Flow at whole day period	Speed at morning peak period	Speed at afternoon peak period	Speed at whole day period
Granger links	71.84%	76.30%	70.69%	87.62%	88.84%	88.94%
Nearest upstream-downstream links	69.48%	73.56%	69.42%	87.77%	88.17%	88.32%
Granger links with upstream-downstream links	69.59%	76.24%	69.55%	88.29%	88.71%	88.92%
All links in the same direction of target links	69.44%	74.15%	70.43%	87.74%	88.16%	88.33%
All links in the road- both directions	70.23%	75.94%	69.85%	88.76%	88.64%	89.15%

Figure 13: Average prediction accuracy percentages of each spatial-temporal based BN model.



(a) Original Y-axis (y-axis bounds: 0% – 90%)



(a) Truncated Y-axis (y-axis bounds: 65% – 90%)

6. Conclusion

This study evaluates the spatial relationship of traffic parameters of various links in a roadway at three different periods of time i.e. morning peak period, afternoon peak period and whole day period. In order to find out spatial connectivity, vector auto-regression based Granger causality model is developed. Then spatial relationship efficiency of Granger causal model is evaluated

by comparing with other spatial connectivity of road links. For this comparison, Bayesian network structures have been adopted in this paper. BN structures are developed based on five different variable selection scenarios of roadway links such as Granger causal links, nearest upstream-downstream links of target link, Granger causal links with nearest upstream-downstream links, all links in the same direction of target link and all links in both directions of the road. This comparison result shows that Granger causality based models provide a good trade-off between model accuracy and model simplicity, suggesting the potential of using Granger causal analysis in guiding variable selection for data-driven traffic prediction models. A further investigation will be carried out to evaluate the performance of the proposed method in a more general setting, focusing on its applications in large-scale networks. Another future research direction includes the consideration of dynamic aspects of spatial dependencies, where the connectivity of road links varies dynamically, in conjunction with the use of dynamic Bayesian networks in order to build models for short-term traffic prediction.

7. References

1. Arnold, A., Liu, Y. and Abe, N., 2007, August. Temporal causal modeling with graphical Granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 66-75). ACM.
2. Bahadori, M.T. and Liu, Y., 2012, October. On Causality Inference in Time Series. In *2012 AAAI Fall Symposium Series*.
3. Cottrell, A. and Lucchetti, R., 2016. GNU regression, econometrics and time-series library. *Computer software*. Retrieved from <http://gretl.sourceforge.net>.
4. Barnett, L. and Seth, A.K., 2014. The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. *Journal of neuroscience methods*, 223, pp.50-68.
5. Blackwell, M., 2008. Multiple hypothesis testing: The F-test. *Matt Blackwell Research*.
6. Dhamala, M., Rangarajan, G. and Ding, M., 2008. Analyzing information flow in brain networks with nonparametric Granger causality. *NeuroImage*, 41(2), pp.354-362.
7. Durbin, J. and Watson, G.S., 1971. Testing for serial correlation in least squares regression. III. *Biometrika*, 58(1), pp.1-19.
8. Engle, R.F. and Granger, C.W., 1987. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pp.251-276.
9. Fernandes, E. and Pacheco, R.R., 2010. The causal relationship between GDP and domestic air passenger traffic in Brazil. *Transportation Planning and Technology*, 33(7), pp.569-581.
10. Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pp.424-438.
11. Granger, C.W., 1980. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2, pp.329-352.

12. Kamarianakis, Y. and Prastacos, P., 2003. Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches. *Transportation Research Record: Journal of the Transportation Research Board*, (1857), pp.74-84.
13. Kim, J. and Wang, G., 2016 Diagnosis and Prediction of Traffic Congestion on Urban Road Networks Using Bayesian Networks. *Transportation Research Record: Journal of the Transportation Research Board* (in press)
14. Li, L., Su, X., Wang, Y., Lin, Y., Li, Z. and Li, Y., 2015. Robust causal dependence mining in big data network and its application to traffic flow predictions. *Transportation Research Part C: Emerging Technologies*, 58, pp.292-307.
15. Pascale, A. and Nicoli, M., 2011, June. Adaptive Bayesian network for traffic flow prediction. In *Statistical Signal Processing Workshop (SSP), 2011 IEEE* (pp. 177-180)
16. Seth, A.K., Barrett, A.B. and Barnett, L., 2015. Granger causality analysis in neuroscience and neuroimaging. *The Journal of Neuroscience*, 35(8), pp.3293-3297.
17. Van De Vijver, E., Derudder, B. and Witlox, F., 2014. Exploring causality in trade and air passenger travel relationships: the case of Asia-Pacific, 1980–2010. *Journal of Transport Geography*, 34, pp.142-150.
18. Weber, C.E., 2001. F-tests for lag length selection in augmented Dickey–Fuller regressions: some Monte Carlo evidence. *Applied Economics Letters*, 8(7), pp.455-458.
19. Zivot, E. and Wang, J., 2006. Vector autoregressive models for multivariate time series. *Modeling Financial Time Series with S-PLUS®*, pp.385-429.