Detecting space-time disease clusters with arbitrary shapes and sizes using a co-clustering approach

Sami Ullah,[1] Hanita Daud,[1] Sarat C. Dass,[1] Habib Nawaz Khan,[2] Alamgir Khalil[3]

[1]Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia; [2]Department of Economics and Management Sciences, University of Science and Technology, Bannu; [3]Department of Statistics, University of Peshawar, Pakistan

Abstract

Ability to detect potential space-time clusters in spatio-temporal data on disease occurrences is necessary for conducting surveillance and implementing disease prevention policies. Most existing techniques use geometrically shaped (circular, elliptical or square) scanning windows to discover disease clusters. In certain situations, where the disease occurrences tend to cluster in very irregularly shaped areas, these algorithms are not feasible in practise for the detection of space-time clusters. To address this problem, a new algorithm is proposed, which uses a co-clustering strategy to detect prospective and retrospective space-time disease clusters with no restriction on shape and size. The proposed method detects space-time disease clusters by tracking the changes in space–time occurrence structure instead of an in-depth search over space. This method was utilised to detect potential clusters in the annual and monthly malaria data in Khyber Pakhtunkhwa Province, Pakistan from 2012 to 2016 visualising the results on a *heat map*. The results of the annual data analysis showed that the most likely hotspot emerged in three sub-regions in the years 2013-2014. The most likely hotspots in monthly data appeared in the month of July to October in each year and showed a strong periodic trend.

Correspondence: Sami Ullah, Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia.
Tel: 0060143074781
E-mail: sami.khan3891@gmail.com

Introduction

Ability to discover significant space-time clusters in spatio-temporal data is essential in many research areas. Cluster detection plays an important role in disease surveillance as it facilitates health officials' efforts to identify targets of possible interest for interventions. In surveillance, the spatio-temporal distribution of disease occurrences in an area is studied with the aim to detect sub-regions and time points where disease occurrences tend to cluster suggesting the potential of an emerging disease outbreak in that area.

A number of space-time statistical methods have been used for disease cluster detection. Amongst these, the scan statistics introduced by Kulldorff *et al*. (1998) is one of the most commonly used approaches. This method detects circular shaped clusters by employing a cylindrical shaped scanning window on the center of each sub-region in the study area. The base of the cylinder is circular or elliptical indicating the geographical space in question, while its height denotes the time variable. At all possible time intervals, it iteratively draws circles around each sub-region with the radius size varying from zero up to a threshold value. For each window, the likelihood ratio statistic is determined and the window with the maximum value of the statistic is chosen as the most likely cluster.

Several studies have been conducted to improve the detection of clusters with non-circular shapes. Iyengar (2004) extended this algorithm to detect space-time clusters with a square-pyramid shape while Neill *et al*. (2005) proposed an algorithm to detect emerging space-time clusters with rectangular shapes. In certain situations, however, rather than adjust to such standardised areas, disease occurrences tend to cluster in very irregular shapes due to the different environmental conditions in the study area. Then the algorithms sometimes detect clusters which include low-risk regions adjacent to high-risk regions due to the geometrical shape of the scanning window. To detect irregularly shaped space-time cluster Takahashi *et al*. (2008) developed the flexible space-time scan statistics, which generates windows with irregular shapes on each sub-region by combining its neighbouring regions. The determination of threshold cluster size in this algorithm is user defined which is suggested to be (10-15) percent of the size of the entire study area. This algorithm extended the scan statistics method to discover hotspots with irregular shape however; it can detect hot spots of smaller sizes only due to the restriction on cluster size. Dong *et al*. (2011, 2012) proposed a grid-based method and Fanaee-T and Gama (2015) proposed the Eigen space method to identify space-time disease clusters with irregular shapes. These algorithms are, however, designed for detecting prospective disease clusters (Neill *et al*., 2005; Dong *et al*., 2011, 2012) and they

