Yang and Japkowicz *J Big Data* (2018) 5:23
https://doi.org/10.1186/s40537-018-0133-8

Journal of Big Data

RESEARCH Open Access

Anomaly behaviour detection based on the meta-Morisita index for large scale spatio-temporal data set

Zhao Yang and Nathalie Japkowicz*

*Correspondence:
japkowic@american.edu
Department of Computer Science, American University, Washington D.C., USA

Abstract

In this paper, we propose a framework for processing and analysing large-scale spatio-temporal data that uses a battery of machine learning methods based on a meta-data representation of point patterns. Existing spatio-temporal analysis methods do not include a specific mechanism for analysing meta-data (point pattern information). In this work, we extend a spatial point pattern analysis method (the Morisita index) with meta-data analysis, which includes anomaly behaviour detection and unsupervised learning to support spatio-temporal data analysis and demonstrate its practical use. The resulting framework is robust and has the capability to detect anomalies among large-scale spatio-temporal data using meta-data based on point pattern analysis. It returns visualized reports to end users.

Keywords: Spatio-temporal data, Point pattern, Data mining, Unsupervised learning, Morisita index

Introduction

Anomaly detection for analysing spatio-temporal data remains a rapidly growing problem in the wake of an ever-increasing number of advanced sensors that are continuously generating large-scale datasets. For example, vehicle GPS tracking, social media, financial network and router logs, and high resolution surveillance cameras all generate a huge amount of spatio-temporal data. This technology is also important in the context of cyber security since cyber data carries with it an IP address which can map to a specific geolocation and a timestamp. Yet, current cybersecurity approaches are not able to process this kind of information effectively. To illustrate this deficiency, consider the scenario of a distributed denial-of-service (DDoS) attack in which the network packets may come from different IP addresses with sparse locations. In such a case, a spatio-temporal analyzing system [1] is required to analyse the spatial pattern of the DDoS attack. Yet, user oriented analytic environments for cyber security with spatio-temporal marks are currently limited to traditional statistical methods like spatial-temporal outlier detection and hotspot detection [2].[1] Furthermore, much of the current work in large scale

[1] A spatial outlier refers to a point whose non-spatial attribute values are significantly different from the values of their spatial neighbors [3]. A hotspot refers to points that show intermittent spatial repetitiveness [4].

analytics focuses on automating analysis tasks, such as detecting suspicious activity in a wide area motion and time interval. But these approaches do not provide analysts of cyber security data with spatio-temporal marks the flexibility to employ creativity and discover new trends in the data while operating over extremely large datasets. Current solutions are prohibitive because they require a multidisciplinary skillset.

One possible solution to performing analytics on such large scale spatio-temporal data is to retrieve the metadata of spatial point patterns [5], and apply metadata processing and storage approaches [6], together with domain knowledge derived by machine learning and statistical means. An added advantage of this method is that meta-data hides the details of the point patterns thus providing privacy while still supporting a variety of analytics.

We, thus, propose a framework for performing analytics with spatio-temporal data that has the following properties:

- Privacy protection: We use a meta analysis of tracking data as an indicator of subjects' behavior. The geolocation of the subject will not be exposed to the system user.
- High scalability: We are able to retrieve the behavior pattern for different amounts of data since the Morisita index provides the scalability adapted to different amounts of tracking data.
- Convenience: We designed a convenient way to map the anomaly event of the cyber threat to the physical threat since the cyber threat can be visualized on the real map.

In more detail, we propose a framework to store and process large-scale spatio-temporal data over a "metadata based point pattern" infrastructure, while providing users with a metadata analysis that hides the details of large-scale spatio-temporal data and provides them with a front-end interface that allows them to run a variety of security checks including outlier detection for a single subject, anomaly group detection, anomaly behavior detection and anomaly event detection. Furthermore, the spatio-temporal data is stored in various data stores. As a result, this framework provides high-performance analytical features, flexibility, and extensibility.

The theoretical contribution and novelty of our work lies in the combination of methods from the areas of spatio-temporal analysis, machine learning and statistical analysis. By extracting relevant methods from these three fields of research, we created an effective and efficient tool for anomaly detection by monitoring the cyber and physical levels, simultaneously.

Background

Spatio-temporal data differs from traditional data since both spatial and temporal attributes are available in addition to the actual measurements/attributes.

In this work, we treat the spatio-temporal data as a time series of spatial data. We are using spatial point patterns as the "snapshot" of spatial data within a specific time frame. The Morisita index has been used as the measure of the spatial point patterns.

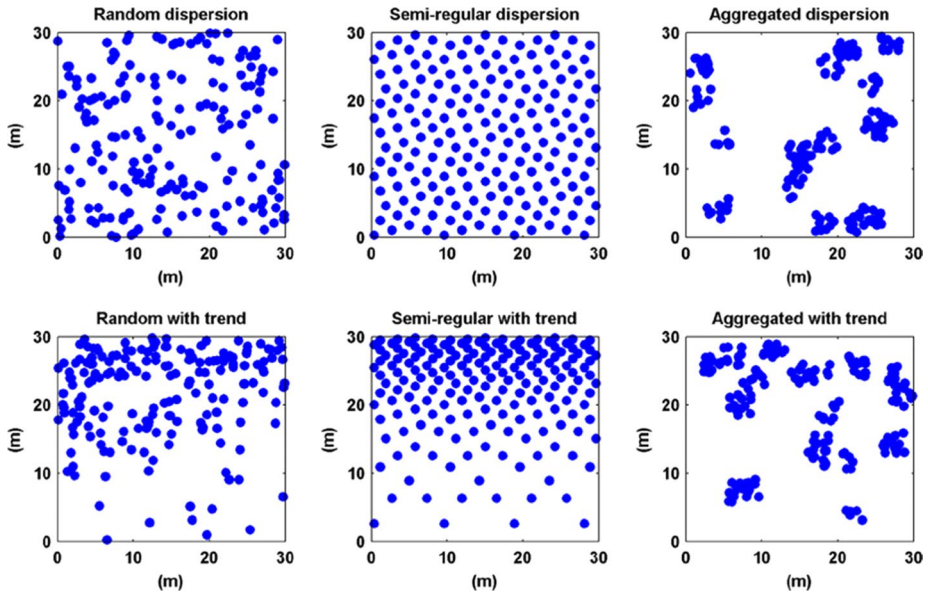

Fig. 1 Spatial point patterns [5]

Spatial point patterns

Spatial point patterns can be stored in a two-dimensional data format to which a variety of analytical methods can be applied to discover useful data patterns in large-scale spatial data. Extracting exact patterns from geospatial data is more complicated than doing so with ordinary data sets because of the nature of geospatial data sources and their associated data structures, which refers to the two or three dimensional data structure.

Figure 1 represents 6 possible spatial distributions of points: random, semi-regular, aggregated, random with a density trend, semi-regular with a density trend, aggregated with a density trend [5].

Common spatial analysis packages use real numbers [e.g. geospatial POIs (Point of Interest) with uncertainty information] like GPS navigation data with errors, categorical values (e.g. fishery production by species) and logical values (e.g. saline water/freshwater) to mark the point patterns [7]. These point patterns are composed of a huge number of points that follow distributions such as those illustrated in Fig. 1 [8]. The region of spatial data can represent a complicated shape, such as an arbitrary polygon or a irregular pixel image pattern. In this work, the spatial data-mining functions are implemented in R.

Morisita index

The Morisita index is a common method for analyzing spatial patterns. It is a statistical measure of dispersion based on the spatial Poisson process. To compute the Morisita index, the function will first calculate the quadrat counts of the spatial point pattern.[2] Then the generated index of spatial aggregation for the previous pattern will be the

[2] The term quadrat in ecology and geography means a plot to isolate a standard size of area to study the distribution of an item over a large area [9].

Morisita index. In more detail, the algorithm divides the spatial domain into Q quadrats of equal size and shape.

Then the algorithm counts the number of points falling in every quadrat. Finally, the $n[i]$ number of points in the ith quadrat will be counted as a vector of values called the Morisita index. The sum of the number of points is represented as the N value.

We can also plot the result of this analysis and the Morisita index of dispersion can calculate the overlap between samples.

The formula used in the analysis package is

$$MI = Q\frac{\sum_{i=1}^{Q} n_i(n_i - 1)}{N(N-1)} \tag{1}$$

n_i: number of points in the ith quadrats, N: total number of points [10].

This formula is based on the assumption that increasing the size of the samples will increase the diversity because it will include different habitats (i.e. different faunas) [11]. The Morisita index is used to compare the similarity between different samples. The advantage of the Morisita index is that it is a vector of data that only varies with size of quadrats, not with population density. For more information about the statistical description of the Morisita index, please see [11].

Related work

In his seminal book "statistics for spatio-temporal data [12]", Cressie et al. characterizes the process of statistical spatio-temporal data analysis in the presence of uncertain and (often) incomplete observations. This work includes prediction in space (interpolation), prediction in time (forecasting), assimilation of observations and mechanistic models and inference on controlling process parameters. The concept of the poisson point process in the book is also the foundation of our research which relies on the Morisita index. However the Morisita index was originally designed for ecological research by Morisita [11]. The method has been implemented by Baddeley et al. [7] in R to analyze spatial point pattern data. Our work intends to extend Baddeley et al. [7] work to spatio-temporal data type by retrieving characteristic value using Morisita index.

Some pilot study in spatial statistics like Kriging [13, 14] are methods of interpolation of spatial data. The values interpolated conform to the Gaussian process. The meta-Morisita index is different as it actually calculates the density of the clusters in each quadrats. Although the two methods appear related, they have different functions. The function of Kriging is to interpolate the predicting points into the insufficient (usually undersized) geospatial data set, which may contains missing points and irregular spatial objects like polygons. These predicted values are generated by the model of spatial autocorrelation. The function of Morisita index is to exploratory analyze the large scale (usually oversized) geospatial data set by different measures (defined by the diameter of the quadrats). That means, the Kriging method predicts the unknown values (making a prediction) [15]. The Morisita index does not insert any extra values into the raw geospatial data set.

Other spatial statistical models have been established for spatio-temporal processes and spatial point processes. Examples of temporal models of point process include Cox

and Isham [16]; Daley and Vere-Jones [17]. Examples of spatial models of point process include Cressie [18]; Diggle [19]; MØller and Waagepetersen [20]. The model of spatio-temporal process is not as well defined a field as the spatial model. Pioneer work in this area includes Diggle [21], Diggle and Gabriel [22]. Literature reviews like Zhuang et al. [23] can be used as references for this field. Some work like Illian et al. [24] introduce models such as goodness-of-fit tests, calculation of summary statistics to process the single realization of the point process.

Our study, however, is functionally closer to clustering work in Machine Learning than the spatio-temporal analyses just mentioned. Performance comparisons will be described in the "Results and discussion" section on quantitative evaluation. Typical clustering algorithms in machine learning including K-means [25], density-based spatial clustering of applications with noise (DBSCAN) [26], expectation maximization (EM) [27][3] are efficient methods to detect the cluster of spatial point patterns.

