Detection of Space-Time Disease Clusters Using A Matrix Factorization Method

Sami Ullah*1, Fahim Raees2, and Zahid Khan3
1Department of Mathematics and Statistics, Institute of Business Management (IoBM), Korangi Creek, Karachi, Pakistan
2Department of Mathematics, NED UET, Karachi, Pakistan
3Department of Mathematics and Statistics, Hazara University Mansehra, Pakistan.
*Correspondence Author: sami.ullah@iobm.edu.pk

Abstract
Space-time cluster detection has important applications in public health management and epidemiology to devise disease prevention strategies and to find the causes of a particular disease outbreak in a country. This study introduced a new method to detect the potential space-time clusters with no restriction on cluster shape and size and further visualize them distinctly on the heat map. The proposed algorithm is based on matrix factorization technique to find the significant components in spatial as well as temporal dimension. Applications to malaria data in Khyber Pakhtunkhwa, Pakistan shows that the proposed method is effective in detecting the potential clusters.

Keywords: disease cluster, singular value decomposition, Heatmap, spatiotemporal data

Introduction
The detection of clusters with unusually high density in space-time dimension plays a central role in disease surveillance and public health management, for example, detecting the outbreak area in a city where the concentration of disease cases is significantly higher for a specific time period. In this situation a cluster is defined to be the area consists of one or more counties, where the concentration of the particular disease cases (inside the area) is significantly higher than the remaining outside area for a specific time period. Such clusters indicate the area and time period of a particular disease epidemic in a country.
A number of statistical methods have been designed for detecting regularly shaped (circular, square or rectangular) space-time clusters(Assunção & Correa, 2009; V S Iyengar, 2005; Vijay S. Iyengar, 2004; M Kulldorff, 2001; Martin Kulldorff et al., 1998; Neill, 2006). In certain situations, where the disease cases tend to bunch in an irregularly shaped area due to some
irregular features in the terrain, these methods are not practically feasible to detect clusters with irregular shapes. In this study, our main concentration is on detecting space-time clusters with irregular shapes. A few of the algorithms have been developed to detect space-time clusters with irregular shapes. A flexible space-time scan statistic in (Takahashi et al., 2008), a grid-based method in (Dong et al., 2012, 2011) space-time permutation scan statistic method in (Costa & Kulldorff, 2014) and the Eigen space method in (Ullah, 2022; Ullah et al., 2017, 2018) were proposed to detect space-time clusters with irregular shapes.

These algorithms have limitations due to complexity and some of them have no system for visualizing the detected clusters, which is very important for the easy understanding of the dynamics of the disease occurrences in space-time dimension.

Addressing these limitations in the above methods, a new method is applied in this study, to identify the potential space-time clusters and to visualize both of them distinctly on the heat map. This algorithm uses the same density measure, Log Likelihood Ratio (LLR) proposed in cylindrical scan statistic method (M Kulldorff, 2001; Martin Kulldorff et al., 1998) but calculated with a different setting; In cylindrical scan statistic LLR is calculated for each possible set of regions at each possible time interval while in this algorithm the LLR is calculated for each region at each discrete time-point (day, month, or year) to make a density matrix. Moreover, in cylindrical method the LLR is replaced by 1 in the window with inside disease risk less than the outside area while in this method, it is replaced by zero, because in case of rare diseases where most of the LLRs in the density matrix are less than one, replacing lower density by one can lead to wrongly detect them as the likely clusters. This algorithm uses singular value decomposition to find the left and right singular vectors of the density matrix. For the purpose of comparison, the singular vectors associated with the largest eigenvalue are considered only because the first principal singular vector explains the major directions and the largest proportion of inertia in the data (Abdi et al., 2010).

The components in the principal left singular vector are associated with the rows (sub-regions) and in the principal right singular vector to the columns (time points) of the density matrix. The contents of rows and columns associated with the significantly higher components in both left and right singular vectors are combined to approximate the most likely clusters. The LLRs of these likely clusters are replaced by their average value in the density matrix and the density measures greater than this average value denotes transient clusters. The density matrix is then visualized on the heat map to visualize the most likely and transient clusters with different colors. The detailed stepwise process is given in the following section. Furthermore, the proposed method was applied to monthly malaria data in the Khyber Pakhtunkhwa province, Pakistan (2014-2016) to detect space-time clusters.

