A Novel Disease Outbreak Prediction Model for Compact Spatial-Temporal Environments

Kam Kin Lao, Suash Deb, Sabu M. Thampi, and Simon Fong

Abstract. One of the popular research areas in clinical decision supporting system (CDSS) is Spatial and temporal (ST) data mining. The basic concept of ST concerns about two combined dimensions of analyzing: time and space. For prediction of disease outbreak, we attempt to locate any potential uninfected by the predicted virus prevalence. A popular ST-clustering software called "SaTScan" works by predicting the next likely infested areas by considering the history records of infested zones and the radius of the zone. However, it is argued that using radius as a spatial measure suits large and perhaps evenly populated area. In urban city, the population density is relatively high and uneven. In this paper, we present a novel algorithm, by following the concept of SaTScan, but in consideration of spatial information in relation to local populations and full demographic information in proximity (e.g. that of a street or a cluster of buildings). This higher resolution of ST data mining has an advantage of precision and applicability in some very compact urban cities. For proving the concept a computer simulation model is presented that is based on empirical but anonymized and processed data.

1 Introduction

Clinical information system is in imperative need for the human society, especially when people experienced some epidemic diseases like severe acute respiratory

Kam Kin Lao · Simon Fong
University of Macau, Taipa, Macau SAR
e-mail: ccfong@umac.mo

Suash Deb
Cambridge Institute of Technology, Ranchi, India
e-mail: suashdeb@gmail.com

Sabu M. Thampi
Indian Institute of Information Technology & Management, Kerala, India
e-mail: sabu.thampi@iiitmk.ac.in

S.M. Thampi, A. Gelbukh, and J. Mukhopadhyay (eds.), *Advances in Signal Processing and Intelligent Recognition Systems*, Advances in Intelligent Systems and Computing 264,
DOI: 10.1007/978-3-319-04960-1_39,

syndrome (SARS), swine flu and enterovirus, etc., which has a high prevalence rate. They outbreak at a very rapid speed, and spread wide and far. There are research papers which advocate developing the clinical decision support system which predicts the time series and space area. However, the efficacy of clinical decision support system is based on the underlying analysis model. Some data related challenges are like: what kind of data attributes the system need to use? How about the scope of data? Is the data useful or not? Which analyzing method is efficient and effective? Any other parameter need to be concerned? What is the trend of disease outbreak? Etc.

Many researchers suggested embedding the clinical decision support system into the GIS (Geographic Information System) as it seems to be more accurate to detect the area whether is in a high prevalence rate and their adjacency areas [1], or even using the ST analyzing method to focus on the analyzing risk of the disease outbreak [2]. Actually these research papers assume the field of analysis is of large terrain or vast piece of land. It is useful for large countries. However, it may not be so applicable for compact urban cities like Macao, Hong Kong, Taipei etc. where the human population is very dense, but they are not necessarily evenly distributed. In the words, the radius approach might not work well in estimating the next infested areas. Different from the other researchers, we introduce a new and simple approach by dividing the city into respective regular polygons, each polygon which is square cell as assumed in this case, is of equal size; and they form grid over the coverage of the city regardless what shape the city is. The number of the square cells to be defined can be selected arbitrarily by the user that depends on the land area and the resolution required.

As a demonstrative case in this paper we use the data simulation of the entorvirus as the experiment part of our research, Macao land and the Tapai land will be separately divided by various numbers of cells and combined for analysis. At the start, the risk of the virus would be evaluated in order to find co-relationship among the areas, the analyzing model will predict how risky of each zone of the city, depends on the risk analyze (some factors may need to reference the previous disease record of the zone and the risk analyzing in the surrounding zones). After locating the high risk areas, the analyzer can group these zones and focus on the relations and/or correlations of them as try to know more about virus and its spread. The associated relationships among the areas are those that have the disease outbreak simultaneously. Technically it will involve using various classifiers of the decision tree and association rules analyzing model.

