K-CUSUM: Cluster Detection Mechanism in EDMON

Prosper K. Yeng, Ashenafi Zebene Woldaregay and Gunnar Hartvigsen

Department of Computer Science, University of Tromsø -The Arctic University of Norway

Abstract

The main goal of the EDMON (Electronic Disease Monitoring Network) project is to detect the spread of contagious diseases at the earliest possible moment, and potentially before people know that they have been infected. The results shall be visualized on real-time maps as well as presented in digital communication. In this paper, a hybrid of K-nearness Neighbor (KNN) and cumulative sum (CUSUM), known as K-CUSUM, were explored and implemented with a prototype approach. The KNN algorithm, which was implemented in the K-CUSUM, recorded 99.52% accuracy when it was tested with simulated dataset containing geolocation coordinates among other features and SckitLearn KNN algorithm achieved an accuracy of 93.81% when it was tested with the same dataset. After injection of spikes of known outbreaks in the simulated data, the CUSUM module was totally specific and sensitive by correctly identifying all outbreaks and non-outbreak clusters. Suitable methods for obtaining a balance point of anonymizing geolocation attributes towards obscuring the privacy and confidentiality of diabetes subjects' trajectories while maintaining the data requirements for public good, in terms of disease surveillance, remains a challenge.

Keywords

Disease Surveillance, KNN, CUSUM, Clustering, Diabetes

1 INTRODUCTION

Electronic Disease Monitoring Network (EDMON) is an ongoing research project in symptoms surveillance at the University of Tromsø - The Arctic University of Norway. One of the main aims of the EDMON project is to detect infectious disease outbreak as early as the incubation stage of infection[1] through detecting infection incidences in people with type 1 diabetes and clustering them based on time and geographical region. The project uses self-recorded health related data from people with type-1 diabetes as input[1]. The project was initiated in response to recent challenges associated with existing disease surveillance systems.

The evolution of disease surveillance started with traditional disease surveillance systems which usually depend on laboratory confirmed results to detect disease outbreak [2]. This results in significant time lag between infection time and the time of detection of infection through laboratory confirmation [2]. The traditional surveillance system was hence improved to syndromic surveillance systems [2], which greatly relied on visible signs and synthons with data sources from emergency department records [3], school absenteeism, work absenteeism, disease reporting systems and over-the-counter medication sales[1, 4]. But delays have been observed between infection time and up to the visible sign and synthons stage[1, 4]. These types of disease surveillance systems do not detect the disease outbreak early enough and their data sources excludes the incubation phase of the infection[1, 4]. They mostly detect disease outbreak after the infected person is at the illness or after terminal stage, thereby increasing the disease burden such as infection rates (IR) and case fatality rates (CFR) [5, 6].

But through the electronic management of diabetes, big data is being generated as a "by-product" which can be processed to detect disease outbreak at an earlier stage in time. Diabetes Mellitus (DM) is related with the lack of insulin secretion (Type 1 Diabetes) or action (Type 2 Diabetes) [7, 8] but it can be treated and its effect mitigated through diet, physical activity, medication, regular screening and treatment for complications [7, 8]. People with diabetes often experience high Blood Glucose (BG) levels during disease infection incidents [7, 8]. EDMON is therefore exploring the idea of type-1 diabetes persons, exhibiting elevated BGs in the midst of other influential factors, for public good in terms of infectious disease surveillance. In EDMON framework, if infected individuals of Type-1 diabetes incidences are detected, which appropriate algorithms would be used in detecting aberrations for disease outbreak detections? What privacy, security and other requirement would be considered? The aim of this study was therefore to address these general questions through an exploratory and prototype implementations of cluster detection mechanism in EDMON. Specific objectives include developing a spatial classifier with a classification error margin of 1% and implementing a temporal method with 1% error margin of sensitivity, and specificity. The timeliness and methods to deal with privacy and location estimation challenges while generating visualization alarm and alert of outbreaks would also be explored.

2 LITERATURE REVIEW

In EDMON, if a person with Type-1 diabetes gets infected with a disease, the infected person is detected through the elevation of their BGs parameters at various points in time and geo-locations. Such infected incidences would be scattered across the surveillance area. Partitioning the surveillance area into manageable cells such as postcodes, and observing excesses of the infection incidences over time and space would reveal outbreak clusters [1, 4]. Clustering methods in disease outbreak detection helps in the identification of environmental factors and spreading patterns linked with certain diseases [9]. Furthermore, the spread of some viruses have been realized among clusters of people through hands resulting in person-to-person transmission [10]. In view of these, cluster detection methods were explored to be used in the disease outbreak detection part of EDMON.

