

# Detection of Statistically Significant Bus Delay Aggregation by Spatial-Temporal Scanning

Xia Wu<sup>1</sup>, Lei Duan<sup>1,2(✉)</sup>, Tinghai Pang<sup>1</sup>, and Jyrki Nummenmaa<sup>3,4</sup>

<sup>1</sup> School of Computer Science, Sichuan University, Chengdu, China  
2013wuxia@gmail.com, pthaike@gmail.com

<sup>2</sup> West China School of Public Health, Sichuan University, Chengdu, China  
leiduan@scu.edu.cn

<sup>3</sup> School of Information Sciences, University of Tampere, Tampere, Finland  
jyrki.nummenmaa@uta.fi

<sup>4</sup> Sino-Finnish Centre, Tongji University, Shanghai, China

**Abstract.** Public bus service plays an indispensable role in modern urban traffic system. With the bus running data, the detection of the statistically significant aggregations of bus delay is useful for optimizing the bus timetable, so that the service quality can be improved. However, previous studies have not considered how to detect bus delay aggregation using statistical hypothesis testing. To fill that gap, this paper considers the detection of bus delay aggregation from bus running data. We present RSTV-Miner, a mining method using statistical hypothesis testing, for detecting statistically significant bus delay aggregation. Our empirical study on real data demonstrates that RSTV-Miner is effective and efficient.

**Keywords:** Bus delay aggregation · Spatial-temporal analysis · Traffic data mining

## 1 Introduction

Public bus, which is one of the most widely used transportation tools for most people, plays an important role in the modern urban traffic system. The on-time performance is a critical factor affecting people's willingness to take buses. Naturally, people prefer to take the buses that have high on-time rate, i.e., the buses run in accordance with the timetable, since people can avoid unnecessary waiting time and estimate the exact time to the destination. Intuitively, the bus drivers can avoid the situation of ahead of schedule by slowing down. However, there are many causes of bus delay [1], for the case of behind schedule, the bus drivers cannot simply catch up time by speeding up due to some limitations such as maximum speed limit and road condition. Thus, the detection of bus delay

---

This work was supported in part by NSFC 61572332, the Fundamental Research Funds for the Central Universities 2016SCU04A22, and China Postdoctoral Science Foundation 2014M552371, 2016T90850.

**Table 1.** Samples of Tampere bus data

Stop code	Latitude	Longitude	Delay	Arrival	Line	Date	Stop name
1500	61.49855	23.73550	60	00:31:31	4Y	2015-9-27	Ammattikoulu
3084	61.47992	23.80464	15	19:21:52	102	2015-10-23	Kuoppamaentie
0098	61.50003	23.73738	-34	00:09:48	3Y	2015-8-27	Savilinna

aggregations, which is the basis for optimizing the bus timetable, can improve the bus service quality.

Recently, the city of Tampere, Finland, released as open data [2] the locations of its buses at every second. In particular, the bus data includes information on, for each bus and each bus stop, whether a bus was on schedule and how much was the delay if it was not on time. Table 1 lists several samples of Tampere bus data. Each record contains the information of a bus arriving at a bus stop. For example, by the first record in Table 1, we can see that a bus of Line 4Y arrives at Stop Ammattikoulu (code: 1500, location: 61.49855°N, 23.73550°E) at 00:31:31 AM on September 27, 2015 with 60s behind schedule. As shown in the last record in Table 1, the value of “Delay” is minus if the bus arrives ahead of schedule, there are also other details not mentioned.

Motivated by the real-world requirement shown above, we try to detect the aggregation of bus delay from the bus running data in this work. Intuitively, it is hard to avoid all bus delays, since the real-world is complex and uncertain. Instead of finding all frequent bus delay cases, we focus on detecting the statistically significant aggregations of delay. Specifically, we apply the spatial-temporal scanning method, which has been verified effective in the early warning of infection diseases outbreak, to the bus running monitoring data, and test the significance level of each aggregation of bus delay in both temporal dimension and spatial dimension. To the best of our knowledge, there is no previous work on detecting the bus delay aggregation using statistical hypothesis testing. We will review the related work systematically in Sect. 3.