cannot always be applied to detect retrospective disease clusters in space-time dimensions, the latter being very important for guiding epidemiological researchers. The Eigen space method lacks the appropriate density measure of disease occurrences and is only valid when a single hotspot exists in the study area; it cannot be adapted for detecting multiple hotspots (Fanaee-T and Gama, 2015). Based on past literature and limitations in the existing algorithms, this study proposes a new method, which uses a co-clustering strategy to detect space-time disease clusters. The likelihood ratio, used in the spatial scan statistic proposed by Kulldorff (1997), is calculated for each sub-region over each time point to allow for space-time cluster detection. The density measures (likelihood ratio scores) can be organised in the form of a data matrix, the rows of which indicate the sub-regions and the columns the time points. The co-clustering technique called Bregman Block Average Co-clustering algorithm (BBAC) is applied to the density matrix to get an optimum co-clustered matrix by the concurrent analysis of rows and columns similarity. The contents of the rows and columns associated with the largest element of the co-clustered matrix are combined to approximate the most likely clusters, while those of the rows and columns related to the second largest element are combined to approximate secondary clusters. This can be visualised on *heat maps* using different colours for the different clusters. The co-clustering algorithm simultaneously partitions rows of a data matrix into clusters based on similarities along all columns, and columns into clusters based on the similarities along all rows. It concurrently assigns rows into row-clusters and columns into column-clusters. This algorithm first initialises m row-clusters and n column-clusters, and then the co-clustered matrix is calculated, where m and n are the desired number of row-clusters and column clusters respectively. The difference of the co-clustered matrix and the original matrix is used as the loss function (Anagnostopoulos *et al.*, 2008). This difference is minimised by repeatedly allocating every row to the nearest row-cluster and every column to the nearest column-cluster until a convergence is seen to obtain an optimal co-clustered matrix. The co-clustering technique BBAC developed by Banerjee *et al.* (2007) allows many loss functions and numerous co-clustering systems in which the summary statistics of rows and columns remain unchanged in the assignment process. This algorithm was applied with I- divergence to the temperature data for spatio-temporal pattern detection (Wu *et al.*, 2015).

The objective of this paper was to present a method, which is more flexible with regard to detection of prospective and retrospective space-time disease cluster with no restriction with regard to cluster shape and size. Moreover, the proposed method can be applied to detect multiple clusters in the available spatio-temporal data on disease occurrences. This paper provides insights that should be useful for the public health community and for epidemiologists in the accurate detection of space-time disease clusters.

Materials and Methods

Health officials and policy makers are interested in detecting the groups of spatial regions and time points where observed disease occurrences are higher than expected, to prioritise the provision of resources. In many countries, the health information systems collect data on disease cases from each sub-region S_i at the discrete time point t_j (days, month or years). Let C^j_i be the observed number of disease cases in region S_i at time point t_j. In the disease cluster detection framework, the observed number of

cases C^j_i is assumed to have a Poisson distribution with some unknown parameter (μ^j_i), $C^j_i \sim$ Poisson μ^j_i. Let p be the probability of disease occurrence inside the region S_i at time t_j and q the probability of disease occurrence in the area outsid S_i at the same time point. For each time point, the null hypothesis H_0 is to be tested as follows: p=q against H_1: p>q where:

$$p = \frac{C^j_i}{n^j_i}, \text{ and } q = \frac{G^j_i - C^j_i}{N^j - n^j_i}$$

The likelihood ratio

The likelihood ratio is calculated under the Poisson model for each region at each discrete time point as:

$$L(S_i, t_j) = \frac{(\frac{C^j_i}{n^j_i})^{c^j_i} \cdot (\frac{G_j - C^j_i}{N^j - n^j_i})^{(G_j - c^j_i)}}{(G_j / N^j)^{G_j}} \qquad \text{Eq. 1}$$

if p>q and $L(S_i, t_j)$=1, otherwise where C^j_i = the total observed cases of disease in sub-region S_i at time t_j; G_j = the total observed cases of disease in the entire study area at a time t_j; n^j_i = the population in sub-region S_i at time point t_j; and N^j = the total population of the entire study area. In this study, log $L(S_i, t_j)$ was used as a density measure instead of $L(S_i, t_j)$. This makes a matrix of density measures log $(L(S_i, t_j))$ in which the rows were regions S_i and the columns the time points. The BBAC algorithm was applied, to detect groups of high-density measures which are close in space as well as in time.

The Bregman Block Average Co-clustering algorithm

The BBAC algorithm with I-divergence as a loss function was applied to calculate an optimum co-clustered density matrix of the original density matrix. The original density matrix is denoted by L(S,T), where S represents the sub-regions with values $(s_1 \dots s_m)$ and T the time points with values $(t_1 \dots t_n)$. The co-clustered density matrix is denoted $L(\hat{S}, \hat{T})$ where $\hat{S}$ takes the values in region-clusters $(\hat{s}_1 \dots \hat{s}_k)$ and $\hat{T}$ the values in time-clusters $(\hat{t}_1 \dots \hat{t}_l)$. The summarised steps of the BBAC algorithm are given in this section. More details and the programming codes for the BBAC-I algorithm have been given by Wu *et al.* (2015).

