A Systematic Review of Cluster Detection Mechanisms in Syndromic Surveillance: Towards Developing a Framework of Cluster Detection Mechanism for EDMON System

Prosper Kandabongee Yeng[a], Ashenafi Zebene Woldaregay[a], Terje Solvoll[b], Gunnar Hartvigsen[a]

[a] Department of Computer Science, University of Tromsø – The Arctic University of Norway, Tromsø, Norway
[b] Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway

Abstract

Time lag in detecting disease outbreaks remains a threat to global health security. Currently, our research team is working towards a system called EDMON, which uses blood glucose level and other supporting parameters from people with type 1 diabetes, as indicator variables for outbreak detection. Therefore, this paper aims to pinpoint the state of the art cluster detection mechanism towards developing an efficient framework to be used in EDMON and other similar syndromic surveillance systems. Various challenges such as user mobility, privacy and confidentiality, geographical location estimation and other factors have been considered. To this end, we conducted a systematic review exploring different online scholarly databases. Considering peer reviewed journals and articles, literatures search was conducted between January and March 2018. Relevant literatures were identified using the title, keywords, and abstracts as a preliminary filter with the inclusion criteria and a full text review were done for literatures that were found to be relevant. A total of 28 articles were included in the study. The result indicates that various clustering and aberration detection algorithms have been developed and tested up to the task. In this regard, privacy preserving policies and high computational power requirement were found challenging since it restrict usage of specific locations for syndromic surveillance.

Keywords:

Syndromic Surveillance, Spatiotemporal Clustering, Smart Phone, Aberration Detection.

Introduction

Late detection of disease outbreak has been a threat to global health security for quite a long time, which cost the world many lives, resources, fear and panic. Case fatality rate (CFR) of pandemic diseases is still in the ascendance. The most recent being Ebola Virus Disease (EVD) in Liberia, West Africa. Apart from global fear and panic, EVD registered over 11000 deaths with national case fatality rate of about 70% and local economic losses of $3-4 billion [1, 2]. Traditional surveillance systems are mostly passive and rely on laboratory confirmations to detect disease outbreak. This has been enhanced to syndromic surveillance systems [3] which largely depends on visible signs and symptoms with data sources including emergency department records [4], school absenteeism, work absenteeism, disease reporting systems and over-the-counter medication sales [5, 6]. Nevertheless, the existing syndromic surveillance systems could not detect the disease outbreak early enough and their data sources, and process excludes the incubation phase of the infection [6]but efforts are being made to bridging the gap [6-9].

Recently, the availability of the internet and ubiquity of systems such as smart phones, tablets, smart watches, laptops and other systems have created greater opportunity for the advancement of diabetes management technologies and this generates big data [10]. In the right mix of cluster detections, big data from self-management of diabetes, internet availability and the prevailing pervasiveness of devices, it is feasible and efficient to detect infectious disease outbreak as early as the incubation stage by using the vulnerability of diabetes patients as a sensor [7]. Detection of disease outbreak at the incubation stage is important for reducing morbidity and mortality through early prevention and control [11-14]. Therefore, the general objective is to conduct a systematic review to determine the state-of-the-art clustering detection method, design and evaluation strategies. Associated challenges such as user mobility, privacy and confidentiality along with estimation of geographical location towards the development of a cluster detection approach for EDMON and other similar syndromic surveillance systems would be pinpointed.

Clustering Approach and Outbreak Detection

Generally, outbreak of diseases could be presented in cluster form either in space, time, or both [15, 16]. Clustering methods in disease outbreak detection helps in the identification of environmental factors and spreading patterns linked with certain diseases [10]. Clustering approach could be roughly categorized as temporal, spatial and spatio-temporal. Spatial clustering uses multi-dimensional vectors with longitudinal and latitudinal coordinates. There are variety of such algorithm such as density-based spatial clustering of applications with noise (DBSCAN) [15-17]. Temporal clustering deals

with data points associated with time [18, 19]. It includes various algorithms such as cumulative summation (CUMSUM) and what is strange about recent event (WSARE) [20-22]. Spatiotemporal clustering occurs when there is the involvement of time and spatial dimension [15-17]. There are variety of strategies including different distance functions [23, 24], importing time to the spatial data, transform spatiotemporal data to the new objects, progressive clustering and spatiotemporal pattern discovery [15, 17]. Aberration detection is mainly performed through thresholding mechanisms including various forms such as number of standard deviation set from the mean (z-score), generalized likelihood ratio (GLR), recurrence interval (RI) and confidence intervals (CI) [25, 26].