To tackle the detection of bus delay aggregation, we need to address two technical challenges. First, how to perform spatial-temporal scanning on the bus running data efficiently. Second, how to perform statistical hypothesis testing for a pair of a given zone and a time interval.

The main contributions of this paper include: (1) introducing a novel data mining problem of bus delay aggregation detection; (2) designing an efficient and effective algorithm for detecting statistically significant bus delay aggregations; (3) conducting extensive experiments using real data to evaluate our proposed algorithm, and demonstrating visually that some interesting results can be found by our algorithm.

The rest of the paper is organized as follows. We formulate the problem of bus delay aggregation mining in Sect. 2, and review related work in Sect. 3. In Sect. 4, we present the critical techniques of our method. We report experiment and case study in Sect. 5, and conclude the paper in Sect. 6.

## 2 Problem Definition

In this section, we give a formal definition for statistically significant bus delay event aggregation. To that end, we will also need to define several concepts concerning spatial-temporal scanning with respect to “time interval” and “zone”.

We start with some preliminaries. We use a series of continuous non-negative integers starting from 0 to denote the time points; we use  $t_{max}$  to denote the maximum time point. Without loss of generality, we assume that the smaller the value, the earlier the time point, and that the interval between any two consecutive time points is a constant.

For brevity, we denote  $[0, t_{max}]$  by  $\tilde{T}$ . A *time interval*  $T$  is a sub-interval of  $\tilde{T}$  (denoted by  $T \sqsubset \tilde{T}$ ) of the form  $T = [T.t_s, T.t_e]$  satisfying  $0 \leq T.t_s < T.t_e \leq t_{max}$ . The *time span* of  $T$ , denoted by  $\|T\|$ , is the number of time points in  $T$ , that is,  $\|T\| = T.t_e - T.t_s + 1$ .

For a given bus, we use  $AT(u)$  and  $ET(u)$  to denote the real arrival time point and the expected arrival time point at Stop  $u$ , respectively. The delay time at Stop  $u$ , denoted by  $\Delta(u)$ , is  $AT(u) - ET(u)$ . Let  $\theta$  be the time threshold. If  $\Delta(u) \geq \theta$ , we say that a *bus delay event* happens at Stop  $u$ .

We denote the *position of bus stop*  $u$  by its geographic coordinates of the form of  $P(u) = (u.lng, u.lat)$ . For two bus stops  $u_1$  and  $u_2$ , the geographical distance between them, denoted by  $Dis(u_1, u_2)$ , is

$$Dis(u_1, u_2) = 2 \times R \times \sin^{-1} \sqrt{(\sin \frac{x}{2})^2 + \cos(u_1.lat \times \frac{\pi}{180}) \times \cos(u_2.lat \times \frac{\pi}{180}) \times (\sin \frac{y}{2})^2} \tag{1}$$

where

$$\begin{aligned} x &= (u_1.lat \times \frac{\pi}{180}) - (u_2.lat \times \frac{\pi}{180}) \\ y &= (u_1.lng \times \frac{\pi}{180}) - (u_2.lng \times \frac{\pi}{180}) \\ R &= 6.37130 \end{aligned}$$

Please note that Eq. 1 is an approximate function to calculate the distance between two objects on the earth.

Let  $\tilde{S}$  be the *study area* that we take into consideration. A *zone*  $S$  is a sub-area of  $\tilde{S}$  (denoted by  $S \sqsubset \tilde{S}$ ), such that at least one bus stop locates within  $S$ . As all observations happen at bus stops, interchangeably,  $S$  will be used to denote a set of bus stops within the zone.

Given zone  $S$  and time interval  $T$ , we define  $\mathcal{N}(S, T)$  to be the number of bus arrival events happened, and  $\mathcal{D}(S, T)$  to be the number of bus delay events observed. Intuitively,  $\mathcal{D}(S, T) \ll \mathcal{N}(S, T)$ .