Step 1: initialisation of the mapping of the region with respect to region-clusters and year to year clusters randomly. This is the essential step needed to analyse the loss of mutual information in the following step. In disease surveillance, only the most likely cluster and the secondary one are of interest, while the rest of the densities can be grouped together as insignificant clusters. Therefore, regional clusters and time-clusters were initialised to group the densities into three types of clusters: the most likely, the secondary and the insignificant clusters.

Step 2: computing the loss in mutual information as:

$$I(S, t) - I(\hat{S}, \hat{T}) = B_1(L(S, T) \| E(S, T)) \qquad \text{Eq. 2}$$

where B1 (.||.)=I-divergence; L(S,T) = the matrix of density measures with i(s,t) as elements; and E(S,T) = the matrix approximation of the original matrix with e (s, t) as elements.

The approximation matrix E(S,T) is calculated as:

$$E(S, T) = RL(\hat{S}, \hat{T})C^t \qquad \text{Eq. 3}$$

where R and C are binary matrices with the sizes m x k and n x l. The matrix R show the membership of sub-regions to region-clusters and the matrix C show the membership of years to the year-clusters. C^t represents the transpose of the matrix C. The loss in mutual information given in Eq 2 can be calculated as:

$$B_1(L(S,T)\|E(S,T)) = \sum_{\hat{s}}\sum_{\hat{t}}\sum_{s\in\hat{s}}\sum_{t\in\hat{t}} l(s,t)\log\left(\frac{l(s,t)}{e(s,t)}\right) \quad \text{Eq. 4}$$

Step 3: decomposing Eq. 4 according to rows and columns of matrix L(S,T) to find the new mapping.

Step 3a: decomposing according to row

$$B_1(L(T|s)\|E(T|\hat{s})) = \sum_{\hat{s}}\sum_{s\in\hat{s}} l(T|s)\log\frac{l(T|s)}{e((T|\hat{s})} \qquad \text{Eq. 5}$$

Step 3b: decomposing according to column

$$B_1(L(S|t)\|E(S|\hat{t})) = \sum_{\hat{t}}\sum_{t\in\hat{t}} l(S|t)\log\frac{l(S|t)}{e((S|\hat{t})} \qquad \text{Eq. 6}$$

Step 4: minimising the loss of mutual information to get the optimal co-clustering. Since Eq. 4 is decomposed to the I divergence according to row and column, individually, then this step is to get the new mapping of every row to row-clusters that minimises Eq. 5.

Step 4a:

$$I = \text{argmin}_{i\epsilon(1....k)}\ B_1(L(T|s)\|E(T|\hat{s}_i)) \qquad \text{Eq. 7}$$

and likewise, getting the new mapping of every column to column-clusters that minimises Eq. 6

Step 4b:

$$j = \text{arg min}_{j\epsilon(1....l)} B_1(L(S|t)\| E(S|\hat{t}_j)) \qquad \text{Eq. 8}$$

Step 5: re-calculating the loss in mutual information according to Eq. 4. If the difference in the loss is less than a fixed threshold value, then the mapping obtained there is the optimal co-clustering end; otherwise repeat steps 2 to 4 to get a new mapping.

Step 6: visualising the resulting co-clustered matrix by a *heat map* in which red colour indicates the most likely space-time cluster and yellow the secondary clusters. The *heat map* is the presentation of a data matrix in which the values are visualised with various colours.

Malaria case study in Pakistan

Malaria prevalence in Pakistan, estimated at 1.6 million occurrences per year (WHO, 2007), is seasonal and the country is prone to epidemics in particular geographical areas. Its geography is subtropical and the economy agriculture-based with the land characterised by vast irrigation networks. The monsoon rains provide a favorable situation for malaria in many parts of the country (Ghanchi *et al*., 2011).

Pakistan is divided into five provinces (Sindh, Balochistan, Punjab, Khyber Pakhtunkhwa and Gilgit Bultistan). Khyber Pakhtunkhwa Province is located in the Northwest sharing 1,100 km of border with Afghanistan. The total area of this province is 74,521 km^2 and the estimated population 26.5 million according to the Bureau of Statistics Khyber Pakhtunkhwa, Pakistan (BOSKP) census of 2012 (BOSKP, 2016). The land area of the province is divided into twenty-five districts and seven federally administered tribal areas (FATA). In Figure 1, the brown colour shows the distributed land area of FATA. The FATA regions are ethnically homogeneous with Khyber Pakhtunkhwa, but not politically connected to the province. However, very recently it has been decided by the federal government of Pakistan to merge these areas, politically and administratively into the province of Khyber Pakhtunkhwa. The climate in this province is of the tropical monsoon type, but most of the districts are situated beyond the tropical zone with relatively high temperatures and a cool, dry winter that runs from December to February with March to June representing the hot and dry season. Summer extends from July to September and is generally rainy. October and November represent the receding monsoon period. The province has two rainy seasons: March to April and the summer monsoon from July to September.