A systematic review was conducted [11], to explore potential methods, evaluation techniques, visualization methods and other dimensions. The systematic review revealed various algorithms that could be used to achieve the spatiotemporal objective of EDMON [11]. Space Time Permutation Scan Statistics (STPSS), CUSUM, K Nearest Neighbor (KNN), K means clustering, WSARE, DBSCAN and Space Scan Statistics (SSS)[11] were some of the algorithms identified. STPSS and CUSUM were found to be the most used algorithms. From the review, STPSS could have been adopted in EDMON-Cluster since STPSS does not require population at risk data to draw the expected baseline value. STPSS dwells on the detected cases to determine the expected count [5]. This approach provides significant trend of baseline data while avoiding inclusion of historical data that is irrelevant to the current period. However, the STPSS algorithm is only efficient on outbreaks that start locally [5]. This suggests that STPSS is not suitable for detecting disease outbreaks which occur simultaneously in the entire surveillance area. STPSS is only efficient on disease outbreaks with higher rate of early symptoms [5]. According to Chen et al. [12],who studied into “Spatial and temporal aberration detection methods for disease outbreaks in syndromic surveillance systems”; spatial scan methods only detect clusters in simple regular shapes such as cylindrical, circular or spherical. The spatial scan algorithms do not also consider prior knowledge such as the impact of the infection rate or size or shape of the outbreak and it is computationally expensive as local cluster search require searching over a large geographical region. Khanita D. et al in their conclusion after evaluating their proposed study on “Symptom-based Data Pre-processing for the Detection of Disease Outbreak”, with time series and K-KNN algorithm [13], indicated the potential for KNN algorithms as an efficient method for syndromic surveillance and suggested for further assessment of the algorithm. Nearest Neighbor and CUSUM were also statistically demonstrated to illustrate their feasibility of monitoring nearest neighbor statistics [14]. When there is an aberration in the surveillance area, the CUSUM can spot this with the mean distances of emerging diseases of various points in the surveillance area [14, 15]. Martin Kulldorff also support this opinion by emphasizing that “efficient disease surveillance will need the parallel use of different methods, each with their own strengths and weaknesses”[5]. Syndromic surveillance system is optimally effective when both spatial and temporal cluster detection methods work in unison to track emerging infectious diseases at an early stage over the surveillance area [12, 15]. Therefore, the combination of KNN and CUSUM was explored in this prototype study.

3 MATERIALS AND METHOD

The main task of this project was to develop an effective spatio-temporal cluster detection method in EDMON system. Synthetic data was simulated for about 12 months period containing 297 diabetes persons with normal period, non-outbreak period, and also certain known detected cases of infections, outbreak period, both in spatial and temporal aspects. First, each individual person was classified into post code area using k-nearness neighbour algorithms. Then each classified clusters were further analyzed into temporal dimension using the cumulative summation (CUSUM) algorithm. The combination of the spatial algorithm (nearness neighbor) and temporal algorithm (CUSUM) formed the spatio-temporal method [16], hereinafter referred to as K-CUSUM. De-identification and one-way hashing technique were adopted for preserving privacy of the subjects involved in the study. A prototype and system development life cycle approach were adopted for the implementation of the system. The output was displayed on maps with indications of the level of aberrations from the baseline mean. Classifiication accuracy, sensitivity, specificity and timeliness of the algorithms were determined. Privacy preserving technique and other performances measures were also evaluated. Generally, the paper is organized as follows; section 3 presents materials and method, section 4 present results, and section 5 discusses principal findings.

3.1 Materials

Simulated dataset was used to test and evaluate the detection performance of the developed cluster detection algorithm. The simulated data was introduced by mimicking the spatial and temporal variables which could be associated with diabetes persons when they check and update their blood glucose dynamics with a mobile application system. The simulated dataset incorporates health status monitoring of 297 number of people with diabetes for a period of 12 months, with each having an average of 3 records of infection status of morning, afternoon and evening and totaling 968, geographical location information within 21 post code centroids for the region under surveillance. The dataset was split and 70% (660) of the data was used as a reference to classify the rest of the 30% (209) unclassified data. The centroid of the postal code was defined in terms of coordinate features of Latitude (Lat) and Longitude (Lon) in Decimal Degree (DD) units as shown in Table 1.

Each synthetic subject in the study was also simulated to contain the location coordinates, date stamp of where and when the infection incidences occurred in the form of Lat, Lon and their respective date and time. The simulated subject’s data also had the infection status (1,0 or -1 as infected, not infected and suspicious respectively) and some personal identifiable features such as names and IDs. For each day, in every hour, the algorithm was to check for new dataset, classify and detect aberrations. What distinguished unclassified simulated dataset from the classified dataset was that the unclassified data set was not categorized into their various geolocations of postcode however, their actual classes were known for evaluation purpose. Each subject with a detection ID (DID), location features of Lat and Lon and temporal feature of time-stamp was classified into their respective response vectors of post codes (Code) area using the KNN algorithm. The classified dataset was also used as a training dataset for the KNN algorithm during classification of new observables in the unclassified dataset.