Without loss of generality, we apply log-likelihood ratio statistic to measure the significance of the spatial-temporal aggregation of bus delay events. Specifically, given zone  $S$  and time interval  $T$ , the likelihood of the happening of bus delay aggregation within  $S$  during  $T$ , denoted by  $\mathcal{L}(S, T)$ , is

$$\mathcal{L}(S, T) = \begin{cases} \mathcal{D}(S, T) \times \log(r_1) + (\mathcal{D}(\tilde{S}, \tilde{T}) - \mathcal{D}(S, T)) \times \log(r_2) \\ -\mathcal{D}(\tilde{S}, \tilde{T}) \times \log \frac{\mathcal{D}(\tilde{S}, \tilde{T})}{\mathcal{N}(\tilde{S}, \tilde{T})} \\ 0 \end{cases}, \begin{matrix} r_1 > r_2 \\ r_1 \leq r_2 \end{matrix} \quad (2)$$

where,

$$r_1 = \frac{\mathcal{D}(S, T)}{\mathcal{N}(S, T)}$$

$$r_2 = \frac{\mathcal{D}(\tilde{S}, \tilde{T}) - \mathcal{D}(S, T)}{\mathcal{N}(\tilde{S}, \tilde{T}) - \mathcal{N}(S, T)}$$

Given a set of bus running records within area  $\tilde{S}$  during time interval  $\tilde{T}$ , the problem of detecting statistically significant bus delay aggregation is to find the pair of  $(S, T)$ , such that  $\mathcal{L}(S, T) = \max \{\mathcal{L}(S', T') \mid S' \subset \tilde{S}, T' \subset \tilde{T}\}$ , and  $\mathcal{L}(S, T)$  is statistically significant.

Table 2 lists the frequently used notations of this paper.

**Table 2.** Summary of notations

Notation	Description
$\theta$	Time threshold
$\tilde{T}$	The maximum time interval
$\tilde{S}$	The study area
$(S, T)$	A pair of a zone and a time interval
$\mathcal{D}(S, T)$	The number of bus delay events happened within $S$ during $T$
$\mathcal{N}(S, T)$	The number of bus arrival events happened within $S$ during $T$
$\mathcal{L}(S, T)$	The likelihood of bus delay aggregation happened within $S$ during $T$

### 3 Related Work

Urban traffic analysis is an important and valuable problem in daily life. Some applications have been applied to help government or company to improve the public transportation service and provide a more convenient traffic experience. For example, bus travel time prediction [3–6] can assist people to get a better schedule when they go out; bus delays prediction [7] can improve the traffic operation. According to our review, there do not exist any works about bus delay aggregation detection.

Aggregation detection is widely studied in disease outbreak monitoring and has attracted extensive attentions from both research and industry. There are large number of studies about the spatial aggregation [8] and temporal-spatial aggregation [9] on the disease detection such as cancer [10], diabetes [11], and sclerosis [12], which is an important method for us to control and prevent the

---

**Algorithm 1.** Framework of RSTV-Miner

---

**Input:** bus data  $D$ , study spatial-temporal area  $(\tilde{T}, \tilde{S})$ , delay threshold  $\theta$ , maximum size of zone  $\alpha$ , Monte Carlo times  $M$ .  
**Output:** a spatial-temporal area  $(S', T')$  with maximum log-likelihood  $\mathcal{L}(S', T')_{max}$ .  
1:  $index\_Rtree(\alpha, \theta, D)$ ; //refer to Section 4.1  
2:  $\{(S, T)\}_{real} \leftarrow ST\_Scan(\tilde{S}, \tilde{T})$ ; //refer to Algorithm 2  
3: **for** each pair  $(S, T)$  in  $\{(S, T)\}_{real}$  **do**  
4:    $\{(S, T)\}_{monte} \leftarrow MonteCarol((S, T), M)$ ; //refer to Section 4.3  
5: **end for**  
6:  $(S', T') \leftarrow \{(S, T)\}_{monte}$  ; //get a zone has maximum log-likelihood  
7: **return**  $(S', T')$ ;

---

disease outbreaks. Especially, more and more applications are adapting cluster detection to deal with some problems, such as terrorism outbreaks [13], air pollution [14], and crime location [15].

Some typical methods are used to deal with aggregation detection problems. Scan statistics are effective tools in these methods [16, 17]. Some works focus on detecting the change in the spatial pattern of a disease [18]. Wallenstein [19] proposed the method to cluster in time by scan statistic. Kulldorff *et al.* [20, 21] proposed a temporal-spatial scanning statistic method to find a time periodic geographical disease.

