Real time traffic flow outlier detection using short-term traffic conditional variance prediction

Article *in* Transportation Research Part C Emerging Technologies · August 2014

Impact Factor: 2.82 · DOI: 10.1016/j.trc.2014.07.005

3 authors, including:

Jianhua Guo
Southeast University (China)
35 PUBLICATIONS **193** CITATIONS
SEE PROFILE

Billy M Williams
North Carolina State University
41 PUBLICATIONS **1,054** CITATIONS
SEE PROFILE

Transportation Research Part C xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Transportation Research Part C

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

Real time traffic flow outlier detection using short-term traffic conditional variance prediction

Jianhua Guo [a,*], Wei Huang [a], Billy M. Williams [b,1]

[a] Intelligent Transportation System Research Center, Southeast University, Nanjing 210096, Jiangsu Province, P.R. China
[b] Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, NC 27695, USA

ARTICLE INFO

Article history:
Received 28 January 2014
Received in revised form 26 June 2014
Accepted 29 July 2014
Available online xxxx

Keywords:
Outlier detection
Intervention analysis
Traffic flow series
Short term traffic forecasting
SARIMA + GARCH
Kalman filter

ABSTRACT

Outliers in traffic flow series represent uncommon events occurring in the roadway systems and outlier detection and investigation will help to unravel the mechanism of such events. However, studies on outlier detection and investigations are fairly limited in transportation field where a vast volume of traffic condition data has been collected from traffic monitoring devices installed in many roadway systems. Based on an online algorithm that has the ability of jointly predict the level and the conditional variance of the traffic flow series, a real time outlier detection method is proposed and implemented. Using real world data collected from four regions in both the United States and the United Kingdom, it was found that outliers can be detected using the proposed detection strategy. In addition, through a comparative experimental study, it was shown that the information contained in the outliers should be assimilated into the forecasting system to enhance its ability of adapting to the changing patterns of the traffic flow series. Moreover, the investigation into the effects of outliers on the forecasting system structure showed a significant connection between the outliers and the forecasting system parameters changes. General conclusions are provided concerning the analyses with future work recommended to investigate the underlying outlier generating mechanism and outlier treatment strategy in transportation applications.

1. Introduction

Outliers and outlier detection are universal for many data processing tasks, covering a wide range of fields such as image processing, econometrical investigation, industrial control, or intrusion detection, to name a few (Hodge and Austin, 2004; Gupta et al., 2013). A good definition of an outlier is *an observation that deviates so much from other observations as to arouse suspicious that it was caused by a different mechanism* (Hawkins, 1980). From this definition, the notion of different mechanism would be essential for delineating an outlier, and this different mechanism would either be system malfunctions to create noisy observations that should be removed in data processing and analysis, or different data generating mechanisms that create valid but extraordinary patterns that should be further investigated. For either case, outlier detection, i.e., the process of identifying such discordant observations, would be the first step of carrying out these outlier-related investigations.

* Corresponding author. Tel.: +86 25 8379 3131; fax: +86 25 8379 2869.
E-mail addresses: jg2nh@yahoo.com (J. Guo), hhhwei2005@126.com (W. Huang), billy_williams@ncsu.edu (B.M. Williams).
[1] Tel.: +1 919 5157813.

http://dx.doi.org/10.1016/j.trc.2014.07.005
0968-090X/

In the field of transportation engineering, over the decades, a vast amount of traffic monitoring and surveillance devices such as inductive loop detectors have been installed on the roadway systems around the globe, and these devices are now continuously collecting traffic characteristic series, representing the evolution of traffic patterns over time on these roadway systems. Consequently, a huge amount of traffic condition data has been archived, which has stimulated a variety of investigations on making use of such data so as to prevent or mitigate the negative effects associated the development of traffic such as congestion and safety issues. It should be pointed out that, in the archival of these traffic condition data, outliers are inevitable, including either erroneous data generated from reasons like device malfunctions or valid data representing extraordinary patterns such as traffic accidents or adverse weather conditions. Therefore, the detection and investigation into the outliers will provide a valuable means of understanding the mechanism of generating such observations and hence supply more insights for supporting the application and analysis of these data.

However, upon the review as in the following section, though there has been a vast literature in the field of general outlier detection and investigation, the outlier related investigations in traffic condition data are seemingly very limited and current works are primarily in the context of short term traffic condition forecasting. Short term traffic condition forecasting is one of the fundamental topics for supporting the development of proactive traffic management and control systems. As noted in Chatfield (1993), short term traffic condition forecasting includes level prediction and prediction interval generation, and in general, the former relies on the modeling of the conditional mean of the traffic condition process and the latter relies on the modeling of the conditional variance of the traffic condition process. It should be noted that for traffic outlier detection, the conditional variance modeling is important to provide the ability of measuring the uncertainty of the process and hence provide the quantitative foundation of designing the outlier detection method. In this regard, as compared to previous studies that assumes conditional variance is constant for traffic condition series, traffic condition series has recently been shown to be heteroscedastic in nature (Guo et al., 2012), and this time-varying second order property of the traffic condition series will be modeled and assimilated into the outlier detection procedure.

In this paper, based on an online short term traffic flow forecasting system that is capable of generating jointly the level forecast and associated time-varying conditional variance, an online outlier detection method is proposed and the effects of the outliers are investigated on the forecasting system. Empirical study is performed using real world data collected from 36 stations from around the world, showing that outliers can be detected and should be assimilated into the forecasting system for adaptively responding to the changing traffic patterns.

In the rest of this paper, first, a literature review section is provided on general outlier detection studies, outlier detection in time series process, and traffic conditional variance modeling. Then, the proposed methodology is presented, followed by empirical investigations including the description of the data, the outlier detection performances, and the effects of outliers on the forecasting system. Finally, conclusions and discussions are provided together with recommendations on future research.

2. Literature review

As discussed previously, outlier detection is a topic that has been investigated in many disciplines, and in this section, first, general outlier detection approaches are briefly discussed, and then the outlier detection in time series process, in particular in vehicular traffic condition time series, is reviewed. Afterward, related work on traffic conditional variance modeling and prediction is presented due to its important role in vehicular traffic flow outlier detection.

2.1. General outlier detection approach

As mentioned in previous discussions, outlier detection has been investigated in a variety of fields and many approaches have been developed. As stated in Angiulli and Pizzuti (2005), given a dataset, the definition of exceptional data should be defined before the exceptional data could be detected. Having this in mind, two approaches have been widely investigated in outlier detection. The first approach can be generally termed as model-based approach in that a model is first identified and applied to describe the data and any data that cannot be described or depart from the model over a certain threshold will be deemed as outliers (Albanese et al., 2014). In this approach, the model could take various forms such as a probability density function, time series model, regression models, etc., depending on the specific applications and datasets. The second approach is generally termed as distance-based approach in that given a specific distance measure (generally the Euclidean distance) the data are deemed as outliers if the data are over a distance away from its nearest neighbors (Angiulli and Pizzuti, 2005; Angiulli et al., 2006, 2013). Distance-based clustering is a typical distance-based approach of classifying data into different groups based on distances of each data to the data groups with the data that cannot be classified into either group deemed as outliers.

In addition to above two general groups of approaches, there have been many other approaches of detecting outliers depending on the specific scenarios and applications. In this section, rather than provide a comprehensive review on all these outlier detection methods, only a brief overview is provided as above to supply necessary background information on outlier detection. Interested readers are referred to Hodge and Austin (2004) and Gupta et al. (2013) for more detailed information on outlier and outlier detection.

