On Detection of Emerging Anomalous Traffic Patterns Using GPS Data

Linsey Xiaolin Pang[a,d], Sanjay Chawla[a], Wei Liu[b], Yu Zheng[c]

[a]*School of Information Technologies, University of Sydney, Australia*
[b]*Dept. of Computer Science and Software Engineering, University of Melbourne, Australia*
[c]*Microsoft Research Asia*
[d]*NICTA, Sydney, Australia*

Abstract

The increasing availability of large-scale trajectory data provides us great opportunity to explore them for knowledge discovery in transportation systems using advanced data mining techniques. Nowadays, large number of taxicabs in major metropolitan cities are equipped with a GPS device. Since taxis are on the road nearly twenty four hours a day (with drivers changing shifts), they can now act as reliable sensors to monitor the behavior of traffic. In this article, we use GPS data from taxis to monitor the emergence of unexpected behavior in the Beijing metropolitan area, which has the potential to estimate and improve traffic conditions in advance. We adapt likelihood ratio test statistic(LRT) which have previously been mostly used in epidemiological studies to describe traffic patterns. To the best of our knowledge the use of LRT in traffic domain is not only novel but results in accurate and rapid detection of anomalous behavior.

Keywords:
Data mining, Mining methods and algorithms, Spatial / Temporal databases

1. Introduction

With the increasing availability of high resolution GPS traces from vehicles in large metropolitan areas, there is an opportunity to infer sophisticated patterns and trends which till now has not been possible [1]. The inferred trends can then be used as input into policy planning across a variety of domains including traffic management, urban planning and environmental monitoring.

In this article, we apply statistical approach on massive taxi location traces, which are explored to extract the outlier traffic pattern in transportation systems. We know thousands of taxis ply the roads of large metropolitan cities like New York, London, Beijing and Tokyo every day. Most taxis are on the road twenty four hours a day with drivers changing shifts. Many of these taxis are now equipped with GPS and their spatio-temporal coordinates are available. Thus if a city is partitioned into a grid then at a given time we can estimate the count of the number of taxis in the grid cells. Over time, the cell counts will settle into a pattern and vary periodically. For example,

Email addresses: qlinsey@it.usyd.edu.au (Linsey Xiaolin Pang),
sanjay.chawla@sydney.edu.au (Sanjay Chawla), wei.liu@unimelb.edu.au (Wei Liu),
yuzheng@microsoft.com (Yu Zheng)

during morning rush hour more taxis will be concentrated in business districts than at other times of the day. Similarly taxi counts near airports will synchronize with aircraft arrival and departure schedules. Occasionally there will be a departure of the cells counts from periodic behavior due to unforeseen events like vehicle breakdowns or one-time events like big sporting events, fairs and conventions.

Our objective is to identify contiguous set of cells and time intervals which have the largest statistically significant departure from expected behavior.

Once such regions and time intervals have been discovered then experts can begin identifying events which may have caused the unexpected behavior. This in turn can help make provisions to manage future traffic behavior. Similar problems appear in many other domains. For example, government healthcare agencies are interested in detecting emergence of disease patterns which deviate from expected behavior.

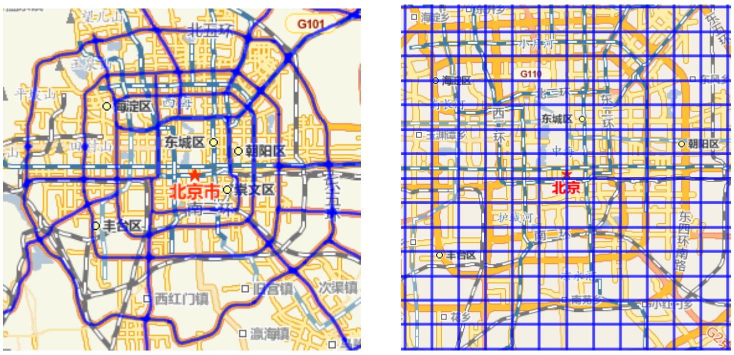
The number of contiguous regions and time intervals is very large. For example, if the spatial grid corresponds to a $n \times n$ matrix and there are T time intervals, then there are potentially $O(n^2T)$ spatio-temporal cells and $O(n^4T^2)$ cubic regions [1]. The huge amount of spatio-temporal data, such as taxi count across different grid regions within different time steps from minutes to hours to days, requires an efficient approach to detect spatial-temporal outliers for predicting abnormal events and implementing traffic control measures in advance. For this motivation, we apply road network of Beijing and partition it into grid to find outliers (Fig. 1).

In a paper of particular relevance to our work, the LRT framework [2] states the computation cost for single statistic value as well as enumerating all the spatial regions to be expensive. To avoid performing statistical computations for every region, it provides a pruning strategy based on classical likelihood test statistic. In this article, we extend the LRT framework to detect abnormal traffic pattern. More specifically, **the contributions** are:

- A general and efficient pattern mining approach for spatio-temporal outlier detection is proposed.
- Persistent and emerging outlier detection statistical models are provided.
- We give our proof that the upper-bounding strategy of LRT is applicable to "persistent" and "emerging" outlier detection models.
- Experiments are conducted on synthetic data to verify the extended pruning approach and show the significant improvement of searching when data set size is large; we also performed real data validation in the detection of emerging taxi count trend due to some major events.

The rest of this article is organized as follows. Section 2 reviews related work. Section 3 illustrates the statistic background and upper-bounding methodology for pruning. Section 4 proposes our approach, in which the statistic detection models are provided. The upper-bounding and pruning mechanism in this framework based on our proof are presented in section 5. Computational complexity is also discussed in this section. Section 6 shows the experiments and case studies. Finally, section 7 concludes this work and section 8 gives the future work.

[1] In this work, $n \times n$ spatial grid and T time intervals are mapped to a three-dimensional grid. The unit cell is in the shape of a cube and every sub-region in the grid is called cubic region



(a) Road Network of Beijing

(b) Grid Map

Figure 1: An example of the traffic network of Beijing. Based on the longitude and latitude, the entire city is partitioned into a grid map. Sub-figure (a) is partitioned into sub-figure (b).

2. Related Work

2.1. Traffic Outliers

Recently, quite a few research projects started to find out traffic patterns and anomalies using taxi trajectories [3, 4, 5, 6]. For instance, to provide a user with the fastest route to a given destination at a given departure time, Yuan et al. [4] mines smart driving directions from historical GPS trajectories of a large number of taxis. It proposes a time-dependent landmark graph to model the properties of dynamic road networks and applies two-stage routing algorithm to find the efficient driving directions. In another work of Yuan et al [3], it presents a cloud-based system to retrieve the fastest driving routes based on traffic conditions and driver behavior. The Cloud builds a model incorporating day of the week, time of day, weather conditions, and individual driving strategies. Using this model, the system predicts the traffic conditions of a future time by given a route and performs a self-adaptive driving direction service for a particular user.

Mining traffic pattern is an important research approach, but detecting the outliers from main traffic flow is also meaningful. For instance, when a traffic incident or jam happens, traffic flow changes suddenly and this will be reflected by outliers. Traffic incidents can be detected through recognizing outliers. Such unusual traffic pattern reflects abnormal traffic streams on road networks and provides useful, important and valuable information. Unknown but potentially important patterns can be forecast by analyzing these outliers. Therefore, the detection of outliers/anomalies from trajectory data can help in sensing abnormal events and plan for their impact to ensure smoother flow of traffic. In Liu et al [5] work, algorithms are presented for discovering spatio-temporal outliers and causal relationships. The discovery of relationships, especially causal interactions, among detected traffic outliers are investigated. Chawla and Zheng et al [7] further diagnose detected traffic anomalies by studying the traffic flows (paths) that lead to an anomaly. In Zheng et al [6] work, it detects flawed urban planning using the GPS trajectories of taxicabs travelling in urban areas. It finds anomalous patterns in a city using taxi trajectories which provides a deeper understanding of the flawed planning. Although these two approaches are used to detect traffic outliers, they are different from our work since we detect traffic outliers using statistical-based approach and we have a different definition for abnormal traffic patterns.