Materials and Methods

A literature search was conducted between January 2018 and March 2018 through Google Scholar, Science Direct, PubMed, IEEE, ACM Digital Library and Scopus. Different key words such as "Spatiotemporal Clustering" /" Syndromic Surveillance", "Syndromic Surveillance"/ "trajectory Clustering", "Real Time"/"Syndromic surveillance"/"Clustering Mechanism", "Cell Phone"/"Syndromic surveillance"/"Clustering", "Mobile Phone"/"Syndromic surveillance"/Clustering", "Smart Phone"/"Syndromic surveillance"/"Clustering", and "Aberration Detection"/"Syndromic Surveillance"/"Clustering" were used. For a better searching strategy, key words were combined using Boolean functions such as 'AND' 'NOT, and 'OR'. Peer reviewed journals and articles were considered. The inclusions and exclusions criteria were developed based on the objective of the study and through rigorous discussions among the authors. Guided with the inclusion and exclusion criteria, basic filtering was done by skimming through the titles, abstracts and keywords to retrieve records which seemed relevant. Duplicates were removed and articles, which seems relevant, based on the inclusion and exclusion criteria, were fully read and judged. Other relevant articles were also retrieved using the reference list of accepted literatures. Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) flow diagram was used to record the article selection and screening [27].

Inclusion and Exclusion Criteria

For an article to be included in the review, the study should be a practically implemented system with cluster detection mechanisms. Practically implemented algorithms were being sorted for because the results of the study were intended to be used for development of a framework and practical implementation of syndromic surveillance system in EDMON and such similar systems. The study did not have sufficient resources to explore into theoretical and unimplemented algorithms for practical implementations hence the need to skew to practically implemented algorithms. The study was also limited to English language as it does not have the required resources needed to evaluate and accommodate participants who do not speak or write English[28]. The publication type included journal articles, conference abstracts and presentations. There were no time restrictions. Any other article outside the above stated scope were excluded

Data Collection and Categorization

The data collection and categorization were developed based on the objective, literature reviews and authors discussions. The categories have been defined exclusively to assess, analyzed and evaluate the study as follows:

Table 13-Data categories and their Definitions

Category	Definition
Clustering and Aberration Detection Algorithm	This category defines the kind of clustering and Aberration detection algorithm which the study has used and implemented.
Type of Clustering Algorithm	This category defines the type of algorithm. The type of algorithms includes spatial, temporal and spatiotemporal algorithms.
Threshold	This category defines the type of threshold used to generate alarms and alerts in the study.
Clustering Category	The clustering algorithms has been categorized [15]. This dimension tags the specified clustering algorithm used to their respective category.
Design method	This category indicates the design method such as prototype, participatory or joint application development, Agile or waterfall model, that has been used in implementing the system.
Evaluation criteria	In this category, the evaluation criteria used in evaluating the algorithms has been specified.
Performance metrics	This category specifies the performance metrics such as sensitivity, specificity, positive predictive value etc., which was used in the evaluation of the algorithms.
Type of Location	Different type of locations are being used in clustering. These include geolocation, postcodes, counties and many others. This category specifies the exact type of location which was used in the system.
Source of Location	The source of location is defined as the location where the type of location information was obtained from.
Nature of Location	The nature of the location is defining the state of the location as static or dynamic nature.
Visualization tool used	This category also records the type of visualization tool used in the implementation of the visualization aspect of the system.

