On Mining Anomalous Patterns in Road Traffic Streams

Linsey Xiaolin Pang[1,4], Sanjay Chawla[1], Wei Liu[2], and Yu Zheng[3]

[1] School of Information Technologies, University of Sydney, Australia
[2] Dept. of Computer Science and Software Engineering, University of Melbourne, Australia
[3] Web Search and Mining Group, Microsoft Research Asia, Beijing, China
[4] NICTA, Sydney, Australia
qlinsey@it.usyd.edu.au, sanjay.chawla@sydney.edu.au,
wei.liu@unimelb.edu.au, yuzheng@microsoft.com

Abstract. Large number of taxicabs in major metropolitan cities are now 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 paper we use GPS data from taxis to monitor the emergence of unexpected behavior in the Beijing metropolitan area. We adapt likelihood ratio tests (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.

Key words: Spatio-temporal outlier, persistent, emerging , upper-bounding

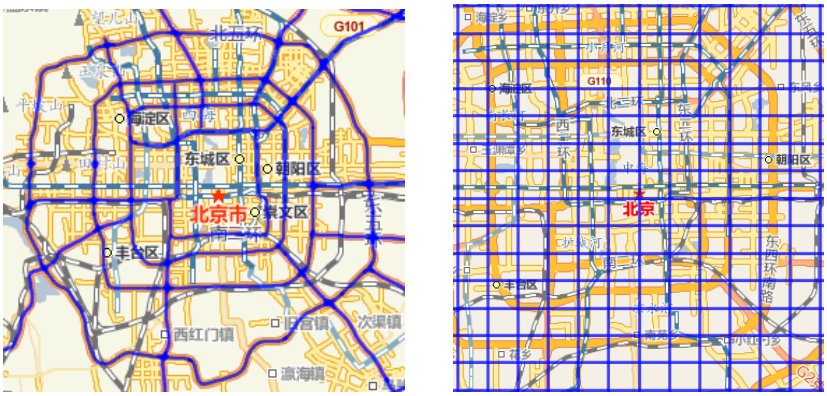
1 Introduction

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, 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 onetime 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. The huge



(a) Road Network of Beijing (b) Grid Map

Fig. 1: An example of the traffic network of Beijing. Based on the longitude and latitude, the entire city is partitioned into a grid map. Subfigure (a) is partitioned into subfigure (b).

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 [15] 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 paper, we extend the LRT framework to detect abnormal traffic pattern. More specifically, **the contributions** are: (1) A general and efficient pattern mining approach for spatio-temporal outlier detection is proposed. (2) In this work, persistent and emerging outlier detection statistical models are provided. (3) We give our proof that the upper-bounding strategy of LRT is applicable to "persistent" and "emerging" outlier detection models. (4) We conducted experiments 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 paper 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 results. Finally, section 7 concludes this work.

2 Background

2.1 Statistical Background

We provide a brief but self-contained introduction for finding the most anomalous region (rectangle) in a spatial setting . 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 distribution $P(X, \theta)$, a null hypothesis H_0 and an alternate hypothesis H_1, the LRT is the ratio

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

where $L()$ is the likelihood function. In a spatial setting the null hypothesis is that the statistical aspects of the phenomenon of interest in a region (that is currently being tested) are no different from their complement. Thus if a region is indeed anomalous then the the alternate hypothesis will most likely be a better fit and the denominator of the λ will have a higher value for θ which is a maximum likelihood estimator. 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[1] (see below). Thus regions whose Λ value likes in the tail of χ^2 distribution are likely to be anomalous.