After applying these two models, some high risk areas and the relationship which the areas almost have the disease outbreak in the same time will be found. The analyzers can concentrate on analyzing the specific characteristic of the areas and deciding which attributes will have the significant relationship between the inflected areas. As the demographic information changes and the risk evaluation suggest, the experiment will vary for different time-series. Finally the analyzers can trace back the source of virus and identify the “flow” of various attributes; and investigate whether the virus has been mutated. This analyzing method will be novel as the part of risk evaluation for detecting among various areas, especially when the forecast of the disease outbreak is changing obviously, the risk index and the associated

relationship can reflect the status of the virus extension. On the other hand, this method could be quite effective and efficient that the users can refer to the spatial and temporal analyzing model to adjust the whole analyzing model, as the learning process to develop a more accurate schema for detecting the disease outbreak.

2 Related Work

Spatial and temporal (Spatio-temporal analyzing, ST) analyzing model is a hot research topic in the last decade, concerning the time and geographic factors to predict the result. The ST analyzing should be based on the spatial analyzing, with an extension part of analyzing the geographic phenomena, which combined with the time sequences. Its purpose is on tracing back or trending the future result. Many researchers advocated their method and framework for the ST analyzing, most of them have a great contribution through their experiment to prove their ST analyzing model is feasible [3, 4, 5]. The details of the classification of the ST data mining task and techniques can be found in [6], ST analyzing can be engaged in various classifiers as clustering [7] and the association rules for ST analyzing [8]. Just like for the traffic jam detection [9], the users can use the ST data about the traffic conditions to simulate the real time traffic surveillance system, warning the drivers which road occurred traffic jam. ST analyzing can also be engaged as the utilization of the land cover change [10], combined with the association rules method, the analyzers can find out which demographic information will impact the utilization of the land reasonably, the relationship between antecedent and consequence can be determined, and the analyzers are able to utilize the result to make the land allocation more scalable.

One of the most popular topics for ST analyzing is applying in the disease outbreak. Many researchers issue various ideas and conducting experiments about it, for their report they are more concerning the serious disease outbreak occurred in a big country, different formulas and external factors like the demographic information, natural disaster like hazard, exposure and vulnerability [11]. As abovementioned, the direction of the ST analyzing is inclined towards the geographic information with the polygon pattern as the original GIS of the country. In fact, it will be bias as the virus extension should not only be analyzing the geographic framework like the predefined map, may be the serious outbreak area is in the edge of the various regions. Nevertheless, how about when the disease outbreak occurred in small urban city like Macao? So far there is no literature on about ST analysis for small cities.

3 Our Proposed Model

Our proposed analysis framework has two parts: locating the risk areas and studying the association between those areas of high risks.

3.1 Risk Analysis via a Spiral Model

The risk of disease outbreak for each specific small cell is computed in a spiral fashion. The computation in terms of risk grades or indices is taken into account of the distance between the adjunct cells and the active cell, the timing, and the strength of the virus dissemination, and other information. For example, a quantitative risk index for a particular cell (called active cell) in a grid is calculated based on the facts about its risk in the previous years, demographic (density, age, race, visitors' traffic) and geographic (climate, number of buildings) factors that will affect the virus dissemination. The risk index would be normalized between 0 and 1 where 0 means it has not ever been infested before. For consideration of risks over certain years, an overall index can be defined by I where:

$$I = (r_{i-1} + r_{i-2} + r_{i-3+\ldots} + r_1) \,/\, \text{number of referenced years} \tag{1}$$

i is the concurrent year need to evaluate, and r is the risk factor of the year. This index I of the particular cell can illustrate the "virus record per history" of the specific zone. It consists of the temporal factors in the analysis when combined with the spatial information. Here, for doing the spatial risk evaluation of each cell in the grid, the coordinates of each grid cell, for the cell i (with coordinate $X = n$, $Y = m$), for estimating the risk index of its adjacency area.