2.2. Outlier Detection Methods

Till now, many anomaly detection techniques have been specifically developed for certain application domains, while others are more generic. The techniques can be categorized as classification-based, distanced-based, clustered-based, and statistical-based, etc. In these surveys [8, 9], they give comprehensive and high-level overview of different outlier detection techniques and some of their applications.
In this article, we focus on the statistic-based approach to detect spatio-temporal outlier. The underlying principle of any statistical outlier detection technique is: "An anomaly is an observation which is suspected of being partially or wholly irrelevant because it is not generated by the stochastic model assumed" [10]. It is based on the key assumption: Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model. Statistical techniques fit a statistical model (usually for normal behavior) to the given data and then apply a statistical inference test to determine if an unseen instance belongs to this model or not. Instances that have a low probability from the applied test statistic are declared as outliers. Scoring techniques are used to assign an anomaly score to each instance in the test data depending on the degree to which that instance is considered an anomaly. Usually, the output of such techniques is a ranked list of outliers. We may choose to either analyze the top few outliers or use a cut-off threshold to select the outliers. Both parametric as well as non-parametric techniques have been applied to fit a statistical model [11, 12]. In our work, we only consider the parametric anomaly detection techniques based on classical likelihood ratio test statistic (i.e. LRT).
In the domain of spatio-temporal applications, most statistic outlier detection approaches proposed so far are on purely spatial searching [13, 14, 15]. Even when considering time aspect, most of the existing work on spatio-temporal outlier detection treats the time dimension simply, either by applying purely spatial outlier detection methods at each time step, or by treating time as another spatial dimension and thus applying spatial outlier detection in one more dimensional space (original spatial dimensions plus time dimension). The disadvantage of the first approach is that by only examining one time step of data at a time, more slowly emerging outliers may not be detected. The disadvantage of the second approach is that less relevant outliers may be detected: those outliers that have constantly existed for a long time, rather than those that are newly emerging [16, 17]. In our following work, we differentiate temporal property with spatial property. More specifically, we investigate the spatial region in two different scenarios in which the temporal property is persistent or otherwise emerging. Persistent temporal data refers to data where its temporal property is consistent over time. Emerging temporal data refers to the data where its temporal property is non-decreasing over time. These two concepts provides more information for practical outlier detections.

Among the various statistic methods for discovering outlier, the spatial and space-time scan statistic, introduced by Kulldorff [18, 19, 20, 21], has been the most widely adopted. However, it is originally designed for Poisson and Bernoulli data. Later on, the different variations of ordinal, exponential and normal models are proposed [22, 23, 24, 25]. They have been implemented in the software (SaTScan) [26]. In the space-time scan statistic of Kulldorff, the key parameter is assumed to be consistent over time. The technique simply applies time as one more dimension. Niell et al. [16] points out the distinct feature of time aspect and proposes a modified test statistic to detect localized and globalized emerging cluster . Tango et al. [27] also proposes a space-time scan statistic based on negative binomial model by taking into account the possibility of nonnegligible time-to-time variation of Poisson mean. Wu et al. [2] proposes a generic framework called LRT for any underlying statistics model. It uses the classic likelihood ratio test (LRT) statistic as a scoring function to evaluate the “anomalousness” of a given spatial region with respect to the rest of the spatial region. Moreover a generic pruning strategy was proposed to greatly reduce the number of likelihood ratio tests. However, it is used for spatial anomaly detection without considering the temporal property. Liu et al. [5] propose an approach to discover casual relationships among spatio-temporal outliers. Here, we only focus on detecting spatial-temporal outliers.

2.3. Performance Issue

Furthermore, performance issue is also a big problem in spatial or spatio-temporal outlier detection. The naive computation of spatial outlier detection is very time-consuming, various strategies have been proposed to speed up the process [18, 28, 29, 30]. These existing methods in the literature are based on Kulldorffs spatial scan statistic and they aim to actually avoid considering all $O(n^4)$ rectangular areas. Also they are only applicable to those relatively simple density measures that are convex or monotonic with respect to the ratio of zone population over entire population and the ratio of zones event count over the entire event count. The LRT framework states the computation cost for single statistic value as well as enumerating all the spatial regions to be expensive [2]. To avoid performing statistical computations for every region, it provides a pruning strategy based on classical likelihood test statistic. In this article, we extend it to be applicable to persistent and emerging outlier detection scenarios.

3. Background

3.1. The Likelihood Ratio Test (LRT)

We provide a brief but self-contained introduction for finding the most anomalous region (rectangle) in a spatial setting. The regions are rectangles mapped onto a spatial grid. We also explain a pruning strategy which can cut down the number of rectangles that need to be checked. The basic tool to find the anomalous region is the Likelihood Ratio Test (LRT).

Given a data set X, the model distribution $f(X, \theta)$, a null hypothesis $H_0 : \theta \in \Theta_0$ and an alternate hypothesis $H_1 : \theta \in \Theta - \Theta_0$, LRT is the ratio

$$\lambda = \frac{\sup_{\Theta_0}\{L(\theta|X)|H_0\}}{\sup_{\Theta}\{L(\theta|X)|H_1\}}$$

where $L()$ is the likelihood function. θ is a set of parameters coming from complete parameter space Θ and null parameter space Θ_0. See detail in [2, 31]. In a spatial setting, the null hypothesis is that the statistical aspect of the phenomenon of interest in a region R (that is currently being tested) are no different from rest of the spatial area (denoted as $\bar{R}$). Thus if a region R is anomalous then the alternate hypothesis will most likely be a better fit and the denominator of λ will have a higher value for the maximum likelihood estimator of θ. A remarkable fact about λ is that under mild regularity conditions, the asymptotic distribution of $\Lambda \equiv -2 \log \lambda$ follows a χ^2_k distribution with k degrees of freedom, where k is the number of free parameters[2]. (See Fig. 2). Thus regions whose Λ value drops in the tail of χ^2 distribution are likely to be anomalous.

[2] If the χ^2 distribution is not applicable then Monte Carlo simulation can be used to ascertain the p-value

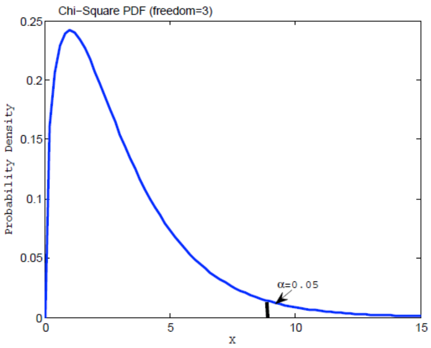


Figure 2: Chi-square distribution with degree of freedom =3

3.2. Constrained Maximum Likelihood Estimation

Constrained maximum likelihood estimation is a set of procedures for the estimation of the parameters of models via the maximum likelihood method with general constraints on the parameters, along with an additional set of procedures for statistical inference [32]. Barlow et al. [33, 34] solved the maximum likelihood estimation based on a reliability growth model, which is applicable to our emerging scenario. It assumes that a system is being modified during K stages of development (in our application, we assume taxi counts in a region to be varying during K time steps). Data consists of x_i successes in n_i trials in stage $i, i = 1, ..., k$. Let p_i be the system reliability at the i-th stage (in our case, p_i is the increasing rate of taxi count at i-th time step). Barlow, et al, obtained the maximum likelihood estimates of $p_1, p_2, ..., p_k$, under the restriction that $p_1 \leq p_2, \leq ..., \leq p_k$. To obtain the maximum likelihood estimates of $p_1, p_2, ..., p_k$ subject to the restriction that $p_1 \leq p_2, ..., \leq p_k$, first form the ratios $x_1/n_1, x_2/n_2, ..., x_k/n_k$. If $x_1/n_1 \leq x_2/n_2, \leq ..., \leq x_k/n_k$, then x_i/x_n is the MLE $\hat{p}_i$ of p_i. If for some $j(j = l, ..., K - l)$, x_j/n_j , combine the observations in the j-th and (j + 1)-st stages and examine the ratios: $x_1/n_1, ..., x_{j-1}/n_{j-1}, x_j + x_{j+1}/n_j + n_{j+1}, x_{j+2}/n_{j+2}, ..., x_k/n_k$, for the $(k - 1)$ stages thus formed. If these ratios are in non-decreasing order, they constitute the MLE's of $p_1 \leq p_2, \leq ..., \leq p_k$ with $\hat{p}_j = \hat{p}_{j+1} =(x_j + x_{j+1})/(n_j + n_{j+1})$. If not, continue the process of combining stages until the ratios are in non-decreasing order. This process need be repeated at most $(k-1)$ times, and the result is independent of the order in which stages are combined to eliminate reversals in the sequence of ratios.