Please cite this article in press as: Guo, J., et al. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. Transport. Res. Part C (2014), http://dx.doi.org/10.1016/j.trc.2014.07.005

2.2. Outlier detection in time series process

Time series analysis is an extensively investigated field in many disciplines and outlier detections has been investigated widely for the time series process. Tsay (1988) discussed four types of outliers defined as: Innovational Outlier (IO), Additive Outlier (AO), Level Shift (LS), and Temporary Change (TC). Tsay actually and more precisely referred to LS and TC occurrences as structural changes. Assuming an outlier occurred at the time index $t = t_1$, the outlier model is as defined in Eq. (1).

$$Y_t = X_t + \omega L(B) I_t(t_1), \tag{1}$$

where

$I_t(t_1)$: the indication function defined as $I_t(t_1) = \begin{cases} 1, t = t_1 \\ 0, t \neq t_1 \end{cases}$;
Y_t: the observed series;
X_t: the background process;
ω: the magnitude of the outlier effect.

For an autoregressive integrated moving average (ARIMA) background process, $L(B)$ is given below for each type of outlier in Eqs. (2)–(5).

$$\text{IO}: \quad L(B) = \frac{\theta(B)}{(1-B)^d \phi(B)}; \tag{2}$$

$$\text{AO}: \quad L(B) = 1; \tag{3}$$

$$\text{LS}: \quad L(B) = \frac{1}{1-B}; \tag{4}$$

$$\text{TC}: \quad L(B) = \frac{1}{1-\delta B}, \ 0 < \delta < 1. \tag{5}$$

where

d: order of the short-term differencing;
B: backshift operator;
$\phi(B)$: short term autoregressive polynomial;
$\theta(B)$: short term moving average polynomial.

Based on above definition, the outlier detection technique based on *intervention analysis* (Box and Tiao, 1975) and *likelihood ratio test* (Fox, 1972) has received great attention. Various strategies to implement this technique have been developed in Tsay (1988), Chang et al. (1988), and Chen and Liu (1993).

In the context of short term traffic condition forecasting, Tight et al. (1993) used two approaches based on residuals obtained from fitting the traffic flow data, i.e., the conventional large residual approach (e.g., greater than 3 standard deviations) and the likelihood ratio statistics approach using weighted average of residuals at and after the point of interest (Tsay, 1988). In addition, Williams (1999) argued that the outliers in the traffic condition series can be exclusively modeled as additive outliers that impose an immediate effect on current observation without affecting ensuing observations, and demonstrated that IOs would create a permanent periodic effect in a seasonal autoregressive integrated moving average (SARIMA) process used to fit the traffic series. Based on above two arguments, a batch *select many* algorithm was developed in Williams (1999) with multiple outlier detection ability. Moreover, Watson et al. (1995) proposed a different approach based on influence statistics, constructed from squared elements of the influence function matrix with each element describing the impact of a pair of observations on the theoretical autocorrelation function at the time lag.

2.3. Traffic conditional variance modeling and prediction

Discussed previously, conditional variance modeling and prediction is important for outlier detection in traffic condition series. Unlike traffic level forecasting that has been investigated extensively over the decades back to 1970s, conditional variance modeling and prediction is only investigated recently for traffic condition series. As shown in Guo et al. (2012), traffic condition series is heteroscedastic in nature, and statistical models have been adopted from the field of financial time series analysis to model this changing conditional variance phenomenon, including primarily the generalized autoregressive conditional heteroscedasticity (GARCH) model (Guo, 2005; Kamarianakis et al., 2005; Tsekeris and Stathopoulos, 2006; Guo et al., 2008; Sohn and Kim, 2009; Karlaftis and Vlahogianni, 2009;Yang et al., 2010; Guo and Williams, 2010), and the stochastic volatility model (Tsekeris and Stathopoulos, 2010). Empirical investigations have shown that GARCH model is workable for generating the time-varying traffic conditional variance and hence the standard deviation that will be exploited in this paper for improving the performance of outlier detection in traffic flow series.

2.4. Summary

In summary, outliers represent unusual events, disturbances, changes, or simply measurement errors in a system, and have been investigated extensively over a wide range of disciplines. In the context of traffic condition series, the

Please cite this article in press as: Guo, J., et al. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. Transport. Res. Part C (2014), http://dx.doi.org/10.1016/j.trc.2014.07.005

investigation of outliers is fairly limited with approaches primarily based on the residuals obtained from filtering the original time series. In these approaches, the process variance was usually used to determine the threshold needed for detecting the outlier. However, the process variance utilized in these algorithms was implicitly assumed to be constant, which does not conform to the heteroscedastic nature of traffic condition series as identified in recent studies. In the literature, conditional variance modeling and prediction has been investigated for traffic condition series with promising results, and based on these advancements, the outliers will be detected in this paper through applying a traffic condition forecasting system that is capable of generating jointly the level and time-varying conditional variances of traffic condition series.

3. Methodology

Note that the purpose of this paper is to propose a traffic flow outlier detection approach based on time-varying conditional variance modeling of the traffic flow series. Therefore, a short term traffic flow forecasting system capable of level forecasting and conditional variance forecasting is briefly revisited for the completeness of this paper. Then the proposed outlier detection approach is presented, and an experimental outlier investigation strategy is described to investigate the performance of the proposed detection approach and the effects of outliers on the forecasting system.

3.1. Traffic flow forecasting approach revisited

Short term traffic flow forecasting is an important topic for traffic management and control systems and many approaches have been proposed in the literature. Recently, a seasonal autoregressive integrated moving average plus generalized autoregressive conditional heteroscedasticity (SARIMA + GARCH) structure is receiving increasing attention for traffic flow forecasting, in which the SARIMA part is used to predict the traffic flow levels and the GARCH part is used to predict the conditional variance. For traffic flow series aggregated at 15-min, the structure is further identified as SARIMA$(1,0,1)(0,1,1)_{672}$ + GARCH(1,1) (Williams, 1999; Williams and Hoel, 2003; Guo, 2005). By further manipulating this structure, this structure can be processed in a cascade mode with three components of a seasonal IMA (integrated moving average) filter, a short term Kalman filter, and a GARCH filter (Guo, 2005; Guo et al., 2008).

The seasonal IMA filter and the short term Kalman filter are constructed to predict the traffic flow level. For traffic flow series aggregated at 15-min time interval, the seasonal IMA filter is described as Eq. (6)

$$\hat{X}_t = \alpha X_{t-672} + (1-\alpha)\hat{X}_{t-672}, \tag{6}$$

where

α: smoothing parameter;
X_t: traffic condition variable at time t;
$\hat{X}_t$: predicted value at time t with $\hat{X}_t = X_t$ for $t = 1,2,\ldots,672$.

For the short term Kalman filter, the state transition equation and the observation equation are defined as Eqs. (7) and (8), respectively.

$$w_t = \Phi w_{t-1} + a_t. \tag{7}$$

$$Y_t = \mathrm{U}_t^T w_t + e_t \tag{8}$$

where

$w_t = (\phi\ \theta)^T$: state variable where ϕ and θ is the parameters of the SARIMA structure;
$\Phi = diag\{\lambda^{-\frac{1}{2}}\}$: state transition matrix, with λ defined as a forgetting factor;
a_t: state noise series;
Y_t: current observation;
U_t: time-varying observation matrix;
e_t: observation noise series.