Zone 1 (n-1,m+1)	Zone 2 (n,m+1)	Zone 3 (n+1,m+1)
Zone 4 (n-1,m)	Active Zone (Zone 5) (n,m)	Zone 6 (n+1,m)
Zone 7 (n-1,m-1)	Zone 8 (n,m-1)	Zone 9 (n+1,m-1)

Fig. 1 Illustration of grids

The risk index of an active cell in consideration of its neighbor adjacent is defined by:

$$R_i = I \times [(R_{z1} + R_{z2} + R_{z3} + R_{z4} + R_{z6} + R_{z7} + R_{z8} + R_{z9}) /9] \tag{2}$$

The computation starts at the top left corner and goes in a spiral fashion around each zone, updating the corresponding risk indices along the way. Considering the history of risk of each area is very important. As an infested area is likely to be infested again, given the same conditions arise. In this spiral model, the factor of distance among the zones would impact the dissemination rate and the outspread. It is assumed that the closer where it is to the infested zone, the more likely the disease will propagate over.

Since square grid is used in partitioning the region of a city into square cells, the level or the gravity of the contagiousness is abstracted into some estimates of distance effects. An example is shown in Fig. 3 where concentric rings are logically laid over the grid, with each ring position at the outer area decreases in contagiousness proportionally.

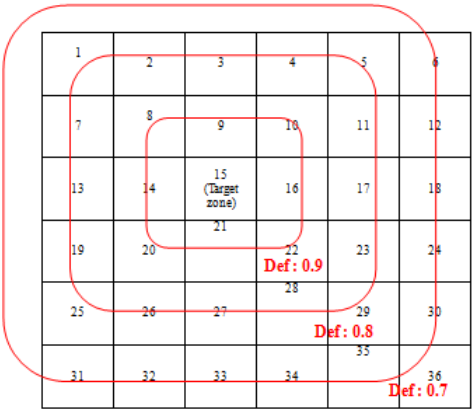

Fig. 2 The effects of the disease are represented by different rings around the zones

In the spiral approach, the "rings" represent various levels of effects measuring from the center (target zone) to how close the designated area by the ring are. In computation, the spiral model generalizes the concept of contagiousness over the distance apart. The spiral index (S_i) to represent this "distance effect" of each area. Moreover, the analyzers can reference from the previous disease outbreaks and calibrated the distance impact index of each level. As in level 1, the distance impacted index is be defined as 0.8 by the user as the previous result illustrated it is not really impacted the specific area significantly. So a moderate factor 0.8 is assumed this time.

In addition to the spatial factors, spatial and temporal analyzing is combining the elements of the time-series with a purpose of predicting the result more reasonable and sensible in consideration of space-and-time. Time series factors are being considered in our model, and it is coined as Seasonality. For instance, Enterovirus is the virus that recognized as it will be disseminated in the middle or later of summer to the beginning of autumn. Actually, for analyzing specific virus, its cycle time should be checked and the seasonal index should be estimated. Given an example as shown in Fig. 3, for the target area 3, we would want to calculate the adjacency areas' risks as well. By considering the time-series factor, the risk (R) value is calculated as

$$D_{(area)} = \{N_{(t,area)}, N_{(t+1,area)}, N_{(t+2,area)}, N_{(t+3,area)} .. N_{(12,area)}\} \quad (3)$$

where $D_{(area)}$ is the data collection of specific area's patients number, and

$$I_{(area)} = \{I_{(t,area)}, I_{(t+1,area)}, I_{(t+2,area)}, I_{(t+3,area)}.... I_{(12,area)}\} \quad (4)$$

where $I = N_{(specific\ month)}/N_{(year)}$. Therefore, the seasonality index is computed as:

$$S_{(area)} = \{S_{(t,area)}, S_{(t+1,area)}, S_{(t+2,area)}, S_{(t+3,area)}.... S_{(12,area)}\} \quad (5)$$

And,

$$R_{(Adjacency\ area)} = \text{Average}\ (D_{(Adjacency\ area)} \times I_{(target\ area)} \times S_{(target\ area)} \times P_{(Adjacency\ area)} \times (A_{(Adjacency\ area)}) \quad (6)$$

where N is the total number of observed case, P is the population density, and A is the spatial index of the surrounding level.