To simplify, we get :

$$\hat{p}_i = Max_{k \geq i} Min_{s \leq i} [\sum_{i=s}^{k} x_i / \sum_{i=s}^{k} n_i]$$

, where $i = 1, ..., k$

3.3. Monte Carlo Simulation

The likelihood ratio, or equivalently its logarithm, can be used to compute a p-value, or compared to a critical value to decide whether to reject null hypothesis (i.e. there is no anomaly in our case) in favor of the alternative hypothesis (i.e. there is anomaly). The probability distribution of likelihood ratio, assuming that the null hypothesis is true, can be approximated by chi-square distribution [35]. But we cannot expect to find the distribution of the likelihood ratio test statistic in closed analytical form and thus Monte Carlo simulation can be performed to obtain p-value in some cases. Therefore, once we have discovered the region with maximum likelihood ratio value, the statistical significance of this region can be derived from the chi-square distribution or by conducting Monte Carlo simulations. To run the simulation test, a large number of replications of data sets are generated under the null hypothesis, for instance, 9999 such replicas are created to perform likelihood ratio test. The test is significant at the 5% level if the likelihood ratio value is among the 500 highest values of the test statistic coming from the replications.

3.4. Upper-bounding Methodology:

The upper-bounding strategy for LRT for anomaly detection was introduced by Wu and Jermaine [2]. The basic observation is that the likelihood value of any given region R under complete parameter space is not greater than the multiplication of the likelihood value of all its non-overlapping sub-regions under null parameter space. Therefore, the log likelihood of any given region R can be upper-bounded. For instance, if a region R is composed of two non-overlapping sub-regions R_1 and R_2, then

$$L(\theta_R|X_R) \leq L(\theta'_{R_1}|X_{R_1}) \times L(\theta'_{R_2}|X_{R_2})$$

It is equivalent to

$$logL(\theta_R|X_R) \leq logL(\theta'_{R_1}|X_{R_1}) + logL(\theta'_{R_2}|X_{R_2})$$

Here θ_R, θ'_{R_1} and θ'_{R_2} are the maximum likelihood estimators under complete and null parameter spaces separately. See figure 3.

The upper-bounding strategy is used to prune non-outliers: *If we replace the likelihood of a region R by the product of the likelihoods of its sub-regions and the new LRT is below the anomalous threshold (i.e. confidence level α), then R cannot be anomalous.*

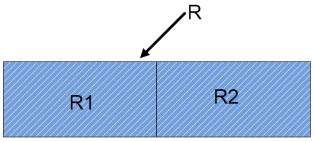


Figure 3: The log likelihood of region R under complete parameter space is upper bounded by the sum of two non-overlapping subregions R_1 and R_2 under null parameter space

3.5. Examples:

3.5.1. Example 1

Using a simple but concrete example, we will now explain how to find anomalous region using traditional LRT computation and the upper-bounding pruning strategy. Consider the 4×4 grid (G) in Figure 4. The number of successes (m_i) independently generated by Poisson model $P_o(b_i p)$ is displayed in each cell c_i. The baseline b_i in each cell c_i is set to 10. The success rate p is 0.5 for the region R and 0.1 for the rest of cells. The significant level is set to α=0.05. We refer the success rate p as the test parameter.

Procedures: For a given region R, traditional LRT calculation involves several steps: maximum likelihood estimator for test parameter of R, $\bar{R}$ and G; likelihood

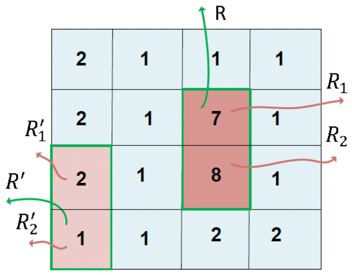


Figure 4: An example of (4 × 4) grid to illustrate the LRT calculation and the upper-bounding pruning methodology

calculation of R, $\bar{R}$ and G; the ratio calculation from the previous two steps. Although the calculation for Poisson distributed data can be simplified into 1EXP statistic model, in order to illustrate the traditional LRT computation, the original steps are carried out as the following :

(1) The likelihood function of each cell i is:

$$f(p|c_i) = \frac{(b_i p)^{k_i} e^{-(b_i p)}}{k_i!} \tag{1}$$

(2) The likelihood of any given region R is, which is composed of cell $c_1, c_2, ..., c_i, ..., c_t$:

$$L(p|R) = \Pi_{c_i \in R} \frac{(b_i p)^{k_i} e^{-(b_i p)}}{k_i!} \tag{2}$$

(3) The MLE_0 of p for a region R (denoted as $\hat{p}$) is calculated as:

$$\hat{p} = (\sum\nolimits_{c_i \in R} k_i)/(\sum\nolimits_{c_i \in R} b_i) \tag{3}$$

Thus $\hat{p}_R = \frac{(7+8)}{(10+10)} = 0.75$. Similarly, $\hat{p}_{\bar{R}}$, $\hat{p}_{R_1}$, $\hat{p}_{R_2}$ and $\hat{p}_G$ are obtained as: 0.14, 0.7 ,0.8 and 0.21.

(4) The Λ of region R is given by

$$\begin{aligned} \Lambda_R &= -2\log(L(p|G) + 2\log L(p|R) + 2\log L(p|\bar{R})) \\ &= -2\log(0.21^{19+15} \times e^{-0.21\times160}) \\ &\quad + 2\log(0.75^{15} \times e^{-0.75\times20}) \\ &\quad + 2\log(0.14^{19} \times e^{-0.14\times140}) \\ &= 20.76 \end{aligned} \tag{4}$$

From above steps, we get the exact log likelihood value of region R: $\log L(p|R) = -19.31$; and the exact log likelihood of R_1 and R_2: $\log L(p|R_1) = -9.49$, $\log L(p|R_2) = -9.78$ separately.

We know the critical value of $\chi^2(\alpha)$=3.84. Obviously, 20.76 is greater than 3.84. Therefore region R is treated as a potential outlier.

Upper-bounding Pruning: We can also verify the upper-bound of region R: The sum of log likelihood of R_1 and R_2 is -19.27, which is greater than the exact log likelihood value of R which was computed as -19.31.

Similarly, For region R', $\log L(p|R'_1) = -1.31$, $\log L(p|R'_2) = -1$, sum of log likelihood of R'_1 and R'_2 is -2.31. It is greater than the log likelihood value of R', which is -2.66. Furthermore, $\Lambda_{R'} = 2.14$, which is smaller than $\chi^2(\alpha)$=3.84. That shows that the upper-bounded LRT value is smaller than the critical value. Using the pruning strategy, the actual value of region R' need not be calculated and can be pruned.

3.5.2. Example 2:

This example illustrates how to estimate maximum likelihood of system reliability in reliability growth model. To detect emerging outlier,we use this way to calculate maximum likelihood estimator.

(1)The number of successes, population and success rate in each time step are shown in Table 1.

Table 1: Reliability Growth Procedure

$TimeStep(i)$	Number of Successes (x$_i$)	Population (n$_i$)	Success Rate (x_i/n_i)
0	20	50	0.400
1	30	70	0.429
2	30	80	0.375
3	20	60	0.333
4	50	60	0.833

(2) To estimate the maximum likelihood of $\hat{P}_i$, the process to get a sequence of non-decreasing ratios is summarized below according to Barlow theorem:

Table 2: The maximum likelihood estimator procedure

i	x$_i$	n$_i$	x_i/n_i	First Calculation	Second Calculation	Third Calculation
1	20	50	0.400	0.400	80/210=0.381	100/260=0.385
2	30	70	0.429	60/150=0.400		
3	30	80	0.375			
4	20	60	0.333	0.333		
5	50	60	0.833	0.833	0.833	0.833

From above procedures, we obtain the maximum likelihood estimates: $\hat{P}_1 = \hat{P}_2 = \hat{P}_3 = \hat{P}_4 = 0.385, \hat{P}_5 = 0.833$.

4. Proposed Statistical Models

Definition 1. KP : It refers to “key parameter”, denoted as $KP\{\theta_1, \theta_2,.., \theta_i , ..,\theta_n\}$. θ_i is a parameter coming from the key parameter set. For instance, in epidemiology, if we are concerned about the trend of the disease rate in a spatio-temporal view, the disease rate is KP. In our application, the variation of taxi count within a period is KP. For simplicity, we only consider one parameter from the key parameter set in our work (denoted as KP).