**Materials and Methods**

**Step-1**: Calculating the Likelihood ratio scores.

Health organizations collect data on disease occurrences from each administrative sub-unit (sub-region) at regular time-points (Days, month, or years) to conduct disease surveillance. The
disease occurrences are assumed to be Poisson distributed over all the sub-regions at each time-point (Martin Kulldorff, 1997). Given the data on disease occurrences and population the likelihood ratio score for each sub-unit over each time-point is calculated under Poisson model as in Eq. (1).

$$L(S_i, t_j) = \left( \frac{c_i^j/n_j^j}{(G_j-n^j_i)^{c_j}} \right)^{c_j}$$

If $$\frac{c_i^j}{n_j^j} > \frac{c_j^j-c_i^j}{N_j-n^j_i}$$ and $$L(S_i, t_j) = 0$$ otherwise

where $$c_i^j$$ is the number of reported cases of a particular disease in sub-region $$S_i$$ in a time period $$t_j$$; $$G_j$$ the total reported cases of a particular disease in the whole study in a time period $$t_j$$; $$n_i^j$$ the population of sub-region $$S_i$$ in the time period $$t_j$$; and $$N^j$$ the population of the whole study area. For computational simplicity, the log likelihood ratio is used as a density measure. The density measures for all sub-regions and time-points can be organized in the form of matrix $$L$$ of size $$n \times m$$ where $$n$$ represents the number of sub-regions and $$m$$ represent the number of time-points.

$$L = \begin{bmatrix}
\log L(1,1) & \cdots & \log L(1,m) \\
\vdots & \ddots & \vdots \\
\log L(n,1) & \cdots & \log L(n,m)
\end{bmatrix}$$

The rows of the density matrix represent the spatial dimension and columns the temporal dimension. Calculating this $$n \times m$$ matrix, our goal is to detect any significantly higher density in spatial dimension as well as in the temporal dimension.

**Step-2:** Singular Value Decomposition (SVD) of the density matrix:

The SVD of the density matrix $$L$$ is a factorization of the form as in Eq. (2)

$$L = U \Sigma V^T$$

where the columns of $$U$$ and $$V$$ consist of the left and right singular vectors, respectively, and $$\Sigma$$ is a diagonal matrix whose diagonal entries are the eigenvalues of the density matrix $$L$$. For comparison, the singular vectors corresponding to the largest eigenvalue are considered only. These principal singular vectors account for major directions and extract the largest amount of inertia in the data. Let $$\bar{A}$$ be the principal left singular vector and $$\bar{B}$$ be the principal right singular vector corresponds to the largest eigenvalue, then $$\bar{A} = (a_1, a_2, a_3, \ldots, a_m)$$ and $$\bar{B} = (b_1, b_2, b_3, \ldots, b_n)$$, where $$a_i$$ represent rows (sub-regions) and $$b_j$$ represent the columns (time points).

**Step-3:** Finding the Significant components in the principal left principal and right singular vectors.
Standardizing vectors, $\vec{A}$ and $\vec{B}$ and finding the components in both vectors with significantly higher values at significance level $\alpha = 0.10$. The components of vectors $\vec{A}$ and $\vec{B}$ that obtain one tailed p-value less than $\alpha$ are considered significantly abnormal.

**Step-4**: Approximating the most likely disease hotspot.

In the case of the significant elements exists in both vectors, the contents of the rows and columns associated with the significant components of both vectors are combined to approximate the most likely space-time cluster, otherwise, no space-time cluster will exist in the study area.

**Step-5**: Replace the likelihood ratio scores of most likely clusters by their average in the density matrix $L$. The LLRs in the matrix $L$ higher than the average LLR of most likely clusters shows transient clusters.

**Step-6**: Display the density matrix $L$ on the heat map to visualize the likely and transient clusters with different colors.