Month	Season	M_1	M_2	M_3	M_4	M_5	M_6
1	Winter	0	1	0	1	0	0
2	Winter	0	1	7	1	0	0
3	Spring	0	0	0	2	0	0
4	Spring	0	0	0	5	5	0
5	Spring	0	0	7	5	13	0
6	Summer	0	0	21	15	13	0
7	Summer	0	3	25	12	18	0
8	Summer	0	4	16	22	7	0
9	Autumn	0	2	27	18	8	0
10	Autumn	0	1	18	0	0	0
11	Autumn	0	0	0	0	8	0
12	Winter	0	1	2	3	0	0

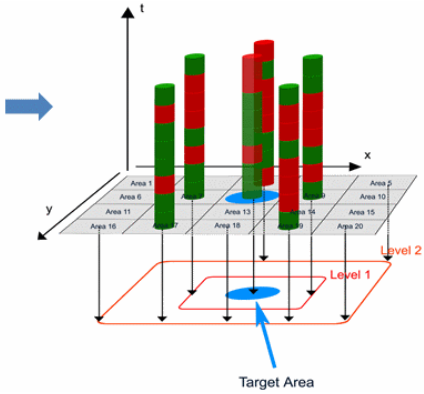

Fig. 3 Illustration of how the seasonality of the Enterovirus outbreaks is incorporated into the spatial-temporal computation model

3.2 *Data Mining the Relationships among the Infested Areas*

In order to extract the potential rules of co-relationship of inflected areas in a small city, some tasks are needed: 1.) Finding decision rules on the likeliness of having the disease outbreaks that happened simultaneously across multiple areas. 2.) Based on the decision rules, applying our pre-defined formulas to the evaluation of each area, so to determine whether the areas belong to those serious areas of disease outbreak or otherwise. 3.) Through the ranking of the evaluation, analyze the demographic information of different ranked areas across various periods. 4.) Predicting the trend line of disease outbreak and the prevalence rate of each zone, with options of mining deeper for the demographic information of co-effected areas.

3.2.1 Decision Rules Generation of Co-inflected Areas

Extracting decision rules from the disease cases is necessary at the beginning of our method. Above all, the city will be divided by the equal cells as a grid.

The disease outbreak in the small city will be simulated and using the data mining software to find out the co-relationship of this area. Two analyzing methods are used to extract the useful rules. They are decision tree and association rule analyzing models. They estimate the degrees of co-relations of the related areas and provide the measures on how reliable the rules are, in terms of Accuracy rate, lift, Confidence, Leverage, and Conviction. Moreover, for each analyzing model, various classifiers and associates rule miner will be applied, (J48, RadomTree, Apriori and HotSpot, etc.) for ensuring the fairness. Default parameters are assumed in each method.

After applying the various models and classifiers, many sets of rules are extracted. The subsequent step is to rank them, and judge on which rules are useful for further processing. In the decision tree model we can use the accuracy rate to decide which rules are acceptable or not. But for the association rules model, four performance results (Lift, Conf, Lev, Conv, in short) are considered. Assuming Lift = L, Conf = Cf, Lev = Lv, and Conv = Cv, Max is the maximum number of the total number of rules we selected, for the number i of Rule, a referenced Score can be calculated as:

$$Si = (L_i/L_{Max})/4 + (Cf_i/Cf_{Max})/4 + (Lv_i/Lv_{Max})/4 + (Cv_i/Cv_{Max})/4 \quad (7)$$

As we are concerning the co-relationship of related inflected area, we opt to filter out any rule disqualified. The workflow of the whole is shown in Fig. 4.