The District Health Information System (DHIS) collects data on disease occurrences in the whole province except in FATA. DHIS has offices in each district. All hospitals in a district report the registered malaria cases to the respective DHIS office on a monthly basis. The offices send the monthly data to the provincial DHIS office. In this study, the malaria case data for twenty-five districts were collected from this provincial DHIS office and the estimated mid-year population of each district was collected for the years 2012-2016 from the annual report of 2016 (BOSKP, 2016). The population distribution of Khyber Pakhtunkhwa is shown in Table 1. Given the number of reported disease cases and

Figure 1. Geographical map of the study area. Red colour represents the location of most likely clusters (2013-14) and yellow the secondary clusters (2016).

 OPEN ACCESS

population for each district S_i in each year t_j from 2012 to 2016, the density measures logs L (S_i, t_j) were calculated, which made a matrix of the size 25×5. The rows in the matrix of the density measure were arranged in descending order according to the total density over all time points, to visualise the dense region on the top of the *heat map*.

To co-cluster this matrix, the BBAC algorithm was applied and the resulted co-clustered matrix visualised on the *heat map* (Figure 2). The detected clusters in the annual data (2012-2016) are displayed on the geographical map of the study area (Figure 1).

The proposed method was applied to the monthly malaria data for each year (2012-2016) to detect space-time clusters in each year and interpret the seasonal pattern in malaria occurrences. The resulting co-clustered matrices of density measure for these years are visualised on the *heat maps* (Figures 3-7). All the steps of the proposed method including visualisation were implemented in MATLAB (https://www.mathworks.com/).

The proposed method was applied to annual and monthly data on malaria occurrences in Khyber Pakhtunkhwa Province to detect the space-time malaria hotspots.

Results and Discussion

The most likely and secondary hotspots were detected in the mid southern part of the province. The sub-matrix in red on the *heat maps* shows the most likely clusters while, the sub-matrix in yellow shows the secondary clusters. The blue colour represents regions with insignificant disease risk. When visualising the clus-
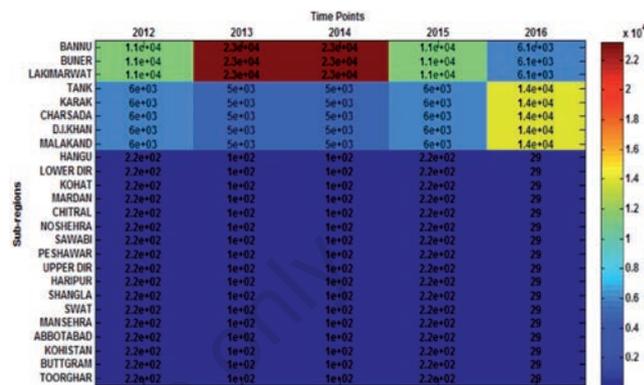

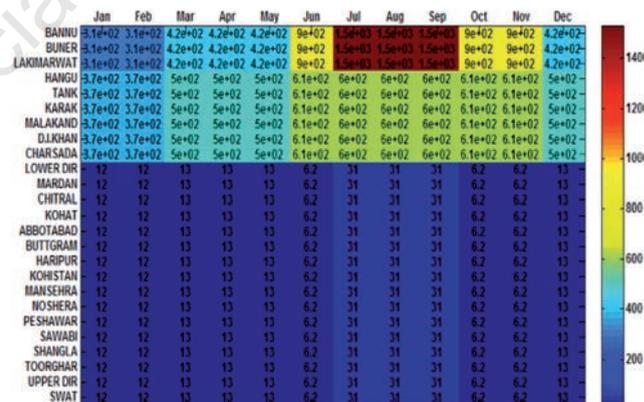**Figure 2. Heat map showing the elements of co-clustered density matrix 2012-2016.**

Table 1. Population (n) in Khyber Pakhtunkhwa by district and year.