Display Report	This category records the type of visual displays (graphs, maps, time series etc.) which were implemented by the various systems in the study.
Design layout	This category records the stages and processes used in the architectural design of the syndromic surveillance system. For example, a layout may consist of data acquisition, clustering and aberration detection and visualization [25]. While other design layout could include privacy preserving mechanisms, machine learning techniques in processing the data and other layers [29, 30].

Literature Evaluation and Analysis

Eligible literatures were assessed, analyzed and evaluated, based on the above defined categories. Analysis was performed on each of the categories (Clustering and Aberration Detection Algorithm, Type of clustering, Threshold, Cluster Category, Location Type, Location Source, Nature of Location Source, Design Method, Evaluation, Visualization Tool, Display Report and Design Layout) to evaluate the state of the art approaches. Percentages of the attributes of the categories were calculated based on the total number of counts (n) of each type of the attribute. It is better to take a note that some studies might use multiple categories, therefore, the number of counts of these categories could exceed the total number of articles of these systems presented in the study.

Results

Relevant Literatures

Through searching in the various online databases, a total of 5,936 records were found. Reading of titles, abstracts, keywords and guided by the inclusion and exclusion criteria, led to an initial exclusion of 5,793 literatures and further removal of duplicates in the record resulted in 125 literatures, which were fully read and judged. After full text reading, a total of 28 articles were included in the study and analysis as shown in figure 3.

Table 1. Type of Clustering Algorithms.

Type of Algorithms	Usage Count	%
Spatial	16	32
Spatiotemporal	19	38
Temporal	15	30

Table 3: Type of Threshold Detection Mechanisms

Threshold	Usage #	%
Confidence Interval (CI)	1	4
Generalized Likelihood Ratio (GLR)	5	19
Incidence Ration (IR)	1	4
Recurrence Interval (RI)	10	38
Z-Score	9	35

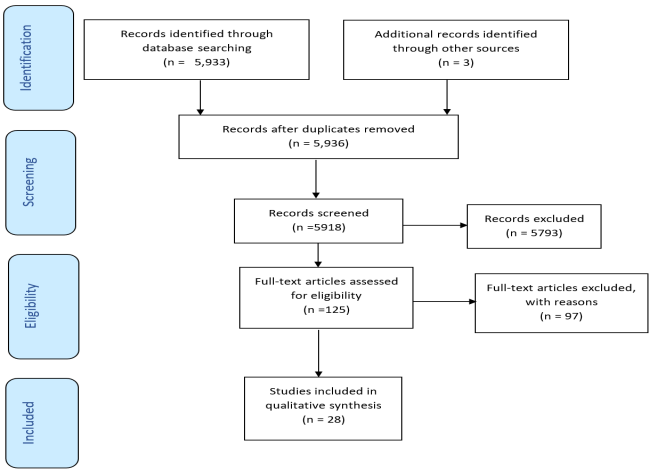

Figure 3: Flowchart of the review process.

Literature Evaluation and analysis

As described earlier, the literatures were assessed, analyzed and evaluated based on the above defined categories. The following section will describe the findings.

I. Types of Clustering Algorithms

Among the three types, namely spatial, temporal and spatio-temporal clustering algorithm, spatio-temporal algorithm is found to be the most preferred approach followed by spatial and temporal algorithm respectively as shown in the table 1.

II. Clustering and Aberration Detection Algorithms

There are a variety of clustering and aberration detection algorithms implemented in the reviewed literatures, where space-time permutations scan statistics is widely adopted followed by cumulative summation, space-time scan statistics and others as shown in the table 2.

Table 4: Categories of Clustering Algorithms.

Algorithm Category	Usage Count	%
Different Distance Function (DDF)	1	3
Importing Time to Spatiotemporal Data (ITTSD)	12	32

Spatiotemporal Pattern Discovery (STPD)	2	5
Threshold base Clustering (TBC)	23	60

Table 2- Clustering and Aberration Detection Algorithms.