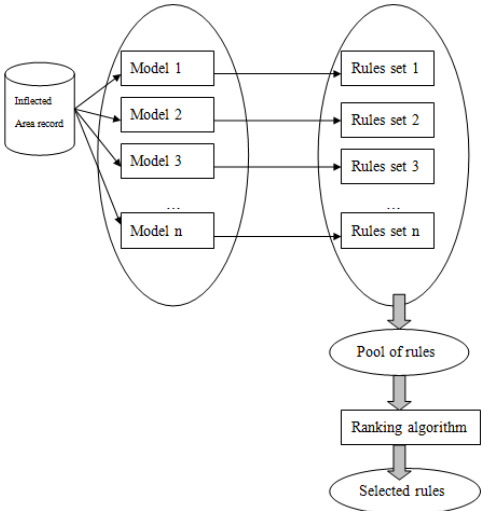

Fig. 4 Concept of rule generation

3.2.2 Analysis with Demographic Information

The likelihood of an area being infested, in addition to its adjacent neighbor areas, is determined by the demographic factors such as transportation, population

density, mobility of the residents, and the vulnerability of the age group of those resides and travelers in the vicinity.

In the previous step, the analyzers can find which areas are relatively more directly impacted and their co-inflected areas as the same group of the disease outbreak. The inflected zone will be extended as like as they have some identical characteristics of these areas (it can also applied in the risk evaluation part of this research paper as finding the characteristics of the high risk areas), just like demographic information by evening the units of the habitant buildings/homes. Combing the analysis results of the demographic and geographic data, we merge information such as how many residents they are born in Macao, their age groups, how long they have stayed in the city, how many schools, human traffic flows etc. in each various area.

After that we will apply various classifiers of the decision tree model for predicting whether the area is highly impacted or potentially a highly impacted area. By using this method, some important attributes of the areas which may significantly impact the prevalence rate will be shown.

3.2.3 Trend-Line for the Prediction of the ST Data

Analyzers can apply this model into various time units as it is flexible to recognize the occurred cases of each area if the disease outbreak happened. Users can define various time unit of the experiment period like day, month, year or even decade. Nevertheless the analyzers can combine with different periods of result to identify the updated status of the virus and decide whether they are mutated or not. Likewise the demographic information is changing over time, and the system can combine with different years' factors to infer the rules, which might have appeared and the cases can be pieced as a sequence. An example is shown in Fig. 5.

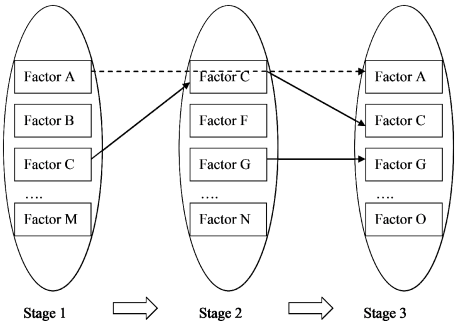

Fig. 5 An example of factors that form a trend

In this example, the factor G is found to have a strong co-relationship of the inflected areas no matter how much the time goes by. It can be illustrated that this factor should be paid more attention as it may be as important as the core factors to impact the disease outbreak. But for factor A it skips one stage and there exists

an association relationship between stage 1 and stage 3. As for this kind of relationship, we call it the "gap" as the interval among different stages for evaluating how strong of these factors through frequency counting in the referenced years. We would observe if such trend has a relation to the prevalence of the virus over the years.

4 Experiment

In order to valid our proposed model, empirical data obtained from the Health Bureau of Macao Special Administrative Region of Macao, are used. The URL to the website is: http://www.ssm.gov.mo/statistic/2012/index.html. A total of 9 years of records which are the number of enterovirus infected patients are obtained in the past nine years. They are, chronologically, 218, 1023, 144, 822,1678, 1023, 1188, 2030 and 1017. A general regression trend indicates that the number critically increased over the years. The r-square value of the regression is 0.4114 which is not a bad fitting of forecast trend and the actual data. The rapid increase in recent year is quite uncommon. Allegedly it is concerned that the virus might have mutated and the disseminated rate probably escalates.