## 4 Design of RSTV-Miner

In this section, we present our method, RSVT-Miner, for mining bus delay aggregation from bus dataset. In general, the framework of RSTV-Miner includes: R-tree index structure, spatial-temporal scan, and statistical test. Technically, the key issues of RSTV-Miner are generation and effective scan of index. Algorithm 1 presents the procedure of our method.

### 4.1 R-Tree Index Structure and Representation

In this study, we partition a spatial study area into  $m \times n$  2-dimensional grids. Each grid covering a set of bus stops within a zone and we get a set of zones  $\tilde{S} = \{S_1, S_2, \dots, S_i\}$ . We generate R-tree index using these zones. Then we use these zones as a minimal unit of space in the following study. In this paper, we need to measure the size of a zone when detecting bus delay aggregation. To address this issue, in our method, each zone is a rectangle and the side length of it is equal to the distance between two nearest bus stops.

To detect the aggregation of bus delay, we need to structure an index for zones. At this moment, using each zone to structure the index and searching for its neighbors is a time-consuming process. In addition, the computation cost is very high. This does not allow us to search all sub-areas of  $\tilde{S}$ . To address this issue, we use R-tree to structure index that will help us to group nearby

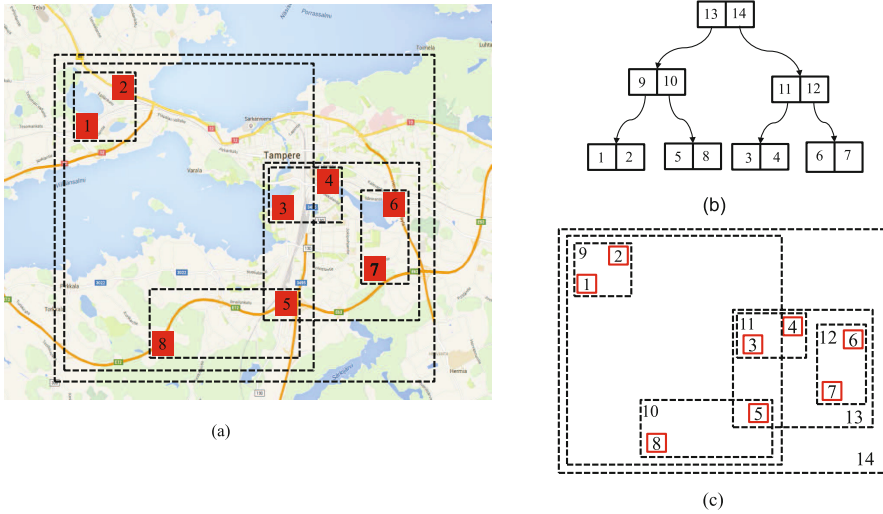


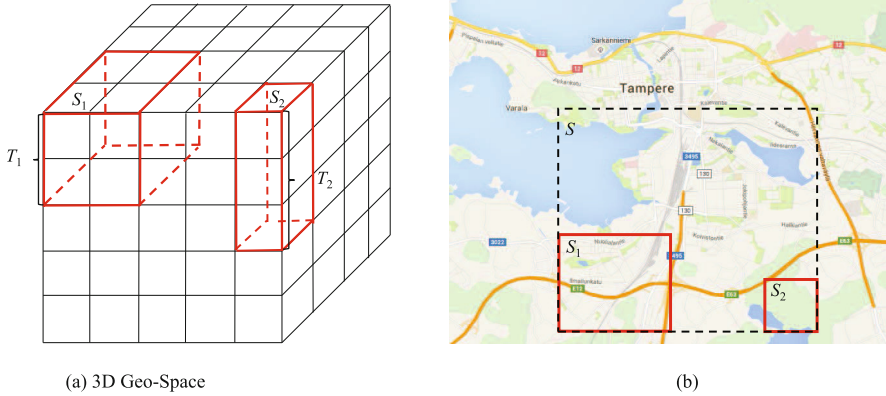
Fig. 1. R-tree index structure demonstrate

zones and retrieve data quickly according to their spatial locations. And to be more statistical significance, when generating an R-tree, we just use zones which overcast at least one bus stop. As demonstrated in Fig. 1, we give an example of R-tree structure in which red rectangles represent the zones overcasting bus stops. We start to build R-tree by using the zones which have partitioned. As illustrate Fig. 1(c), we use R-tree structure, so that a spatial search requires to visit only a small number of nodes.