In the above definition, the seasonal IMA filter is to capture the seasonal effect on the level of the traffic condition series. In this paper, the weekly pattern is used with the seasonal order determined as $4 \times 96 = 672$ time intervals for 15-min time interval. The purpose of the short term Kalman filter is to pick up the local effect on the traffic flow levels left after applying the seasonal IMA filter. Through the consecutive processing of the seasonal IMA filter and this short term Kalman filter, the level of the traffic condition series will be captured and removed, with the random noise left in the residual series. Note that the noise series is uncorrelated by definition, while the squared residual series will be correlated, indicating the heteroscedasticity of traffic flow series, and this phenomenon will be modeled and predicted using the GARCH filter.

The purpose of the GARCH filter is to model and predict the time-varying conditional variance of the traffic flow process. For the GARCH filter, the observation equation and state transition equation are defined as Eqs. (9) and (10), respectively.

$$\varepsilon_t^2 = \begin{pmatrix} 1 & \varepsilon_{t-1}^2 & \eta_{t-1} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha \\ \beta \end{pmatrix} + \eta_t, \tag{9}$$

Please cite this article in press as: Guo, J., et al. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. Transport. Res. Part C (2014), http://dx.doi.org/10.1016/j.trc.2014.07.005

$$\begin{pmatrix} \alpha_0 \\ \alpha \\ \beta \end{pmatrix}_t = diag\left\{\lambda^{-\frac{1}{2}}\right\} \begin{pmatrix} \alpha_0 \\ \alpha \\ \beta \end{pmatrix}_{t-1} + z_t \tag{10}$$

where

ε_t: residual series;
η_t: observation noise series;
z_t: state noise series.

Therefore, combined, the seasonal IMA filter, the short term Kalman filter, and the GARCH filter constitute an integrated forecasting system. Since the two state space models can be solved recursively using the well-known Kalman recursion (Kalman, 1960), and the seasonal IMA filter is also recursive in nature, this forecasting system can be implemented in real time. Furthermore, square root transformation is used in this paper when performing the online short term traffic condition forecasting. Interested readers are referred to Guo (2005) and Guo et al. (2008) for more information on this short term traffic flow forecasting system.

3.2. Proposed outlier detection approach

Conventionally, the outlier detection for time series modeling is conducted in an iterative manner such that the effect of the outliers can be precisely identified and removed, as in Tsay (1988) and Box et al. (1994, 2008). However, in the context of online short term traffic condition forecasting where the incoming data point will be processed instantly at its arrival, this iterative procedure is not easy to be incorporated and handled. Therefore, in this paper, the proposed outlier detection strategy is straightforward such that an outlier at time index t is claimed when Eq. (11) holds.

$$\frac{\hat{e}_t}{\hat{\sigma}_t} > c \tag{11}$$

Table 1
Overview of traffic flow data.

Region	Highway	Station	Lane	Start	End	Total Length	Missing	Percent missing (%)
UK	M25	4762a	4	9/1/1996	11/30/1996	8736	300	3.43
UK	M25	4762b	4	9/1/1996	11/30/1996	8736	300	3.43
UK	M25	4822a	4	9/1/1996	11/30/1996	8736	552	6.32
UK	M25	4826a	4	9/1/1996	11/30/1996	8736	483	5.53
UK	M25	4868a	4	9/1/1996	11/30/1996	8736	354	4.05
UK	M25	4868b	4	9/1/1996	11/30/1996	8736	367	4.20
UK	M25	4565a	4	1/1/2002	12/31/2002	35040	1922	5.49
UK	M25	4680b	4	1/1/2002	12/31/2002	35040	1625	4.64
UK	M1	2737a	3	2/13/2002	12/31/2002	30912	1454	4.70
UK	M1	2808b	3	2/13/2002	12/31/2002	30912	1349	4.36
UK	M1	4897a	3	2/13/2002	12/31/2002	30912	2422	**7.84**
UK	M6	6951a	3	1/1/2002	12/31/2002	35040	622	1.78
MD	I270	2a	3	1/1/2004	5/5/2004	12096	622	5.14
MD	I95	4b	4	6/1/2004	11/15/2004	16128	927	5.75
MD	I795	7a	2	1/1/2004	5/5/2004	12096	503	4.16
MD	I795	7b	2	1/1/2004	5/5/2004	12096	503	4.16
MD	I695	9a	4	1/1/2004	5/5/2004	12096	717	5.93
MD	I695	9b	4	1/1/2004	5/5/2004	12091	707	5.85
MN	I35W-NB	60	4	1/1/2000	12/31/2000	35136	878	2.50
MN	I35W-SB	578	3	1/1/2000	12/31/2000	35136	945	2.69
MN	I35E-NB	882	3	1/1/2000	12/31/2000	35136	1899	5.40
MN	I35E-SB	890	3	1/1/2000	12/31/2000	35136	1734	4.94
MN	169-NB	442	2	1/1/2000	12/31/2000	35136	633	1.80
MN	169-SB	737	2	1/1/2000	12/31/2000	35136	1124	3.20
MN	I35W-NB	60	4	1/1/2004	12/31/2004	35136	249	0.71
MN	I35W-SB	578	3	1/1/2004	12/31/2004	35136	31	**0.09**
MN	I35E-NB	882	3	1/1/2004	12/31/2004	35136	375	1.07
MN	I35E-SB	890	3	1/1/2004	12/31/2004	35136	371	1.06
MN	169-NB	442	2	1/1/2004	12/31/2004	35136	298	0.85
MN	169-SB	737	2	1/1/2004	12/31/2004	35136	35	0.10
WA	I5	ES-179D_MN_Stn	4	1/1/2004	6/29/2004	17298	318	1.84
WA	I5	ES-179D_MS_Stn	3	1/1/2004	6/29/2004	17298	335	1.94
WA	I5	ES-130D_MN_Stn	4	4/1/2004	9/30/2004	17516	168	0.96
WA	I5	ES-130D_MS_Stn	4	4/1/2004	9/30/2004	17516	383	2.19
WA	I405	ES-738D_MN_Stn	3	7/1/2004	12/29/2004	17406	381	2.19
WA	I405	ES-738D_MS_Stn	3	7/1/2004	12/29/2004	17406	380	2.18

Please cite this article in press as: Guo, J., et al. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. Transport. Res. Part C (2014), http://dx.doi.org/10.1016/j.trc.2014.07.005

Table 2
Number of outliers by hour of day for Choice A.