Algorithm	Usage Count	%
Risk-Adjusted Support Vector Clustering(RSVC)	1	2
Bayesian Spatial Scan Statistics (BSSS)	1	2
Cumulative Summation (CUMSUM)	8	16
DBSCAN	2	4
Exponentially Weighted Moving Average (EWMA)	1	2
Flexible Space-Time Scan Statistic(FSTSS)	1	
kernel Density	3	6
K-means clustering	1	2
K-NN with Haversian distance (KNearness)	1	2
Log-Linear Regression(LLR)	3	6
Moving Average(MA)	2	4
(Shewhart Chart (P Chart)	1	2
Pulsar Method(PM)	1	2
Recursive Least Square(RLS)	2	4
Risk Adjusted Nearest Neighbor Hierarchical Clustering (RNNH)	1	2
Small Area Regression and Testing (SMART),	1	2
Statistical Process Control(SPC)	2	4
Space Scan Statistics(SSS)	3	6
ST-DBSCAN	1	2
space-time permutation scan statistic(STPSS)	9	18
Space-time Scan Statistics(SSS)	5	10
What is Strange About Recent Event (WSARE)	1	2

III. Threshold Detection Mechanisms

Aberration detection is mainly performed using thresholding mechanisms and in this regard, there are various types of approaches implemented in the reviewed literatures. To this end, Recurrence Interval (RI) is the most widely adopted strategies followed by Z-score, GLR and others as shown in the table 3.

IV. Categories of Clustering Algorithms (CCA)

There are various categories of clustering algorithms, from which threshold-based clustering is the most widely adopted as shown in table 4.

V. Design, Evaluation Methods and Performance Metrics

The reviewed literatures have used various evaluation strategies, among which simulation with historical data stood out as the most widely adopted approach as shown in table 6. The performance metrics which were mostly used are sensitivity (44%) and specificity (36%) as shown in table 5. Prototype and participatory designed were used in the study. Out of 5 systems which disclosed their design methods, 4 of them used participatory approach.

Table 5- Performance Metrics

Performance Metric	Usage Count	%
Sensitivity	11	44
Specificity	9	36
Timeliness	2	8
Consistency	1	4
Correlation	1	4
Positive Predictive Value	1	4

Table 6- Evaluation Method

Evaluation Type	Usage Count	%
Simulation with Historical Data	12	80
Comparison with Known Outbreak	2	13
Power of Cluster Detection Test	1	7

VI. Location Type & Nature, and Source of Location

The literatures have used variety of location type, nature and source as shown in the table 7-9. In this regard, majority of the study used static location (79%) and the rest used dynamic location (21%). Moreover, the study exploited various address such as Geocode (50%), Zip Code (46%) and County (4%). Furthermore, various source of locations has been explored such as Patient Health Record (64%), Mobile Device (14%), TCP/IP (11%), County (4%), and School Address (4%).

Table 7- Location Type

Type of Location	Usage Count	%
Geocode	14	50
Zip Code	13	46
County	1	4

Table 8- Nature of Location Type

Nature of Location	Usage Count	%
Static	22	79
Dynamic	6	21

Table - Source of Location

Source of Location	Usage Count	%
Patient Health Record	18	64
TCP/IP	3	11
Mobile Device	4	14
County	1	4
School Address	1	4

VII. Visualization Tools and Visual Displays

Clustering and aberration detection mechanisms in diseases outbreak needs to be backed up with excellent visualization tools and display to facilitate a quick response from the concerned bodies on the exact timing and place. In this regard, the reviewed literatures have adopted various kinds of tools, among which ArcGIS (24%), Google Map API (22%), TwiInfo (22%) as shown in table 10. Moreover, as to the displaying mechanisms, map (47%) is the most widely used followed by time series (27%), graphs (23%) as shown in table 11.