For the population, we assume a parameter called human mobility that is the sum of averaged tourists number over that area and the number of original habitants. In 2012, Macao population is 582,000. On the other hand, the total tourists' number is 28,082,292, especially for 13,577,298 tourists will stay overnight. In average there are at least 37,198 tourists stay in Macao every day. Such numbers are approximately divided into different areas in the grid model.

The objective is by conducting an experiment for the enterovirus surveillance, to predict the likely areas to be infested in the near future. Refer to the infected case of various areas in referenced period and its related population, the index is called the 'basic risk index" (*Rbi*) which calculates by the formula we develop. It illustrates how risky an area will be infected in accordance to their inflected history. The formula of *Rbi* is defined as:

$$Rbi = (N_i / N) \times (I / P) \tag{8}$$

where N_i is the number of years (can be months, or even years) that there are detected of infected cases in the target area. N is the surveillance period of time for the analysis. I is the number of infected cases that had been observed in the target area. P is the target area population in terms of human mobility; basically it will be assumed as how many people spend most of time here, regardless of tourist or permanent residents. After *Rbi* is calculated, the spiral model (as in Fig. 2) will be applied to calculate the new risk index which concerning the geographic position and other factors. The idea of spiral model is centered on the target area as the center point, then the effects of the adjacency areas which surround the target area are estimated accordingly with various levels of effect. Each level is inputted with a user-defined value as a parameter, such as level 1=0.9 or 0.7, level 2=0.6 and so

on. Later on the level parameter will be multiplied by the risk index of the corresponding area that it associates with. Here assuming there are some residents who are already infested with enterovirus infection in the surveillance period, N_i. Further, each area will be set as a target area in turn, and the spiral computation traverses to re-calculate the risk index with reference to the other adjacent areas. In our case there are 30 areas in the grid of the city. There will be 30 target areas to be considered in turn by the spiral model to analyze the related risk index (S_r). Finally all the 30 risk indices would be computed in preparation for the next step – that is to judge according to some user-defined rules on whether this area is deemed as risky area.

For example, area 1 is set as the target area. By utilizing the spiral model to calculate the risk index in relation to its adjacent areas, we use the data set D_r = { S_{r1}, S_{r2}, S_{r3}, S_{r4}, S_{r5} ... etc }. Then we use the ranking method to select the several highest risk area. The types of risks are predefined as the following Table 1.

Table 1 Descriptions of the levels of risks

	Over average Risk	Under average Risk	
Over or reach 4/6	Breaking Point, High disseminated rate	Target risk > 0	Target risk = 0
		Safe but potentially risky	Safe and observed area
Under 4/6	Breaking Point, Low disseminated rate	Safe and observed area	Safe and observed area

The computed results assist the user to determine this area is a center point of disease outbreak or potentially risky areas. Moreover, the risk standard as assumed in Table 1 is calculated by the averaging the risk index (A_R) of the area in the selected level and the mean of this list of number (M_R), the formula goes like this: the risk standard (S_R) = (A_R + M_R) / 2. From the computed results, analyzer can study specific target areas by their disease breaking point and the propagation rate, potential intermediate or even safety zone.

The operation steps of the spiral approach are as follow:

Step 1: Set each area as the target area,

Step 2: Compute the influential risk from the adjacent areas for the target area.

Step 3: By referring to the Table 1, identify the area type.

Step 4: Move to the next target area. Repeat step 1 until all the area is coved.

Through the spiral model the risk index for each area of the city is calculated. The corresponding color code of risk by Table 1 applies. The left hand side of Fig. 6 is an extract of Macao map over which is a translucent layer. The colored boxes simply indicate the number of residents contracted with enterovirus infection. However, the right side of Fig. 6 is a predicted risk distribution computed by the spiral model. It shows also the potential areas which will likely be infested, the next potential disease out-breaking point or safe zones, etc. Therefore personnel from CDC can well fine-tune their resources and pay attention to the predicted next outbreak areas.