Region	Year	Station	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
UK	1996	4762a	4	1	2	0	0	0	7	61	70	65	51	52	35	34	25	33	34	37	30	45	36	20	6	7
UK	1996	4762b	15	13	4	8	17	24	39	43	60	67	89	91	68	63	51	54	63	68	69	120	130	84	64	35
UK	1996	4822a	5	3	0	0	0	0	8	53	59	45	48	45	28	24	20	26	27	40	33	49	34	11	6	4
UK	1996	4826a	6	3	0	0	0	1	5	56	50	46	49	48	26	27	21	32	28	45	45	55	42	13	5	6
UK	1996	4868a	23	19	6	4	7	12	34	86	96	102	104	104	87	72	53	63	75	94	113	143	129	102	70	39
UK	1996	4868b	11	3	1	0	2	9	31	54	62	80	93	87	49	35	29	27	39	53	70	89	90	51	33	26
UK	2002	4565a	5	3	2	4	3	34	115	121	109	94	80	60	43	45	57	37	42	52	46	38	44	33	22	7
UK	2002	4680b	13	12	5	5	9	13	68	101	114	100	56	42	40	53	32	44	60	90	77	69	68	32	26	17
UK	2002	2737a	5	3	2	1	4	25	64	100	89	108	69	92	88	87	70	76	88	68	87	72	100	92	54	21
UK	2002	2808b	39	16	10	14	24	88	77	69	84	104	87	79	78	97	96	119	132	124	147	131	184	161	98	68
UK	2002	4897a	11	11	8	5	8	29	41	78	77	72	54	96	62	59	69	47	64	53	57	49	70	81	47	18
UK	2002	6951a	70	66	62	63	62	57	82	138	152	97	102	102	97	123	122	125	125	129	144	137	130	119	101	89
MD	2004	2a	0	0	0	0	0	1	6	2	1	3	0	0	1	0	17	5	15	15	14	21	3	0	0	4
MD	2004	4b	15	14	7	7	6	15	13	13	17	20	18	7	2	8	9	14	10	26	11	21	24	36	45	41
MD	2004	7a	2	2	0	0	8	24	53	49	46	27	14	20	10	6	5	9	9	9	10	5	3	0	3	5
MD	2004	7b	2	4	4	0	0	10	30	32	24	12	9	5	1	2	0	4	4	5	8	6	5	1	3	4
MD	2004	9a	22	0	0	1	0	0	3	4	14	11	7	1	1	0	1	4	1	2	4	3	0	0	1	14
MD	2004	9b	6	19	5	4	0	20	56	54	48	48	25	11	4	13	25	27	19	21	23	13	11	12	11	6
MN	2000	60	28	16	13	8	8	30	49	87	109	110	84	57	74	81	72	68	83	93	84	62	33	48	32	38
MN	2000	578	19	0	0	1	11	38	71	97	111	77	64	44	45	47	42	43	33	28	33	30	32	28	23	12
MN	2000	882	7	5	4	4	4	27	55	90	95	35	14	11	3	2	3	12	18	17	17	12	9	12	9	14
MN	2000	890	24	2	4	2	2	14	33	27	26	7	9	7	6	5	10	44	53	72	34	23	31	23	56	50
MN	2000	442	23	7	4	4	6	28	45	59	48	34	21	19	24	27	32	76	82	85	80	43	36	50	25	24
MN	2000	737	6	5	3	1	15	56	107	195	176	122	59	41	32	25	26	33	36	74	48	40	24	24	19	11
MN	2004	60	20	22	19	12	10	30	50	59	70	59	44	28	34	25	30	31	34	50	50	54	24	22	25	41
MN	2004	578	11	10	2	0	2	38	83	113	99	89	31	34	34	30	22	8	10	17	24	20	12	16	28	19
MN	2004	882	5	1	0	0	0	30	68	115	73	30	3	2	2	1	8	8	8	11	10	8	8	7	9	8
MN	2004	890	11	2	0	0	0	7	20	27	21	2	1	4	3	7	17	32	44	54	39	9	15	33	47	58
MN	2004	442	25	18	16	14	10	30	46	52	52	40	25	22	17	19	33	52	61	61	68	37	22	22	34	33
MN	2004	737	9	6	4	4	8	36	68	146	145	92	48	40	33	22	23	23	34	61	33	32	22	21	19	19
WA	2004	ES_179D_MN_Stn	37	37	36	36	14	14	15	16	15	13	12	12	9	9	16	26	19	17	17	19	9	9	15	14
WA	2004	ES_179D_MS_Stn	10	8	8	2	6	16	35	80	67	43	10	11	7	6	4	4	1	3	0	2	6	0	0	4
WA	2004	ES_130D_MN_Stn	301	221	195	156	151	187	232	278	301	286	268	287	291	279	252	277	403	397	325	282	227	327	415	372
WA	2004	ES_130D_MS_Stn	21	10	5	7	13	38	51	87	84	79	43	60	64	76	57	98	105	91	82	53	32	31	31	18
WA	2004	ES_738D_MN_Stn	224	159	105	94	119	160	220	239	217	207	193	193	203	200	239	277	320	321	281	242	212	205	239	248
WA	2004	ES_738D_MS_Stn	16	3	0	3	24	53	75	87	101	59	30	16	15	16	10	14	19	20	15	2	6	8	12	14

Where

$\hat{e}_t$: traffic flow residual series at time index t after applying the seasonal IMA filter and the short term Kalman filter;
$\hat{\sigma}_t$: conditional standard error of the traffic flow series at time index t computed using the GARCH filter;
c: prescribed critical value.

Note that in this strategy, first, the conditional variance generated from the GARCH filter is time-varying as compared with the constant process variance that is commonly used in outlier detection. Second, the critical value c in this paper is set to 2.5, which is relatively low compared with the conventional value of 3 so as to provide more power of detecting outliers. Considering the outliers indicates potentially extraordinary patterns occurring in traffic streams, this added outlier detection power will help to identify and save these patterns for further investigation.

3.3. Experimental outlier investigation

In addition to outlier detection, a comparative experimental investigation is designed to show the effects of the outliers on the forecasting system. In the experiment, when an outlier is claimed at a certain time index, two possible choices regarding the treatment of the detected outlier are designed, namely, Choice A and B. For Choice A, the outlier will not be used to update the forecasting system, and for Choice B, the outlier will be used to update the forecasting system. In this study, these two choices are implemented to show the outlier detection performances under these two choices; in addition, effects of the outliers on the forecasting system are also studies through investigating the impact of the outliers on the forecasting system parameters. Note that the outlier detection critical value c is set to the same for the two choices.

4. Empirical study

In this section, the real world data used in this study is described and the empirical results are presented, including the outlier detection performances and the effects of outliers on the forecasting system structure.

Table 3
Level of outlier clustering for Choice A.

Region	Year	Station	G1	G2	G3	G4	G5	G6	G7	G8	G9	Outlier percentage (%)
UK	1996	4762a	225	55	29	10	8	8	2	1	7	9.75
UK	1996	4762b	410	125	36	18	22	15	6	3	19	19.93
UK	1996	4822a	204	50	21	9	10	4	2	4	3	8.45
UK	1996	4826a	223	53	17	15	7	4	5	3	4	9.06
UK	1996	4868a	416	126	61	27	15	19	6	8	30	24.36
UK	1996	4868b	298	102	38	18	12	7	5	5	14	15.24
UK	2002	4565a	216	52	27	9	5	7	4	7	36	3.32
UK	2002	4680b	183	66	29	19	7	2	8	4	37	3.47
UK	2002	2737a	426	129	55	29	13	6	4	5	25	5.07
UK	2002	2808b	559	181	72	32	17	11	8	5	36	7.36
UK	2002	4897a	323	85	53	14	14	8	4	1	24	4.04
UK	2002	6951a	227	77	37	23	14	8	8	8	42	7.55
MD	2004	2a	26	11	7	1	1	2	0	0	2	1.07
MD	2004	4b	115	36	13	4	3	5	2	2	6	2.83
MD	2004	7a	108	14	12	4	1	1	1	0	7	3.16
MD	2004	7b	38	4	2	1	1	0	1	0	8	1.74
MD	2004	9a	24	9	4	4	2	2	0	0	0	0.93
MD	2004	9b	68	24	13	3	3	2	3	3	15	4.77
MN	2000	60	190	76	28	17	19	6	4	3	31	4.13
MN	2000	578	90	38	19	3	6	1	3	4	26	2.80
MN	2000	882	90	15	6	6	2	4	2	2	16	1.45
MN	2000	890	110	31	18	2	3	2	1	4	18	1.70
MN	2000	442	152	33	11	11	6	2	4	5	26	2.66
MN	2000	737	124	45	23	20	13	6	9	0	34	3.56
MN	2004	60	95	38	19	6	2	9	2	4	21	2.55
MN	2004	578	132	55	13	7	2	4	3	0	26	2.27
MN	2004	882	66	18	7	1	4	4	2	1	14	1.25
MN	2004	890	96	36	18	11	2	1	1	2	11	1.37
MN	2004	442	106	33	17	11	2	1	3	2	15	2.44
MN	2004	737	126	39	25	13	3	3	5	3	29	2.86
WA	2004	ES_179D_MN_Stn	50	15	5	4	1	1	1	0	12	2.85
WA	2004	ES_179D_MS_Stn	77	12	6	8	4	3	1	1	8	2.18
WA	2004	ES_130D_MN_Stn	1069	373	221	138	86	67	48	29	156	43.29
WA	2004	ES_130D_MS_Stn	419	115	52	27	6	1	7	1	16	7.97
WA	2004	ES_738D_MN_Stn	823	276	138	65	54	24	21	21	121	33.25
WA	2004	ES_738D_MS_Stn	96	29	14	9	5	0	2	4	19	4.02