Table 10- Visualization Tools

Visualization Tool	Usage Count	%
ArcGIS	3	24
Google Map API	2	22
TwiInfo	2	22
OpenStreetMap	1	11
JFreeChart	1	11

Table 11- Visual Displays

Visual Display	Usage Count	%
Maps	14	47
Time Series	7	27
Graphs	8	23
Color Indicators	1	3

VIII. Design layout

The design layout identified in the study have been abbreviated and defined as follows;

DCADAA: This layout consists of obtaining Data first. Then Clustering and Aberration detection are done, followed by generating Alarms to create Alerts of aberrations [20]. DCAVAA: A visualizing module is built in addition, to processes defined in DCADAA [29]. DCTCAVAA: In addition to DCAVAA layer defined above, this layer has data cleaning and transformation features. DCFADAA: In addition to DCADAA, this layout does data filtering or categorizing the data into some defined groups either manually or by employing machine learning techniques. DPVCAAA: In addition to DCAVAA layout, this layout has privacy preserving mechanisms such as anonymization and pseudonymizing [31, 32]. RDPVCAAA: On top of DPVCAAA layout, there is an additional module which for real time data process[31] [29, 31]. TDCAVVAA: In addition to DCAVAA, this layout, tracks user's movement to obtain the data. This is followed with validating the data before Clustering and Aberration detection.[29, 30].

Table 12- Design Layout.

Abbreviation	Usage Count	%
DCADAA	12	55
DCAVAA	1	4.5
DCTCAVAA	3	14
DCFADAA	2	9
DPVCAAA	2	9
RDPVCAAA	1	4.5
TDCAVVAA	1	4.5

Discussion

The general objective is to use a systematic review to assess the state-of-the-art clustering algorithms and other features of systems, which can be used to develop an effective and efficient cluster detection mechanism in EDMON and other similar syndromic surveillance systems. A summary of the most

used approaches and categories are given in the table 13 below;

Table 13: Summary of the most used approaches

Category	Most Used
Clustering Algorithm	Space Time Permutation Scan Statistics
Type of Clustering	Spatiotemporal type
Threshold	Recurrence Interval
Algorithm Category	Threshold base Clustering
Design Method	Participatory Design
Evaluation Method	Simulation with historical data
Performance Metric	Sensitivity
Type of Location	Geocode
Source of Location	Patient Health Record
Nature of Location Source	Static
Visualization Tool Used	ArcGIS
Displayed Output	Maps
Layout	DCADAA

STPSS is one of the spatiotemporal algorithms which is used by most of the syndromic surveillance systems in detecting disease outbreak. Space and time of potential disease outbreak detection is a very efficient method since health management can plan for such potential outbreaks. Health management would know where and when to allocate resources to potential outbreak areas. Another reason of its high usage count could be that the algorithm does not require population at risk data to draw the expected baseline value. But it dwells on the detected cases to determine the expected count [14]. This approach provides significant trend of baseline data while avoiding inclusion of historical data that is irrelevant to the current period. STPSS unlike most of the algorithms does not draw its baseline data (expected cases) from inaccurate population at risk, a control group, or other data that provide information about the geographical and temporal distribution of the underlying population at risk. Such baseline data are inaccurate because there exist significant geographical variation in health-care utilization data due to differences in disease prevalence, health care access and consumer behavior [14]. Unlike spatiotemporal algorithm, spatial algorithms would only indicate where aberrations would occur. This makes planning difficult for health management since it will be difficult to know when to implement health interventions having known potential places for disease outbreak. Sometimes, spatial algorithms are implemented together with temporal algorithms [33]. This gives the surveillance system the spatiotemporal properties. The most used thresholds for aberration detection in spatiotemporal algorithms was Recurrence Interval (RI). This could be as a result that the combination of RI and Monte Carlo Replication helps to easily determine and set specificity of the system [34]. The Monte Carlo simulation is a probability module which is often used with RI or GLR on a cluster to draw a threshold and to determine the likelihood occurrence of a cluster by chance within a specified period for which the analysis is repeated in a regular basis. For instance, in a daily analysis, if the Monte Carlo replication was set to 999 with statistically significant signal of p value < 0.001, the RI would be 1000 days since in disease surveillance the RI is the inverse of the p value. [34]. This implies that, for each 1000 day, the expectation of false alarms would be an average of one false signal per 1000 days or 2.7 years and the RI would be set to the number of days of the baseline data[35].