PSTO Model (Persistent Spatio-Temporal Outlier Model): It is used to detect persistent spatio-temporal outliers. The null hypothesis H_0 assumes that the KP is consistent for all regions over time. The alternative hypothesis H_1 assumes that KP has a higher value in region $r_i \in R$ than the value outside of region $r_j \in G\text{-}R$ (i.e. $\bar{R}$), but the value in region $r_i \in R$ is consistent over time. We calculate the likelihood ratio test as follows:

$$D(R) = \begin{cases} \frac{\Pi_{r_i \in R} L(\theta_r|X_R) \Pi_{r_i \in \bar{R}} L(\theta_{\bar{r}}|X_{\bar{r}})}{\Pi_{r_i \in G} L(\theta_G|X_G)} & \text{for } \theta_r \geq \theta_{\bar{r}} \text{ ,} \\ 1 & \text{otherwise.} \end{cases}$$

This formula is the classical LRT statistic. We first calculate the MLE of θ_r and $\theta_{\bar{r}}$ to maximize the numerator and the MLE of θ_G to maximize the denominator. Then the ratio is the score we use to evaluate the "anomalousness" of a given spatio-temporal region.

ESTO Model (Emerging Spatio-Temporal Outlier Model): This model is used to detect emerging spatio-temporal outliers. The null hypothesis H_0 assumes that the KP is consistent for all regions over time. The alternative hypothesis H_1 assumes that KP is non-decreasing with every time step over region $r_i \in R$ and higher than $r_j \in \bar{R}$. We calculate the likelihood ratio test as follows:

$$D(R) = \begin{cases} \frac{Max_{\theta_{\bar{r}} \leq \theta_{t_{min}} \leq \ldots \leq \theta_T} \Pi_{r_i \in R} L(\theta_r^t|X_r^t) \Pi_{r_i \in \bar{R}} L(\theta_{\bar{r}}^t|X_{\bar{r}}^t)}{\Pi_{r_i \in G} L(\theta_G^t|X_G^t)} & \text{for } \theta_{\bar{r}} \leq \theta_{t_{min}} \leq \ldots \leq \theta_T, \\ 1 & \text{otherwise.} \end{cases}$$

This formula is derived from the classical LRT statistic and designed for the emerging scenario. User needs to find a solution to maximize the numerator with the increasing KP. For instance, Barlow [34] provide an approach to solve the constrained maximum likelihood estimation on the reliability growth model in which the relative risk is non-decreasing over time. Or EM algorithm can be performed to estimate the key parameter.

5. Upper-bounding Strategy and Pruning Mechanism for Proposed Framework

5.1. Upper-bounding Strategy

(1) In $PSTO$ model, the upper-bounding strategy explained in section 2.2.2 can be extended directly to spatio-temporal dimension.

(2) In $ESTO$ model, KP is assumed to vary at different time step; we show below that the upper-bounding strategy is still applicable to this model.

Theorem 1. *Let region $R = R_{t1} \cup R_{t2}$, for non-overlapping time interval $t1$ and $t2$, we have:*

$$L(\theta_R|X_R) \leq L(\theta'_{R_{t1}}|X_{R_{t1}}) \times L(\theta'_{R_{t2}}|X_{R_{t2}}) \tag{5}$$

, where $\theta_R = \theta_{R_{t1}} \cup \theta_{R_{t2}}$ and $X_R = X_{R_{t1}} \cup X_{R_{t2}}$

Proof. We know $L(\theta_R|X_R) = L(\theta_{R_{t1}}|X_{R_{t1}}) \times L(\theta_{R_{t2}}|X_{R_{t2}})$. Using the LRT upper-bounding basic concepts, we know that $\theta_{R_{t1}}$ is chosen under more strict complete parameter space and $\theta'_{R_{t1}}$ is chosen under loosen null parameter space. That means performing MLE_0

on a sub-interval of R has loosen the constraints comparing with performing MLE_1 on R. Thus, we have $L(\theta_{R_{t1}}|X_{R_{t1}}) \leq L(\theta'_{R_{t1}}|X_{R_{t1}})$ and $L(\theta_{R_{t2}}|X_{R_{t2}}) \leq L(\theta'_{R_{t2}}|X_{R_{t2}})$. Therefore, $L(\theta_R|X_R) \leq L(\theta'_{R_{t1}}|X_{R_{t1}}) \times L(\theta'_{R_{t2}}|X_{R_{t2}})$

Theorem 2. *Let region $R = R1 \cup R2$, for non-overlapping spatial region R1 and R2,we have:*

$$L(\theta_{R1}, \theta_{R2}|X_{R1}, X_{R2}) \leq L(\theta'_{R1_{t1}}, \theta'_{R1_{t2}}|X_{R1_{t1}}, X_{R1_{t2}}) \times L(\theta'_{R2_{t1}}, \theta'_{R2_{t2}}|X_{R2_{t1}}, X_{R2_{t2}}) \qquad (6)$$

,where R, $R1$, $R2$ are composed of (t1,t2) time steps respectively. Here we just use two time steps to illustrate. It is applicable to any t time steps.

Proof. For each time step i, we have: $L(\theta_{R_{ti}}|X_{R_{ti}}) \leq L(\theta'_{R1_{ti}}|X_{R1_{ti}}) \times L(\theta'_{R2_{ti}}|X_{R2_{ti}})$

$$\begin{aligned} L(\theta_{R1}, \theta_{R2}|X_{R1}, X_{R2}) &= L(\theta_{R1}|X_{R1}) \times L(\theta_{R2}|X_{R2}) \\ L(\theta_{R1_{t1}}, \theta_{R1_{t2}}|X_{R1_{t1}}, X_{R1_{t2}}) &= L(\theta_{R1_{t1}}|X_{R1_{t1}}) \times L(\theta_{R1_{t2}}|X_{R1_{t2}}) \\ L(\theta_{R2_{t1}}, \theta_{R2_{t2}}|X_{R2_{t1}}, X_{R2_{t2}}) &= L(\theta_{R2_{t1}}|X_{R2_{t2}}) \times L(\theta_{R2_{t2}}|X_{R2_{t2}}) \end{aligned}$$

Therefore we get

$$L(\theta_{R1}, \theta_{R2}|X_{R1}, X_{R2}) \leq L(\theta'_{R1_{t1}}, \theta'_{R1_{t2}}|X_{R1_{t1}}, X_{R1_{t2}}) \times L(\theta'_{R2_{t1}}, \theta'_{R2_{t2}}|X_{R2_{t1}}, X_{R2_{t2}})$$

From theorem 1 and theorem 2, we know that the upper-bounding strategy is applicable to emerging model ($ESTO$).

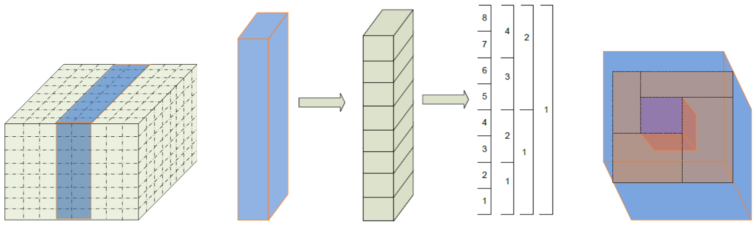
5.2. Pre-computation and Pruning Mechanism

5.2.1. Pre-computation for region R:

We recursively split the region into two sub-regions of the same size, starting from the biggest cuboid enclosed by two planes from time view, ending at the lowest resolution of the spatial-temporal grid. Fig. 5b shows the split approaches for a sub-cuboid highlighted as blue from the temporal dimension in a 8 × 8 × 8 grid (Fig. 5a). The likelihood of any given region can be upper-bounded by this pre-computed set via the tiling of LRT.

5.2.2. Pre-computation for the complement of region R (i.e. $\bar{R}$):

By considering all of the intersection points, we connect each intersection point on the 3-dimensional grid with the eight corners of the grid. This produces eight diagonals, each of which creates one cuboid in the pre-computed set. Since there are $O(n^4)$ intersection points, there are $O(n^4)$ cuboids in the pre-computed set. After we get the pre-computed set, for any given region $\bar{R}$, we use the radial and sandwich methods in LRT to get the upper-bounded likelihood value of $\bar{R}$. These two methods produce six non-overlapping sub-cuboid regions for $\bar{R}$ separately. Radial method is performed by elongating the sides of a region R until the sides hit the grid borders using clock-wise counter clock-wide order. Sandwich method is performed by elongating two parallel sides of R in both directions until they reach the borders of the grid. See detail of these methods in 2-dimensional grid [2]. In 3-dimensional view, twelve times tiling is involved. Fig. 5c shows the tiling in radial way.



(a) R (b) R (c) $\bar{R}$

Figure 5: Pre-computation of any given spatial-temporal region R and tiling of $\bar{R}$. Sub-figure (a) shows a $8 \times 8 \times 8$ spatial-temporal grid; sub-figure(b) shows: one of the cuboids from spatial pre-computed set is split from temporal dimension and results in 15 smaller cuboids. Sub-figure(c) is the radial method to tile $\bar{R}$

5.2.3. Computational Complexity

In the brute-force approach, there are a total of $O(n^6)$ regions that need to be searched. Our approach reduces the cost by precomputing two likelihood data set with size of $O(n^4)$. The likelihood of every region is upper-bounded and the real likelihood is calculated only for a number of regions. Furthermore, In our implementation, we have already ranked the top-k regions according to the likelihood ratio values. Therefore, the performance wont be affected no matter which significance testing method is being applied.