Please cite this article in press as: Guo, J., et al. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. Transport. Res. Part C (2014), http://dx.doi.org/10.1016/j.trc.2014.07.005

4.1. Data

Real world traffic flow data collected from 36 stations on four highway systems, including the motorway system in the United Kingdom and the metropolitan freeway systems of Maryland, Minnesota, and Washington State in the United States, are used in this study. An overview of the data are shown in Table 1. For these traffic flow data, 15-min interval was selected for aggregating the data according to Edie (1963) in that all the vehicles crossing the detection zone are counted for each 15-min interval. Missing values are imputed based on the SARIMA$(1,0,1)(0,1,1)_{672}$ model using the back forecasting technique within an iterative framework (Box et al., 1994). For each iteration step, an estimate of the SARIMA model and prediction based on the estimated model will be conducted on the normal series, followed by the same estimation and prediction procedure conducted on the reversed series, termed as back forecasting. In the prediction step within each iteration step, missing values will be replaced by the predictions at corresponding time interval, and the iteration will stop when the difference between each pair of iterations will fall within a predefined tolerance level. When performing the analysis, these data are normalized by number of lane to provide a fair comparison. For more information on the data, please refer to Guo (2005).

4.2. Outlier detection performance

In this section, the outlier detection performances are presented in terms of hour of day and level of outlier clustering. For outlier detection performance, the number of detected outliers with respect to each hour of day and each level of outlier clustering are shown in Tables 2 and 3 for Choice A and Tables 4 and 5 for Choice B, respectively. Note that for Tables 2 and 4, the cells in each table are highlighted according to the number of outliers detected for the hour shown in the headline of the tables, namely, no highlight for the number of outliers between 0 and 4, light gray for the number of outliers between 5 and 15, and deep gray for the number of outliers greater than 15. For Tables 3 and 5, nine groups are used to represent the level of outlier clustering, i.e., G1 indicates a single outlier, G2 indicates the number of cases when two outliers are detected consecutively (two-outlier cluster), G3 indicates the number of cases when three outliers are detected consecutively (three-outlier cluster), etc.

From the outlier detection results by hour of day for both choices (Tables 2 and 4), it can be seen that both choices yield more outliers for hours corresponding to high traffic levels than those corresponding to low traffic levels, which is normal since intuitively traffic pattern is more changeable when traffic is operating at high levels and hence creating more

Table 4
Number of outliers by hour of day for Choice B.

Region	Year	Station	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
UK	1996	4762a	0	0	1	0	0	0	1	14	22	27	11	4	2	1	4	4	3	5	2	7	0	0	0	0
UK	1996	4762b	0	1	0	0	0	0	0	1	3	4	7	3	7	7	4	5	6	12	8	13	5	2	2	0
UK	1996	4822a	1	0	0	0	0	0	0	14	13	11	14	4	2	1	2	3	4	7	7	10	0	0	0	0
UK	1996	4826a	1	0	0	0	0	0	0	15	7	13	11	5	3	1	4	2	4	11	10	8	3	0	0	0
UK	1996	4868a	0	0	0	0	0	0	2	7	10	10	5	4	2	2	1	2	14	18	12	9	1	4	0	0
UK	1996	4868b	0	1	0	0	0	0	0	4	7	7	6	5	0	7	4	3	4	7	11	17	6	1	0	2
UK	2002	4565a	2	1	2	5	2	7	32	17	23	14	14	17	17	14	11	5	8	15	11	5	5	2	4	2
UK	2002	4680b	1	0	0	1	0	1	17	13	15	17	17	15	19	16	18	16	22	41	31	20	10	2	3	0
UK	2002	2737a	0	1	0	0	0	1	15	18	12	16	12	28	22	12	16	17	18	10	19	11	19	20	10	3
UK	2002	2808b	6	0	0	1	4	34	11	5	1	8	12	15	12	19	14	19	31	21	28	35	50	32	28	21
UK	2002	4897a	3	2	0	0	0	1	10	14	9	10	12	17	23	15	20	13	14	8	11	8	15	18	14	6
UK	2002	6951a	4	3	0	0	1	1	8	21	18	12	22	21	25	37	24	19	23	28	45	24	25	17	7	10
MD	2004	2a	0	0	0	0	0	1	2	2	1	2	0	0	1	0	14	1	5	4	6	8	1	0	0	2
MD	2004	4b	5	2	0	1	1	8	9	7	5	4	6	1	0	0	5	3	6	10	1	6	4	9	8	11
MD	2004	7a	0	0	0	0	0	2	9	6	3	0	0	3	0	0	0	0	0	0	0	1	0	0	1	0
MD	2004	7b	1	0	0	0	0	2	3	3	3	0	1	1	1	0	0	0	3	3	5	1	0	0	4	2
MD	2004	9a	7	4	0	0	0	0	2	0	7	5	2	0	2	0	1	1	1	2	2	1	0	0	1	3
MD	2004	9b	2	7	0	2	1	0	2	2	10	3	1	0	1	8	0	1	1	4	1	0	0	0	1	0
MN	2000	60	3	0	0	0	1	15	22	16	7	23	10	11	4	11	10	18	17	16	9	9	3	7	7	8
MN	2000	578	12	0	0	1	4	30	35	27	24	23	6	10	1	1	0	5	1	3	6	4	0	2	3	6
MN	2000	882	2	0	0	0	4	19	28	39	28	7	4	3	1	0	2	4	3	3	6	2	0	4	4	5
MN	2000	890	10	0	2	0	1	6	13	10	3	1	7	4	4	5	3	22	31	30	8	6	12	5	30	22
MN	2000	442	8	0	0	0	2	15	17	19	4	2	1	1	5	5	10	27	34	23	27	8	1	8	8	6
MN	2000	737	2	1	0	0	3	18	31	47	45	36	3	3	3	0	4	5	6	18	10	1	1	2	5	3
MN	2004	60	2	2	0	0	1	11	14	10	11	13	10	10	8	7	13	10	9	16	18	17	3	4	9	8
MN	2004	578	2	1	0	0	0	13	18	24	21	28	10	9	5	7	6	3	6	5	5	5	3	4	11	4
MN	2004	882	0	0	0	0	0	9	26	20	24	6	2	2	1	0	1	1	1	1	1	2	2	2	3	3
MN	2004	890	3	0	0	0	0	2	7	6	2	1	0	2	1	2	8	12	13	12	13	3	10	21	26	27
MN	2004	442	1	0	0	0	0	8	14	6	3	3	2	6	5	3	8	16	19	29	24	2	1	3	7	5
MN	2004	737	0	0	0	0	1	10	16	43	33	21	7	8	2	4	6	2	7	24	8	3	2	6	2	3
WA	2004	ES_179D_MN_Stn	15	1	0	1	0	1	2	3	1	2	4	1	1	1	6	14	7	6	6	5	0	3	7	3
WA	2004	ES_179D_MS_Stn	5	0	0	0	1	5	14	30	11	12	1	1	2	0	1	2	0	2	0	2	0	0	0	0
WA	2004	ES_130D_MN_Stn	14	1	1	2	0	5	11	14	8	3	9	5	8	10	6	7	25	18	9	5	1	1	15	10
WA	2004	ES_130D_MS_Stn	11	1	0	0	3	13	14	21	7	7	5	10	14	12	19	32	26	21	17	7	8	5	4	2
WA	2004	ES_738D_MN_Stn	9	0	0	1	0	6	7	1	0	1	0	2	4	4	6	12	14	11	10	7	0	1	5	2
WA	2004	ES_738D_MS_Stn	3	0	0	2	8	14	14	11	10	11	7	2	2	6	4	2	5	0	0	0	0	1	0	2