CUMSUM is a temporal algorithm which was mostly used together with special algorithms. Its ease easy and efficiency might have accounted for the high usage[36].About 60% of the algorithms were classified to be Threshold Based Category (TBS) [15]. This corresponded to relatively high usage of spatiotemporal algorithms. Most of these algorithms employed cylindrical risk regions to detect clusters. The radius formed the area of the map, while the height represented the time. The radius and time were varied to some upper bound thresholds. Participatory design was majorly used while simulation with historical data was mostly used to evaluate the clusters in most of the algorithms. Sensitivity and specificity were the most used performance metrics in the evaluation. This could be the case because users were possibly much interested in a system with reduced false alarms rate. In terms of location, geocodes of census track or hospitals and zip codes were mostly used as location points for the clustering algorithms. These records were mostly retrieved from patient health records. Dynamic nature of the sources of location were of low count. The low count could have been due to the undeveloped and difficulties associated in acquiring and processing dynamic nature of location source data for syndromic surveillance. Also, the stringent inclusion and exclusion criteria on practically implemented syndromic surveillance systems might have accounted for the low count of dynamic nature of location sources. Furthermore, privacy preserving polices and high computational time requirement prohibited the use of exact location of persons for syndromic surveillance. Exact locations such as house numbers and tracking of individuals were only used for group data at the zip code or county level. Information on the exact place of infection is also vital for early prevention and control of morbidity and mortality. But these limitations often hamper the accuracy of information on place of infection since the information collected often relates much to the place of notification which is usually far from place of infection [37, 38]. Also, systems which provided text space for users to indicate their location had some limitations. Users did not indicate proper locations or addresses so their locations could not be geocoded. This resulted in limited sample size [32, 39].

ArcGIS was mostly used to display graphs in this review. It is possible that maps were majorly displayed because it can be used to represent both spatial and spatiotemporal data.

This could have accounted for their high usage of 34% and 47% in their respective categories. In the system design layout category, most of the systems were interested in obtaining data from various sources first. Clustering and Aberration detection were done, followed by generating Alarms to create Alerts of aberrations. This was abbreviated to (DCADAA) for ease of data processing. Tracking for data, acquiring data in real time, privacy preserving mechanisms, filtering and data cleaning were some of the layout processes employed in few of the systems studied. The low rate of tracking persons for data sources could be due to legal, privacy and ethical reasons. Low count of filtering and data cleaning could be due to implementation challenges as machine learning algorithms and natural language processing tools are used for effectiveness. Privacy preserving mechanism is also very vital of which all the systems should have implemented [31]. But the low count rate could have been due to low enforcement of privacy preserving laws in data processing.

The Study Limitations

There is a limitation resulting from impact and study design[40].The study was specifically focused on practically implemented algorithms in relation to syndromic surveillance using clustering mechanisms. The inclusion and exclusion criteria were very specific and stringent on practically implemented syndromic surveillance systems. Therefore, there is the tendency of missing out some algorithms which were not practically implemented in syndromic surveillance systems. For instance, despite an exhaustive search in combination with the search keys, "Cell Phone", "mobile phone" and "Smart Phone", there were limited information regarding mobile phone base trajectories clustering used in syndromic surveillance.

Conclusion

The aim of this review was to derive the state-of-the-art clustering algorithm and its associated design and evaluation methods from practically implemented syndromic surveillance systems. The study revealed Space-Time Permutation Scan Statistics as the most implemented algorithm. The uniqueness and efficiency of STPSS is that its baseline or expected count is based on its detected cases within a defined geographical distance (cylinder radius) and area or temporal window (cylinder height). This approach provides significant trend of baseline data while avoiding inclusion of historical data that is irrelevant to the current period. This algorithm can be used in EDMON and other similar syndromic surveillance systems that are aiming towards implementing state-of-the-art cluster detection mechanism. Temporal and spatial algorithms can also be combined to achieve efficient space time result. This study has also provided wide data categorization, ranging from design of the system to the display of reports. Therefore, we foresee these results might foster the development of effective and efficient cluster detection mechanisms in EDMON and other similar syndromic surveillance systems.