The advantage of K-means and its derivatives K-mediods [28], CLARANS (Clustering Large Applications based on RANdomized Search) [29], K-modes [30], ISODATA (Iterative Self-Organizing Data Analysis Technique) [31], FCM (fuzzy-c-means) [32] is scalability for large data and efficiency. However the k value is difficult to predict. The expectation maximization (EM) also has disadvantages—the convergence is slow and not capable to provide estimation of the asymptotic variance-covariance matrix of the maximum likelihood estimator (MLE). The density based algorithms like DBSCAN (density-based spatial clustering of application with noise) and their derivatives [26], GDBSCAN [33], DBRS [34], ST-DBSCAN [35], OPTICS (ordering points to identify the clustering structure) [36] have some advantage like to detect the outlier efficiently. However it is hard to set the global parameter. Also DBSCAN is not precise enough to measure the clusters adjacent to each other (neck problem).

We now describe how machine learning and clustering has previously been applied to spatio-temporal analyses and contrasts these works with our proposed approach. Yang et al. [2] highlights the demands of analysing human mobility data and detecting hot spots. The author proposes a framework to identify human mobility hotspots that represent the status of human mobility in local areas and group these hotspots into different classes by clustering their temporal signatures. Their work focuses on converting spatio-temporal data to convergent hotspot and dispersive hotspot. Clustering analysis follows based on the temporal characteristics. Their work focuses on the trajectory of human mobility. It models behaviour but it does so in a short time window such as one hour. Izakian et al. [37] considers Fuzzy C-means (FCM) as a conceptual and algorithmic setting to deal with the problem of anomaly detection. Their work is also based on small size of time series which contains only 10 data points. Our work uses 1 day as the time to calculate the behaviour indicator, which can represent the pattern of the subject over a long period of time and large scale data set.

Birant et al. [38] proposes a three-step approach: clustering, checking spatial neighbours, and checking temporal neighbours to detect spatio-temporal outliers in large databases. Cheng et al. [39] proposes a multiscale approach to detect the spatio-temporal

[3] EM is a much broader statistical estimation method than simple clustering algorithm. It is a general modelling process which can be applied to clustering.

outliers by evaluating the change between consecutive spatial and temporal scales. Their work is based on classification, aggregation, comparison, verification. These two works both focus on detecting the spatio-temporal outliers in large data set. They do not analyse the behaviour pattern as our framework does.

Saligrama et al. [40] proposes a novel graph-based statistical notion called MAX-LCS (local neighborhood-based composite scores) that unifies the idea of temporal and spatial locality. Their work focuses on detecting local anomalies but not for large scale group detection, as in our work.

Young et al. [41] proposes scalable time-series models so that geographically aggregated call volume can accurately identify the onset of major events when the approximate time and location of the event is known. Their work is based on known event. Our work, which is based on unsupervised learning, does not require any prerequisite information.

Liu et al. [42] presents a new data-driven framework for a spatiotemporal feature extraction scheme built on the concept of symbolic dynamics for discovering and representing causal interactions. The extracted spatiotemporal features are then used to learn system-wide patterns via a restricted Boltzmann machine (RBM). Their work is implemented on an energy system with intelligent sensing and control systems. The limit of their work is based on anomalies of sensor data. They don't design the behaviour indicator for tracking data like our work.

Capdevila et al. [43] discusses the mining events using the twitter data. The author proposes the Warble, which is a new probabilistic model and learning scheme. Their work focuses on event detection by probabilistic model. The origin of our work is based on spatial analysis. It is a different way to analyze geospatial data. Anagnostopoulos et al. [44] also discusses the twitter data. However this paper focused on targeted outdoor advertising.

Pappalardo1 et al. [45] highlights the Ditras (DIary-based TRAjectory Simulator), which is a framework to simulate the spatio-temporal patterns of human mobility. The author proposes the framework to identify human mobility by diary and trajectory generators. Their work focuses on the statistical properties of real trajectories. Our work is not based on the trajectories but on the characteristic value from the statistical model.

The framework design

The purpose of this section is to present the essential elements of our study: the characteristics of the method on which our work is based; the data sets we employed to demonstrate the usefulness of these characteristics and the framework we built to exploit the basic method to the fullest. In more detail, we first describe the way in which the Morisita index can be used to detect anomalous events on a synthetic example. We then describe the data sets on which we conducted our actual study. The first one of these data sets contains a large number of taxi trajectories over a week. It is a physical security problem chosen to illustrate the usefulness of our network on a large data set. Most of our subsequent analysis was conducted on that data set. The second data set is based on Twitter data and contains spatio-temporal information about tweets. As such, it is a cyber dataset and shows how our framework allows cyber and physical security to be considered jointly, as it should be in order to improve real safety. However, the individual

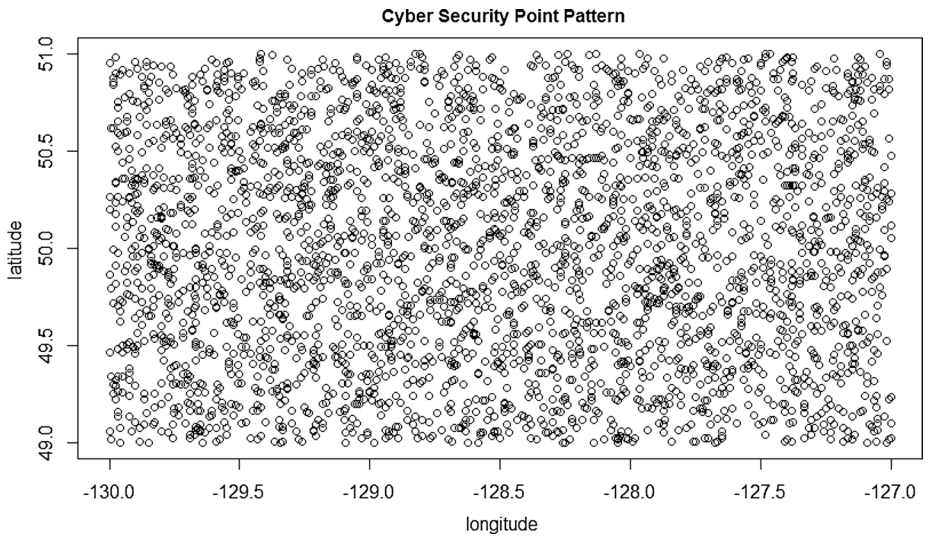

Fig. 2 Crowd with tracking device uniformly distributed in metro area

Twitter user information is not available from the data set which restricts the type of analysis that can be conducted on it. This is why the Taxi data set was used to demonstrate the versatility of our approach. The third part of this section introduces our framework and explains its functionality.

In this work, we use the spatstat [7] package which is capable of analysing three or more dimensional point pattern datasets. This spatial analysis package supports a variety of statistical analysis methods such as model fitting, spatial data sampling, and statistical formulation. The particular method used in this work is the Morisita index value which was presented in "Morisita index" section. Other models will be explored in future work.

In this study, we calculate the highest Morisita index value as the behaviour indicator in a long time slot (1 day) to avoid sampling bias. If the data is mostly distributed uniformly, the Morisita index value falls between 0 and 1. However, if the data is in clumped distribution then the Morisita value falls between 1 and n [10]. n is the maximum value of the Morisita index which means the density of points is > 1 suggest clustering.

Morisita index as the real-time clustering indicator

In this study, we use the Morisita index method to process the point pattern before detecting an anomalous event. The Morisita index is designed to determine the density of point patterns according to their statistical characteristics; the method is extensively applied to classify and visualize data with geolocation. We chose this method to indicate the point patterns of crowds of people and mark each situation using the corresponding Morisita value to indicate the density of the crowd. Thus, the process includes methods from computational statistics. Given that we want to indicate that these point patterns belong to different social events according to the statistical characteristics of their density values, the method will return Morisita values as a density value. For example, the density value of a downtown social event yields a large Morisita value, whereas the density value of a suburban social event yields a small value. For each social event, the

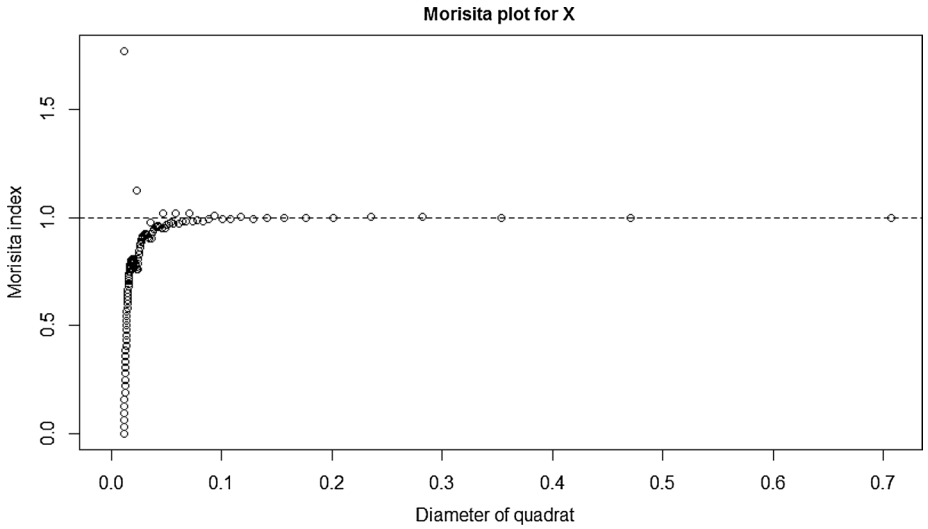

Fig. 3 Morisita plot shows crowd with tracking device gathering in the downtown area

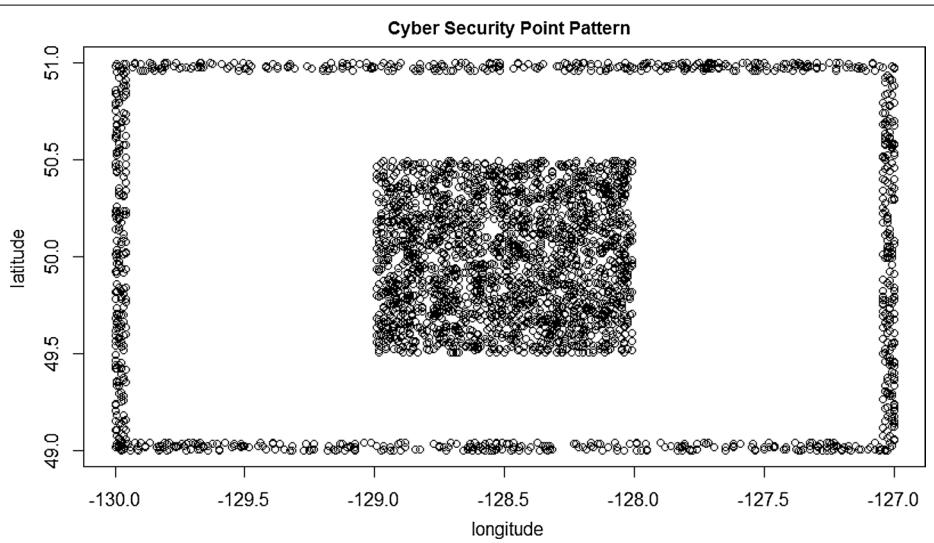

Fig. 4 Morisita plot shows that crowd with tracking device uniformly distributed in metro area

density value is correlated to the Morisita value in indicating the social event. This is illustrated in Figs. 2 and 3.