Lat (DD)	Lon (DD)	Code	CID
69.55799	19.33103	9027	1
69.57781	18.55499	9106	2
69.627957	18.915001	9006	3
69.63077	19.04736	9020	4
69.63702	17.981	9110	5
69.640574	18.927288	9007	6
69.64225	18.90889	9013	7
69.65079	18.95493	9008	8
69.251024	18.54714	9272	9

Table 1 Simulated Centroid of post codes of study area

The data was manually generated by first, creating estimated decimal degree coordinates (DDC) of centroids of postcode areas using google GPS coordinate lookup system[17]. The DDC were then varied randomly to create artificial locations for the fictitious data subjects. The random variation of the DDC could overlap to other post code areas which might introduce some errors. Therefore, carefulness was being taken not to introduce large degree of variations of the DDC.

3.2 System development tools used

The system implementation relied on various programming and messaging tools including Python 3.6, Leaflet.js 4.2, visual Studio Code 1.3.3 and Twilio SMS Python QuickStart. Python was used as a programming and data manipulation language. The leaflet.js was used to display the map for visualization alarm and alert of the surveillance system and the Twillio API was used to generate SMS alerts. Visual Studio Code was used as a source code editor for the python.

3.3 Methods

Partitioning the region of interest into different small equal cells[18] and assigning the diabetes subjects to the respective cells in which the diabetes status variables were captured could have been used in the study. Additionally, the diabetes persons could manually record the post code addresses during capturing of their diabetes statuses. However, patitioning the surveillance region is deemed expansive manually recording of the post code addresses would be an extra burdon and inconvinincing to the diabetes subjects[18].

The KNN was used as classification algorithm to classify infected persons into various postcodes areas. A dynamic odd value of the number of data points which are more close to the data point to be classified, (K), was determined after computing the Euclidean distances between the unclassified data point and the referenced or classified data point as shown in eqn (1) and eqn(2). CUSUM was also used in this study as a temporal algorithm to detect aberrations with baseline and observable occurrences using the z-score.

$$d_{x,y}^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 \qquad \text{.. eqn (1)}$$

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \qquad \text{.. eqn(2)}$$

The formula used to express CUSUM is as follows;

$$CuSum_t = \sum_1^t e_t \qquad \text{.. eqn(3)}$$

The e in eqn(3) represents the observed number of events minus the reference value (the baseline), while the t represents the time associated. Conventionally, the CUSUM value is initialized to zero[19]. A positive result indicates a change above expectation, zero outcome signifies a period when the observed number of events are the same as the expected number while a negative value of the result indicates that events have fallen below expected levels[19, 20]. Early Aberration Reporting Systems in CUSUM (EARS) are C1-MILD(C1), C2-MEDIUM(C2) and C3-ULTRA(C3) [20-21]. The C1 method depends on a conventional alarm level of Cl=2[22]. This means in the C1 algorithm, the current detected value is greater than the baseline means with an addition of three standard deviations (Z score) which has been calculated based on the past 7days of historical data [22]. The baseline of most syndromic surveillance systems depends on 3 to 5 years long historical data [20,23]. But developments in biological attacks in the United State and higher case fatality rates, consisted the need to develop efficient syndromic surveillance systems which are independent on baselines with long term historical data[20,23]. In a study which compared aberration detection methods with simulated data [23], the aberration detection algorithms with short term duration baseline data (C1, C2 and C3) were as effective as the methods requiring long term historical data in terms of specificity, sensitivity and timeliness [23]. C1, C2 and C3 algorithms have also been developed to accommodate daily and seasonal variations. Their mean and standard deviations were based on a week's (7 days) information which were computed in the same season [24]. To this end, the baseline in this study was chosen to be 7 days of past detected infections of diabetes persons in K-CUSUM.

The classification accuracy of the KNN was computed by determining the proportion of the correctly classified test sets. The performance of the developed CUSUM was evaluated using a confusion metric to asses the sensitivity (Se) and specificity (Sp) while the detection time was determined by considering average time used in the surveilance structure [24].

4 RESULTS

4.1 Framework and design considerations

The proposed framework for the detection of clusters based on a spatio-temporal detection algorithm is given in Figure 1. The framework was developed after a thorough assessment of the state-of-the art cluster detection system found in the literature [11]. Accordingly, the framework incorporates various units; Input data, Pre-processing, Clustering and Aberration detection, Visualization, Alarm and Alerts. The system accepts input data containing different features such as geocodes, infection status, date and time stamp. The input data then goes through cleaning and data conversion into the appropriate data format such as xml and comma separated values (csv). The suitable clustering algorithms are then applied yielding outputs such as alarms and alerts, maps and other visualization output as shown in figure 4.0.