The process of outlier detection is shown in Algorithm 1. The inputting parameters are: data grid (G), probability density function (f), maximum likelihood estimation function under different parameter space (MLE_0, MLE_1), likelihood function (L), number of top regions to be returned (K) and the significance level (α). In this process, step 1 and 2 perform pre-computations; Step 5 to step 8 obtains the upper-bounded likelihood value of current cuboid for each iteration. During each iteration, the chi-squared distribution is applied to prune normal regions. Finally, it outputs top-k anomalous regions.

6. Experiments, Results and Analysis

We report on experiments conducted where we have used Algorithm 1 to test for accuracy, pruning ability and performance. K was set to 1. In section 6.1 and section 6.2 all experiments were carried out on synthetic data. In section 6.3, we demonstrate the usefulness of our approach on a real data set.

6.1. Results on Synthetic Data

We tested four variants of the outlier detection:

(1) brute-force persistent spatio-temporal outliers (bpsto)

(2) brute-force emerging spatio-temporal outliers (besto)

(3) pruning-based persistent spatial-temporal outliers (ppsto)

(4) pruning-based emerging spatio-temporal outliers (pesto)

Algorithm 1: Top k spatio-temporal outlier detection

```
Input: G, $MLE_0$, $MLE_1$, L, k and $\alpha$
//G: a spatial-temporal grid ; L: likelihood function;
//MLE: maximum likelihood estimator;
//k:number of outliers to be detected; $\alpha$:  critical value;
Output: top-k anomalous spatio-temporal regions.

Pre-compute the $O(n^4)$ cuboids for upper-bounding any given cuboid R;
Pre-compute the $O(n^3)$ cuboids for upper-bounding any given cuboid $\bar{R}$;
Let $\theta_0 = MLE_0(f(G))$;
for Each cuboid R in the grid do
    Get the upper-bounded value for log L($\theta_R|X_R$);
    Get the upper-bounded value for log L($\theta_{\bar{R}}|X_{\bar{R}}$);
    Combine the results of above steps to get an upper bound for $\Lambda_R$;
    Check upper-bounded value of $\Lambda_R$ from chi-square distribution;
    if The $\Lambda_R$ is in the $\alpha$ level and less than the kth best;
    then
        Prune R;
    end
    else
        Compute real $\Lambda_R$ ;
    end
    if $\Lambda_R$ is in the top k;
    then
        Remember R
    end
end
Output top-k regions;
```

We generated data set on a grid size varying from (4×4×4) to (128×16×16). Fifty separate trials were carried out for each scenario (see below) and we measured three aspects: (a) pruning rate (b) accuracy, and (c) running time. The significance level was set at $\alpha = 0.05$.

6.1.1. Scenario I

The null hypothesis holds. The baseline b_c is generated relatively uniformly by a Normal distribution ($\mu = 10^4, \sigma = 10^3$) and a fixed success rate p of 0.001. The number of successes k_c is generated from Po ($b_c p$). Results are shown in Table 3.

6.1.2. Scenario II

The null hypothesis holds. The only difference with scenario I is that the data in a random selected cuboid area with size of (5×4×3) is generated by a Normal distribution with different parameter setting ($\mu = 10^5, \sigma = 5 \times 10^3$). Results are shown in Table 3.

6.1.3. Scenario III

The alternative hypothesis holds. It is similar to the null model except that the data of a randomly selected cuboid area of size (5 × 4 × 3) is generated from a Poisson distribution with $p = 3, 6, 9, 18, 36$ for emerging case and $p = 3$ for persistent case.

$Test$	$Pruning(\%)$	$Accuracy(\%)$
$4 \times 4 \times 4$	100	no false alarm
$8 \times 8 \times 8$	100	no false alarm
$16 \times 16 \times 16$	99.9	0.1 false alarm

(a) Scenario *I*

$Test$	$Pruning(\%)$	$Accuracy(\%)$
$4 \times 4 \times 4$	100	no false alarm
$8 \times 8 \times 8$	99.99	0.01 false alarm
$16 \times 16 \times 16$	100	no false alarm

(b) Scenario *II*

Table 3: Average Pruning Rate and Accuracy in Scenario *I* and *II*

$Test$	$16 \times 16 \times 16$	$32 \times 16 \times 16$	$64 \times 16 \times 16$	$32 \times 32 \times 32$	$128 \times 16 \times 16$
ppsto (%)	95.27	97.35	97.64	97.47	96.74
pesto (%)	98.37	98.46	98.69	99.11	99.23

Table 4: Average Pruning Rate in Scenario *III*

The data not within the cuboid area was also from a Poisson distribution with $p = 1$. Results are shown in Table 4.

6.1.4. Scenario IV

The alternative hypothesis holds. It is similar to scenario III except that data of a randomly selected cuboid area of size $(5 \times 4 \times 3)$ was generated from a Poisson distribution with $p = 10, 50, 250, 1250, 6250$ for emerging case and $p = 10$ for persistent case. Results are shown in Table 5.

6.2. Evaluations on Synthetic Data

6.2.1. Analysis on Scenario I and II

The results of Scenario I and II show that we achieve a high pruning rate and no false alarm is generated even when we perturb the distribution of one region. This is as expected and demonstrates that the algorithm is well calibrated. By a high pruning rate we mean that we can rule out the outliers by just checking the LRT upper bound derived from the tiling. If the upper bound value is less than the critical value then the true LRT value of the region cannot be anomalous.

6.2.2. Analysis on Scenario III and IV

For Scenario III and IV the anomalous regions were correctly identified while maintaining a high pruning rate. Also there were no regions declared as false positives.

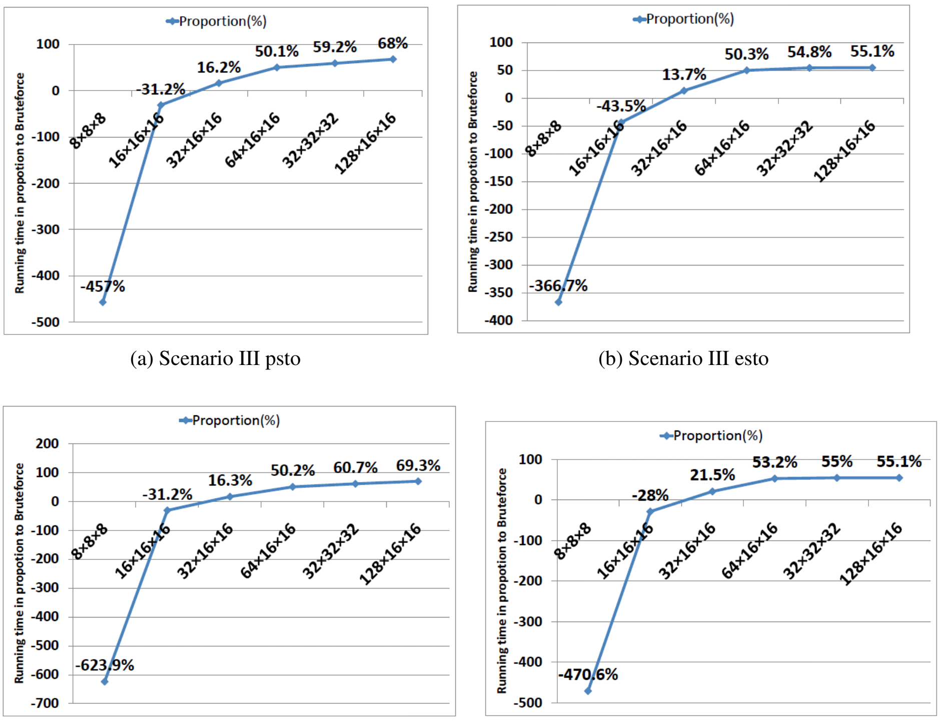
6.2.3. Analysis on Running Time

I Proportion of Running Time:

We analyze the running time with and without pruning for Scenario III and IV. We plot out the proportion of running time of pruning approach relative to brute-force approach in Fig. 6, which is the computation of running time of (brute-force -pruning)/brute-force. Also, the proportion percentage is displayed on these line graphs. It shows that as the size of the spatial and temporal region increase, the effect of pruning becomes prominent. For the largest data tested, the pruning mechanism resulted in a savings of nearly 50% compared to the brute-force approach.