District	2012	2013	2014	2015	2016
Abbotabad	1,110,000	1,151,000	1,172,000	1,193,000	1,215,000
Bannu	1,011,000	1,019,000	1,048,000	1,080,000	1,110,000
Buner	811,000	888000	922,000	957,000	994,000
Buttgram	200,000	210,000	217,000	228,000	239,000
Charsada	1,500,000	1,558,000	1,603,000	1,649,000	1,696,000
Chitral	443,000	461,000	473,000	485,000	497,000
Hangu	485,000	505,000	522,000	539,000	556,000
Haripur	935,000	955,000	975,000	997,000	1,019,000
Karak	693,000	693,000	716,000	739,000	763,000
Kohat	880,000	904,000	934,000	964,000	995,000
Kohistan	459,000	479,000	479,000	480,000	480,000
Lower Dir	1,100,000	1,182,000	1,222,000	1,264,000	1,307,000
Malakand	731,000	738,000	763,000	789,000	815,000
Mansehra	1,603,000	1,639,000	1,678,000	1,719,000	1,760,000
Mardan	2,267,000	2,267,000	2,335,000	2,405,000	2,478,000
Noshera	1,310,000	1,336,000	1,375,000	1,415,000	1,456,000
Peshawar	3,292,000	3,392,000	3,513,000	3,638,000	3,768,000
Sawabi	1,520,000	1,583,000	1,630,000	1,678,000	1,728,000
Shangla	690,000	700,000	723,000	747,000	771,000
ToorGhar	227,000	227,000	247,000	257,000	266,000
Upper Dir	810,000	862,000	886,000	911,000	936,000
D.I. Khan	1,300,000	1,373,000	1,418,000	1,464,000	1,511,000
Lakkimarwat	767,000	777,000	802,000	827,000	853,000
Swat	2,010,000	2,056,000	2,125,000	2,197,000	2,271,000
Tank	350,000	376,000	388,000	400,000	413,000

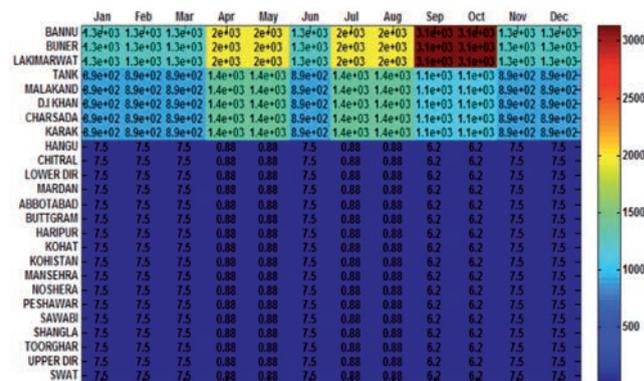**Figure 3. Heat map for the co-clustered density matrix 2012.**

Figure 4. Heat map for the co-clustered density matrix 2013.

ters in the annual data on the geographical map of the province (Figure 1), it is obvious that the regions in the secondary cluster were spatial neighbours of the most likely hotspots, but they emerged at different points in time. For the purpose of disease surveillance, we were only interested to detect regions with unusually high disease occurrences, *i.e.* the most likely and secondary hotspots. From the results, it is evident that malaria occurrences tended to cluster in three sub-regions (Bannu, Buner and Lakimarwat) for the years 2013-2014 and in five sub-regions (Tank, Karak, Charsadda, D.I. Khan and Malakand) for the year 2016. The sub-regions in the detected clusters exhibited a significant rate of malaria occurrences.

To justify the spatial dimension of the most likely cluster in the annual data, the average malaria incidence rate per hundred thousand population in the years 2013-2014 for each district was displayed on a bar chart (Figure 8A), in which the highest malaria rate was found in the three regions (Bannu, Buner and Lakki Marwat). To confirm the temporal dimension of the most likely cluster, the malaria incidence rate in the combined three most likely districts for each year was also displayed on a bar chart (Figure 8B), which showed the years 2013 and 2014 to have the highest malaria rates in the study period.

Similarly the malaria incidence rate in the secondary cluster period (2016) for each district was displayed on a bar chart (Figure 9A) and the malaria incidence rate in the five secondary districts for each year as well (Figure 9B). Figure 9A shows the five districts (Tank, Karak, Charsadda, D.I. Khan and Malakand) to be the peak regions and Figure 9B the year 2016 to be the peak year. The bar charts in Figures 8 and 9 confirmed the detected clusters in the annual data.