Figure 2 shows that a crowd with tracking devices is uniformly distributed in the metro area. Figure 4 shows the Morisita plot corresponding to Fig. 2. In this plot, the Morisita value falls into the [0, 1] range. That means that the point pattern is close to the uniform distribution. Figure 2, therefore, shows that the people with tracking devices are uniformly distributed.

Figure 5 shows a crowd with tracking devices gathering in the downtown area. Figure 3 shows the Morisita plot corresponding to Fig. 5. The highest value of the Moritisa plot is the highest density of the point pattern. The maximum value of Morisita index is close to 8. It shows that people with tracking devices are gathering in some area.

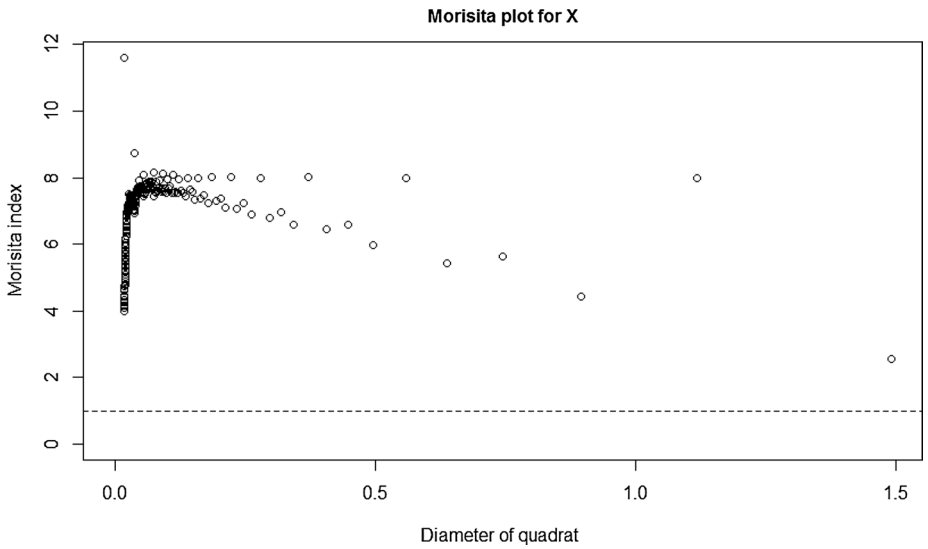

Fig. 5 Crowd with tracking device gathering in the downtown area

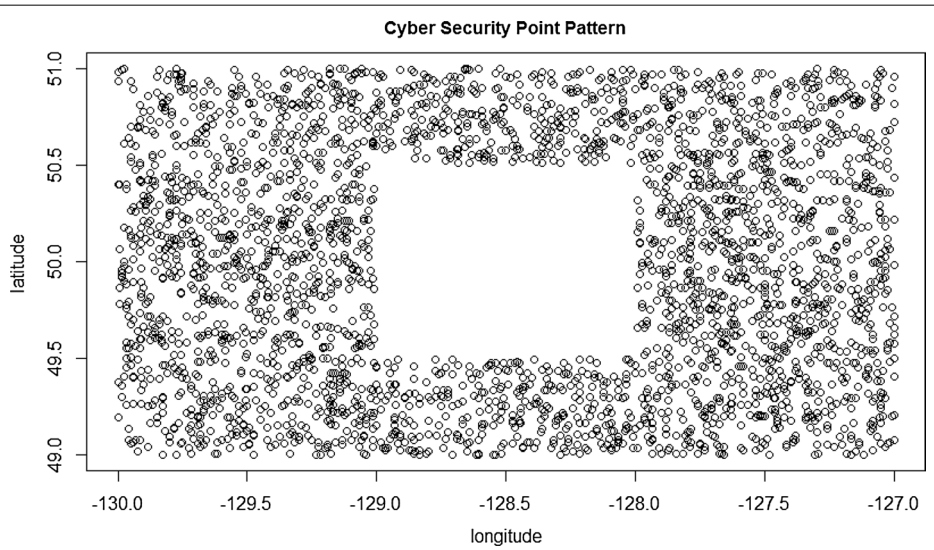

Fig. 6 Morisita plot about crowd gathering in the suburban area

For the downtown case, the maximum of Morisita index values climbs up to 8, which means our raw data was clumped together. For the suburban case, the maximum of Morisita index values falls down to 1.5. The value is significantly smaller than the maximum value 8 in Fig. 3. The result means that the point pattern in this case was not as clumped as in the downtown case.

Figure 7 shows a crowd with tracking devices gathering in the suburban area. Figure 6 shows the Morisita plot corresponding to Fig. 7. The highest value of the Moritisa plot is the highest density of the point pattern. The maximum value of Morisita index is close to 1.5.

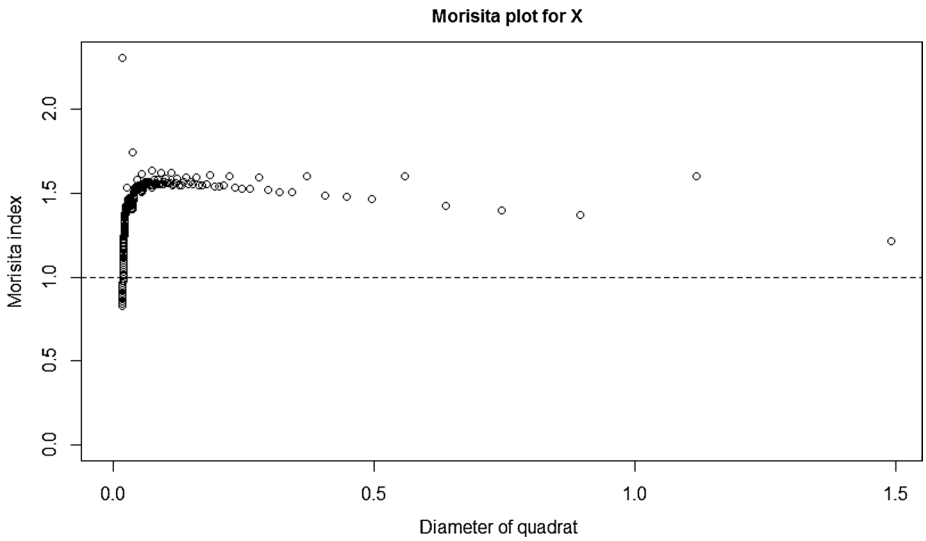

Fig. 7 Morisita plot about crowd gathering in the suburban area

It also shows that the crowd with tracking devices in Fig. 7 was not as clumped as in Fig. 5.

The Morisita index is superior for comparing the similarities between different samples. If the data is mostly uniformly distributed, the Morisita index value falls between 0 and 1. However, if the data is in a clumped formation, then the Morisita value falls between 1 and N [7]. N is the highest positive number that indicates the highest degree of density. In this case, we use our framework to get the distribution and point pattern of the large-scale tracking data with spatio-temporal marks. Users should check the Morisita value as the real time indicator to determine if it is a gathering event or not. The gathering can be physical or cyber (with IP address mapping). This example shows the function that can help users find social events from point patterns.

Data sets

The first data set is a sample of the T-Drive trajectory dataset from Microsoft Research [46, 47] that contains a trajectory of 10,357 taxis from 02/02/2008 to 02/08/2008. The total number of points in this dataset is about 15 million and the total distance of the trajectories reaches 9 million km. Taxi drivers are experienced drivers who can usually drive around the metro area. The taxis with tracking devices are mobile sensors probing the behaviour pattern of the subject. So, the taxi tracking record contains the information of both the spatio-temporal pattern and their behaviour patterns.

The second dataset is a twitter data set. It contains data derived from spatio-temporal information of tweets originated from the city of Milan during the months of November and December 2013 [48]. There is no content of the tweets included in this data set.

The simulated data set was used to indicate the behavior of the Morisita index on specific spatial point patterns. The taxi and twitter data set were used to show the Morisita index values generated by the spatio-temporal data.

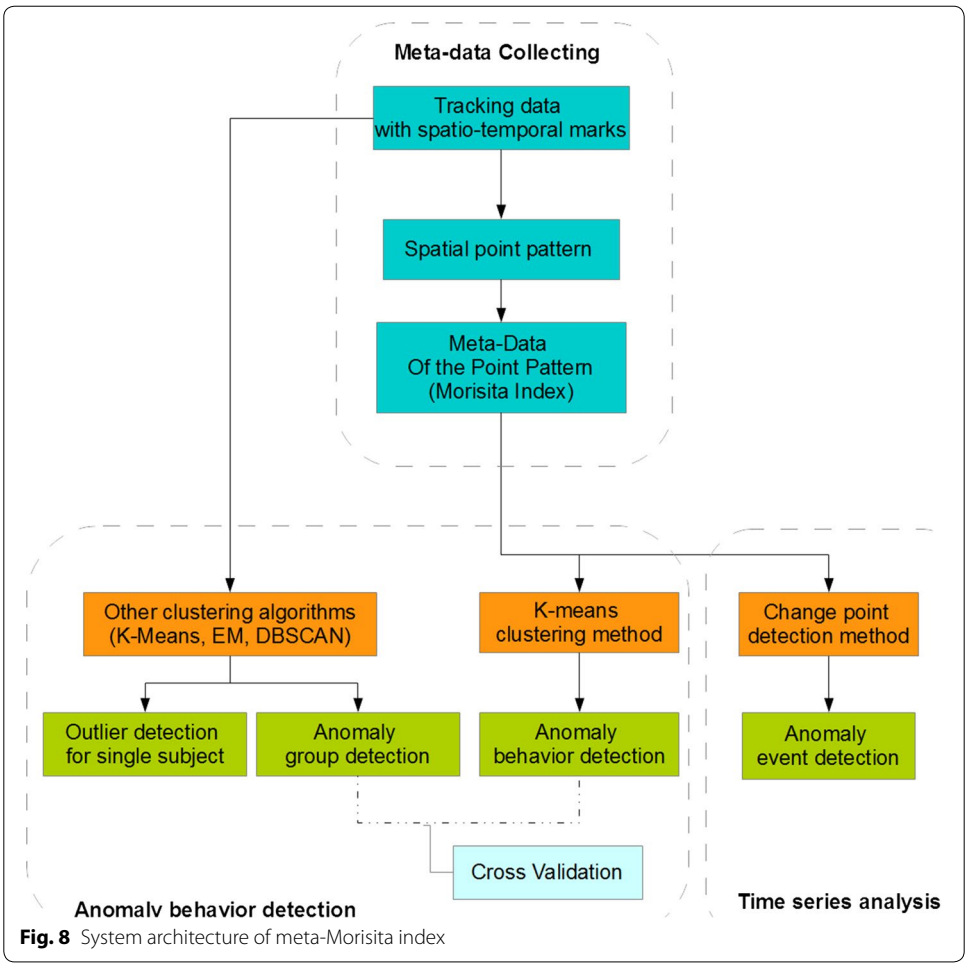

Fig. 8 System architecture of meta-Morisita index

The taxi and twitter datasets are both typical large scale spatio-temporal datasets. The statistical anomaly event detection and interpretation of new trends and spatio-temporal pattern changes in sequences of social and political events are hidden behind the large scale data. The taxi data is a typical tracking data set with spatio-temporal marks. The twitter data is a typical social media data set with spatio-temporal marks.