Test	$16 \times 16 \times 16$	$32 \times 16 \times 16$	$64 \times 16 \times 16$	$32 \times 32 \times 32$	$128 \times 16 \times 16$
ppsto (%)	79.27	97.51	97.77	97.22	96.68
pesto (%)	95.57	97.40	96.78	94.70	95.23

Table 5: Average Pruning Rate in Scenario *IV*



(a) Scenario III psto

(b) Scenario III esto

(c) Scenario IV psto

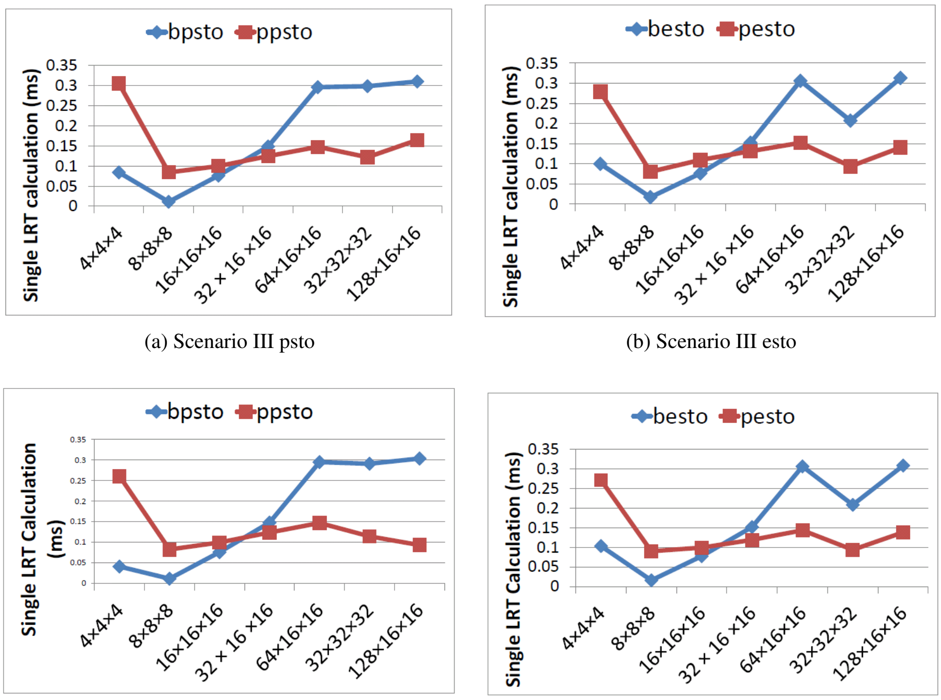
(d) Scenario IV esto

Figure 6: The proportion of running time of pruning vs. brute-force approach. It shows that outlier pruning searching is significantly improved when the dataset size starts from $32 \times 16 \times 16$ in these four different scenarios.

II Single LRT Calculation Cost: We have also calculated the cost of a single likelihood calculation as the dimension of the grid size increases. The results are show in Fig. 7. For the $8 \times 8 \times 8$ data set, the cost of the likelihood calculation using the brute-force approach is 0.01ms while with pruning it increases to 0.08ms. However, for the larger data sets (e.g., $128 \times 16 \times 16$) the cost of a single likelihood calculation goes from 0.30ms for the brute-force approach to around 0.16ms with pruning. Another observation is that the cost of the single likelihood calculation is nearly similar for data sets of the same size but different dimensions, for example $128 \times 16 \times 16$ and $32 \times 32 \times 32$.

III Components Running Cost: We have also analyzed and compared the running of the different components both for the brute-force and pruning approaches. The results are shown in Fig. 8. The brute-force approach has the following components:

(1) The cost of the likelihood calculation for each region R (R Computation).



(a) Scenario III psto

(b) Scenario III esto

(c) Scenario IV psto

(d) Scenario IV esto

Figure 7: The single likelihood ration calculation cost of pruning vs. brute-force approach. It shows that outlier pruning searching is significantly improved when the dataset size starts from 32 × 16 × 16 in these four different scenarios.

(2) The cost of the likelihood calculation for the complement of each region R, denoted as $\bar{R}$ ($\bar{R}$ Computation).

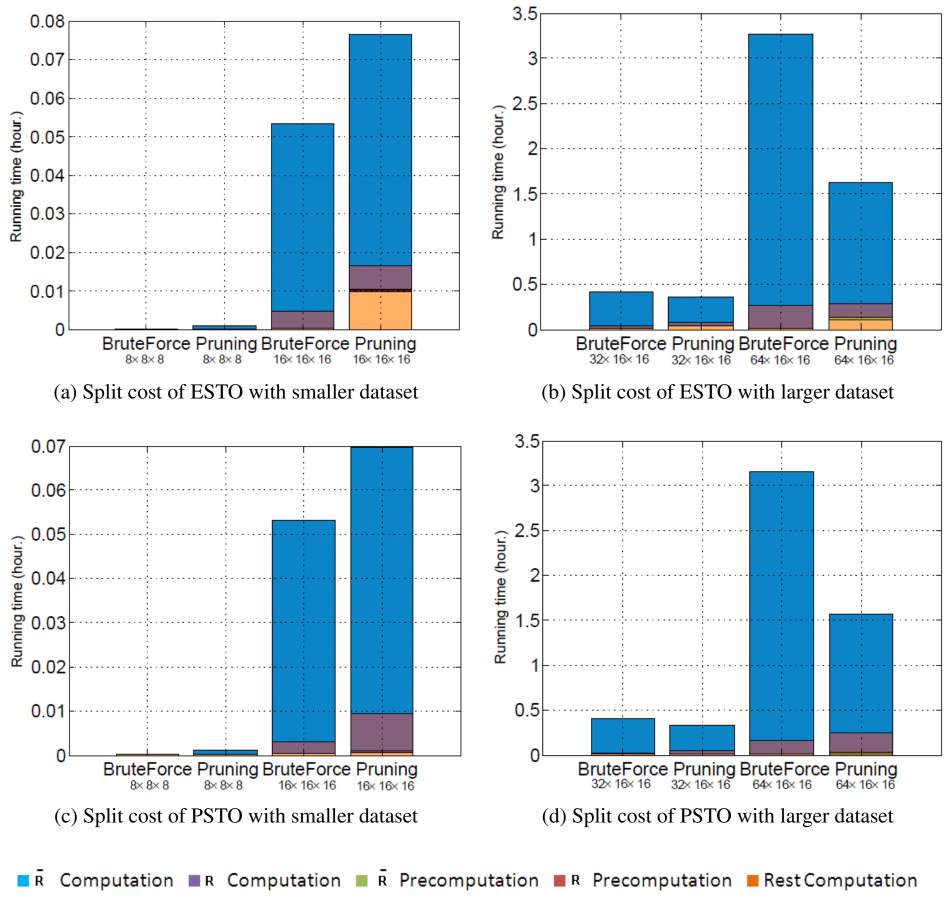
The pruning approach is more complex and involves the following components:

(1) The cost of computing the likelihood for each element of the tiling set T_R. This will be used to upper bound the likelihood value for an arbitrary spatio-temporal region. (R pre-computation)

(2) The cost of computing the likelihood for each element of the tiling set $T_{\bar{R}}$ ($\bar{R}$ pre-computation).

(3) The cost of upper-bounding the likelihood of R. This involves first expressing R as a union of subregions and then each subregion as a union of tiles from T_R.

(4) The cost of upper-bounding the likelihood of $\bar{R}$ ($\bar{R}$ Computation). This involves first expressing $\bar{R}$ as union of subregions and then each subregion as a union of tiles from $T_{\bar{R}}$. Each $\bar{R}$ region can be expressed as a union of six subregions and there are two types of tiling methods: sandwich and radial. We calculate the likelihood value using both tiling methods and then select the tightest upper-bound.
As is clear from Fig. 8, $\bar{R}$ computation is the most expensive part of the calculation. However, as the data set size increases, the overheads of the tiling give way to its more efficient reuse resulting in considerable savings.