### 4.2 Spatial-Temporal Scan

Spatial-temporal clustering is a process of grouping objects based on their spatial and temporal similarity. We need a statistical data to express clustering degree for detecting bus delay aggregation. To address this issue, we use log-likelihood ratio to test the spatial and temporal similarity in Tampere. We propose a approach to analyze spatial-temporal data at a higher level of abstraction by grouping the data according to its similarity into meaningful clusters. Given a spatial-temporal study area  $(S, T)$ , we can calculate log-likelihood ratio for a zone  $\mathcal{L}(S, T)$ .

*Example 1.* Given a spatial-temporal study area  $(\tilde{S}, \tilde{T})$ . Figure 2 presents an example to illustrate the problem definition. As shown, we build the spatial-temporal area as the 3D space  $C$ . Given two spatial-temporal area  $(S_1, T_1)$ ,  $(S_2, T_2)$  in  $C$ . To calculate  $\mathcal{L}(S_1, T_1)$  and  $\mathcal{L}(S_2, T_2)$  for spatial-temporal area  $(S_1, T_1), (S_2, T_2)$  by Eq. 2, and if the result is  $\mathcal{L}(S_1, T_1) > \mathcal{L}(S_2, T_2)$ ,  $(S_1, T_1)$  can be considered it has a higher clustering degree than  $(S_2, T_2)$ . If we check all the subspace in  $C$ , we can find the largest outlier space cluster which has a biggest log-likelihood ratio. We can consider it has the bus delay aggregation on this spatial-temporal area.



**Fig. 2.** Illustration of bus delay aggregation

We apply spatial-temporal scan to Tampere bus data by R-tree index, as illustrated in Fig. 1 (b). In this method, we use R-tree to retrieve the spatial dimension. To attach temporal dimension with spatial dimension, we divide temporal area into time slots. There are many combinations of spatial-temporal entries  $(S, T)$ . It is time consuming to retrieve all the combinations, so we just retrieve the spatial data attached consecutive time slots in rush hour at Tampere. We only scan the region indexed by R-tree that side length is no more than maximum pre-set value, so that we never retrieve more than 50% of total testing area at risk. Algorithm 2 gives an implementation of spatial-temporal scanning. We propose this method as following: (1) retrieve R-tree to find every space dimension zone; (2) attach all time slots with each zone, and use log-likelihood to test this spatial-temporal area clustering degree; (3) repeat the two steps for each zone and generate a collection set of spatial-temporal clusters.

---

**Algorithm 2.** Spatial-temporal scanning

---

**Input:** R-tree index for  $\tilde{S}$ , temporal area  $\tilde{T}$

**Output:** spatial-temporal set  $\{S, T\}$  has maximum log-likelihood

- 1: **for** each  $S_i$  in  $\tilde{S}$  **do**
  - 2:     **for** each  $T$  in  $\tilde{T}$  **do**
  - 3:         calculate  $\mathcal{L}(S_i, T)$ ;
  - 4:     **end for**
  - 5:      $(S_i, T) \leftarrow \max(\{\mathcal{L}(S_i, T)\})$ ;
  - 6:      $\{(S, T)\}_{max} \leftarrow addToList((S_i, T))$ ;
  - 7: **end for**
  - 8: **return**  $\{(S, T)\}_{max}$ ;
-

### 4.3 Statistical Test

To ensure our model is statistic significance, we use Monte Carol model to simulate the result. Poisson distribution [22] is widely used to model underlying distribution of spatial-temporal data. We use it to simulate spatial-temporal data of bus delay.