Meta-Morisita index architecture

Figure 8 shows the system architecture of the meta-Morisita index based framework. As shown in the figure, we propose a framework able to handle three different tasks: outlier detection for a single subject, anomaly detection for a group of subjects and anomalous social event detection. The outlier detection for a single subject means that the system can detect some anomaly behaviour for a single person, for example, frequent credit card charges in an anomalous location. The anomaly detection for a group of subjects means that the system can identify anomalous subjects by their behaviour, for instance, the person who posts larger amounts of tweets than other people. The anomalous social event detection means that the system can detect social events, such as, network flow burst.

We will now describe each of the components of the flow chart of Fig. 8.

During data processing, we extract the highest Morisita value of taxi drivers for each day from the dataset. The Morisita value which is the indicator of density value can be treated as the indicator of behaviour.

First, to identify the outlier, a box plot is employed to generate an outlier list based on the box plot of Morisita values. Second, we employ a clustering method to classify taxi drivers into several groups according to the statistical characteristics of the Morisita value. A K-means algorithm is designed to extract the specific locally convergent and dispersive drivers from the meta-data of point patterns. Finally, we execute a time series analysis using a change point detection algorithm to extract the change point of human convergent and dispersive behaviours from the meta-data of point patterns. We now describe the data processing and anomaly detection parts of our framework in more detail.

Data preprocessing

For a taxi driver, we can construct an individual's behaviour pattern by analysing the point pattern in the time sequence, where the longitude, latitude, and timestamp of a mobile tracking subject and the pattern represent the time when the spatio-temporal points are updated. For two adjacent Morisita values of the records, the time interval is 1 day. We can identify the density of the point pattern during the interval. The Morisita value as the behaviour indicator can be extracted from the spatio-temporal data within every day. For example, the first Morisita index value was collected on 02/02/2008, and the second record was collected on 02/03/2008; we can extract one time series of Morisita value during 02/02/2008–02/08/2008. The time attributes of the two adjacent records are considered to be the time slot. Thus, we can extract the flow matrices (time slots) of Morisita value for 1 week. We use the time series of the Morisita index values to denote the behaviour pattern of taxi drivers in 1 week. For each day, we calculate the Morisita value of each taxi driver, which represents the behaviour pattern of the taxi drivers' moves from day to day during that week. We define the time series of the Morisita value as the largest difference observed during that week. The time series thus provides a behaviour pattern for the taxi driver during the week. The variation of Morisita values during a week can reveal the potential function of the anomaly detection. The clustering analysis of the Morisita values can be used as in the following section to identify convergent and dispersive behaviours.

Anomaly behaviour detection

Box plot method for outlier detection in a single subject In order to detect anomalies in a single subject's behavior, we used a box plot and studied the outliers in that plot.

Clustering method for anomaly group detection For anomaly group detection, we used a clustering method, namely k-means [49]. k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. In this work, k-means was used with the value $k = 4$ which provided the best visualization. Since the

Table 1 Saptio-temporal tracking record of taxi driver

Taxi_ID	TimeStamp	Longitude	Latitude
39	2008-02-02 13:37:30	116.29369	39.92272
39	2008-02-02 13:40:17	116.28015	39.92321
39	2008-02-02 13:45:17	116.28065	39.92330
39	2008-02-02 13:49:15	116.28012	39.92327
39	2008-02-02 13:54:15	116.30334	39.92250
39	2008-02-02 13:59:15	116.31801	39.93740

box plot approach could also be used to detect anomalies in a group situation, we validated the results obtained by k-means with the output of the box-plot.

Time series analysis for social event detection In order to detect social events, we performed time series analysis using the PELT (Pruned Exact Linear Time) [50, 51] algorithm which is a change point detection method. The PELT algorithm is a multiple change point method that is both computationally efficient and flexible in its application [52]. It has been shown that under certain conditions, especially when the number of change points is increasing linearly with n, the computational efficiency of PELT is $O(n)$. As a result, we can detect change points efficiently from the time series of the behaviour pattern of our data sets.

Experimental methods

We applied our full framework to the taxi driver data set and illustrated its usefulness on the Twitter data. The first two parts of this section describe our analysis on individual taxi drivers and groups of taxi drivers, respectively. The third part is an initial illustration of our framework on the twitter data set.

The "Experimental methods" section concerns the application of the meta-Morisita index. The "Results and discussion" section concerns the performance analysis of our meta-analysis method based on the Morisita index. The "Experimental methods" section describes the proposed process and demonstrates its effectiveness. The "Results and discussion" section evaluates our meta-Morisita based approach by comparing its performance to that of the most popular clustering algorithms from machine learning.

Behaviour pattern for taxi driver

Table 1 shows spatio-temporal data from taxi driver records. The data was obtained from Microsoft Research [46, 47].

The first column in Table 1 is the ID of the Taxi. The second column is the timestamp of the tracking data. The third and fourth columns are the coordinates of the Taxi. The data was collected in 5 min intervals.

Table 2 shows the information from Table 1 mapped onto the Morisita representation. The first column in Table 2 is the ID of the Taxi. The second column is the highest Morisita value of tracking data per day for each driver. The third column is the date.

In this first study, we use the box plot method to process the metadata of point patterns before detecting the anomalies. The metadata values are designed to determine the abnormal values according to their statistical characteristics; the method is extensively

Table 2 Morista value generated by tracking data per day

Taxi_ID	Morisita_Value	Date
39	450	02/02/2008
39	13,000	02/03/2008
39	2100	02/04/2008
39	29,000	02/05/2008
39	6800	02/06/2008
39	1100	02/07/2008
39	1100	02/08/2008

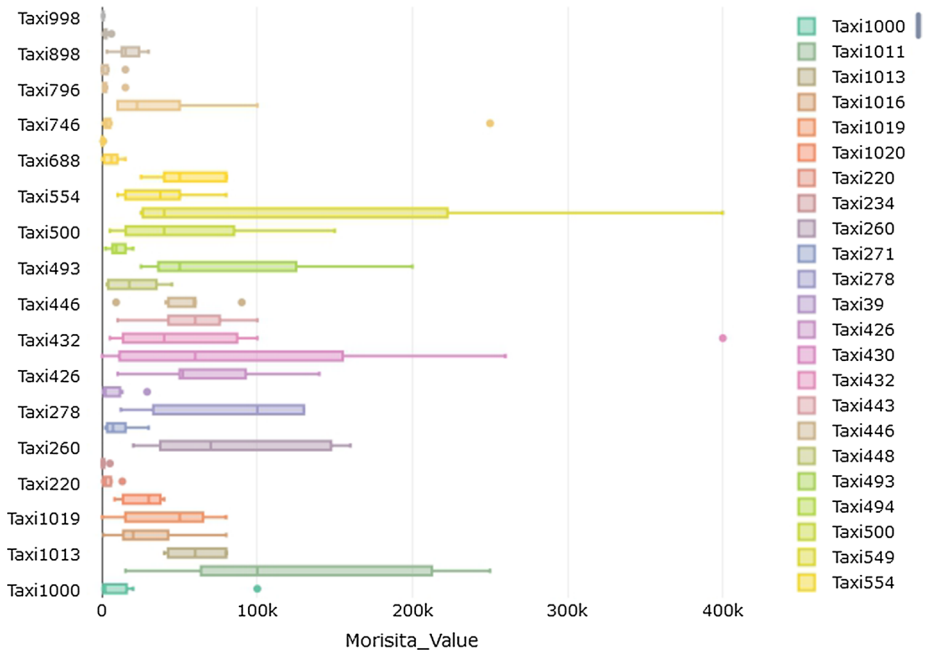

Fig. 9 Box plot about metadata of tracking records of 30 taxi drivers

applied to classify and visualize the abnormal points in the metadata. We chose this method to indicate the outliers of the metadata of the point patterns and marked each outlier using the corresponding metadata value. Figure 9 shows the box plot of the tracking metadata of 30 taxi drivers.

Given that we want to indicate these outliers according to the statistical characteristics of their density values, the method will return Morisita values as density values. The density value indicates that the taxi driver displayed anomalous behaviour on the day the outlier was recorded. From the box plot, we can find the outliers, such as Taxi 549 on 02/07/2008. The Morisita value is extremely high on that day. That means Taxi 549 displayed anomalous behaviour on that day. Taxi 549, Taxi 493, Taxi 1011, and Taxi 430 have box plots that are different from other taxi drivers.

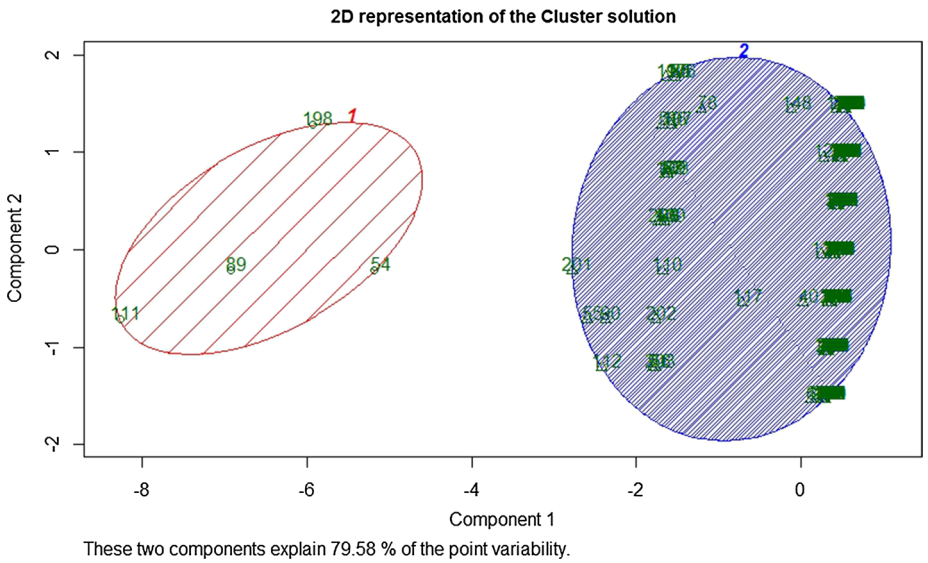

Fig. 10 2D representation of clustering about 30 taxi drivers

Statistical analysis for meta-data of spatio-temporal data from taxi drivers

As mentioned in "Anomaly behaviour detection" section, anomaly detection for a group of drivers was conducted using clustering and time series analysis based on meta-data of point patterns, where the output from the meta-data of point patterns is converted to a simple time series data using a Morisita value. We used K-means as the unsupervised learning method to cluster the group of drivers. Also, we use change point detection to detect the change point of the behaviour pattern.

In particular, we used PELT (Pruned Extract Linear Time) [50] as the change point detection approach. PELT was used to detect the exact moment of breakout when the algorithms report a change in distribution (if at all), along with precision, recall, and the F-measure.

Anomaly detection for the behaviour pattern of taxi drivers

In this framework we chose K-means to cluster the group of drivers. Figure 10 shows the 2D representation of a clustering plot displaying the behaviour patterns of 30 taxi drivers using the K-means algorithm. Figure 11 shows the four-cluster plot displaying the behaviour patterns of 30 taxi drivers using K-means algorithms.