Table 5
Level of outlier clustering for Choice B.

Region	Year	Station	G1	G2	G3	G4	G5	G6	G7	G8	G9	Outlier percentage (%)
UK	1996	4762a	72	18	0	0	0	0	0	0	0	1.61
UK	1996	4762b	52	5	3	1	0	0	1	1	0	1.34
UK	1996	4822a	68	8	3	0	0	0	0	0	0	1.38
UK	1996	4826a	71	9	3	0	0	0	0	0	0	1.46
UK	1996	4868a	73	5	3	1	0	0	1	0	0	1.53
UK	1996	4868b	56	11	1	0	1	1	0	0	0	1.37
UK	2002	4565a	157	24	10	0	0	0	0	0	0	0.71
UK	2002	4680b	222	29	5	0	0	0	0	0	0	0.89
UK	2002	2737a	200	25	3	3	0	0	0	0	1	0.97
UK	2002	2808b	255	64	5	1	1	0	0	0	0	1.41
UK	2002	4897a	202	16	3	0	0	0	0	0	0	0.84
UK	2002	6951a	284	43	4	2	1	0	0	0	0	1.20
MD	2004	2a	42	1	2	0	0	0	0	0	0	0.50
MD	2004	4b	85	5	0	0	0	0	1	0	1	0.79
MD	2004	7a	16	3	1	0	0	0	0	0	0	0.25
MD	2004	7b	22	4	1	0	0	0	0	0	0	0.33
MD	2004	9a	29	6	0	0	0	0	0	0	0	0.41
MD	2004	9b	45	1	0	0	0	0	0	0	0	0.47
MN	2000	60	161	19	5	1	0	0	0	0	1	0.69
MN	2000	578	133	11	2	2	1	2	0	1	1	0.62
MN	2000	882	88	11	4	4	1	0	2	0	1	0.51
MN	2000	890	149	18	9	2	3	0	0	0	0	0.71
MN	2000	442	160	19	8	1	1	0	0	0	0	0.70
MN	2000	737	181	26	2	2	0	0	0	0	0	0.75
MN	2004	60	163	16	2	0	1	0	0	0	0	0.62
MN	2004	578	155	15	0	0	1	0	0	0	0	0.57
MN	2004	882	74	5	2	1	0	1	1	0	0	0.32
MN	2004	890	149	9	0	1	0	0	0	0	0	0.52
MN	2004	442	137	11	2	0	0	0	0	0	0	0.50
MN	2004	737	189	8	1	0	0	0	0	0	0	0.63
WA	2004	ES_179D_MN_Stn	67	10	1	0	0	0	0	0	0	0.59
WA	2004	ES_179D_MS_Stn	70	8	1	0	0	0	0	0	0	0.58
WA	2004	ES_130D_MN_Stn	117	14	2	1	3	1	0	0	1	1.21
WA	2004	ES_130D_MS_Stn	184	20	3	0	1	0	0	1	1	1.67
WA	2004	ES_738D_MN_Stn	75	11	2	0	0	0	0	0	0	0.67
WA	2004	ES_738D_MS_Stn	59	14	3	2	0	0	0	0	0	0.68

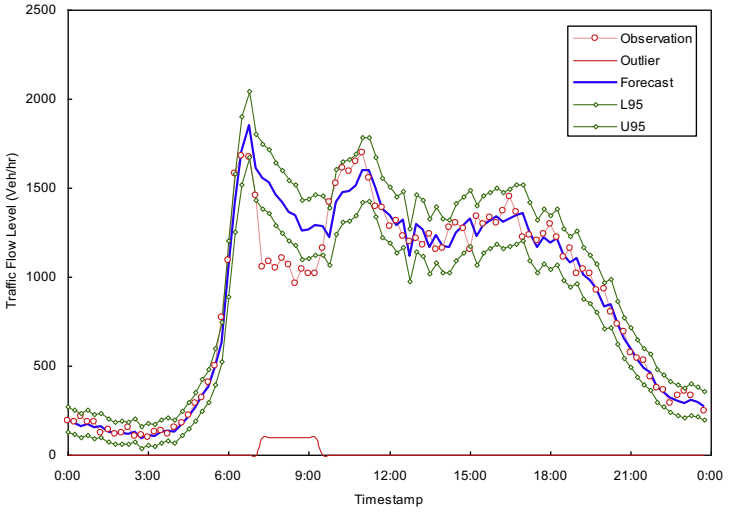

Fig. 1. Outlier detection and treatment effect for Choice A.

extraordinary patterns in the traffic stream data. However, a pronounced difference is that Choice A yields far more outliers than Choice B, which indicates an over-detection for Choice A than Choice B. From the outlier detection results by level of clustering for both choices (Tables 3 and 5), it can be seen that Choice A generates more clustered outliers than Choice B, and the percentage of outliers for Choice A is much higher than the percentage of outliers for Choice B, which is in agreement with the observation from Tables 2 and 4.

Please cite this article in press as: Guo, J., et al. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. Transport. Res. Part C (2014), http://dx.doi.org/10.1016/j.trc.2014.07.005

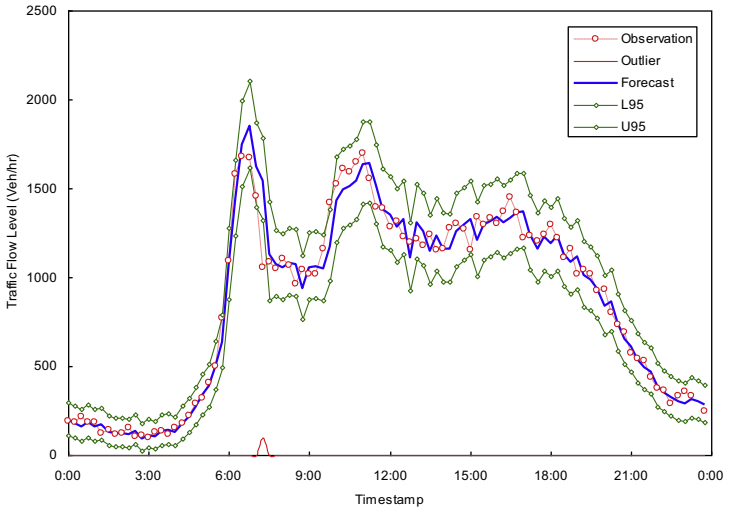

Fig. 2. Outlier detection and treatment effect for Choice B.

