scientific reports

OPEN

Novel Eigen space method for multiple Spatiotemporal rare diseases clusters detection: a case study of waterborne disease

Muhammad Fayyaz[1], Alamgir[1], Sami Ullah[2], Hameed Ali[3], Abdulrahman Obaid Alshammari[4], Zeineb Klai[5]✉ & Bilal Himmat[6]✉

The development of robust and efficient analytical tools for informed decision making, mainly in epidemiological contexts, remains a persistent challenge. This study presents an enhanced algorithm designed to accurately detect vulnerable spatiotemporal hotspots associated with unexpected disease outbreaks. We introduce an improved novel Multi-EigenSpot algorithm by systematically integrating the functionalities of both EigenSpot and its Multi-HotSpot extension. The EigenSpot algorithm effectively identifies single spatiotemporal clusters, it is unable to detect multiple hotspots. The Multi-EigenSpot algorithm overcomes this limitation through an iterative process of cluster detection and removal. However, challenges persist regarding computational efficiency and sensitivity in identifying rare clusters. To address these limitations, we propose an efficient Novel Multi-EigenSpot algorithm. This method is designed to detect multiple irregularly shaped, rare spatiotemporal clusters with significantly improved computational performance. Furthermore, the proposed algorithm integrates heatmap visualizations to enhance the interpretability of detected clusters. We evaluated our method using monthly waterborne disease surveillance data from Khyber Pakhtunkhwa, Pakistan (January - December 2024), comparing its performance against both the original EigenSpot and Multi-EigenSpot algorithms. Empirical results demonstrate the proposed algorithm's superior performance in accurately identifying multiple spatiotemporal clusters. Beyond public health surveillance, this algorithm is readily adaptable to diverse domains, including crime analysis, environmental hazard detection, and other applications requiring spatiotemporal clustering.

Keywords Waterborne diseases, Spatiotemporal cluster detection, Multi-EigenSpot algorithm, Singular value decomposition, Rare event detection, Epidemiological hotspot analysis, Heatmap

Access to clean water is a fundamental human right, essential for sustaining life and health. Waterborne diseases, primarily affecting the gastrointestinal tract, pose a major global public health challenge, arising from diverse pathogenic organisms including viruses, bacteria, and parasites (WHO, 2023). Inadequate Water, Sanitation, and Hygiene (WASH) infrastructure remains a primary contributor to the global disease burden[1], with associated deficiencies causing approximately 1.5 million deaths. In low- and middle-income countries (LMICs), nearly 69% of diarrheal mortality is attributed to inadequate WASH facilities (WHO, 2023). Children under five years of age are disproportionately affected, with waterborne infections persisting as a leading cause of morbidity and mortality in this demographic. Notably, an estimated 1.4 million diarrheal deaths in 2019 could have been prevented through effective WASH interventions[2–5].

In Pakistan, only 39–41% of the population has access to safely managed drinking water, while approximately 68% have access to basic sanitation services. Subsequently, waterborne diseases (WBD) account for about 30–50% of all diseases and up to 40% of deaths. UNICEF reports over 53,000 annual child deaths from diarrhea

[1]Department of Statistics, University of Peshawar, Peshawar, Pakistan. [2]School of Mathematics Statistics and Mechanics, Beijing University of Technology, Beijing, China. [3]Department of Mathematics, Statistics & Computer Science, The University of Agriculture Peshawar, Peshawar, Pakistan. [4]Department of Mathematics, College of Science, Jouf University, 72388 Sakaka, Saudi Arabia. [5]Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Arar, Saudi Arabia. [6]Department of Software Engineering, Faculty of Computer Science, Sayed Jamaluddin Afghani University (SJAU), Kunar, Afghanistan. ✉email: Zeineb.klai@nbu.edu.sa; bilalhimmat@sjau.edu.af

linked to poor WASH systems. Water and sanitation problems also carry out major economic burdens, estimates suggest annual losses amounting to PKR 343 billion (≈ USD 1.5 billion). In Khyber Pakhtunkhwa, endemic water contamination is intensified by old infrastructure, sewage leaks, and insufficient treatment systems. Nearly 80% of water samples in some areas have been found unsafe for consumption, contributing to outbreaks of cholera, typhoid, hepatitis, and bloody-diarrhea[6].