As a result of K-means clustering, Taxi 549, Taxi 493, Taxi 1011, Taxi 430 were grouped as anomalies. The box plot was used to validate the cluster analysis. The K-means results were shown to match the results of the box plot.

Figure 12 shows the spatial distribution of Taxi 549 and Taxi 234 in the city of Beijing between 02/02/2008 and 02/08/2008 . The pink lines represent the road network. The red dots are the location of taxis. Taxi 549 is marked as anomaly taxi driver as the result of K-means analysis. The median values of Morisita index are 8000 and 400 for Taxi 549 and 243 in Fig. 9. Taxi 549's tracking record are more clumped than normal taxis in the upper east area (airport). When looking at the spatial distribution it is clear that Taxi 549 behaves differently from normal driver Taxi 234.

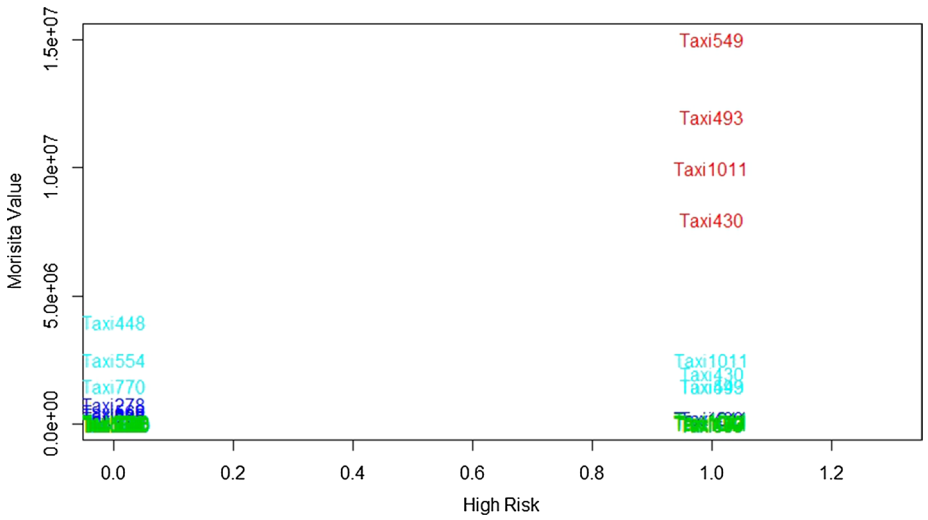

Fig. 11 4 Cluster k-means clustering about 30 taxi drivers

Fig. 12 Spatial distribution of two taxi drivers in Beijing city between 02/02/2008 and 02/08/2008 display at the real map. The pink lines are the read road network. The red dots are the location of taxis

Change point detection for groups of taxi drivers

In statistical analysis, change detection or change point detection tries to identify times when the probability distribution of a stochastic process or time series changes. In general the problem concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes. In this framework we choose the PELT (Pruned Exact Linear Time) [50] method to detect the change point of the behaviour patterns of taxi drivers.

Figure 13 shows the change point value of the time series of Morisita values between 02/02/2008–02/08/2008. The vertical red lines mark the change point.

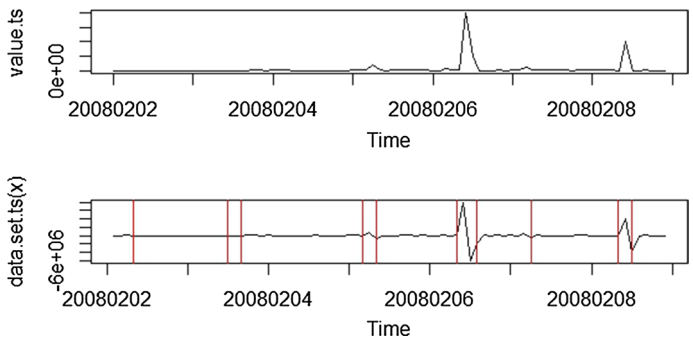

Fig. 13 Change point detection between 02/02/2008–02/08/2008

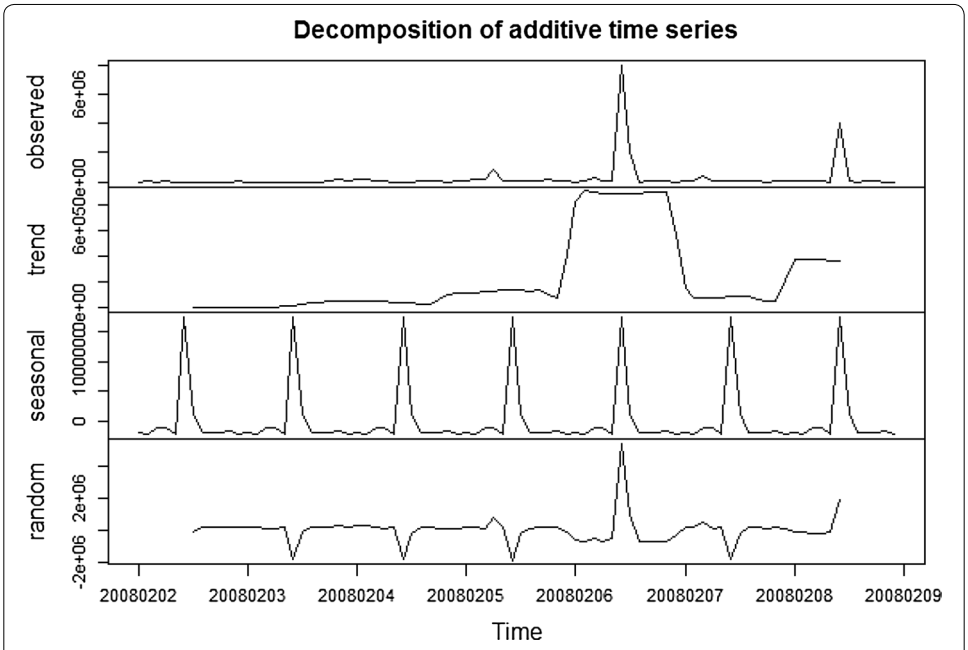

Fig. 14 Decomposition of additive time series between 02/02/2008–02/08/2008

Figure 14 shows the decomposition of the additive time series between 02/02/2008–02/08/2008.

As a result of change point detection, Fig. 13 shows that 02/06/2008 is the date at which the maximum change point in the behaviour patterns of taxi drivers was observed. It turns out that 02/06/2008 was the Eve of the Chinese New Year. This may explain the high density value of the point patterns, which may have been caused by the high traffic volume during the holiday. Figure 14 also shows that every morning represents a change point in the behaviour patterns of taxi drivers. These two observations validate the usefulness of our approach since they show that the events detected by our approach correspond to actual events. In the future, this approach could be used to detect spontaneous gatherings resulting from incidents such as accidents, natural disasters or spontaneous demonstrations and could help alert emergency services and security patrols faster, thus providing greater security.

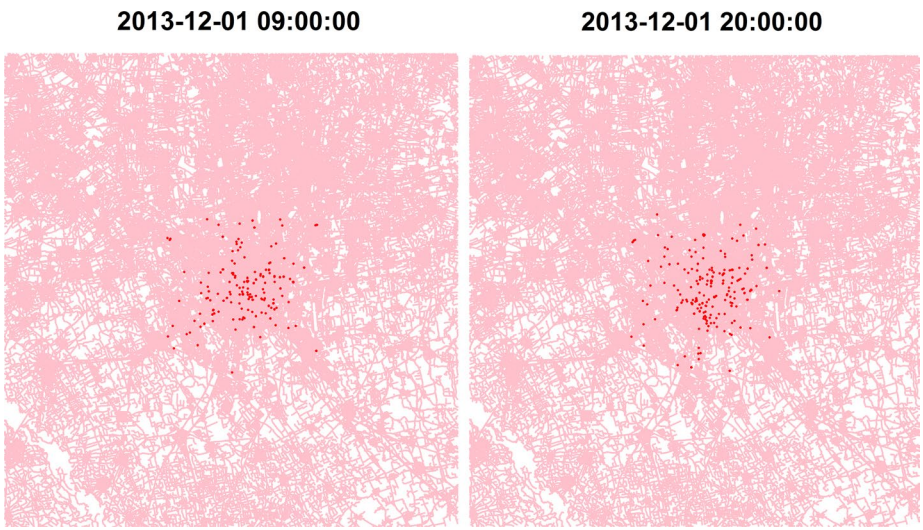

Fig. 15 Location of every tweets in Milan city at 9 a.m. and 8 p.m. at 12/01/2013 display at the real map. The pink lines are the read road network. The red dots are the location of twitter users

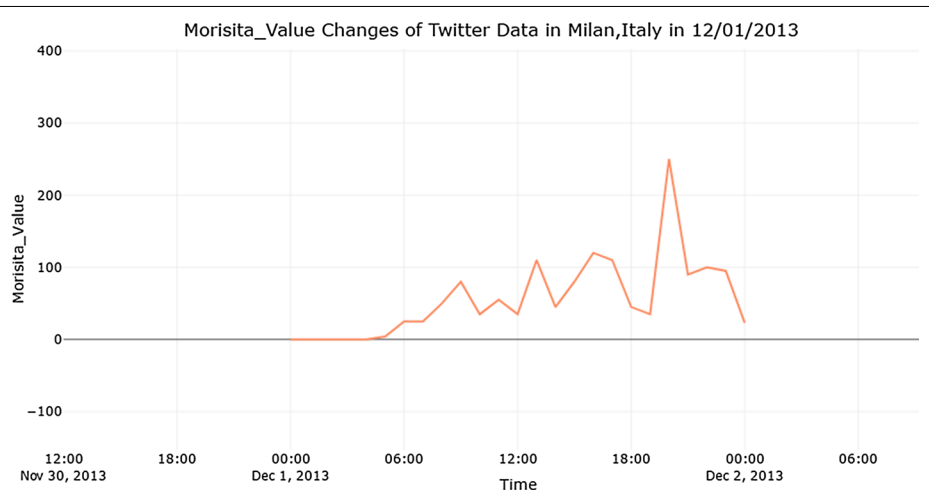

Fig. 16 Line chart of Morisita value of twitter data in Milan, Italy in 12/01/2013

Analysis for twitter data

We now turn to the analysis of the second data set, the twitter data set. Figure 15 shows the spatial distribution of tweets in Milan, Italy at 9 a.m. and 8 p.m. on 1 December 2013 [53]. The pink lines represent the road network. The red dots are the location of twitter users.

Figure 16 shows the line chart of changes in Morisita values over time for this twitter data on 12/01/2013. The line chart in Fig. 16 shows that the Morisita values is higher at night than during the day time. This may be explained by the fact that at night (see Fig. 15, right plot), twitter users do not move around as much, but instead stay put in a concentrated area of the city whereas during the day (see Fig. 15, left plot), they are more dispersed and, therefore, tweet from various parts of the city. This, once again, is not

Fig. 17 Location of every tweets in Milan city at 12/02/2013 and 12/07/2013 display at the real map. The pink lines are the read road network. The red dots are the location of twitter users

particularly novel and useful information here, but it illustrates how gatherings of Twitter users in a single location can be detected, which could be useful for security reasons.