The proposed method is explained in the flowchart in [Error! Reference source not found.].
Results

Malaria cluster detection in Khyber Pakhtunkhwa, Pakistan
The proposed method was applied to detect potential space-time clusters in the monthly data on malaria from January-2014 to December-2016 in Khyber Pakhtunkhwa province, Pakistan. Khyber Pakhtunkhwa, the northwestern province of Pakistan shares a very long border with the neighboring country Afghanistan. The province consists of twenty-five districts and seven Federally Administered Tribal Areas (FATA). The FATA regions are ethnically homogeneous with Khyber Pakhtunkhwa, but not politically connected to the province. The total area of this province is 74,521 km² with an estimated population 26.5 million (BOSKP, n.d.). The climate of this province is of the tropical monsoon type, but most of the districts are situated beyond the tropical zone with relatively high temperatures and dry winter. The rainy summer season runs from July to September. October and November represent the receding monsoon period. From December to February is the dry cool winter and March to April representing the spring season. Malaria incidence is periodic and affected by climatic conditions; therefore, it is essential to analyze the monthly case data for explaining the periodic trend in malaria incidence.

The malaria case data for each of twenty-five districts from January-2014 to December-2016 were collected from the provincial office of District Health Information System (DHIS), Khyber Pakhtunkhwa. All hospitals in a district report the registered malaria cases to the respective district office of DHIS on a monthly basis and these offices further report the monthly data to the provincial DHIS office. The Population Data for each district came from Bureau of statistics Khyber Pakhtunkhwa (BOSKP, n.d.). The malaria case data and population data were processed in the proposed algorithm.

The results showed four districts (Bannu, Buner, Karak, Lakimarwat) to be the potential malaria clusters occurred in (Aug-Sep) in the year 2014, (July- Sept) in 2015 and September in 2016 with average LLR = 1.7 e^{0.3} as shown in Figure 2.
Figure 2: The heat map shows the potential clusters in Khyber-Pakhtunkhwa.

Three districts (Charsada, Malakand, and D.I Khan) were detected to be the transient clusters each for one month October, July, and August respectively in the year 2016. Each of these districts exhibited a high risk for one month only and in the same month, no other district appeared to be the significant cluster. The location and size of the likely and transient malaria clusters were displayed on the geographical map of Khyber Pakhtunkhwa province with light-red and dark-red color respectively (Figure 3).
Figure 3: The geographical map shows the location and size of the likely and transient clusters in Khyber-Pakhtunkhwa.

In general, the proposed method detected two likely malaria clusters appeared three times during the years 2014-2016 as shown with light red color in Figure 3. One cluster detected in the south of the province consists of three adjacent districts (Bannu, Karak, Lakimarwat) and the other cluster appeared in the middle, comprising of one district (Buner), was an isolated likely cluster.

**Discussion**

The application of the proposed algorithm to the malaria data detected four districts to be the most likely clusters appeared in the summer monsoon each year showed a strong seasonal trend. All the most likely clusters emerged in the rainy summer season each year. Hot wet season creates a fertile breeding atmosphere for mosquitoes which transfer the malaria parasite into the human host. These results are consistent with the previous study in (Ibrahim et al., 2014) and (Khan et al., 2012) conducted in district Buner and Bannu respectively in which the highest malaria case concentration observed in the summer monsoon and the lowest in the winter season. It is obvious from Figure 3 that most of the likely regions were adjacent to FATA regions, which were found to be the high malaria risk region due to the heavy influx of Afghan
refugees (Hussain et al., 2016). The presence of Internally Displaced People (IDP) from FATA regions due to the recent military operation against terrorists may cause malaria burden in these districts. In addition, refugee camps are very congested and insanitary, which may contribute to the malaria spread in the neighboring regions.

Conclusion
In this study, a matrix factorization strategy was proposed for detecting space-time disease clusters. The proposed approach can detect prospective and retrospective space-time clusters and clearly visualize on the heat map. This method is efficient in terms computational cost for detecting shape-free and size-free clusters; however, it has limitation to detect the secondary clusters if exists in the data. This paper provides useful insights for the public health community and epidemiologists in the quick detection of disease clusters in space-time dimension. The results of such analysis at country level provides guidance for proper planning and decision making in disease surveillance and crimes monitoring departments.

Acknowledgement
We are thankful to the editorial team for their support.

References
Detection of Space-Time Disease Clusters Using A Matrix Factorization Method