The burden of WBD in Pakistan emphasizes the pressing need for systemic WASH interventions, improving water quality, sanitation access, hygiene education, and infrastructure monitoring, to save lives and reduce economic and health consequences. Several epidemiological studies on waterborne diseases have been carried out in Khyber Pakhtunkhwa, focused on the epidemiological characteristics and risk factor analysis of these diseases[7–9]. However, the proposed work develops algorithms for accurate detection of waterborne disease clusters to enhance epidemiological decision-making. Spatiotemporal cluster detection is central to epidemiological research, revealing disease burden distributions and underlying health determinants. By analyzing population level patterns across regions and time periods, epidemiologists gain insights into transmission dynamics and prioritize high-risk areas. Such analyses typically require examining ecological, socio demographic, and infrastructural factors associated with elevated prevalence[10,11]. Cluster mapping delineates hotspots and informs targeted resource allocation. Identifying spatiotemporal disease clusters is pivotal for strengthening public health surveillance and guiding effective interventions. Health agencies routinely collect spatiotemporal case data to monitor disease dynamics and mitigate outbreaks. Systematic analysis of these patterns enables detection of localized incidence surges, facilitating timely resource deployment. A cluster is formally defined as a spatial or temporal domain where observed cases significantly exceed expected counts[12,13]. While diverse statistical techniques detect regularly shaped clusters, scan statistics have emerged as the predominant method, particularly for circular clusters[14–16]. These methods employ a cylindrical scanning window traversing the study area: the base defines a circular/elliptical spatial zone, while the height represents the temporal dimension, capturing both persistent and emerging clusters. As the window expands from minimum to maximum radius, overlapping regions are evaluated via likelihood ratio tests comparing observed versus expected cases under spatial randomness. The window maximizing the test statistic and identifies the most likely cluster, indicating significant disease incidence elevation, an approach proven effective in surveillance contexts[17–19]. Nevertheless, traditional scan statistics struggle to detect irregularly shaped clusters, especially in geographies constrained by natural boundaries (rivers, mountains) or urban landscapes, where circular/elliptical windows inadequately capture disease dispersion. Advanced methods detecting arbitrary-shaped clusters address this limitation[20–23], but often rely on restrictive distributional assumptions (Poisson or Gaussian), limiting applicability to complex modern datasets. Moreover, scan-statistic algorithms inherently require strict parametric assumptions, impairing performance when these assumptions are violated, particularly for nontraditional or structurally complex data. In addition, these algorithms are designed to identify regular-shaped clusters and are less efficient for irregular-shaped clusters. These algorithms require high-quality data, making them vulnerable to noise and outliers[24].

The EigenSpot algorithm was developed by Fanaee-T and Gama[25], as a nonparametric, eigenspace based algorithm capable of detecting disease clusters without presuming any particular data distribution, quality, or cluster shape. Yet it can identify only a single hotspot, rendering it inadequate for uncovering multiple high-risk cluster over space and time. To overcome this, Sami Ullah et al.[26] proposed a generalized Multi-EigenSpot algorithm that, like its predecessor, relies on eigenspace techniques, but substitutes expected case counts for population data as its baseline, thereby enabling the detection of several spatiotemporal clusters. Eigenspace methods have since gained widespread application from data mining and signal processing to information retrieval powering innovations, such as Google's search engine, famously explained in "The $25 000 000 000 Eigenvector"[27], and behind the BellKor team's 2008 Netflix Prize winning use of singular value decomposition in collaborative filtering[28].

The pioneering work by Fanaee-T and Gama[25] and Ullah et al.[26], introduced eigenspace methods to epidemiology, marking the first application of these techniques to disease cluster identification. However, related to clustering and anomaly detection, hotspot detection (also called outbreak or event detection) is distinct. Such as clustering partitions an entire dataset into groups, anomaly detection flags unexpected individual instances, and hotspot detection pinpoints areas of statistically significant deviation from a defined baseline.

All of the above-mentioned approaches involve scanning the entire space, being computationally laborious and time-consuming. The computing time for spatial scan statistics is given as $O\left(N^3\right)$, whereas the computation time for space-time scan statistics is given as $O\left(N^4\right)$. Several recent initiatives have been undertaken to minimize this complexity. Spatial scan statistics approach that is more efficient, requiring just $O\left({}^{1}/_{\epsilon} * N^2 * log_2\left(N\right)\right)$. Under optimal conditions, the minimal complexity for Sat Scan has not yet reached below O(N^3) because it has not yet achieved that level. This enormous processing cost makes it almost impossible to employ them in applications that are used in the real world or with large-scale datasets. In terms of time complexity, the Eigenspot and Multi Eigenspot methods are both considered to be $O\left(KN^2\right)$ or $O\left((mn)^2\right)$ [25]. This is significantly faster than scanning methods, yet still presents challenges for high-dimensional data due to the super-linear growth in computational time.