The pictorial view of the clustering mechanism as shown in Figure 2 basically accept input from the unclassified data. KNN algorithm was then applied on the unclassified data to cluster it into the near centroid of postcodes.

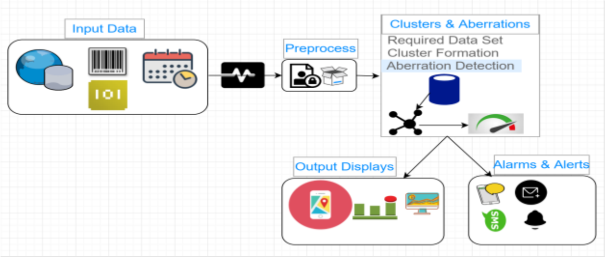


Figure 1 Layout of Framework

Eventually, the classified dataset was grouped and displayed on the map based on the centroid post code. Excess of observations per post code was then determined using the CUSUM algorithm.

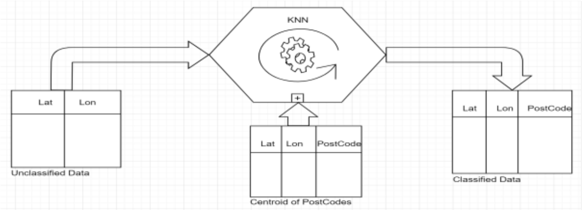
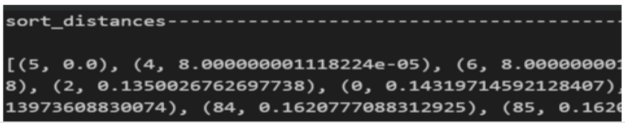


Figure 2 Clustering mechanism

4.2 Classification

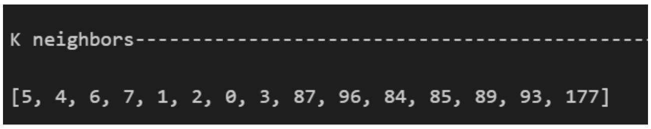
Three modules were involved in the clustering of the infected individuals around the centroids of each postcode area. These include the K of KNN, computation of the Euclidian distance and the determination of the class based on the K factor. The K factor was determined by calculating the odd integer value of the square root of the total number of the classified data[25-26]. The Euclidean distance of each unclassified infected individuals was computed by using their respective geolocation coordinates (Lat, Lon), to each of the coordinates of the infected classified persons' location. The Euclidean distances were obtained, and these distances were sorted in ascending order as shown in Figure 3 and 4.. So, the first K number of the shorter distances were obtained as shown in figure 3 and 4. In an instance, K was determined to be 15, so the first 15 shorter distances were obtained as shown in Figure 4 and 5.



```
sort_distances----------------------------------------

[(5, 0.0), (4, 8.000000001118224e-05), (6, 8.00000000
8), (2, 0.1350026762697738), (0, 0.14319714592128407),
13973608830074), (84, 0.1620777088312925), (85, 0.162
```

Figure 3 Sorted K Nearest Distances

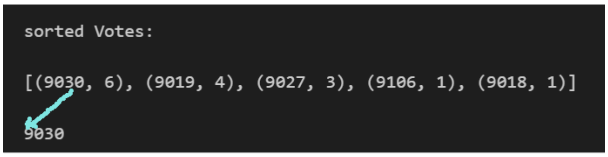


```
K neighbors---------------------------------------------

[5, 4, 6, 7, 1, 2, 0, 3, 87, 96, 84, 85, 89, 93, 177]
```

Figure 4 Sorted IDs of K Nearest data points

After the selection of the K number of data points which were closer to various classified data points, the K data points further 'voted' or were categorized and tagged to various postcodes or classes based on their proximity to the centroid coordinates of the simulated postcodes as shown in Figures 5 and 6. The final counts of votes or tagged K number of data point distances to each postcode area, were declared and the post code with the higher number of K data points was declared as shown in figure 5 . In demonstrating with the synthetic data, 40% of the 15-total number of K were closer to the postcode, 9030 as shown in Figures 5 and



```
sorted Votes:

[(9030, 6), (9019, 4), (9027, 3), (9106, 1), (9018, 1)]

9030
```