References

[1] WHO. *Ebola Virus Disease*. 2017 June 2017 [cited 2018 20/01/2018]; Available from: http://www.who.int/mediacentre/factsheets/fs103/en/.

[2] Daulaire, N.M., *Global Health Security*. 2018.

[3] Hope, K., et al., *Syndromic surveillance: is it a useful tool for local outbreak detection?*, in *J Epidemiol Community Health*. 2006. p. 374-5.

[4] Choi, J., et al., *Web-based infectious disease surveillance systems and public health perspectives: a systematic review.* BMC Public Health, 2016. **16**(1): p. 1238

[5] Nie, S., et al., *Real-Time Monitoring of School Absenteeism to Enhance Disease Surveillance: A Pilot Study of a Mobile Electronic Reporting System*, in *JMIR Mhealth Uhealth*. 2014.

[6] Woldaregay, A.Z., et al. *EDMON-A Wireless Communication Platform for a Real-Time Infectious Disease Outbreak De-tection System Using Self-Recorded Data from People with Type 1 Diabetes*. in *Proceedings from The 15th Scandinavian Conference on Health Informatics 2017 Kristiansand, Norway, August 29–30, 2017*. 2018. Linköping University Electronic Press.

[7] Heffernan, R., et al., *Syndromic surveillance in public health practice, New York City.* Emerg Infect Dis, 2004. **10**(5): p. 858-64,15200820,

[8] Jacquez, G., *Spatial Clustering and Autocorrelation in Health Events | SpringerLink.* 2018

[9] Woldaregay, A., et al., *An Early Infectious Disease Outbreak Detection Mechanism Based on Self-Recorded Data from People with Diabetes.* Studies in health technology and informatics, 2017. **245**: p. 619-623

[10] Wang, H. and U.o.S.C.-. Columbia, *Pattern Extraction From Spatial Data - Statistical and Modeling Approches*. 2014, University of South Carolina.

[11] MedicineNet, *Modeling Infectious Diseases in Humans and Animals*. 2017.

[12] Study.com. *Progress of Disease: Infection to Recovery - Video & Lesson Transcript | Study.com*. 2018; Available from: http://study.com/academy/lesson/progress-of-disease-infection-to-recovery.html.

[13] Marshall, J.B., et al., *Prospective Spatio-Temporal Surveillance Methods for the Detection of Disease Clusters.* 2009

[14] Martin Kulldorff, R.H., Jessica Hartman, Renato Assunção, Farzad Mostashari, *A Space–Time Permutation Scan Statistic for Disease Outbreak Detection.* 2005

[15] Fanaee-T, H., *Spatio-Temporal Clustering Methods Classification (PDF Download Available)*, in *Doctoral Symposium on Informatics Engineering*. 2012.

[16] P.N. Tan, Vipin Kumar, and M. Steinbach, *Cluster Analysis: Basic Concepts and Algorithms.* 2005

[17] Birant, D. and A. Kut, *ST-DBSCAN: An algorithm for clustering spatial–temporal data.* Data & Knowledge Engineering, 2007. **60**(1): p. 208-221