The excessive number of outliers detected in Choice A indicates a potential over detection. On reflection, a first thought on the reason of this over detection for Choice A could be the result of the relative small critical value of 2.5 used in this paper. However, using the same critical value, Choice B generates number of outliers at an acceptable level, indicating the critical value selected in this paper is not likely the reason for causing the excessive outliers detection in Choice A. In fact, the percentages of outliers detected for Choice B show a workable selection of the critical value. In the following, this over-detection phenomenon of Choice A can be further investigated and shown through investigating the detailed outlier detection performance over a typical day for the two choices.

Using data from station 4565a in UK for year 2002 as the typical example, the detailed outlier detection performances are shown in Figs. 1 and 2 for Choice A and Choice B, respectively. From Fig. 1, Choice A detects an outlier at around 8 AM. However, as defined, Choice A discards the outlier and generates forecasts according to the normal traffic pattern without updating the forecasting system. In this case, the forecast keeps going with the normal traffic pattern while the real traffic is going down with the changed traffic pattern. As a result, a series of outliers is detected until the pattern contained in the forecasting system coincides with the pattern of the real traffic. Obviously, this is dangerous considering the possibility of no coincidence between the two patterns.

In contrast, Choice B responds differently from Choice A. As defined, Choice B detects the same outlier at the same time as Choice A. But unlike Choice A, this outlier is assimilated to update the forecasting system, and the updated forecasting system deviates from its normal path and tracks the changed traffic pattern, generating forecasts according to the new traffic pattern. As a result, only one outlier is detected for Choice B when the traffic pattern change is first encountered around 8 AM. Therefore, Choice A tends to detect outliers with high clustering rate, resulting in the phenomenon of over-detection.

Considering outliers as the extraordinary traffic pattern occurring on the roadway, the above investigation shows that it is desirable to assimilate the outliers into the forecasting system to ensure the system adaptability of detecting and tracking the traffic patterns changes. In other words, the information contained in the outliers should be imparted into the forecasting system, as conducted in Choice B. This finding substantiates the aforementioned statement that different mechanism is essential for outlier: in this specific case, the detected outlier at 8 AM indicates a change of traffic flow data generation mechanism, and this mechanism might be a recurrent phenomenon or non-recurrent phenomenon, either of which should be reported to the traffic management center to further indentify the nature of this changed traffic condition pattern. In the case of traffic accident, this ability added through outlier detection will serve as a workable means of detecting and hence responding to such incident.

4.3. Outlier effects on forecasting system structure

Previous investigation shows that Choice B is preferable for outlier detection with information contained in the detected outliers assimilated to adjust the forecasting system. In this section, the outliers detected through Choice B are further studied by investigating the effects of these outliers on the forecasting system parameter changes. In doing so, first, the change of each parameter of the forecasting system across each time interval was computed and the 95th percentile of the changes is found for each station as a threshold of indicating significant forecasting system structure change; afterwards, using these thresholds and for each detected outlier, if the change of each parameter is greater than the corresponding threshold of this parameter, a connection between this outlier and forecasting system structure change is deemed as established. As a result,

Please cite this article in press as: Guo, J., et al. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. Transport. Res. Part C (2014), http://dx.doi.org/10.1016/j.trc.2014.07.005

Table 6
Connection between system structure changes and outliers.

Region	Year	Station	Short-term Kalman filter (%)		GARCH filter (%)		
			ϕ	θ	α_0	α	β
UK	1996	4762a	57	63	100	6	84
UK	1996	4762b	71	73	100	18	98
UK	1996	4822a	71	77	100	8	85
UK	1996	4826a	76	76	100	7	88
UK	1996	4868a	65	73	100	8	91
UK	1996	4868b	57	65	100	7	100
UK	2002	4565a	66	73	100	18	80
UK	2002	4680b	60	62	100	11	78
UK	2002	2737a	65	72	100	9	51
UK	2002	2808b	62	71	100	11	57
UK	2002	4897a	74	70	100	9	50
UK	2002	6951a	65	64	100	11	60
MD	2004	2a	82	74	100	12	86
MD	2004	4b	72	79	100	5	76
MD	2004	7a	76	84	100	4	100
MD	2004	7b	67	70	100	12	88
MD	2004	9a	76	78	93	12	22
MD	2004	9b	62	72	83	9	100
MN	2000	60	64	62	100	11	56
MN	2000	578	66	65	100	22	30
MN	2000	882	79	70	100	27	77
MN	2000	890	72	70	100	16	51
MN	2000	442	68	60	100	20	56
MN	2000	737	75	62	100	23	70
MN	2004	60	69	60	100	11	31
MN	2004	578	62	61	100	9	32
MN	2004	882	79	70	100	21	49
MN	2004	890	66	64	100	8	29
MN	2004	442	65	55	100	13	24
MN	2004	737	70	56	100	11	33
WA	2004	ES_179D_MN_Stn	53	70	96	7	70
WA	2004	ES_179D_MS_Stn	66	87	89	7	87
WA	2004	ES_130D_MN_Stn	70	74	100	14	73
WA	2004	ES_130D_MS_Stn	58	61	100	9	60
WA	2004	ES_738D_MN_Stn	74	64	100	16	83
WA	2004	ES_738D_MS_Stn	80	70	100	18	85

for each parameter, a percentage can be computed by dividing the total connections by the total number of outliers. As can be expected, the higher the percentage, the more pronounced the effects of the outliers on the forecasting system structure.

The result of outliers on forecasting system parameters is listed in Table 6. From Table 6, it can be seen that outliers do have a high chance of connecting to the forecasting system structure changes. On reflection, this is not surprising since the forecasting systems should adjust its structure to accommodate the extraordinary patterns indicated by outliers, and the significant traffic flow pattern changes indicated by the outliers will in turn drive the parameters of the short term forecasting system changing significantly to adapt to the traffic flow pattern change. This further supports the argument on assimilating outliers into the system structure

5. Conclusions and discussions

Outlier is universal for many data analysis related applications and many investigations have been proposed in related fields. However, the investigation on the outliers are fairly limited for transportation field where a vast volume of traffic condition data has been collected from traffic monitoring devices such as inductive loop detectors, video based traffic detection detectors, etc. that are installed in many roadway systems. Outliers in the traffic condition series, regarded in general as discordant observations, usually represent uncommon events occurring in the roadway systems, such as traffic incident, weather change, or other disturbances such as device malfunctions, and outlier detection and investigation will help to unravel the mechanism of such underlying reasons.

In this paper an online outlier detection method was proposed and investigated. This outlier detection method is based on an online short term traffic flow forecasting algorithm that has the ability of jointly generating a traffic flow series level prediction and a time-varying estimate of the series conditional variance. Using real world data collected from 36 stations from four regions in both the United States and the United Kingdom, it was found that outliers can be effectively detected using the proposed detection strategy, and through a comparative experimental investigation, it was shown that the information contained in the outliers should be assimilated into the forecasting system so as to enhance the ability of the system to adapt

Please cite this article in press as: Guo, J., et al. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. Transport. Res. Part C (2014), http://dx.doi.org/10.1016/j.trc.2014.07.005

to the changing traffic patterns. The effects of the outliers on the online forecasting system structure were also investigated to show that assimilation of the outliers does significantly alter the parameters of the forecasting system.