We can get the clustering spatial-temporal area  $(S, T)$  with maximum  $\mathcal{L}(S, T)$  by our spacial-temporal scan method, written as  $\mathcal{L}(S, T)_{real}$ . In order to validate that the result can not be simulated, we must test the significance of the result. For every zone, we calculate the expectation of bus delay. And then we generate the simulated delay of each zone by poisson distribution with delay expectation. Then we use our spacial-temporal scan method to get all the  $\mathcal{L}(S, T)$  values in this simulate data, denoted as  $\mathcal{L}(S, T)_{simu}$ . We conduct this test and get the result  $\{\mathcal{L}(S, T)\}_{simu}$ . If the test is significant at 5 percent level, which means  $\mathcal{L}(S, T)_{real}$  ranks top 5 percent of  $\{\mathcal{L}(S, T)\}_{simu}$ , we can conclude that  $\mathcal{L}(S, T)_{real}$  satisfies the statistic significant.

## 5 Experiments

In this section, we report a systematic empirical study using Tampere bus data set to verify the bus delay aggregation result by RSTV-Miner. All experiments were conducted on a PC computer with an Intel Core i7-3770 3.40 GHz CPU, and 16 GB main memory, running Windows 7 operating system. All algorithms were implemented in C++.

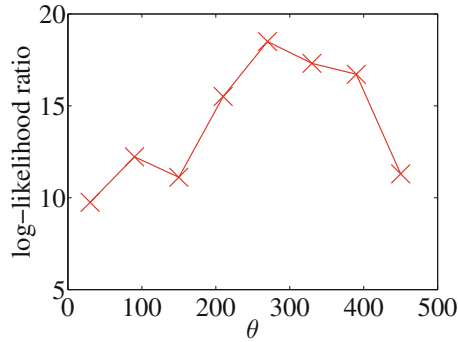
To mine bus delay aggregation of Tampere, we scan every working day of data set. When travel the R-tree, we abandon the grid whose area is more than half of testing space. Using R-tree index to reduce the time of traverse can make an efficient way to spatial-temporal scanning. We use Monte Carol method to validate the output of model with simulation times  $M = 100$ . By comparison, if the real result area larger than the top-5 simulated result, the result is statistic significance.

At the beginning of our experiment, we need to confirm the threshold  $\theta$  which estimates whether a bus delay event happened. In our experiment, we find when we change  $\theta$ , spatial-temporal clustering degree change. To set  $\theta$ , Fig. 3 presents some results on the Tampere map. When  $\theta = 270$ , we can get maximum log-likelihood at the same spatial-temporal testing area.

When traveling R-tree index for each zone, we set a maximum value of side length as  $\alpha$  to avoid travel the zone which is more than half of Tampere at risk. As listed in Table 3, Spatial area changes with  $\alpha$  when we fixed other parameters, study on 14th August, 9:00~13:00. When  $\alpha > 120$ , the result zone area is nearly half of the Tampere area and the finding zone area does not change.

As listed on Table 4, we list results on consecutive days, from 26th August to 28th August, these results have been checked by Monte Carol simulation. The bus delay aggregation of specify spatial-temporal area has a larger degree. To mine the bus delay aggregation, the most clustering spatial-temporal area, listed in Table 4, has the higher log-likelihood compared to other spatial-temporal area





**Fig. 3.** Threshold  $\theta$  variation with log-likelihood ratio test

**Table 3.** Result zone area variation with  $\alpha$

$\alpha$	Side length of zone	Result zone area	$\mathcal{L}$
30	3 km	3.75 km <sup>2</sup>	14.8657
60	6 km	14.28 km <sup>2</sup>	29.0666
90	9 km	62.15 km <sup>2</sup>	11.9916
120	12 km	62.15 km <sup>2</sup>	14.1069
150	15 km	167.50 km <sup>2</sup>	18.3807
180	180 km	167.50 km <sup>2</sup>	17.6174
210	210 km	167.50 km <sup>2</sup>	45.2287
240	240 km	167.50 km <sup>2</sup>	35.8306

on the same day. We can find that these zones attached to spatial-temporal area are nearby the center square, a bridge, some traffic hubs and narrow places of Tampere, as illustrated in Fig. 4 which presents the results from 26th August to 28th August. These results are in accordance with the previous studies [1] and our general knowledge of Tampere traffic.