[18] Hutwagner, L., et al., *Comparing Aberration Detection Methods with Simulated Data*, in *Emerg Infect Dis*. 2005. p. 314-6.
[19] Chan, T.C., Y.C. Teng, and J.S. Hwang, *Detection of influenza-like illness aberrations by directly monitoring Pearson residuals of fitted negative binomial regression models*, in *BMC Public Health*. 2015.
[20] Kleinman, K.P., et al., *A model-adjusted space-time scan statistic with an application to syndromic surveillance.* Epidemiol Infect, 2005. **133**(3): p. 409-19,15962547,2870264.
[21] Kulldorff, M., *A spatial scan statistic.* http://dx.doi.org/10.1080/03610929708831995, 2007
[22] Chen, D., et al., *Spatial and temporal aberration detection methods for disease outbreaks in syndromic surveillance systems.* http://dx.doi.org/10.1080/19475683.2011.625979, 2011
[23] Khokhar, S. and A.A. Nilsson. *Introduction to Mobile Trajectory Based Services: A New Direction in Mobile Location Based Services*. in *Wireless Algorithms, Systems, and Applications*. 2009. Berlin, Heidelberg: Springer Berlin Heidelberg.
[24] Jeung†, H., et al., *Discovery of Convoys in Trajectory Databases.* 2008
[25] Sharip, A., *Preliminary Analysis of SaTScan's Effectiveness to Detect Known Disease Outbreaks Using Emergency Department Syndromic Data in Los Angeles County*. 2006.
[26] Kajita, E., et al., *Harnessing Syndromic Surveillance Emergency Department Data to Monitor Health Impacts During the 2015 Special Olympics World Games.* Public Health Rep, 2017. **132**(1_suppl): p. 99s-105s,28692391,PMC5676508.
[27] PRISMA. *PRISMA*. 2018; Available from: http://www.prisma-statement.org/.
[28] Omicsonline. *Inclusion and Exclusion Criteria and Rationale*. 2018; Available from: https://www.omicsonline.org/articles-images/2157-7595-5-183-t001.html.
[29] Ali, M.A., et al., *ID-Viewer: a visual analytics architecture for infectious diseases surveillance and response management in Pakistan.* Public Health, 2016. **134**: p. 72-85,26880489,
[30] Groeneveld, G.H., et al., *ICARES: a real-time automated detection tool for clusters of infectious diseases in the Netherlands.* BMC Infect Dis, 2017. **17**(1): p. 201,28279150,PMC5345172.
[31] GDPR, E. *EU GDPR Information Portal*. 2018; Available from: http://eugdpr.org/eugdpr.org.html.
[32] Yan, W., et al., *ISS--an electronic syndromic surveillance system for infectious disease in rural China.* PLoS One, 2013. **8**(4): p. e62749,23626853,PMC3633833.
[33] Khanita Duangchaemkarn Varin, C.P., Wiwatanadate, *Symptom-based data preprocessing for the detection of disease outbreak - IEEE Conference Publication*, in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2017: Seogwipo. p. 2614-2617.
[34] Takahashi, K., et al., *A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring.* International Journal of Health Geographics, 2008. **7**(1): p. 14
[35] Yih, W.K., et al., *Evaluating real-time syndromic surveillance signals from ambulatory care data in four states.* Public Health Rep, 2010. **125**(1): p. 111-20,20402203,PMC2789823.
[36] Hutwagner, L., et al., *The bioterrorism preparedness and response Early Aberration Reporting System (EARS).* J Urban Health, 2003. **80**(2 Suppl 1): p. i89-96,12791783,PMC3456557.
[37] Cesario, M., et al., *Time-based Geographical Mapping of Communicable Diseases - IEEE Conference Publication.* 2012
[38] Qi, F. and F. Du, *Tracking and visualization of space-time activities for a micro-scale flu transmission study.* International Journal of Health Geographics, 2013. **12**(1): p. 6
[39] Nicholas Thapen, et al., *DEFENDER: Detecting and Forecasting Epidemics Using Novel Data-Analytics for Enhanced Response.* 2016
[40] Edanz Group Japan K.K., *Writing Point: How to Write About Your Study Limitations Without Limiting Your Impact | Edanz Editing.* 2015

Address for correspondence:

Prosper Kandabongee Yeng,
MSc (Information and Network Security)
Department of Computer Science
University of Tromsø - The Arctic University of Norway
Realfagbygget Hansine Hansens vei 54 Breivika
Tromsø, 9019
Norway
Phone: 47 96992748
Email: prosper.yeng@gmail.com/pye000@post.uit.no