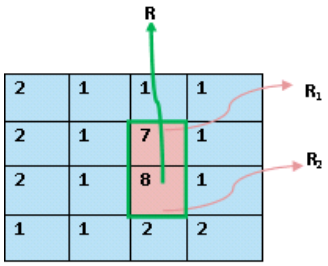


Fig. 2: An example of (4×4) grid to illustrate the LRT calculation and the upper-bounding methodology

2.1.1 Upper-bounding Methodology: The basic idea of upper-bounding methodology in LRT [15] is: if region R is composed of two non-overlapping regions R_1 and R_2, then $L(\theta_R|X_R) \leq L(\theta'_{R_1}|X_{R_1}) \times L(\theta'_{R_2}|X_{R_2})$, where θ_R is calculated by performing $MLE_1(R, f(G))$; θ'_{R_1} and θ'_{R_2} are computed directly by performing $MLE_0(f(R_1))$ and $MLE_0(f(R_2))$. And the formula is equivalent to: $logL(\theta_R|X_R) \leq \log L(\theta'_{R_1}|X_{R_1}) + \log L(\theta'_{R_2}|X_{R_2})$. Therefore, the log likelihood of any given region R can be upper-bounded.

2.1.2 Example: We generate a grid of 4×4 in Table 2. The number of successes (k_c) generated by poisson model $P_o(b_c p)$ is displayed in each cell. The baseline b_c in each cell is set to be 10. The success rate of p is 0.5 for the region R and p is 0.1 for the rest of cells. The significant level α=0.05. Here we refer the success rate of p as test parameter.

(1) The likelihood function of each cell is: $f(p|c) = \frac{(b_c p)^k e^{-(b_c p)}}{k!}$

(2) The likelihood of any given region R is: $L(p|R) = \Pi_{c_i \in R} \frac{(b_{c_i} p)^{k_i} e^{-(b_{c_i} p)}}{k_i!}$.

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

(3) The MLE_0 of p for a region R (denoted as $\hat{p}$), which is composed of cell $c_1, c_2, ..., c_t$, is calculated as: $\hat{p} = \frac{\sum_{i=1}^{t} k_i}{\sum_{i=1}^{t} b_i}$. 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) Λ of region R are calculated by the above definition:

$$\Lambda_R = (-2)\log(\frac{0.75^{15} \times e^{-0.75\times 20} \times 0.14^{19} \times e^{-0.14\times 140}}{0.21^{19+15} \times e^{-0.21\times 340}}) = 20.79$$

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, 20.79 is greater than 3.84. Therefore region R is treated as a potential outlier. As a verification of the upper-bounding strategy, we can see that 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 with -19.31.

3 Proposed Framework

3.1 Preliminaries

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 concern about the trend of the disease rate in a spatio-temporal view, the disease rate is KP. For simplicity, we only consider one parameter from the key parameter set in our work (denoted as KP).*

Definition 2. *PSTO (Persistent Spatial-Temporal Outlier): The KP is consistent throughout its duration.*

Definition 3. *ESTO (Emerging Spatial-Temporal Outlier): The KP is non-decreasing throughout its duration until it reaches the peak.*

Definition 4. *Spatio-Temporal Data Cube: A cube is the minimum spatial-temporal unit with relation to spatial and temporal perspective. The data in each cube is based on a statistical model characterized by a probability density function.*

Definition 5. *Temporal Unit: It is the time length that we apply test statistic on every time step.*

Definition 6. *Temporal Interval: It is the period that the user defines to detect outlier.*

3.2 Statistical Models

Definition 7. *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-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:*

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

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.

Definition 8. *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 [2] 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.

4 Upper-bounding Strategy and Pruning Mechanism for Proposed Framework

4.1 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.

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

$$L(\theta_R | X_R) \leq L(\theta'_{R_{t1}} | X_{R_{t1}}) \times L(\theta'_{R_{t2}} | X_{R_{t2}}) \tag{1}$$

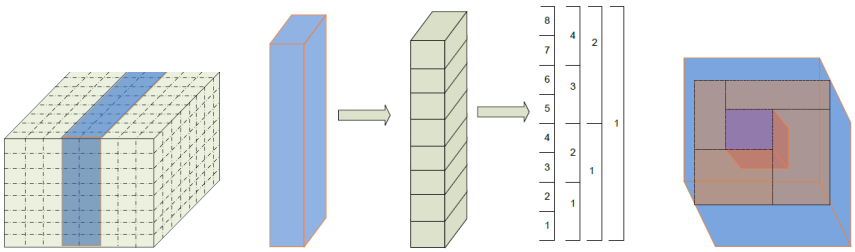
, where $\theta_R = \theta_{R_{t1}} \cup \theta_{R_{t2}}$ and $X_R = X_{R_{t1}} \cup X_{R_{t2}}$