Figure 5 Voting results of infected individuals

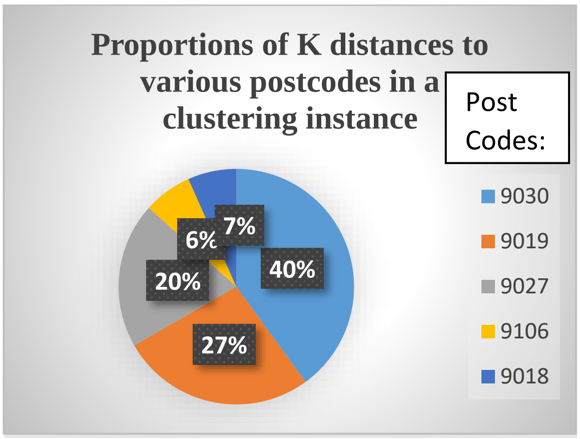


Figure 6 percentages of nearness of data point

4.3 CUSUM Aberration Detection

Having obtained infections data of 7 days baseline, 7 days observations, mean and standard deviation of the baseline per postcode area variables in the CUSUM calculation, a function was therefore developed as shown in Figure 7. From figure 7, if the observed count value was less than three times of the standard deviation plus the average cumulative value within the postcode area in the defined time frames, the cluster shows green to indicate no detections of aberrations as shown in figure 9. However, if the observed count value was more than three times of the standard deviation plus the baseline cumulative value within the postcode area in the defined time frames[22], the cluster showed red, to indicate detections of aberrations or outbreak of infected individuals as shown in figure 8 and 10. A cluster showed yellow if the observed count value was equal to three times of the standard deviation plus the baseline cumulative value within the postcode area in the defined time frames[22].

```
#function get color for aberrations
def Color_for_aberration_detection():
if observed_count < 3*(standDev)+(baseline_count):
mycolor='green'
elif observed_count > 3*(standDev)+(baseline_count):
mycolor='red'
elif observed_count == 3*(standDev)+(baseline_count):
mycolor='yellow'
```

```
return mycolor
```

Figure 7 Aberration detection function

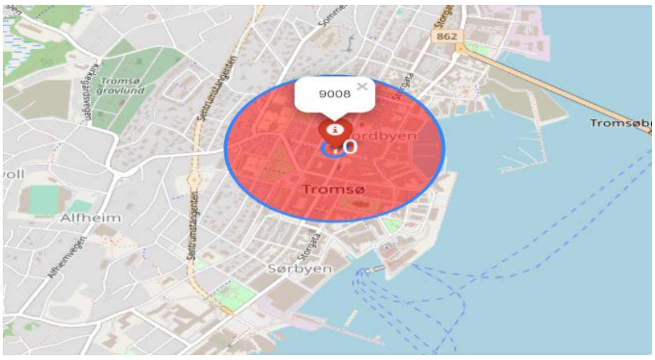


Figure 8 Outbreak cluster

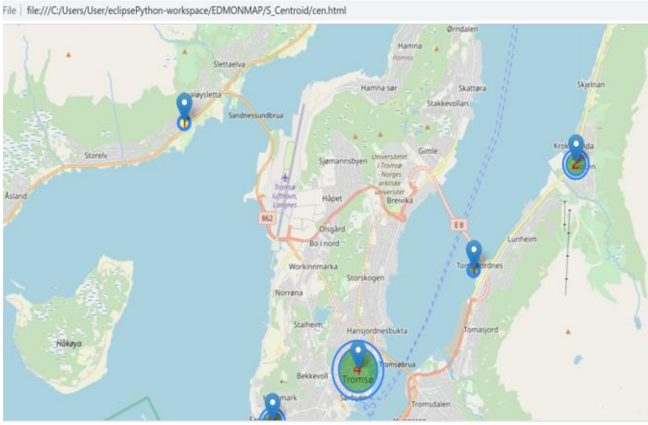

File | file:///C:/Users/User/eclipsePython-workspace/EDMONMAP/S_Centroid/cen.html
Hamna
Skjelnan
Storelv
Sandnessundbrua
Tromsø
Kroken
Tromsdalen
Håkøya
E8
862