**Table 4.** Spatial-temporal scanning result

Fig	Date	Time interval	$(lat_1, lng_1, lat_2, lng_2)$	$\mathcal{L}$
(a)	26th Oct, 2015	10:30~11:00	(61.5192, 23.6172, 61.5747, 23.6797)	13.8912
(b)	26th Oct, 2015	16:30~17:30	(61.4987, 23.6092, 61.5327, 23.6997)	29.0666
(c)	27th Oct, 2015	10:30~12:00	(61.4977, 23.7022, 61.5197, 23.8457)	11.9916
(d)	27th Oct, 2015	18:00~18:30	(61.4987, 23.6092, 61.5327, 23.6997)	14.1069
(e)	28th Oct, 2015	10:30~11:00	(61.5137, 23.6127, 61.5747, 23.7037)	18.3807
(f)	28th Oct, 2015	16:00~19:00	(61.4957, 23.5957, 61.5227, 23.6457)	17.6174

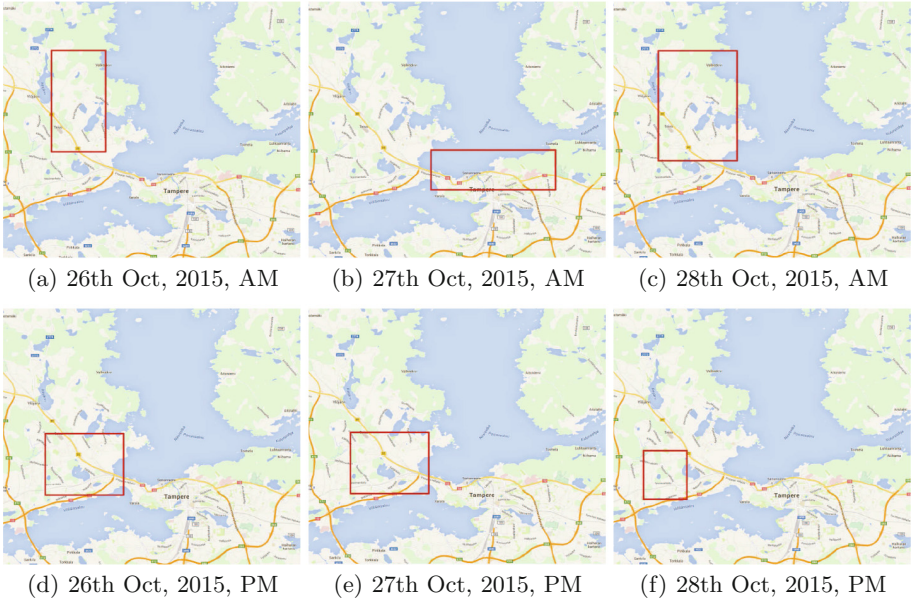


Fig. 4. Spatial-temporal area detection

## 6 Conclusions

In this paper, we study a problem of mining bus delay aggregation which is useful for optimizing the bus delay timetable. To mine these clusters of bus delay and present the result sets concisely, we propose an algorithm called RSTV-Miner. RSTV-Miner has three components: R-tree index structure, spatial-temporal scan and statistical test. We evaluate our method using the open data set released by city of Tampere, Finland recently. Our experiment verifies the bus delay aggregation in spatial-temporal area by RSTV-Miner. And it is useful for traffic schedule structure to find the traffic congestion place in Tampere.

In our work, we apply spatial-temporal scanning in our algorithm that help us to learn the aggregation of bus delay which is useful for traffic schedule construction. We use statistical hypothesis test to detect the clustering degree of bus delay. Mining this interest problem in our work, we find the aggregation of bus delay in both temporal dimension and spatial dimension. After finding the aggregation, we obtain many clusters of bus delay on spatial-temporal area using RSTV-Miner. When scanning area, we partition study area into zones, which omits some details about bus stops and affects the efficiency of algorithm. This lead us to do more future work on mining bus delay. In the future, we could also do more data mining on experiment results to predict bus delay and traffic congestion.