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

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

$$L(\theta_{R1}, \theta_{R2} | X_{R1}, X_{R2}) \leq L(\theta'_{R1_{t1}}, \theta'_{R1_{t2}} | X_{R1_{t1}}, X_{R1_{t2}}) \times L(\theta'_{R2_{t1}}, \theta'_{R2_{t2}} | X_{R2_{t1}}, X_{R2_{t2}}) \tag{2}$$

,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.



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

Fig. 3: Precomputation of any given spatial-temporal region R and tiling of $\bar{R}$. Subfigure (a) shows a $8 \times 8 \times 8$ spatial-temporal grid; subfigure(b) shows: one of the cuboids from spatial precomputed set is split from temporal dimension and results in 15 smaller cuboids. Subfigure(c) is the radial method to tile $\bar{R}$

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

4.2 Precomputation and Pruning Mechanism

4.2.1 Precomputation 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. 3b shows the split approaches for a sub-cuboid highlighted as blue from the temporal dimension in a $8 \times 8 \times 8$ grid (Fig. 3a). The likelihood of any given region can be upper-bounded by this pre-computed set via the tiling of LRT.

4.2.2 Precomputation 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 precomputed set. Since there are $O(n^4)$ intersection points, there are $O(n^4)$ cuboids in the precomputed set. After we get the precomputed set, for any given region $\bar{R}$, we use the radial and sandwich methods in LRT to get the upper-bounded likelihood value of $\bar{R}$. It involves of a total of twelves times tiling in 3-dimension view. Fig. 3c shows the tiling in radial way.

4.2.3 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

Algorithm 1 Top k spatio-temporal outlier detection

Input: a spatial-temporal grid G, f, MLE_0, MLE_1, L, k and α
Output: top-k anomalous spatio-temporal regions.

1: Precompute the $O(n^4)$ cuboids for upper-bounding any given cuboid R.
2: Precompute the $O(n^3)$ cuboids for upper-bounding any given cuboid $\bar{R}$
3: Let $\theta_0 = MLE_0(f(G))$.
4: **for** Each cuboid R in the grid **do**
5: Get the upper-bounded value for log L($\theta_R|X_R$)
6: Get the upper-bounded value for log L($\theta_{\bar{R}}|X_{\bar{R}}$)
7: Combine the results of step 3, 5, 6 to an upper bound for Λ_R
8: Check upper-bounded value of Λ_R from chi-square distribution
9: **if** The Λ_R is in the α level and less than the kth best **then**
10: Prune R
11: **else**
12: Compute real Λ_R;
13: **if** Λ_R is in the top k, **then**
14: Remember R
15: **end if**
16: **end if**
17: **end for**

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 won't be affected no matter which significance testing method is 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 precomputations; 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.

5 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 this section all experiments were carried out on synthetic data. In Section 5.3, we will demonstrate the usefulness of our approach on a real data set.

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)

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 1: 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 2: Average Pruning Rate in Scenario III

5.1 Evaluations on Synthetic Data

We generated data set on a grid size varying from $(4 \times 4 \times 4)$ to $(128 \times 16 \times 16)$. 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 $\alpha = 0.05$.

5.1.1 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 1.

5.1.2 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 $(5 \times 4 \times 3)$ is generated by a normal distribution with different parameter setting ($\mu = 10^5, \sigma = 5 \times 10^3$). Results are shown in Table 1.

Analysis: The results of Scenario I and II show that the we achieve a high pruning rate and almost no false alarm 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.