The most likely clusters in the monthly data for each year occurred in the same three regions (Bannu, Lakki Marwat and Buner) (Figures 3-6). None of these regions appeared to be a likely cluster in the year 2016 as in Figure 7, which shows that the most likely malaria hotspots had changed location in 2016. In the years 2012, 2014 and 2015, the most likely cluster occurred in the months of July, August and September. In 2013, it occurred in September and October and in 2016 in August and September. The most likely hotspots occurred in the summer monsoon season (July-October) each year showing a strong seasonal trend. In this study, no cluster appeared in the dry winter season (November-February) similar to the result of the study conducted in Buner District by Ibrahim *et al*. (2014), in which the highest number of malaria cases was recorded in July and the lowest in January. It is evident from the results that malaria occurrences tended to occur in the hot, wet season, which provides a favorable environment for mosquito breeding.

Out of 25 districts, eight were shown to have had unusually high malaria occurrences. Seven of these districts were geographically adjacent to FATA. Indeed, people from FATA and Afghanistan were found to be the malaria carriers in Pakistan (Kazmi *et al*., 2001; Hussain *et al*., 2016). Most of the Afghan refugees and the Internally Displaced People (IDPs) from FATA were settled in these neighbouring districts due to military operations. The heavy influx and continued presence of Afghan refugees and IDPs from FATA may have contributed to the high malaria occurrences and added to the malaria burden in these districts. The high malaria occurrences in the districts of Malakand and Charsada can be partially attributed to floods that affect these districts in the summer monsoon season.

Conclusions

A new method has been proposed which uses a co-clustering approach to detect disease hotspots with no restrictions with regard to cluster shape and size. This method can detect prospective and retrospective disease clusters with no low-risk region included in the clusters. The proposed algorithm is valid for detecting multiple clusters existing in the disease data. The prospective and retrospec-
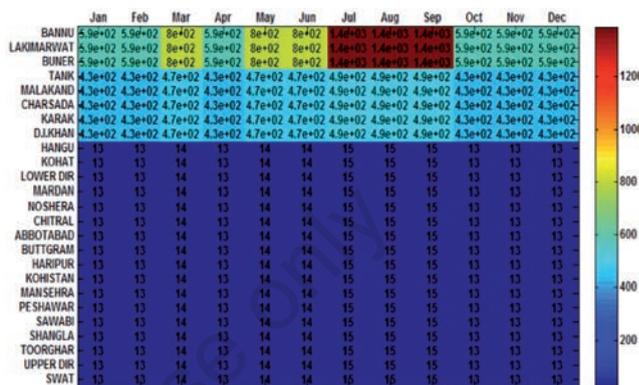

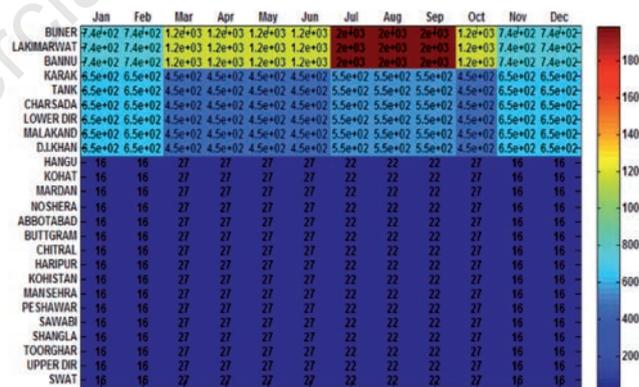Figure 5. Heat map for the co-clustered density matrix 2014.

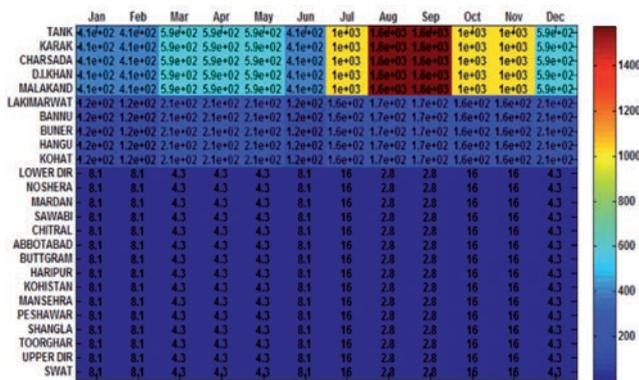Figure 6. Heat map for the co-clustered density matrix 2015.

Figure 7. Heat map for the co-clustered density matrix 2016.

tive malaria clusters in Khyber Pakhtunkhwa, detected by the proposed method, all exhibited significant malaria risk, while no low-risk regions were included in the clusters (Figures 8 and 9).