Figure 9 Sample map presentation

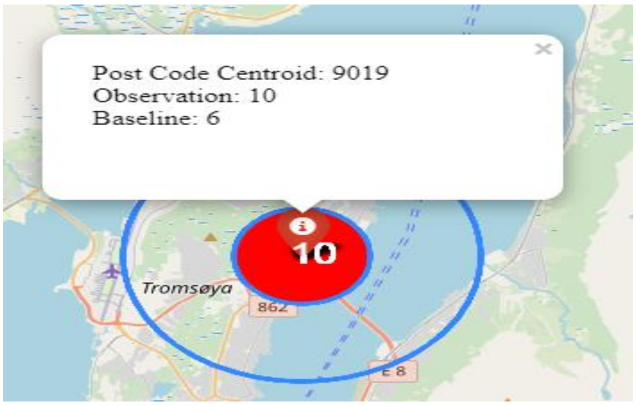

Post Code Centroid: 9019
Observation: 10
Baseline: 6

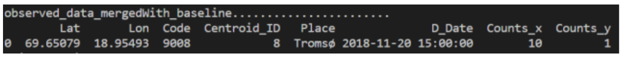
Figure 10 Details of outbreak cluster

4.3 Assessment of KNN

The effectiveness of the KNN algorithm which was implemented in this study (K-KUSUM), was initially assessed with simulated infectious data containing location features with known targets or classes. The algorithm was trained with the entire dataset and was tested with the same dataset. All the features were correctly predicted to be the true classes. To overcome over fitting, under fitting and class imbalance issues, 660 training and 209 testing datasets of 70% : 30% were randomly simulated [28, 29] and evaluated with the algorithm. 99.52% of the test dataset was accurately classified. The same datasets were tested with Scikit Learn KNN algorithm which resulted in 93.81% classification accuracy.

4.4 Assessment of CUSUM

In the CUSUM evaluation in K-CUSUM, the baseline values of past one-week infections, were compared with the observed values of current one week while taking into consideration, the thresholding of the standard deviations of the baseline values



```
observed_data_mergedWith_baseline........................
       Lat       Lon  Code  Centroid_ID   Place             D_Date  Counts_x  Counts_y
0  69.65079  18.95493  9008            8  Tromsø 2018-11-20 15:00:00        10         1
```

Figure 11: Observed and corresponding baseline

In the post code area of 9008 in figure 10, the observed count (counts_x)which was 10 infected individuals was indeed more than the average baseline value (counts_y) (1) in addition to three times the standard deviation (0) of the baseline. This indicates excess of infected individuals in the 9008-post code area which indicates a possible outbreak as shown in figure 4.3.3. Out of about 40 spikes of simulated infections which were injected, 13 of them which were spikes of outbreaks were all identified as outbreaks and the remaining 27 which were not outbreak were indeed determined by the CUSUM algorithm as either green or yellow clusters signifying no outbreaks.

5 DISCUSSION

A prototyping approach was used to explore, developed and assessed K-CUSUM with simulated synthetic data. The purpose was to determine the suitability of the hybrid of KNN and CUSUM algorithm towards empirical implementation of a cluster detection mechanism in electronic disease monitoring network (EDMON) project. The KNN was evaluated with 209 test datasets of which 208 records, representing 99.52% were correctly classified with simulated training set of 660. The CUSUM algorithm in this study was also able to accurately identify all spike injects of infected person's data as either outbreak clusters or non-outbreak clusters. The entire surveillance time was estimated to be 12.5 minutes with the input data.

The prototyping approach was highly useful in K-CUSUM since determining disease outbreak at the pre-symptomatic stage in EDMON is a novel area. As the certainty of the requirements for EDMON-Clustering at the unset was not clear, the iterative, try-and-error-approach of prototyping was ideal to systematically reveal the needed requirement out of the initial fuzzy and unclear visibility of the study.

The simulation of the synthetic data in K-CUSUM was quite useful since the algorithm needed to be tested and results evaluated with data to assess performance and robustness regarding erratic data requirements [29]. Apparently, actual data or semi-synthetic data could be used in the assessment however, there are regulatory hurdles and stringent privacy laws across the globe [30] protecting the sensitive healthcare data which cannot be toyed with. To succeed in implementing this prototype despite these challenges, synthetic data was an obvious choice since it served as a playground or surveillance range which can be manipulated in different ways to test the scalability and robustness of new algorithms without transgressing on privacy laws [31].

The KNN algorithm in the K-CUSUM demonstrated high accuracy by correctly classifying 99.52% of the tested dataset with error margin of 0.48%. A further test with another KNN algorithm in Scikit Learn with the same training and test dataset showed that the KNN in K-

CUSUM performed better as the Scikit-Learn KNN had 93.81% classification accuracy with higher margin of error of 6.19%. Disease surveillance systems which relies on geographical location of each detection point with the aim of aggregating the detections in smaller spatial units such as the zip codes for aberration detection, can easily rely on KNN with distance measures. In EDMON, the infection persons (unclassified or unknown classes) are geographically located on their respective latitude and longitude coordinates. If other detections of infectious persons have reference of post code in their geolocations, the Euclidean distances between the unclassified infected person and the referenced subjects with labeled post codes can be computed with their geocodes. What remains a hurdle is to locate a balance point of using geocodes of the surveillance subjects for detecting disease outbreak to safeguard the health of the entire community while maintaining privacy of the subjects. Much as one-way hashing and deleting anonymization techniques used from the recommended techniques of GDPR was effective in shielding sensitive data of the subjects, suitable methods for effective anonymization of location data remains a challenge since the geocodes in this experiment are required for computation of the Euclidean distances.