5.1.3 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 $(5 \times 4 \times 3)$ is generated from a Poisson distribution with $p = 3, 6, 9, 18, 36$ for emerging case and $p = 3$ for persistent case. The data not within the cuboid area was also from a Poisson distribution with $p = 1$. Results are shown in Table 2.

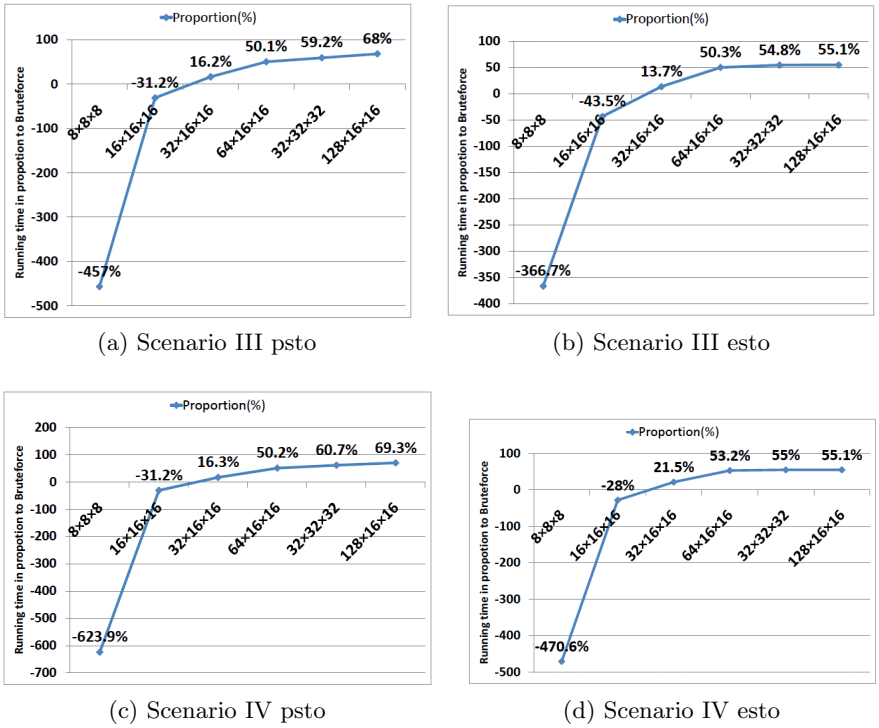
5.1.4 Scenario IV The alternative hypothesis holds. It is similar to scenario III except that data of a randomly selected cuboid area of size $(5 \times 4 \times 3)$ was generated from a Poisson distribution with $p = 10, 50, 250, 1250, 6250$ for emerging case and $p = 10$ for persistent case. Results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

Analysis: 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.

5.1.5 Running Time We analyze the running time with and without pruning for Scenario III and IV. Figure 4 shows that as the size of the spatial and temporal region

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 3: Average Pruning Rate in Scenario *IV*