Fig. 6 Left: Observed cases of diseases over a city map. Right: Predicted next outbreak locations

5 Conclusion

In this paper we present two major processes for analyzing the risk evaluation and finding the co-relationship of areas which have occurrences of disease outbreak at the same time. We proposed a framework which combines in use of spatial and temporal factors for predicting disease outbreak in a small city. The method can be extended to detecting the earth quake, risk of the hotel occupancy rate and the other analyzing of finding the evaluation of which will concern the spatial and temporal factors. In order to validate the proposed model, an experiment is conducted for analyzing the spatial and temporal factors on empirical data of enterovirus infection in Macao. Through this method, certain areas can be identified as risky zones that have high spreading rate. Our model is quite different from the traditional spatial-temporal analyzing methods, such as it will not do the surveillance for the area which is not polygon as the city size like to be but for the equivalent "grid", Without the limitation of the polygon of the city size, the analysis is more flexible as the analyzer can do the investigation to evaluate the risk relationship among the target area and its adjacent areas. By using this grid scheme and the spiral calculation method, we can calculate the risk index of each area, and be able to identify areas that are of high or low risks.

References

1. Boulos, M.: Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. International Journal of Health Geographics 3(1), 1 (2004)
2. Raheja, V., Rajan, K.S.: Risk Analysis based on Spatio-Temporal Characterization - a case study of Disease Risk Mapping. In: Proceedings of the First ACM SIGSPATIAL International Workshop on Use of GIS in Public Health, pp. 48–56 (2012)
3. Ward, M.P.: Spatio-temporal analyzsis of infectious disease outbreaks in veterinary medicine: clusters. Hotspots and Foci, Vet Ital 43(3), 559–570 (2007)
4. Si, Y.L., Debba, P., Skidmore, A.K., Toxopeus, A.G., Li, L.: Spatial and Temporal Patterns of Global H5N1 Outbreaks. The International Archives of the Photogrammetry. In: Remote Sensing and Spatial Information Sciences, Beijing, vol. XXXVII (B2), pp. 69–74 (2008)
5. Pathirana, S., Kawabata, M., Goonatilake, R.: Study of potential risk of dengue disease outbreak in Sri Lanka using GIS and statistical modeling. Journal of Rural and Tropical Public Health 8, 8–17 (2009)
6. Yao, X.: Research Issues in Spatio-temporal Data Mining. In: Geographic Information Science (UCGIS) Workshop on Geospatial Visualization and Knowledge Discovery, Lansdowne, Virginia, pp. 1–6 (2003)
7. Sato, K., Carpenter, T.E., Case, J.T., Walker, R.L.: Spatial and temporal clustering of Salmonella serotypes isolated from adult diarrheic dairy cattle in California. J. Vet Diagn Invest. 13(3), 206–212 (2001)
8. Shu, H., Dong, L., Zhu, X.Y.: Mining fuzzy association rules in spatio-temporal databases. In: Proc. SPIE 7285, International Conference on Earth Observation Data Processing and Analysis (ICEODPA), 728541 (2008), doi:10.1117/12.815993.
9. Jin, Y., Dai, J., Lu, C.-T.: Spatial-Temporal Data Mining in Traffic Incident Detection. In: SIAM Conference on Data Mining, Workshop on Spatial Data Mining, pp. 1–5. Bethesda, Maryland (2006)
10. Mennis, J., Liu, J.W.: Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change. Transactions in GIS 9(1), 5–17 (2005)
11. Raheja, V., Rajan, K.S.: Risk Analysis based on Spatio-Temporal Characterization a case study of Disease Risk Mapping. In: Proceedings of the First ACM SIGSPATIAL International Workshop on Use of GIS in Public Health, pp. 48–56 (2012)