The CUSUM was also evaluated in this prototype by injecting spikes of simulated synthetic data. A total of 13 outbreak spikes of data which were injected were accurately identified by the CUSUM algorithms as outbreak with example shown in figure 4.3.3 and about 27 spikes which were injected as non-outbreak were all detected by the algorithm as either green or yellow non-outbreak clusters with an instance shown in figure 6.0.1. CUSUM is generally known to be very sensitive in the disease surveillance system. The current prototype results have further confirmed the effectiveness of CUSUM for aberration detection if adopted in EDMON. Therefore, the great performance of the hybrid of KNN and CUSUM is deemed ready for further assessment with empirical data for real implementation in EDMON.

The main output of the framework includes timely alerts through alarms and visualizations of detected aberrations. From the studies, various visualization tools for output displays such as bar charts, pie charts, graphs and maps have been realized. Guided with the results of the systematic review [11], ArcGIS, Leaflet-Open Source or Google Map tool was used to implement the visualization module such as what was used in Google flu trend visualization and Flu near you [32]. This visual display would mainly be mapped with other displays such as time series and graph. The maps would indicate where and when clustering and aberrations occur. Leaflet map was chosen for the prototype due to it being open source, less expensive and does not require license to use [33]. The short messaging service (SMS) was created with a trial version of an application development interface (API) known as Twilio [34]. The Twilio API was selected based on cost, ease and flexibility of use.

6 CONCLUSION AND FUTURE WORKS

K-CUSUM was explored in EDMON with prototyping method to cover the gap created by existing infectious disease surveillance system. This was a combination of KNN and CUSUM algorithm to form a spatiotemporal method. Each algorithm was assessed with simulated data.

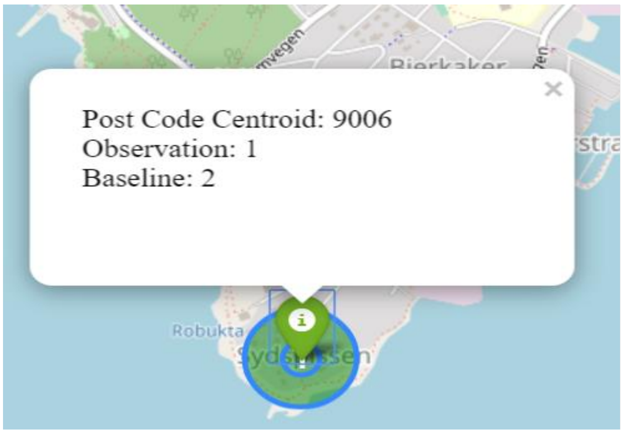

Post Code Centroid: 9006
Observation: 1
Baseline: 2

Figure 12 Non-Outbreak Cluster

The KNN registered 99.52% classification accuracy with less than 1% margin of error. The CUSUM algorithm was also able to correctly identify all outbreak and non-outbreak spikes of infection injects. Base on the results, the K-CUSUM can further be assessed with empirical data for adoption in EDMON as effective cluster detection mechanism. Suitable methods for obtaining a balance point of anonymizing geolocation attributes towards obscuring the privacy and confidentiality of diabetes subjects' trajectories while maintaining the data requirements for public good in terms of disease surveillance, remains a challenge. In future consideration, unsupervised learning methods can be explored for cluster detection in EDMON since gathering adequate training data can be challenging.

7 REFERENCES

[1] Woldaregay, A.Z., E. Årsand, A. Giordanengo, 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.

[2] Hope, K., D.N. Durrheim, E.T. d'Espaignet, et al., Syndromic surveillance: is it a useful tool for local outbreak detection? in J Epidemiol Community Health. 2006. p. 374-375.

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

[4] Nie, S., E. Lau, S. Lawpoolsri, et al., Real-Time Monitoring of School Absenteeism to Enhance Disease Surveillance: A Pilot Study of a Mobile Electronic Reporting System, in JMIR Mhealth AND Uhealth. 2014, JMIR. p. 1-10.

[5] Martin Kulldorff, R.H., Jessica Hartman, Renato Assunção,Farzad Mostashari, A Space–Time Permutation Scan Statistic for Disease Outbreak Detection. PLOSMedicine, 2005. 2(3): p. 126-224.

[6] WHO. Ebola | Ebola virus disease. WHO 2019 2019-06-13 09:57:28 [cited 2019 14-06-19]; Available from: https://www.who.int/ebola/en/.

[7] Diabetes Research and Wellness Foundation. Illness And Diabetes. 2018; Available from: https://www.diabeteswellness.net/sites/default/files/Illness%20and%20Diabetes.pdf.

[8] Casqueiro, J. and C. Alves, Infections in patients with diabetes mellitus: A review of pathogenesis. Indian J Endocrinol Metab, 2012. 16(7): p. 27-36.