Figure 17 shows the spatial distribution of tweets in Milan, Italy at 12/02/2013 (Monday) and 12/07/2013 (Saturday) [53]. The pink lines represent the road network. The red dots are the location of twitter users.

Figure 18 shows the line chart of changes in Morisita value over time for this twitter data between 12/01/2013 and 12/12/2013. The line chart in Fig. 18 shows that the Morisita value is higher during the weekend than during the working day. The Morisita value on the Monday (12/02/2013 and 12/09/2013) were local minima. This may be explained by the fact that on weekends (see Fig. 17, right plot), twitter users do not move around as much, but instead stay put in a concentrated area of the city whereas during the working day (see Fig. 17, left plot), they are more dispersed and, therefore, tweet from various parts of the city. This, once again, is not particularly novel and useful

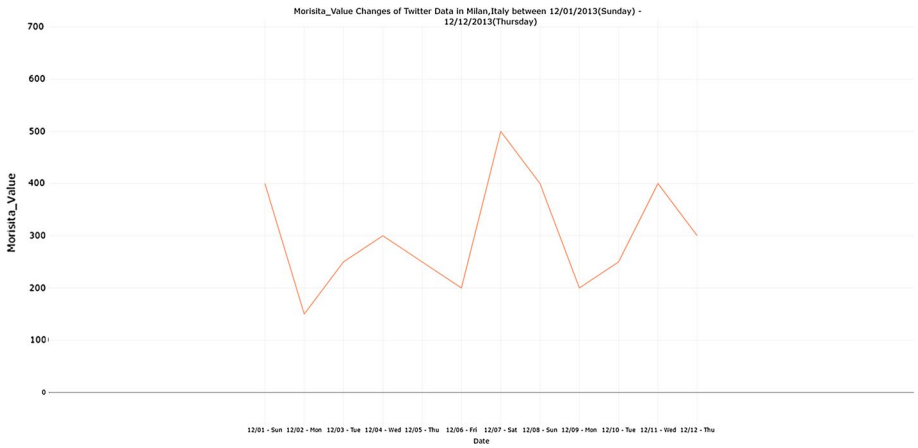

Fig. 18 Line chart of Morisita value of twitter data in Milan, Italy between 12/01/2013 and 12/12/2013

information here, but it illustrates how gatherings of Twitter users in a single location can be detected, which could be useful for security reasons.

The user information has been masked in the raw data by the original distributor due to privacy reason. We can, therefore, not analyse individual twitter user based on the meta-Morisita index method. So we use Taxi data as substitute. The report based on a single Taxi user can be found in "Statistical analysis for meta-data of spatio-temporal data from taxi drivers" section. Using the same learning method as the one used on Taxi data, the anomaly in twitter users could be detected efficiently provided that data on individual users is made available.

Results and discussion

Once again, given the lack of individual twitter data, we could not perform a full quantitative analysis on that data set. Instead, we performed a full quantitative analysis of the Taxi data set. Our experimental environment is based on an eight-core server equipped with Intel(R) Xeon(R) CPU E5-1630 v4 @ 3.70 GHz and 16 GB memory. The version of operating system is Ubuntu 16.04.1.

Time analysis

In this section we recorded the time analysis result of the performance evaluation of different algorithms on our Taxi database.

Figure 19 shows the elapsed time of different analysis methods. The four different methods in the comparison are K-means, density-based spatial clustering of applications with noise (DBSCAN), the expectation maximization (EM) Clustering algorithm and the meta-Morisita index algorithm.

There is no significant difference when processing small scale data such as the data containing less than 10^6 points. The meta-Morisita index , however, obtains better performance when dealing with large scale data such as when the data contains 10^8 points. Thus, the experiment shows that the meta-Morisita index obtains better performance than traditional clustering methods for large scale spatio-temporal data. This is because

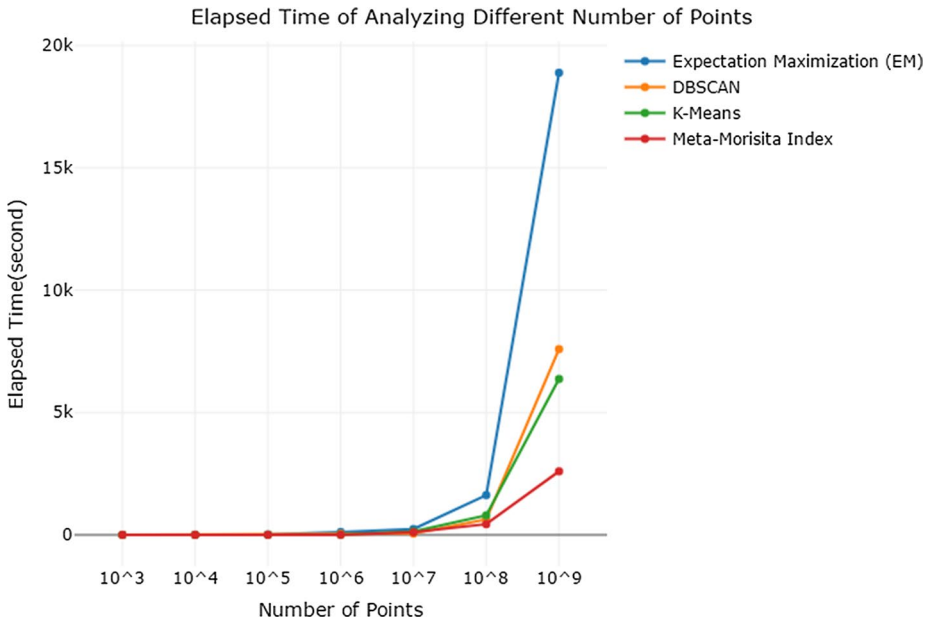

Fig. 19 Performance Comparison among meta-Morisita index, K-means clustering, density-based spatial clustering of applications with noise (DBSCAN) and expectation maximization (EM) Clustering algorithm

the meta-Morisita index retrieves the characteristic value of different time windows for the whole spatio-temporal data set. Then the machine learning algorithm runs on the meta-data only, which significantly reduces the computational complexity. On the other hand, the other clustering algorithms like K-means compute their results directly based on the entire large scale data set. The remaining question concerns the quality of the results obtained by the Meta-Morisita index. We now explore this question by comparing the results obtained by each of the methods just considered to a reference method.

Reference method

Map algebra [54] is a basic set-based algorithm that manipulates the geospatial data. Several algebraic operations like addition, subtraction, etc. can be performed on two or more raster layers of similar dimensions. The output of the map algebra primitive operations is a new raster layer (map). Map algebra operations work on four different classes: local, focal, global and zonal. The operations on raster cells and pixels are local operations. The operations on the entire layer are focal operations. The operations on the cells which have the same value are zonal operations. In Geographic Information Systems (GIS), map algebra is implemented by script or procedure. All the operations are displayed on the map. Map algebra calculates the exact number of incidents (here, Taxi occurrences, but could also be, number of tweets, etc.) that occur in a particular location. While map algebra can give us exact results, it is not practical to use in large scale analyses, which is why alternative methods, such as clustering methods and meta-Morisita analysis, were sought and map algebra only used as a reference for a small sample of data as a sanity check. In this section we use map algebra as the referential method

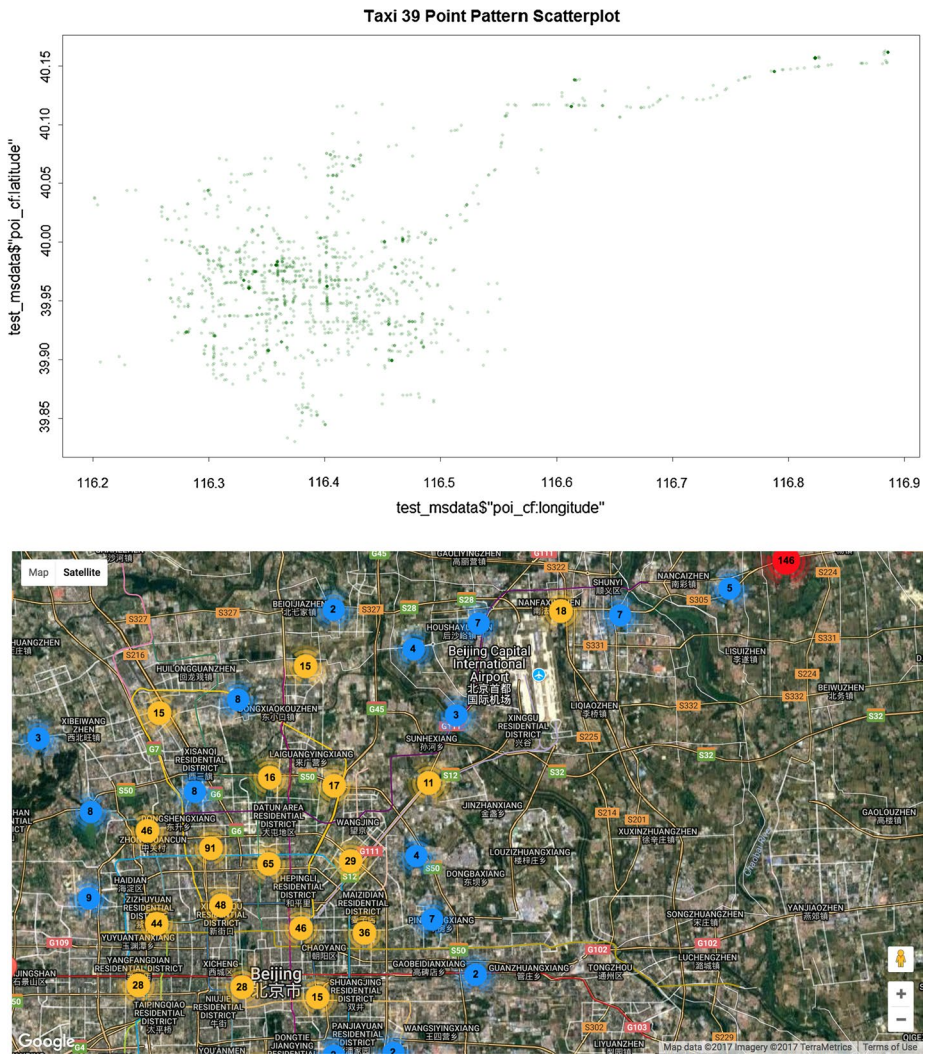

Fig. 20 Taxi 39 tracking record. The number represents the count of the points which generated by map algebra

to detect clusters in the Taxi data. The number of the clusters will be recorded as the reference value.

Figure 20 shows the point pattern of Taxi 39 on 02/07/2008. This figure is generated by map algebra operations. The behaviour of the taxi driver can be visualized by the density value of the spatio-temporal tracking record. In this map the density value has been displayed with color. Red corresponds to a high density value, and blue corresponds to a low density value. In this case we manually count the number of clusters of size greater than or equal to 3. This count is recorded in Table 3 as the reference value.

Accuracy, precision, recall evaluation

In this section we recorded the accuracy, precision and recall rate of the four algorithms previously considered.