## References

1. Syrjärinne, P., Nummenmaa, J., Thanisch, P., Kerminen, R., Hakulinen, E.: Analysing traffic fluency from bus data. *IET Intell. Transport Syst.* **9**(6), 566–572 (2015)
2. Syrjärinne, P., Nummenmaa, J.: Improving usability of open public transportation data. In: 22nd ITS World Congress, pp. 5–9 (2015)
3. Mazloumi, E., Rose, G., Currie, G., Sarvi, M.: An integrated framework to predict bus travel time and its variability using traffic flow data. *J. Intell. Transp. Syst.* **15**(2), 75–90 (2011)
4. Padmanaban, R., Divakar, K., Vanajakshi, L., Subramanian, S.: Development of a real-time bus arrival prediction system for indian traffic conditions. *IET Intell. Transport Syst.* **4**(3), 189–200 (2010)
5. Yu, B., Lam, W.H., Tam, M.L.: Bus arrival time prediction at bus stop with multiple routes. *Transp. Res. Part C: Emerg. Technol.* **19**(6), 1157–1170 (2011)
6. Chien, S.I.J., Kuchipudi, C.M.: Dynamic travel time prediction with real-time and historic data. *J. Transp. Eng.* **129**(6), 608–616 (2003)
7. Abdelfattah, A., Khan, A.: Models for predicting bus delays. *Transp. Res. Rec.* **1**(1623), 8–15 (1998)
8. Kulldorff, M.: A spatial scan statistic. *Commun. Stat.-Theory Methods* **26**(6), 1481–1496 (1997)
9. Takahashi, K., Kulldorff, M., Tango, T., Yih, K.: A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *Int. J. Health Geographics* **7**(1), 1 (2008)
10. Michelozzi, P., Capon, A., Kirchmayer, U., Forastiere, F., Biggeri, A., Barca, A., Perucci, C.A.: Adult and childhood leukemia near a high-power radio station in Rome, Italy. *Am. J. Epidemiol.* **155**(12), 1096–1103 (2002)
11. Green, C., Hoppa, R.D., Young, T.K., Blanchard, J.: Geographic analysis of diabetes prevalence in an urban area. *Soc. Sci. Med.* **57**(3), 551–560 (2003)
12. Sabel, C., Boyle, P., Löytönen, M., Gatrell, A.C., Jokelainen, M., Flowerdew, R., Maasilta, P.: Spatial clustering of amyotrophic lateral sclerosis in finland at place of birth and place of death. *Am. J. Epidemiol.* **157**(10), 898–905 (2003)
13. Gao, P., Guo, D., Liao, K., Webb, J.J., Cutter, S.L.: Early detection of terrorism outbreaks using prospective space-time scan statistics. *Prof. Geogr.* **65**(4), 676–691 (2013)
14. Zou, B., Peng, F., Wan, N., Mamady, K., Wilson, G.J.: Spatial cluster detection of air pollution exposure inequities across the united states. *PLoS ONE* **9**(3), e91917 (2014)
15. Helbich, M., Leitner, M.: Evaluation of spatial cluster detection algorithms for crime locations. In: Gaul, A., Geyer-Schulz, A., Schmidt-Thieme, L., Kunze, J. (eds.) *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, pp. 193–201. Springer, Heidelberg (2010)
16. Naus, J., Wallenstein, S.: Temporal surveillance using scan statistics. *Stat. Med.* **25**(2), 311–324 (2006)
17. Naus, J.I.: The distribution of the size of the maximum cluster of points on a line. *J. Am. Stat. Assoc.* **60**(310), 532–538 (1965)
18. Rogerson, P.A., Yamada, I.: Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Stat. Med.* **23**(14), 2195–2214 (2004)

19. Wallenstein, S.: A test for detection of clustering over time. *Am. J. Epidemiol.* **111**(3), 367–372 (1980)
20. Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., Mostashari, F.: A space-time permutation scan statistic for disease outbreak detection. *PLoS Med.* **2**(3), e59 (2005)
21. Kulldorff, M.: Prospective time periodic geographical disease surveillance using a scan statistic. *J. R. Stat. Soc.: Ser. A (Stat. Soc.)* **164**(1), 61–72 (2001)
22. Ahrens, J.H., Dieter, U.: Computer methods for sampling from gamma, beta, poisson and binomial distributions. *Computing* **12**(3), 223–246 (1974)