Detection of such types of disease hotspots provides useful insights about the space-time pattern in disease occurrences for health officials and policy makers. On the basis of this information, public health officials can prioritise the allocation of resources, *e.g.*, funds, medical equipment, staff and activation of disease control programmes at exact times and in exact geographical locations. Similar policy recommendations have been suggested by Khan *et al.* (2016). Based on the results of annual data analysis, the secondary clusters emerged in 2016 and still exist, which needs the

possible interventions of health officials on the priority basis to take control of the malaria outbreaks. In addition, the proposed approach can guide epidemiologic research on finding factors which cause specific disease outbreaks in a geographical area. The method also helps epidemiologists to test the hypotheses about the cause of an epidemic and provide sufficient information about the transmission dynamics of a particular disease. This technique is computationally less intensive; however, the results are influenced by specifying the number of regional clusters and time-clusters in the algorithm for large data sets. Future work is required to guide the proposed method to finding the optimum number of region-clusters and time-clusters in the data.
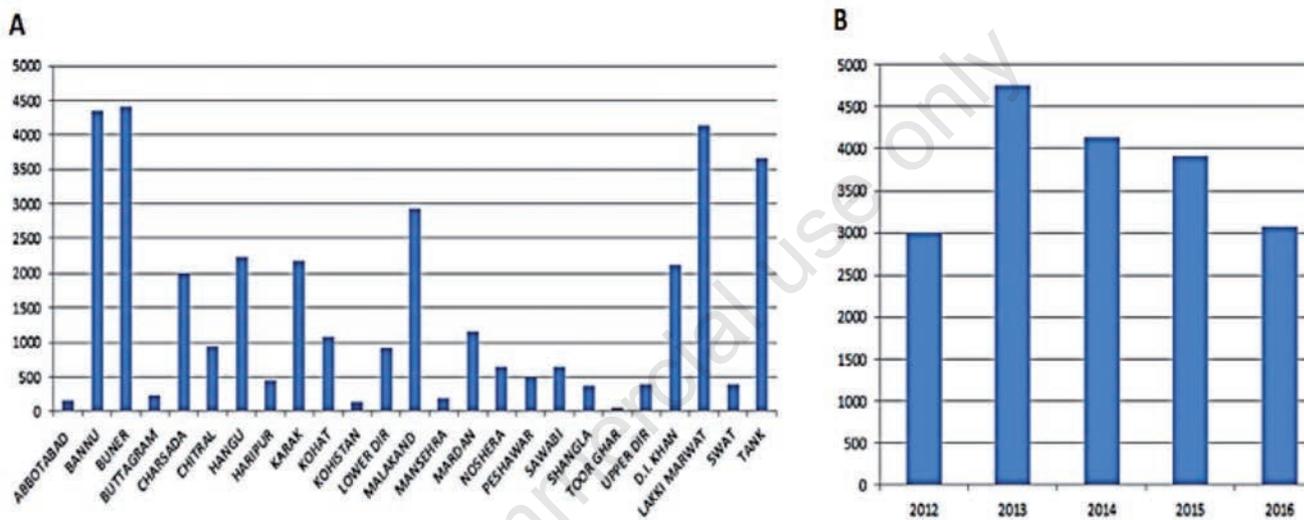

Figure 8. Bar chart showing the average malaria incidence rate in the years (2013-2014) for each district (A), and the malaria incidence rate in three most likely regions for each year in the study period (B).
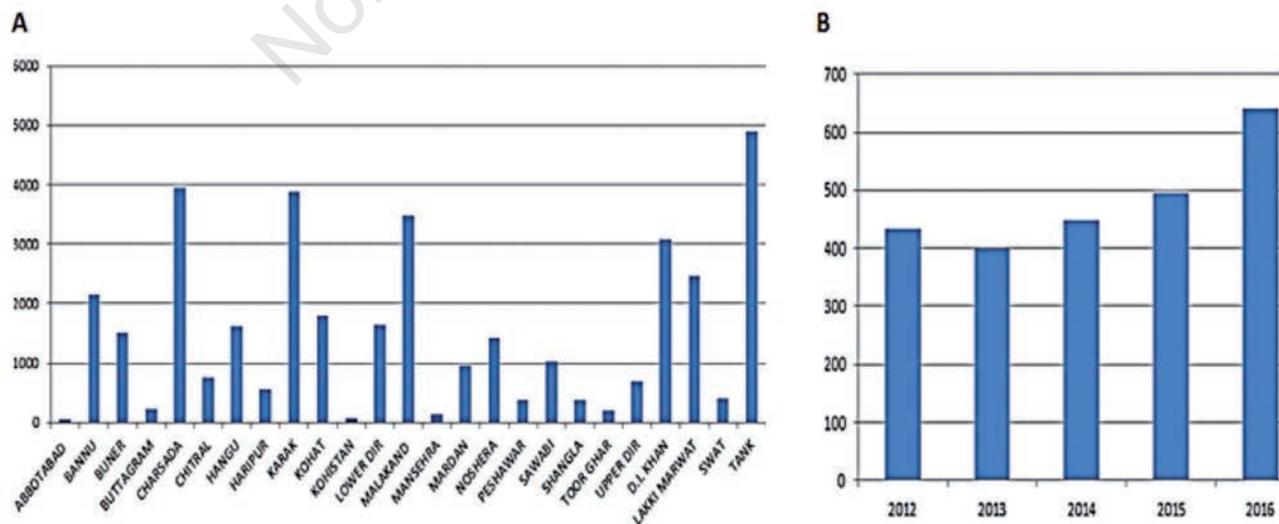