Table 3 Number of cluster identified in the sample, cluster size ≥ 3

Taxi_ID	Map algebra	K-means	DBSCAN	Expectation maximization(EM)	Meta-Morisita
39	40	52	133	22	41
220	34	60	103	13	35
234	34	19	88	55	37
260	23	16	77	35	24
271	23	83	174	27	24
278	16	14	61	48	22
426	27	24	92	44	29
430	21	23	69	25	25
493	32	39	74	42	33
549	33	46	109	21	34

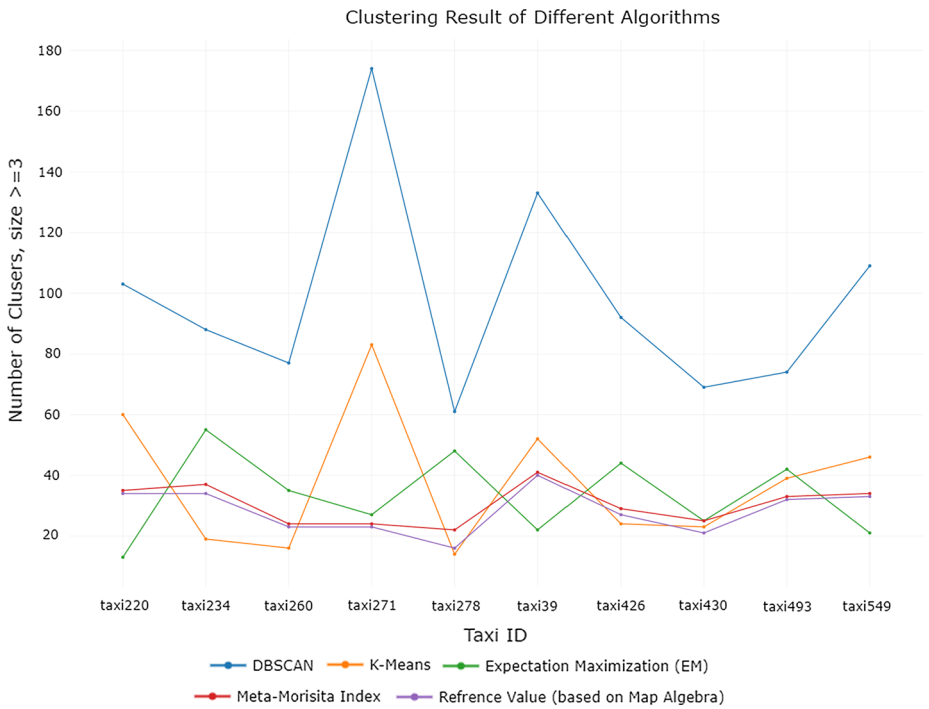

Fig. 21 The number of the clusters being detected by meta-Morisita index, K-means clustering, DBSCAN and EM Clustering algorithm. The size of the cluster is great or equal to 3

The first column in Table 3 is the ID of the Taxi. The other entries represent the number of clusters which have been detected by different algorithms. The result are also visualized in Fig. 21. The assumption is that the closer the match in number of clusters detected, the closer the actual match is between detected and reference clusters.

The first column in Table 4 is the evaluation metric under consideration. The other entries in the table are the results obtained for these metrics by the four methods under consideration. The result was also visualized in Figure 22.

Figure 21 shows the number of clusters (size ≥ 3) obtained by different analysis methods. The four different methods in the comparison are K-means, DBSCAN, EM

Table 4 Evaluation result of different clustering algorithms

Evaluation method	K-means	DBSCAN	Expectation maximization (EM)	Meta-Morisita
Accuracy rate	68.09	23.56	70.18	83.23
Precision rate	75.27	28.88	85.24	93.09
Recall rate	90.46	100.00	89.40	100.00

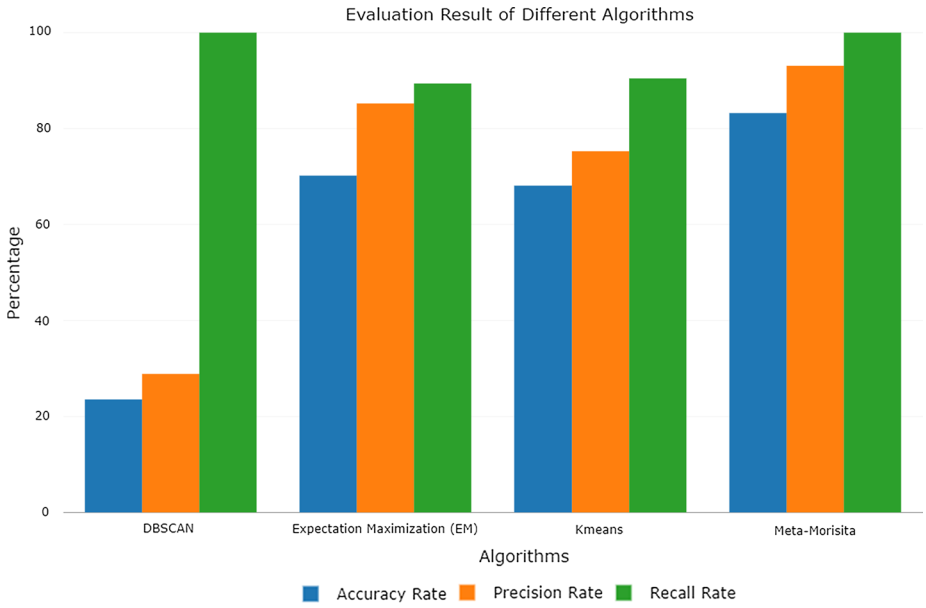

Fig. 22 Evaluation comparison among K-means clustering, DBSCAN and EM Clustering algorithm

Clustering and meta-Morisita index. The reference value for the number of clusters for each taxi is shown by the purple curve. The meta-Morisita value is shown in red. The figure clearly shows that the red curve is the closest match to the purple curve in Fig. 21.

Figure 22 shows the evaluation of different analysis methods. Once again, the four different methods in the comparison are K-Means, DBSCAN, EM Clustering and meta-Morisita Index. The K-means and EM obtain similar results. They both yield a high number of false negatives (FN). DBSCAN is a little different. It detects all the clusters but obtains too many False Positives (FP).

Map-Algebra algorithms [54] have been used as the reference value since they manually compute occurrences in the same location. The clusters in meta-Morisita index are detected by the number of points falling into the same quadrat. As a result we can see that the meta-Morisita index significantly improves the accuracy, precision and recall rate compared with other learning algorithms. The reason is that the meta-Morisita index algorithm comes from spatial statistics which is similar to map algebra. Furthermore, the Euclidean distance used in the machine learning algorithms caused instability in the clustering result for different data samples. The corresponding concept in Morisita index is the smallest diameter of the quadrat, which is a constant value for geospatial data with the same significant numbers. For example,

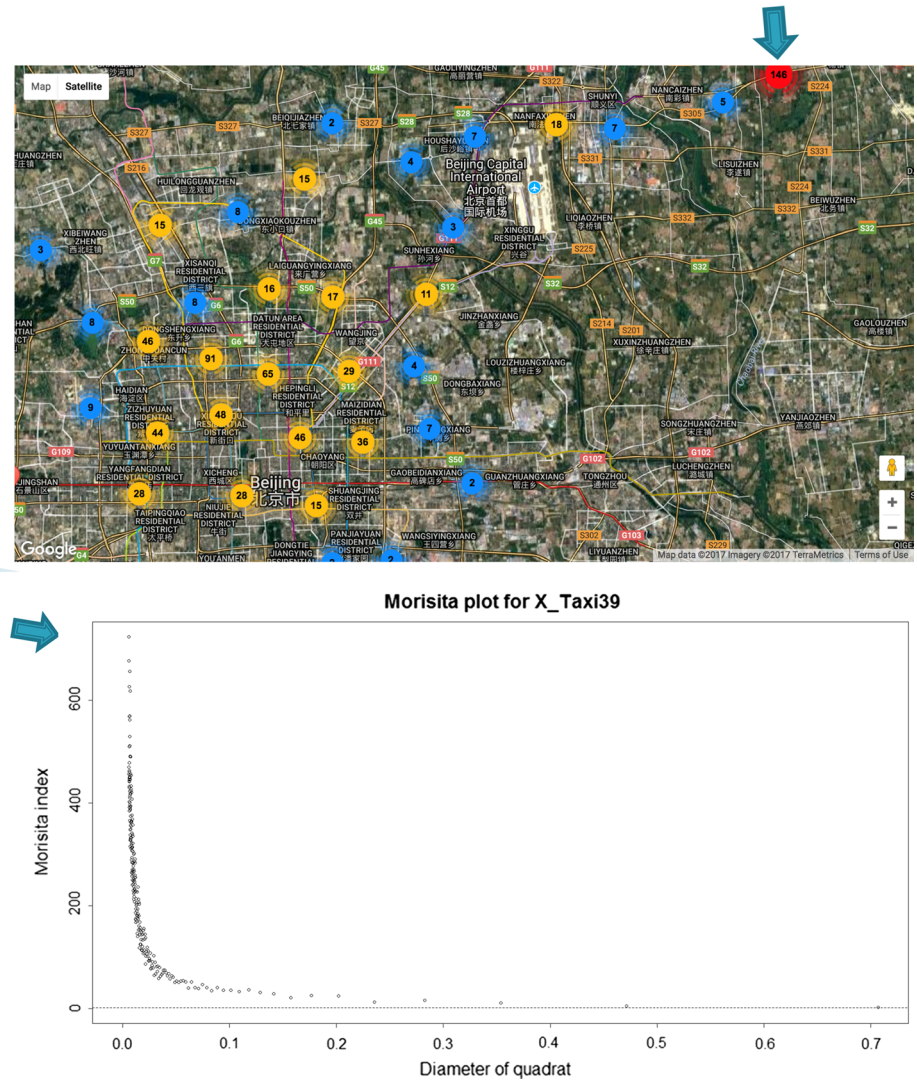

Fig. 23 The relationship between the point density and the Morisita index in large area. The point has highest Morisita value in the Morisita plot corresponds to the location which have highest density in the map

coordinates (in WGS84 geodetic datum) with four significant values like (Longitude: 116.2936, Latitude 39.9227) , the scale of the data is about 10 m. That constant measure ensures that the meta-Morisita value is stable for different data samples.

Figure 23 shows the relationship between the point pattern and the Morisita index. The outlier is the cluster which contains 146 points and is displayed in red on the picture. This cluster has been detected by the Morisita index. However the Morisita index is based on random sampling. As a result, duplicate clusters have also been erroneously detected by the Morisita index which suggests that more than one cluster of this type is present. Please note that the errors made by the meta-Morisita index occurred only in the in the vicinity of the outlier area where the density of the cluster is too high for several clusters to form.

Conclusions and future work

The availability of large-scale spatio-temporal data sets (e.g., social media, vehicle or cellphone tracking data, financial network log) provides the opportunity and challenge to study behaviour patterns to better understand the interactions between cyber and physical events. In this paper, we explore tracking data by investigating the spatio-temporal patterns of taxi drivers and twitter users. A brief work flow is proposed to identify and extract spatio-temporal patterns of outliers based on meta-data of the tracking data. Two case studies of Beijing, China and Milan, Italy are employed to test the proposed method; multiple typical spatio-temporal convergent and dispersive patterns are identified in the large area. We discuss the spatio-temporal distribution of these patterns in different functional areas to obtain better knowledge of the behaviour pattern.