(a) Scenario III psto

(b) Scenario III esto

(c) Scenario IV psto

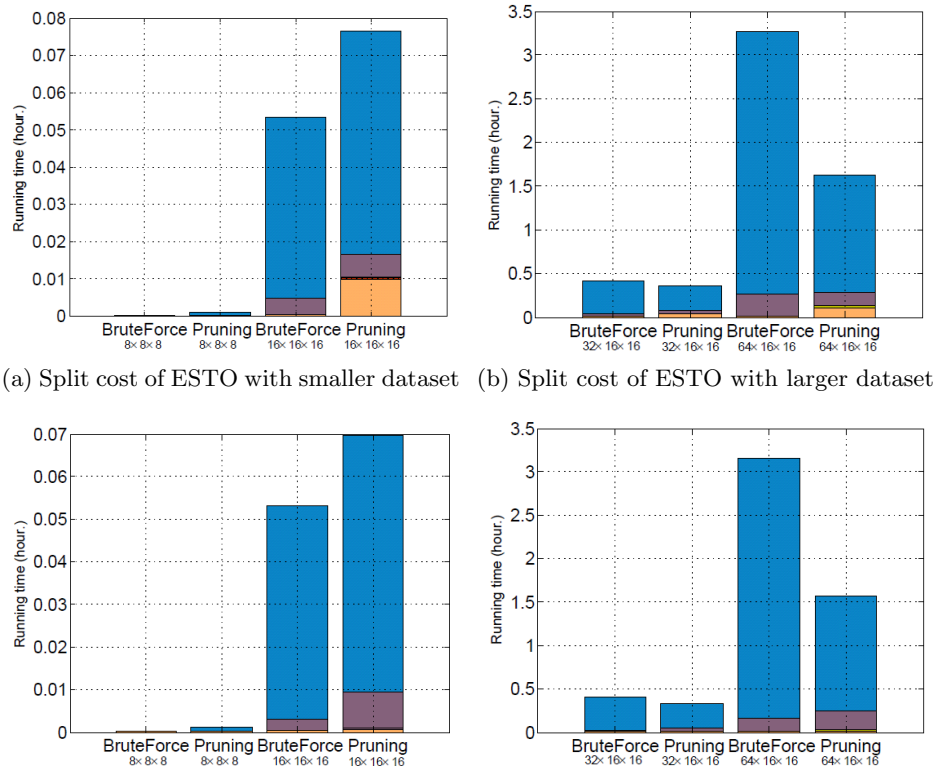
(d) Scenario IV esto

Fig. 4: 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.

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. We have also calculated the cost of a single likelihood calculation as the dimension of the grid size increases. 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 does 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 $32 \times 32 \times 32$.

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 Figure 5. The brute-force approach has the following components:

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



(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

Fig. 5: 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.

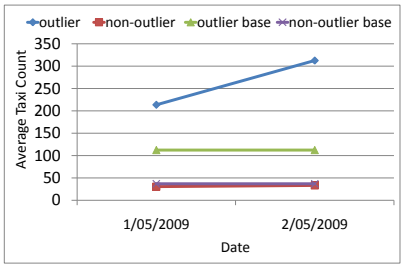
The pruning approach is more complex and involves the following components:

(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 precomputation)
(2) The cost of computing the likelihood for each element of the tiling set $T_{\bar{R}}$ ($\bar{R}$ precomputation).
(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 Figure 5, this is the most expensive part of the calculation. However, as the data set size

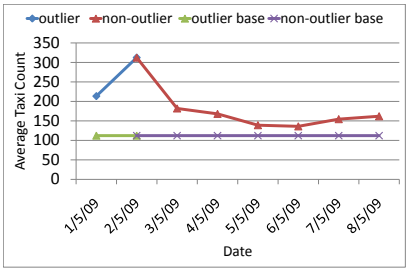
increases, the overheads of the tiling give way to its more efficient reuse resulting in considerable savings.

5.2 Case Study: Beijing Taxi GPS Data

We illustrate the use of the Pesto method on a real data set [4], [8], [3]. 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. We search for the most anomalous region and the time period and then provide a possible explanation for the anomaly.

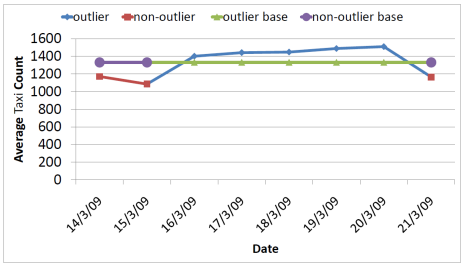


(a) The average taxi counts within outlier regions vs. non-outlier regions from 01/05/2009 to 02/05/2009



(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

Fig. 6: 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.

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.

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 truncated between May 1st

(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)

Fig. 7: Outlier Locations from our two case studies on Beijing Map

and May 3rd. 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. 6a, 6b, 7a. We can see that the number of taxis increased from 1st May to 2nd May and then decreased from 3rd of May onwards.