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

[10] J. Barker, D.S., S.F. Bloom, Spread and prevention of some common viral infections in community facilities and domestic homes Journal of Applied Microbiology, 2001. 91(1): p. 7-21.

[11] Yeng, P.K., A.z. Woldergay, T. Solvoll, et al. A systematic review of cluster detection mechanisms in syndromic surveillance: Towards developing a framework of cluster detection mechanisms for EDMON system in Scandinavian Conference on Health Informatics. 2018. Aalborg, Denmark: Linköping University Electronic Press, Linköpings universitet.

[12] Chen, D., J. Cunningham, K. Moore, 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.

[13] Duangchaemkarn K.,et al., 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, IEEE: Seogwipo. p. 2614-2617.

[14] Rogerson, P.A., Surveillance systems for monitoring the development of spatial patterns. STATISTICS IN MEDICINE, 1997. 16(18): p. 2081–2093

[15] Rogerson, P.A., Spatial Surveillance and Cumulative Sum Methods, in Spatial and Syndromic Surveillance for Public Health, A.B. Lawson and K. Kleinman, Editors. 2005, John Wiley & Sons, Ltd. p. 269.

[16] Fanaee-T, H., Spatio-Temporal Clustering Methods Classification. 2012.

[17] gps-coordinates. GPS coordinates, latitude and longitude with Google Maps. 2019 [cited 2019 27-05-2019]; Available from: https://www.gps-coordinates.net/#.

[18] Yang, X., & Abraham O. Fapojuwo. (2015). Performance analysis of hexagonal cellular networks in fading channels - Yang - 2016 - Wireless Communications and Mobile Computing - Wiley Online Library. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/wcm.2573. doi:10.1002/wcm.2573

[19] O'Brien, S.J. and P. Christie, Do CuSums have a role in routine communicable disease surveillance? Public Health, 1997. 111(4): p. 255-8.

[20] Hutwagner, L., W. Thompson, G.M. Seeman, et al., The bioterrorism preparedness and response Early Aberration Reporting System (EARS). J Urban Health, 2003. 80(2 Suppl 1): p. i89-96.

[21] Groeneveld, G.H., A. Dalhuijsen, C. Kara-Zaitri, 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.

[22] Watkins, R.E., S. Eagleson, B. Veenendaal, et al., Applying cusum-based methods for the detection of outbreaks of Ross River virus disease in Western Australia, in BMC Med Inform Decis Mak. 2008. p. 37.

[23] Hutwagner, L., T. Browne, G.M. Seeman, et al., Comparing aberration detection methods with simulated data. Emerging infectious diseases, 2005. 11(2): p. 314-316.

[24] Josseran, L., A. Fouillet, N. Caillère, et al., Assessment of a Syndromic Surveillance System Based on Morbidity Data: Results from the Oscour® Network during a Heat Wave, in PLoS One. 2010.

[25] Watkins, R.E., S. Eagleson, B. Veenendaal, et al., Applying cusum-based methods for the detection of outbreaks of Ross River virus disease in Western Australia. BMC Medical Informatics and Decision Making, 2008. 8(1): p. 37.

[26] Gil-García, R. and A. Pons-Porrata. A New Nearest Neighbor Rule for Text Categorization. in Progress in Pattern Recognition, Image Analysis and Applications. 2006. Berlin, Heidelberg: Springer Berlin Heidelberg.

[27] Cochran, W.G., Sampling Techniques 3rd ed. 1977, New York: JOHN WILEY & SONS. 433.

[28] Liu, H. and M. Cocea, Semi-random partitioning of data into training and test sets in granular computing context. Granular Computing, 2017. 2(4): p. 357-386.

[29] Jafarpour Khameneh, N., Machine Learning for Disease Outbreak Detection Using Probabilistic Models. 2014, UNIVERSITÉ DE MONTRÉAL: PolyPublie. p. 96.

[30] Beredskapsdepartementet, J.-o., Proposition 56 LS (2017–2018)/Act on the processing of personal data (the Personal Data Act). 2018, regjeringen.no.

[31] Burgard, J.P., J.-P. Kolb, H. Merkle, et al., Synthetic data for open and reproducible methodological research in social sciences and official statistics. AStA Wirtschafts- und Sozialstatistisches Archiv, 2017. 11(3): p. 233-244.

[32] Flu Near You. Flu Near You. 2019 [cited 2019 25-05-2019]; Available from: https://flunearyou.org/#!/.

[33] Leafletjs. Leaflet — an open-source JavaScript library for interactive maps. 2019; Available from: https://leafletjs.com/.

[34] Twilio. Twilio SMS Python Quickstart - Send & Receive SMS. 2019; Available from: https://www.twilio.com/docs/sms/quickstart/python?utm_source=docs&utm_medium=social&utm_campaign=guides_tags.