In the paper "Statistical Modeling: The Two Cultures" [55], Breiman compared the data and algorithm modeling cultures. The framework we presented is the combination of spatial statistics and machine learning. The Morisita index is a data modeling approach for spatial data in statistics. However the original Morisita index does not provide the ability to learn. In our work, the Morisita index (data model) has been used to generate the characteristic value of raw data, then learning approaches (algorithm model) were applied to meta data generated by the Morisita index. Our framework thus combines both the advantage of machine learning and spatial statistics. At the same time meta-Morisita index prevents the high computational complexity of current clustering algorithms applied to spatial data.

The findings derived from this study provide insights about the location, time, intensity of the taxi drivers in Beijing and twitter users in Milan, which is helpful for mining behaviour pattern and surveillance for cyber-physical subjects. The identified patterns can help government agencies and urban administrations make targeted adjustments to monitoring cyber-physical events with high anomaly activity as a way to improve the efficiency of the methods used to maintain security in society. In addition, the findings can be used as a reference for understanding subject behaviour. For example, if we only know the spatio-temporal distribution of the active areas of a city, it is possible to have a general understanding of daily subject behaviour and dispersion in other cities according to the discussion in "Statistical analysis for meta-data of spatio-temporal data from taxi drivers" section.

In the future, we will use spatio-temporal statistical models to analyse intelligence information about the behaviour of each subject, to determine the spatio-temporal interactions among different areas of the city, and to explore the behaviour patterns among different social roles, which can provide in-depth knowledge regarding the interactions between subjects and their social roles.

Our plan is to collect tracking data with different social roles, such as Teacher, Police Officer, UPS drivers, etc. We will train the system on each social role separately in order to learn behavior patterns from each category and help us detect behaviors that do not fall into expected categories and may be considered suspicious.

In general, the framework provides a universal solution for spatio-temporal analytic tasks beyond the meta-data of spatial point pattern, which is needed for statisticians and researchers. The Morisita index was selected as the indicator of daily behaviour from spatial point pattern. Multiple analytic methods have been used as efficient statistical

computing approaches to accommodate multiple spatial-temporal data sources and data schemas. The statistical analyses beyond the meta-data of spatial point pattern, which work to reduce the computational complexity for large scale spatio-temporal data, are flexible enough to be added to existing spatio-temporal data warehouse systems. Using this framework is a more convenient, flexible, and scalable way for data analysts and statisticians to process and analyse large-scale cyber security data with spatio-temporal marks.

Authors' contributions
ZY and NJ conceived of and designed this study. ZY analyzed and drafted the manuscript. NJ gave many valuable suggestions about the machine learning algorithms and help editing the language of the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
The data sets supporting the results of this article are included within the article

Consent for publication
All authors have approved the manuscript and agree with its submission to the journal.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable.

Publisher's Note

Received: 23 March 2018 Accepted: 26 June 2018
Published online: 11 July 2018

References

1. Shrestha A, Zhu Y, Manandhar K. Nettimeview: applying spatio-temporal data visualization techniques to ddos attack analysis. LNCS. 2014;8887:357–66.
2. Yang X, Zhao Z, Lu S. Exploring spatial-temporal patterns of urban human mobility hotspots. Sustainability. 2016;8(7):674.
3. Chen D, Lu C-T, Kou Y, Chen F. On detecting spatial outliers. GeoInformatica. 2008;12:455–75.
4. Brimicombe AJ. Cluster detection in point event data havingtendency towards spatially repetitive events. In: Proceedings of the 8th international conference on geocomputation, Ann Arbor; 2005.
5. Hijbeek R, Koedam N, Khan MNI, Kairo JG, Schoukens J, Dahdouh-Guebas F. An evaluation of plotless sampling using vegetation simulations and field data from a Mangrove forest. PLoS ONE. 2013;8(6):e67201.
6. Yang Z. Spatial data mining analytical environment for large scale geospatial data. Ph.D. thesis, University of New Orleans. University of New Orleans Theses and Dissertations. 2284. 2016. http://scholarworks.uno.edu/td/2284
7. Baddeley A, Rubak E, Turner R. Spatial point patterns: methodology and applications with R. Boca Raton: CRC Press; 2015.
8. Ioup E, Yang Z, Barré B, Sample J, Shaw KB, Abdelguerfi M. Annotating uncertainty in geospatial and environmental data. IEEE Internet Comput. 2015;19:18–27.
9. Stiling P. Ecology: global insights and investigations. New York: McGraw-Hill Education; 2011.
10. Berthelsen KK, Jalilian A, van Lieshout M-C, Rajala T, Schuhmacher D, Waagepetersen R. Spatstat quick reference guide. http://spatstat.org/resources/spatstatQuickref.pdf
11. Morisita M. Measuring of the dispersion and analysis of distribution patterns. Memoires of the Faculty of Science, Kyushu University, Series E. Biology. 1959;2:215–35.
12. Cressie N, Wikle CK. Statistics for spatio-temporal data. New York: Wiley; 2011.
13. Wahba G. Spline models for observational data. Philadelphia: SIAM; 1990.
14. Koziel S. Accurate modeling of microwave devices using kriging-corrected space mapping surrogates. Int J Numer Model. 2011;25:1–4.
15. ArcMap 10.3. How kriging works. Redlands: ESRI. 2016. http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm
16. Cox DR, Isham V. Point processes. Boca Raton: Chapman and Hall; 1980.

17. Daley DJ, Vere-Jones D. An introduction to the theory of point processes volume I: elementary theory and methods. Berlin: Springer; 2003.
18. Cressie NAC. Statistics for spatial data, revised edition. New York: Wiley; 2015.
19. Diggle PJ. Statistical analysis of spatial point patterns. London: Hodder Education Publishers; 2003.
20. MØller J, Waagepetersen RP. Statistical inference and simulation for spatial point processes. Boca Raton: CRC Press; 2003.
21. Diggle PJ. Spatio-temporal point processes, partial likelihood, foot and mouth disease. Stat Methods Med Res. 2006. https://doi.org/10.1191/0962280206sm454oa.
22. Diggle PJ, Kaimi I, Abellana R. Partial-likelihood analysis of spatio-temporal point-process data. Biometrics. 2009. https://doi.org/10.1111/j.1541-0420.2009.01304.x.
23. Zhuang J, Ogata Y, Vere-Jones D. Stochastic declustering of space-time earthquake occurrences. J Am Stat Assoc. 2002;97:369–80.
24. Illian J, Penttinen A, Stoyan H, Stoyan D. Statistical analysis and modelling of spatial point patterns. New York: Wiley; 2008.
25. MacQueen J. Some methods for classification and analysis of multivariate observations. Proc Fifth Berkeley Symp Math Stat Prob. 1967;1:281–97.
26. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD'96 Proceedings of the second international conference on knowledge discovery and data mining. Cambridge: AAAI Press; 1996. pp. 226–231.
27. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc. 1977;39:1–38.
28. Park HS, Jun CH. A simple and fast algorithm for k-medoids clustering. Exp Syst Appl. 2009;36:3336–41.
29. Ng RT, Han J. Clarans: a method for clustering objects for spatial data mining. IEEE Trans Knowl Data Eng. 2002;14:1003–16.
30. He Z, Xu X, Deng S. Attribute value weighting in k-modes clustering. 2007.
31. Merzougui M, Nasri M, Bouali B. Isodata classification with parameters estimated by evolutionary approach. In: 2013 8th International Conference on intelligent systems: theories and applications (SITA); 2013.
32. Bezdek C, Ehrlich J, Full W. Fcm: The fuzzy c-means clustering algorithm. Comput Geosci. 1984;10:191–203.
33. Sander J, Ester M, Kriegel H-P, Xu X. Density-based clustering in spatial databases: the algorithm gdbscan and its applications. Data Mining Knowl Dis. 1998;2:169–94.
34. Wang X, Hamilton HJ. Dbrs: A density-based spatial clustering method with random sampling., Lecture notes in computer science book seriesBerlin: Springer; 2003.
35. Birant D, Kut A. St-dbscan: an algorithm for clustering spatialtemporal data. Data Knowl Eng. 2007;60:208–21.
36. Ankerst M, Breunig MM, Kriegel H-P, Sander J. Optics: ordering points to identify the clustering structure. In: Proceeding SIGMOD '99 proceedings of the 1999 ACM SIGMOD international conference on management of data. 1999.
37. Izakian H, Pedrycz W. Anomaly detection and characterization in spatial time series data: a cluster-centric approach. IEEE Trans Fuzzy Syst. 2014;22:1612–24.
38. Birant D, Kut A. Spatio-temporal outlier detection in large databases. J Comput Inf Technol. 2006;14:291–7.
39. Cheng T, Li Z. A multiscale approach for spatio-temporal outlier detection. Trans GIS. 2006;10:253–63.
40. Saligrama V, Zhao M. Local anomaly detection. In: Proceedings of the 15th international conference on artificial intelligence and statistics (AISTATS), vol. 22. La Palma. 2012.
41. Young WC, Blumenstock JE, Fox EB, McCormick TH. Detecting and classifying anomalous behavior in spatiotemporal network data. New York: KDD-LESI; 2014.
42. Liu C, Ghosal S, Jiang Z, Sarkar S. An unsupervised spatiotemporal graphical modeling approach to anomaly detection in distributed cps. In: Proceedings of the 7th international conference on cyber-physical systems, 1. Vienna. 2016.
43. Capdevila J, Cerquides J, Torres J. Mining urban events from the tweet stream through a probabilistic mixture model. Data Mining Knowl Discov. 2017;93:58–68.
44. Anagnostopoulos A, Petroni F, Sorella M. Targeted interest-driven advertising in cities using twitter. Data Mining Knowl Discov. 2017;32(3):737–63.
45. Pappalardo L, Simini F. Data-driven generation of spatio-temporal routines in human mobility. Data Mining Knowl Discov. 2017;91:511–24.
46. Yuan, J., Zheng, Y., Xie, X., Sun, G.: Driving with knowledge from the physical world. In: The 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'11 (2011)
47. Yuan J, Zheng Y, Zhang C, Xie W, Xie X, Sun G, Huang Y. T-drive: driving directions based on taxi trajectories. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, GIS'10. 2010.
48. di Milano D-P, SpazioDati. Social pulse—Milano. Harv Dataverse. 2015;12:1. https://doi.org/10.7910/DVN/9IZALB.
49. Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory. 1982;28:129–37.
50. Killick R, Eckley LA. Changepoint: an R package for changepoint analysis. J Stat Softw. 2014;58:1–9.
51. Killick R, Fearnhead P, Eckley IA. Optimal detection of changepoints with a linearcomputational cost. J Am Stat Assoc. 2012;107:1590–8.
52. Lesmeister C. Changepoint analysis of time series? Tech Rep. 2013. https://www.r-bloggers.com/changepoint-analysis-of-time-series/
53. Center of Computational Communication of Nanjing University. Case study of spatial analysis: spatial point pattern analysis **(In Chinese)**. https://site.douban.com/146782/widget/notes/15468638/note/337537003/.
54. Longley PA, Goodchild M, Maguire DJ, Rhind DW. Geographic information systems and science. 3rd ed. New York: Wiley; 2010.
55. Breiman L. Statistical modeling: the two cultures. Stat Sci. 2001;16(3):199–231.