The analyses in this paper lead to the following general conclusions. First, it is clear that utilization of the predicted time-varying conditional variance is the unique feature that differentiates the proposed traffic flow outlier detection approach from previous outlier detection approaches. In the conventional approaches, due to the lack of investigation into the second order conditional moment of traffic flow series, time invariant conditional variance has been commonly assumed and exploited for conducting outlier detection. This assumption of time invariant conditional variance renders it impossible for these conventional outlier detection approaches to adapt to the temporally changing traffic patterns, resulting in both missed outliers and false detections. In this sense, considering that the modeling of the second order conditional moment of traffic flow series is still at its early stage, the development of outlier detection approaches will naturally be affected by the advancement of the second order conditional moment modeling.

Second, it should be noted that although the strategy of assimilating the outliers is recommended for performing the short term traffic condition forecasting (Choice B), this it does not mean that ignoring the detected outliers (Choice A) provides no information or is of no use for other applications. In fact, if each cluster identified through Choice A is treated as a single temporal deviation, such clusters could be investigated to determine whether or not they represent a specific induced temporary traffic pattern change that could be mitigated by traffic management strategies. Careful investigation of such outlier clusters would be necessary because outlier clusters detected through Choice A might indicate a permanent pattern changes or system measurement errors rather than temporary induced pattern changes.

Finally, considering that outliers and outlier investigation in traffic condition series are still at an early stage, future work is recommended as follows. First, it is recommended that the underlying generating mechanism be investigated to deepen understanding of the properties of outliers in traffic condition series. In doing so, the traffic condition data collection systems should incorporate an ability of registering detailed information on the transient series altering events, such as vehicle crashes, adverse weather, etc. Detailed time-based archiving of such information would enable the detected outliers to be matched with these events allowing in depth study of how the events generate series outliers. In addition, as discussed in the Choice A versus Choice B discussion above, further research is needed on tailoring specific outlier detection and treatment strategies to specific transportation management and control applications.

Acknowledgments

The authors would like to thank the Minnesota Department of Transportation, the Washington State Department of Transportation, the Maryland Department of Transportation, and the United Kingdoms Highways Agency for providing the data used in this study. This research is supported by the Natural Science Foundation of China under Grant No. 71101025, the National Key Technology R&D Program under Grant No. 2011BAK21B01, and the Fundamental Research Funds for the Central Universities.

References

Albanese, A., Pal, S.K., Petrosino, A., 2014. Rough sets, kernel set, and spatiotemporal outlier detection. IEEE Trans. Knowl. Data Eng. 26 (1), 194–207.
Angiulli, F., Pizzuti, C., 2005. Outlier mining in large high-dimensional data sets. IEEE Trans. Knowl. Data Eng. 17 (2), 203–215.
Angiulli, F., Basta, S., Pizzuti, C., 2006. Distance-based detection and prediction of outliers. IEEE Trans. Knowl. Data Eng. 18 (2), 145–160.
Angiulli, F., Basta, S., Lodi, S., Sartori, C., 2013. Distributed strategies for mining outliers in large data sets. IEEE Trans. Knowl. Data Eng. 25 (7), 1520–1532.
Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis: Forecasting and Control, third ed. Prentice-Hall Inc., Upper Saddle River, NJ 07458.
Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 2008. Time Series Analysis: Forecasting and Control, fourth ed. John Wiley & Sons Inc, Hoboken, New Jersey.
Box, G.E.P., Tiao, G.C., 1975. Intervention analysis with applications to economic and environmental problems. J. Am. Stat. Assoc. 70 (349), 70–90.
Chang, I., Tiao, G.C., Chen, C., 1988. Estimation of time series parameters in the presence of outliers. Technometrics 30, 193–204.
Chatfield, C., 1993. Calculating interval forecasts. J. Business Econ. Stat. 11 (2), 121–135.
Chen, C., Liu, L., 1993. Joint estimation of model parameters and outlier effects in time series. J. Am. Stat. Assoc. 88, 284–297.
Edie, L.C., 1963. Discussion on traffic stream measurements and definitions. In: Proceedings of Second International Symposium on the Theory of Traffic Flow. OECD, Paris, France, pp. 139–154.
Fox, A.J., 1972. Outliers in time series. J. Roy. Stat. Soc.: Ser. B (Methodol.) 34 (3), 350–363.
Guo, J., 2005. Adaptive Estimation and Prediction of Univariate Traffic Condition Series (doctoral dissertation). North Carolina State University, Raleigh, NC.
Guo, J., Huang, W., Williams, B.M., 2012. Integrated heteroscedasticity test for vehicular traffic condition series. ASCE J. Transp. Eng. 138 (9), 1161–1170.
Guo, J., Williams, B.M., 2010. Real time short term traffic speed level forecasting and uncertainty quantification using layered Kalman filters. Transportation Research Record, vol. 2175. Transportation Research Board of the National Academies, Washington, DC, pp. 28–37.
Guo, J., Williams, B.M., Smith, B.L., 2008. Data collection time intervals for stochastic short-term traffic flow forecasting. Transp. Res. Rec. 2024, 18–26.
Gupta, M., Gao, J., Aggarwal, C.C., Han, J., 2013. Outlier detection for temporal data: A survey. IEEE Trans. Knowl. Data Eng. 25 (1), <http://doi.ieeecomputersociety.org/10.1109/TKDE.2013.184>.
Hawkins, D., 1980. Identification of Outliers. Chapman and Hall.
Hodge, V.J., Austin, J., 2004. A survey of outlier detection methodologies. Artif. Intell. Rev. 22 (2), 85–126.
Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Eng. D 82, 35–45.
Kamarianakis, Y., Kanas, A., Prastacos, P., 2005. Modeling traffic volatility dynamics in an urban network. Transp. Res. Rec. 1923, 18–27.
Karlaftis, M.G., Vlahogianni, E.I., 2009. Memory properties and fractional integration in transportation time series. Transp. Res. Part C 17, 444–453.
Sohn, K., Kim, D., 2009. Statistical model for forecasting link travel time variability. ASCE J. Transp. Eng. 135 (7), 440–453.
Tight, M.R., Redfern, E.J., Watson, S.M., Clark, S.D., 1993. Outlier detection and missing value estimation in time series traffic count data: Final Report of SERC Project GR/G23180. Institute of Transport Studies, University of Leeds, Working Paper 401.
Tsay, R.S., 1988. Outlier, level shifts, and variance changes in time series. J. Forecasting 7, 1–20.
Tsekeris, T., Stathopoulos, A., 2006. Real-time traffic volatility forecasting in urban arterial networks. Transp. Res. Rec. 1964, 146–156.

Please cite this article in press as: Guo, J., et al. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. Transport. Res. Part C (2014), http://dx.doi.org/10.1016/j.trc.2014.07.005

Tsekeris, T., Stathopoulos, A., 2010. Short-term prediction of urban traffic variability: stochastic volatility modeling approach. ASCE J. Transp. Eng. 136 (7), 606–613.
Watson, S.M., Redfern, E., Clark, S., Tight, M., Davies, N., 1995. An influence method for outlier detection applied to time series traffic data. J. Appl. Stat. 22 (1), 135–149.
Williams, B.M., 1999. Modeling and forecasting vehicular traffic flow as a seasonal stochastic time series process (PhD dissertation). University of Virginia.
Williams, B.M., Hoel, L., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA: theoretical basis and empirical results. ASCE J. Transp. Eng. 129 (6), 664–672.
Yang, M., Liu, Y., You, Z., 2010. The reliability of travel time forecasting. IEEE Trans. Intell. Transp. Syst. 11 (1), 162–171.

Please cite this article in press as: Guo, J., et al. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. Transport. Res. Part C (2014), http://dx.doi.org/10.1016/j.trc.2014.07.005