Case II: All (8×8) grids were tested. The temporal unit is from $3:15:00pm$ to $4:30:00pm$ every day for 8 days. Twelve days of data was used to calculate the baseline.

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 13/0/20093, the 11th National People's Congress (i.e. NPC) was held in Beijing, which is the annual meeting of the highest 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 traveling on Tonghuihe North road increased until most of the deputies left Beijing. The results are shown in Fig. 6c, 6d, 7b.

6 Related Work

Among the various methods for discovering outlier, the spatial and space-time scan statistic, introduced by Kulldorff [9–11, 14], 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 [12, 6, 7, 13]. They have been implemented in the software (SaTScan)[1]. In the space-time scan statistic of Kulldorff, the key parameter is assumed as consistent over time. The technique simply applies time as one more dimension. Niell et al. [5] points out the distinct feature of time aspect and proposes a modified test statistic to detect localized and globalized emerging cluster . Tango et al. [16] 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. [15] 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 that can greatly reduce the number of likelihood ratio tests. However, it is used for spatial anomaly detection without considering the temporal property. Wei et al. [3] propose an approach to discover casual relationships among spatio-temporal outliers. Here, we only focus on detecting spatial-temporal outliers.

7 Conclusions

In this paper, we proposed an efficient pattern mining approach catered for spatio-temporal traffic data, which is able to detect "persistent outliers" and "emerging outliers". We also derived an upper-bounding strategy for the two statistic models. Experiments show that the performance of computational time is greatly improved when the dataset size is very large, and we can still find the correct outliers. 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.

References

1. http://www.SatScan.org.
2. R.E Barlow, E.M Scheuer. Reliability Growth During A Development Testing Program. *Technometrics*, 1966.
3. W. Liu, Y. Zheng, S. Chawla, J. Yuan and X. Xie. Discovering spatio-temporal causal interactions in traffic data streams. In *KDD '11 17th SIGKDD conference on Knowledge Discovery and Data Mining*, pages 1010–1018, 2011.
4. Z. Chen, H. T. Shen, X. Zhou, Y. Zheng, and X. Xie. Searching trajectories by locations: an efficiency study. In *Proceedings of the 29th ACM SIGMOD International Conference on Management of Data (SIGMOD'10)*, pages 255–266, 2010.
5. 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)*, pages 218–227.
6. I. Jung, M. Kulldorff and AC. Klassen. A spatial scan statistic for ordinal data. *Stat Med*, pages 1594–1607, 2007.
7. I. Jung, M. Kulldorff and OJ. Richard. A spatial scan statistic for multinomial data. *Stat Med*, pages 1910–1918, 2010.
8. 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 Conference on Advances in Geographical Information Systems*, pages 99–108, 2010.
9. M. Kulldorff. A spatial scan statistic. *Comm. in stat.: Theory and Methods*, pages 1481–1496, 1997.
10. M. Kulldorff. *Spatial scan statistics: models, calculations, and applications.* In J. Glaz and N. Balakrishnan, editors, Scan Statistics and Applications,Birkhauser, 1999.
11. M. Kulldorff and N. Nagarwalla. Spatial disease clusters: detection and inference. *Statistics in Medicine*, pages 799–810, 1995.
12. L. Huang, M. Kulldorff, and D. Gregorio. A Spatial Scan Statistic for Survival Data. *International Biometrics Society*, pages 109–118, 2007.

13. L. Huang, R. Tiwari, M. Kulldorff, J. Zou, and E. Feuer. Weighted normal spatial scan statistic for heterogenous population data. *American Statistical Association*, 2009.
14. M. Kulldorff, W. Athas, E. Feuer, B. Miller, and C. Key. Evaluating cluster alarms: a space-time scan statistic and cluster alarms in los alamos. *American Journal of Public Health*, 88(9):1377–1380, 1998.
15. 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)*, pages 887–896.
16. T. Tango, K. Takahashi, and K. Kohriyama. A SpaceTime Scan Statistic for Detecting Emerging Outbreaks. *International Biometrics Society*, pages 106–115, 2010.