Figure 9. Bar chart showing the malaria incidence rate in the year 2016 for each district (A), and in 5 secondary districts for each year in the study period (B).

References

Anagnostopoulos A, Dasgupta A, Kumar R, 2008. Approximation algorithms for co-clustering. Proceedings of the twentyseventh ACM SIGMODSIGACTSIGART Symposium on Principles of Database Systems 60:201-10.

Banerjee A, Dhillon I, Ghosh J, Merugu S, Modha DS, 2007. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. J Mach Learn Res 8:1919-86.

BOSKP, 2016. Development statistics of Khyber Pakhtunkhwa 2016. Available from: http://kpbos.gov.pk/development-statistics.php

Dong W, Zhang X, Jiang Z, Sun W, Xie L, Hampapur A, 2011. Detect irregularly shaped spatio-temporal clusters for decision support. In Proceedings of 2011 IEEE International Conference on Service Operations, Logistics Informatics 231-236.

Dong W, Zhang X, Li L, Sun C, Shi L, Sun W, 2012. Detecting irregularly shaped significant spatial and spatio-temporal clusters. In Proceedings of the 2012 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, pp. 732-43.

Fanaee-T H, Gama J, 2015. Eigenspace method for spatiotemporal hotspot detection. Expert Syst 32:454-64.

Ghanchi NK, Ursing J, Beg MA, Veiga MI, Jafri S, Mårtensson A, 2011. Prevalence of Resistance associated polymorphisms in *Plasmodium falciparum* field isolates from Southern Pakistan. Malar J 10:18.

Hussain I, Qureshi NA, Afzal M, Shaheen N, Ali A, Ashraf A, 2016. Prevalence and distribution of human *Plasmodium infection* in federally administrative tribal areas of Pakistan. Acta Parasitol 61:537-43.

Ibrahim SK, Khan S, Akhtar N, 2014. Epidemiological finding of malaria in district Buner Khyber Pakhtunkhwa, Pakistan. World J Med Sci 11:478-82.

Iyengar VS, 2004. On detecting space-time clusters. Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1:587-92.

Kazmi JH, Pandit K, 2001. Disease and dislocation: the impact of refugee movements on the geography of malaria in NWFP, Pakistan. Soc Sci Med 52:1043-55.

Khan HN, Razali RB, Shafie AB, 2016. Modeling determinants of health expenditures in Malaysia: evidence from time series analysis. Front Pharmcol 7:69.

Kulldorff M, 1997. A spatial scan statistic. Commun Stat-Theor M 26:1481-96.

Kulldorff, Martin, Athas WF, Feuer EJ, Miller BA, Key CR, 1998. Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. Am J Public Health 88:1377-80.

Neill DB, Moore AW, Sabhnani M, Daniel K, 2005. Detection of emerging space-time clusters. Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 218-27.

Takahashi K, Kulldorff M, Tango T, Yih K, 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. Int J Health Geogr 7:14.

WHO, 2007. Strategic plan for malaria control and elimination in the WHO Eastern Mediterranean region 2006-2010. DocumentWHOEM/MAL/340/E/02.07/1000. Available from: http://www.emro.who.int/dsaf/dsa741.pdf

Wu X, Zurita-Milla R, Kraak M-J, 2015. Co-clustering geo-referenced time series: exploring spatio-temporal patterns in Dutch temperature data. Int J Geogr Inf Sci 8816:624-42.