(a) Split cost of ESTO with smaller dataset

(b) Split cost of ESTO with larger dataset

(c) Split cost of PSTO with smaller dataset

(d) Split cost of PSTO with larger dataset

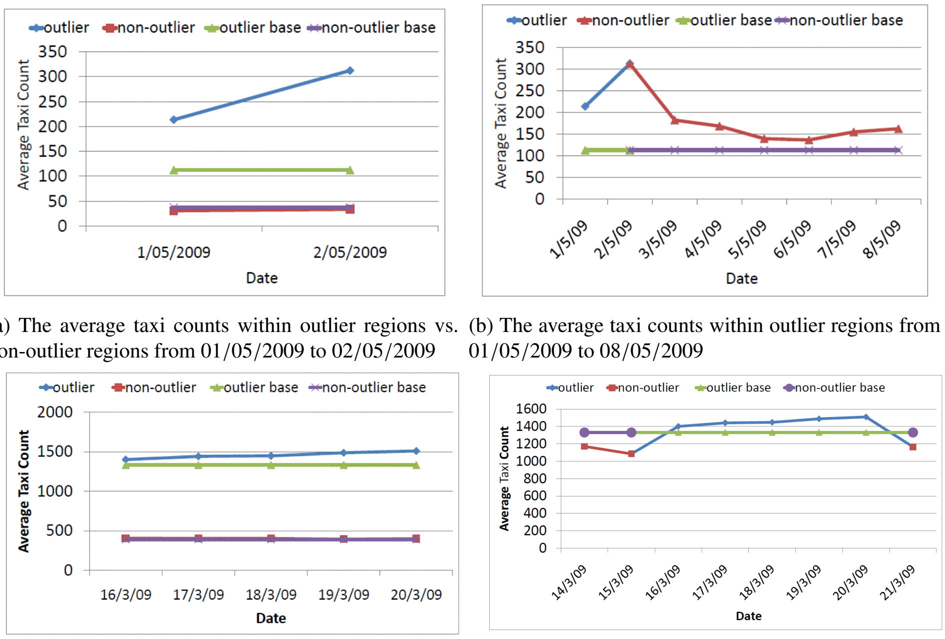
Figure 8: The running time of comparable parts of brute-force vs. pruning approach in scenario III. It shows that the pruning searching is faster with the larger dataset. Although the tiling of every $\bar{R}$ takes longest time in pruning searching, the cost is small compared to the likelihood calculation of every $\bar{R}$ in brute force searching.

6.3. Case Studies: Beijing Taxi GPS Data

We illustrate the use of the Pesto method on a real data set [4, 5]. The data set consists of three months of GPS trajectories collected from 33,000 taxis in Beijing between 01/03/2009 and 31/05/2009. The road network of Beijing is split into grid. The taxi counts of each cell which is identified by column index and row index are provided in a text file. The time frequency of monitoring the taxi counts is 15 minutes. To deal with real data, we only need to calculate the total taxi counts in each cell within a given period, then this pre-processed data can be directly loaded into our algorithm. In this section, we search for the most anomalous emerging region within a specified time period and then provide a possible explanation for the anomaly.

6.3.1. Case I:

All (8×8) grid were tested between $9:00:00am$ and $10:00:00am$ for sixteen days. We choose 20 days of data to calculate the baseline probabilities.



(a) The average taxi counts within outlier regions vs. non-outlier regions from 01/05/2009 to 02/05/2009

(b) The average taxi counts within outlier regions from 01/05/2009 to 08/05/2009

(c) The average taxi counts within outlier regions vs. non-outlier regions from 16/03/2009 to 20/03/2009

(d) The average taxi counts within outlier regions from 14/03/2009 to 21/03/2009

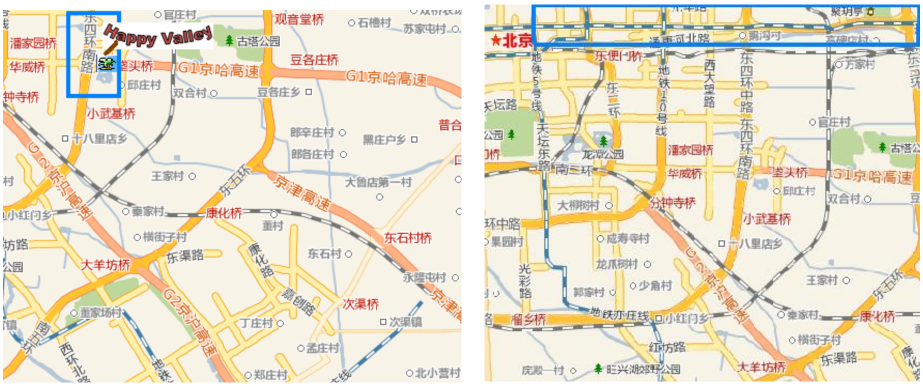
Figure 9: Comparison of outlying and non-outlying regions in $8 \times 8 \times 8$ grid. It shows: (a) the average taxi counts within outlier regions is non-decreasing compared to non-outlier regions which share the same emerging period with outlier. (b) the average taxi counts within outlier regions throughout the emerging period is non-decreasing compared to the outlier regions for the rest of the period.

Result I: The period from 01/05/2009 to 02/05/2009 emerged as a top outlier at the position of (0, 1) and (1, 1) on the grid. This period corresponds to the Labor day public holiday ("Golden Week") . Usually the holiday duration is seven days (from May 1st to May 7th) , but starting from 2009, the holiday period was shorten to between May 1st and May 3rd inclusive. To celebrate the holidays it appears that many people visited Happy Valley, the biggest amusement park in Beijing. The 3rd International Fashion festival was also held in that location. Our results coincide with the fact that taxis enjoy good business on public holidays and there is usually an increase in the number of taxis near tourist spots. The results are shown in Fig. 9a, 9b, 10a. We can see that the number of taxis increased from 1st May to 2nd May and then decreased from 3rd of May onwards.

6.3.2. Case II:

All (8×8) grids were tested between $3 : 15 : 00pm$ to $4 : 30 : 00pm$ for 8 days. We use 12 days of data to calculate the baseline probabilities.

Result II: The region highlighted as blue on the map was detected as an emerging outlier from 16/03/2009 to 20/03/2009. It is one of the city express road called Tonghuihe North Road. From 01/03/2009 to 13/03/20093, the 11th National People's Congress (i.e. NPC) was held in Beijing, which is the annual meeting of the highest



(a) The region highlighted with blue borders on the map is the outlier region of Case I. The icon shows the exact location of Happy Valley.

(b) The region highlighted with blue borders is the outlier of Case II. It is the city express road of Beijing. (i.e. Tonghuihe North Road)

Figure 10: Outlier Locations from our two case studies on Beijing Map

legislative body of the People's Republic of China. Nearly 3000 deputies from all over China attended the Congress. During this period, the traffic authorities in Beijing imposed temporary restriction measures on vehicles to control traffic flow. Most people choose to take bus or subway instead of driving or taking taxi to commute to work. The number of taxi travelling on Tonghuihe North road increased until most of the deputies left Beijing. The results are shown in Fig. 9c, 9d, 10b.

To investigate more, we set $k = 5$ in our case studies. We found the other top 4 outliers have big overlap on the top 1 outlier region. These outliers have similar spatial area and spanning time period. It verifies that the emerging outlier can be correctly located.

7. Conclusion

In this article, an efficient pattern mining approach was proposed to cater for spatio-temporal traffic data, which is able to detect "persistent outliers" and "emerging outliers". We proposed two statistical models , which encompass the generic features of anomalous patterns. We also derived an upper-bounding strategy for the two statistic models supporting for fast outlier detection. Our comprehensive experiments show that the performance of computational time is greatly improved when the dataset size is large, and we can still find the correct outliers. We also carefully analyzed the tiling scheme and the upper-bounding strategy in the synthetic experiments. In our case studies, our model is able to detect regions with emerging number of taxis that can be validated by known major traffic events.

Our approach is applicable to a wide variety of contexts. For instance, in weather forecast models, it can be used to detect emerging weather pattern which raises possibility of dry and warm climate. These climates may have great impact on infectious disease. Investigating such weather variables associated with infectious diseases can help anticipating future epidemics, and early warning system can be developed for surveillance and interventions. In gene expression models, discovering emerging patterns is helpful for diagnosis and understanding correlation of gene expression profiles to disease states in a significant way. They are useful for capturing interactions among genes, finding signature patterns for disease subtypes, and generating potential disease treatment plans, etc. In finance models, mining emerging business trend from different geographical regions can provide insight for identifying some profitable investments. It can also be applied in emerging event identification, intrusion detection, etc.

8. Future Work

From the above discussion,we already observed the cost of pruning approach has nearly above 50% speed-up with respect to the naive algorithm as the grid size increases. It allows the computation to be performed for large-scale applications. However, we also noticed that it still needs almost one to two days to find the top outlier even with pruning approach in our synthetic experiments with grid size of $128\times16\times16$ or $32 \times 32 \times 32$. When the datasets are not modestly sized, the scalability is still not good. This performance might not be acceptable in real detection applications.

We see that the scalability of outlier detection has become important as the amount of data for analysis has been increasing greatly. For dealing with large datasets, it is important to both parallelize the algorithms, and implement them to execute efficiently. Fortunately, the LRT scanning algorithm is highly parallelizable, each sub-grid computation is independent of each other and the whole grid can be partitioned into equal parts and distributed over the multiple processors. The GPU computing or GPGPU (i.e. General-Purpose computation on Graphics Processing Units) has become a new trend for researchers to do general purpose scientific and engineering computation by the use of GPU. It enables dramatic increases in computing performance by harnessing the power of the GPU and is starting to play a significant role in large-scale modeling. Due to our highly parallelizable algorithm, the technology of graphics processing unit (GPU) and compute unified device architecture (CUDA), which are ideal for massive data parallelism, might be considered in our implementation to accelerate the spatio-temporal exploring and analyzing processes. As part of our future work, parallelization of pruning approach will be pursued to achieve faster outlier detection.

References

[1] Y. Zheng, X. F. Zhou, Computing with spatial trajectories, Springer, 2011.

[2] M. Wu, X. Song, C. Jermaine, S. Ranka, J. Gums, A LRT framework for fast spatial anomlay detection, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09), pp. 887–896.

[3] J. Yuan, Y. Zheng, X. Xie, G. Sun, Driving with knowledge from the physical world, in: Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'11), pp. 316–324.

[4] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, Y. Huang, T-Drive: Driving directions based on taxi trajectories, in: Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2010), pp. 99–108.

[5] W. Liu, Y. Zheng, S. Chawla, J. Yuan, X. Xie, Discovering spatio-temporal causal interactions in traffic data streams, in: KDD '11 17th SIGKDD conference on Knowledge Discovery and Data Mining, pp. 1010–1018.

[6] Y. Zheng, Y. Liu, J. Yuan, X. Xie, Urban computing with taxicabs., in: UbiComp11, September 17-21, 2011, Beijing, China, pp. 89–98.

[7] S. Chawla, Y. Zheng, J. Hu, Inferring the root cause in road traffic anomalies, in: IEEE International Conference on Data Mining (ICDM 2012).

[8] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, in: Acm Computing Surveys, volume 41, pp. 1–58.

[9] V. J. Hodge, J. Austin, A survey of outlier detection methodologies, in: Artificial Intelligence Review, volume 22, pp. 85–126.

[10] F. J. Anscombe, I. Guttman, Rejection of outliers, in: Technometrics, volume 2, pp. 123–147.

[11] E. Eskin, Anomaly detection over noisy data using learned probability distributions, in: In Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, pp. 255–262.

[12] M. Desforges, P. Jacob, J. Cooper, Applications of probability density estimation to the detection of abnormal conditions in engineering., in: In Proceedings of the Institute of the Mechanical Engineers, volume 212, pp. 687–703.

[13] R. Ng, J. H. Clarans, A method for clustering objects for spatial data mining, IEEE Trans. Knowl. Data Eng. 14 (2002) 1003–1016.

[14] J. S. M. Ester, H.-P. Kriegel, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, Data Mining and Knowledge Discovery 2 (1998) 226–231.

[15] W. Wang, J. Yang, R. R. M. Sting, Sting: A statistical information grid approach to spatial data mining, in: VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece, Morgan Kaufmann, 1997.

[16] D. B. Neill, A. W. Moore, M. Sabhnani, K. Daniel, Detection of emerging space-time clusters, in: Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05), pp. 218–227.

[17] D. B. Neill, A. W. Moore, Detection of emerging space-time clusters:prior work and new directions, Technical Report, Carnegie Mellon University, 2004.

[18] M. Kulldorff., A spatial scan statistic, Comm. in stat.: Theory and Methods (1997) 1481–1496.

[19] M. Kulldorff., Spatial scan statistics: models, calculations, and applications, In J. Glaz and N. Balakrishnan, editors, Scan Statistics and Applications,Birkhauser, 1999.

[20] M. Kulldorff, N. Nagarwalla, Spatial disease clusters: detection and inference, Statistics in Medicine (1995) 799–810.

[21] M. Kulldorff, W. Athas, E. Feuer, B. Miller, C. Key, Evaluating cluster alarms: a space-time scan statistic and cluster alarms in los alamos, American Journal of Public Health 88 (1998) 1377–1380.

[22] L. Huang, M. Kulldorff, D. Gregorio, A spatial scan statistic for survival data, International Biometrics Society (2007) 109–118.

[23] I. Jung, M. Kulldorff, A. Klassen., A spatial scan statistic for ordinal data, Stat Med (2007) 1594–1607.

[24] I. Jung, M. Kulldorff, O. Richard, A spatial scan statistic for multinomial data, Stat Med (2010) 1910–1918.

[25] L. Huang, R. Tiwari, M. Kulldorff, J. Zou, E. Feuer, Weighted normal spatial scan statistic for heterogenous population data, American Statistical Association (2009).

[26] http://www.satscan.org (2008).

[27] T. Tango, K. Takahashi, K. Kohriyama, A spacetime scan statistic for detecting emerging outbreaks, International Biometrics Society (2010) 106–115.

[28] J. H. Friedman, N. I. Fisher, Bump hunting in high-dimensional data, Stat. and Comp. 9 (1999) 123–143.

[29] D. B. Neill, A. W. Moore, A fast multi-resolution method for detection of significant spatial disease clusters, NIPS (2003) 651–658.

[30] V. Chandola, A. Banerjee, V. Kumar, Anomaly Detection: A survey, Acm Computing Surveys 41 (2009) 1–58.

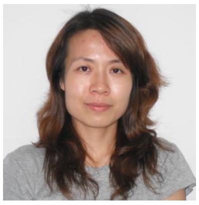
[31] http://en.wikipedia.org/wiki/likelihood-ratio-test (2011).

[32] R. Schoenberg, Maximum likelihood estimation and conservative confidence interval procedures in reliablity growth and debugging problems, Computational Economics 10 (1997) 251–266.

[33] R. E. Barlow, F. Proschan, E. M. Scheuer, Maximum likelihood estimation and conservative confidence interval procedures in reliablity growth and debugging problems (1966).

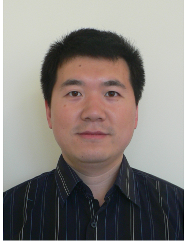
[34] R. E. Barlow, E. M. Scheuer., Reliablity growth during a development testing program, Technometrics 8 (1966) 53–60.

[35] http://en.wikipedia.org (2011).



Linsey Xiaolin Pang is currently studying Ph.D. in school of information technologies at University of Sydney. She completed her master degree from National University of Singapore. Her research interests include data mining, outlier detection and parallel computing.

Sanjay Chawla is a full professor in the School of Information Technologies, University of Sydney. He received his Ph.D. in 1995 from the University of Tennessee, Knoxville, USA. His research work has appeared in leading data mining journals and conferences including ACM TKDD, Machine Learning, IEEE TKDE, DMKD, ACM SIGKDD, IEEE ICDM, SDM, and PAKDD. He serves on the editorial board of Data Mining and Knowledge Discovery and is Program Co-Chair of PAKDD 2012. He has received four best paper awards in the last five years - most recently at the IEEE International Conference in Data Mining (ICDM) 2010.



Wei Liu is a research fellow at the University of Melbourne, Australia. He obtained his Ph.D. degree from the University of Sydney in 2011, in the area of class skewness and adversarial learning. His work has appeared in several conferences and journals including KDD, SDM, ECML/PKDD, PAKDD and Machine Learning journal. His research interests include data mining, machine learning and bioinformatics.

Yu Zheng is a lead researcher from Microsoft Research Asia (MSRA). He is an IEEE senior member and ACM senior member. His research interests include location-based services, spatio-temporal data mining, ubiquitous computing, and mobile social applications. He has published over 50 referred papers as a leading author at high-quality international conferences and journals, such as SIGMOD, SIGKDD, ICDE, WWW, AAAI, and IEEE TKDE, where he has received 3 best paper awards as well as 1 best paper nominee. These papers have also been featured by top-tier presses like MIT Technology Review multiple times. In addition, he has been serving over 30 prestigious international conferences as a chair or a program committee member, including ICDE, KDD, Ubicomp, and IJCAI, etc. So far, he has received 3 technical transfer awards from Microsoft and 20 granted/filed patents. In 2008, he was recognized as the Microsoft Golden Star. He joined MSRA in 2006 after received a Ph.D. degree in EE from Southwest Jiaotong University.