This study covers the following three-fold gap in the literature:

1. It is not appropriate for detecting clusters of rare diseases.
2. In nations like Pakistan, a zero count frequently denotes unrecorded data or a lack of data availability. However, a zero that falls between two high counts is misclassified as a disease cluster, which results in erroneous cluster identification.

3. Lastly, the EigenSpot methods are computationally costly on large-scale spatiotemporal matrices, especially when repeated singular value decomposition (SVD) calculations are required for multiple clusters.

The novelty of this works lies in the following ballots:

- To develop an efficient approach for identifying spatiotemporal clusters of rare disease with linear complexity in both spatial and temporal dimensions.
- To present a novel method designed for identifying spatiotemporal clusters in rare diseases and characterized by its efficiency in time complexity.
- To integrate singular value decomposition spare (SVDs) instead of SVD to manage false positive detection of cluster.
- To provide robust alternative of classical tools for finding abnormal components.

The paper is organized into the following sections:

- Section "Methodology" presents the study area, data sources, and the proposed methodology, detailing the novel EigenSpace algorithm based on SVDs, Z-control charts, and heatmap visualization.
- Section "Results and discussions" describes the implementation steps and computational procedures of the algorithm, including matrix formation, anomaly detection, and iterative updates.
- Section "Performance evaluation" discusses the results and findings from the application of the proposed method to typhoid disease data in Khyber Pakhtunkhwa, highlighting detected clusters and comparing performance with existing approaches.
- Section "Discussion and conclusion" evaluates the computational efficiency of the proposed method.
- Section 6 concludes the study with key insights, limitations, and future directions.

Methodology

Materials and methods

Approach

The analytical procedures were implemented in MATLAB R2017, where the proposed Novel EigenSpace Method was systematically compared against the baseline algorithms. For spatial visualization, the Pakistan administrative boundary shapefile was obtained from the Humanitarian Data Exchange (HDE)[29] portal (https://data.humdata.org/dataset/cod-ab-pak)[30]. The Khyber Pakhtunkhwa shapefile was extracted from this dataset, and the study area as well as the cluster distribution maps (Figs. 1, 9 and 10, and 11) were generated using QGIS[31] Version 3.34 Firenze (QGIS.org, 2025, https://qgis.org).

Study area

This study is conducted in Khyber Pakhtunkhwa (KP), a northern province of Pakistan distinguished by varied topography, including mountainous terrains and alluvial plains. The geographic coordinates of Khyber Pakhtunkhwa is 34.9526205° N latitude and 72.331113° E longitude. Home to approximately 40.85 million people, KP exhibits considerable disparities in access to safe drinking water and sanitation. The province experiences recurrent waterborne disease outbreaks, particularly during the monsoon season, largely due to the contamination of water sources[32]. These health vulnerabilities are intensified by deficient infrastructure, unregulated urban expansion, and climatic fluctuations. Given these conditions, KP serves as a critical setting for examining spatial distribution and identifying spatiotemporal clusters of waterborne diseases. Figure 1 displays the study area map of Khyber Pakhtunkhwa, generated by the authors using QGIS software with administrative boundary data.

Data collection

The dataset used in this study was acquired from the Directorate General (DG) Health Services, Khyber Pakhtunkhwa, and is available at https://www.nih.org.pk/phb/weekly-bulletin. The data includes 35 districts and 12 months specific records of WBD taken for the year (2024). Supplementary data, Population estimates were extracted from the 2017 national census records to support demographic normalization and computation of expected disease counts in the spatiotemporal analysis.

Computationally efficiency

The algorithms proposed by Fanaee-T and Gama[25] and Ullah et al.[26] implement the EigenSpace framework by integrating conventional SVD techniques for dimensionality reduction and matrix factorization within the spatiotemporal data structure. However, standard SVD is computationally expensive, particularly for large and sparse datasets. To address the computational and structural limitations essential in EigenSpot and Multi-EigenSpot, the proposed methods integrate an advanced variant SVDs. Design for efficient processing of high dimensional and sparse matrices, SVDs significantly improves decomposition accuracy and scalability, making it particularly effective for analyzing rare disease datasets with limited cases.

The proposed algorithm integrates the following three methodological components:

- **SVDs**: Employed to extract the principal left and right singular vectors (LSV and RSV) from the K and E matrices for dimensionality reduction.
- **Robust Z-Control Chart**: Employed to identify abnormal components in differences vectors.

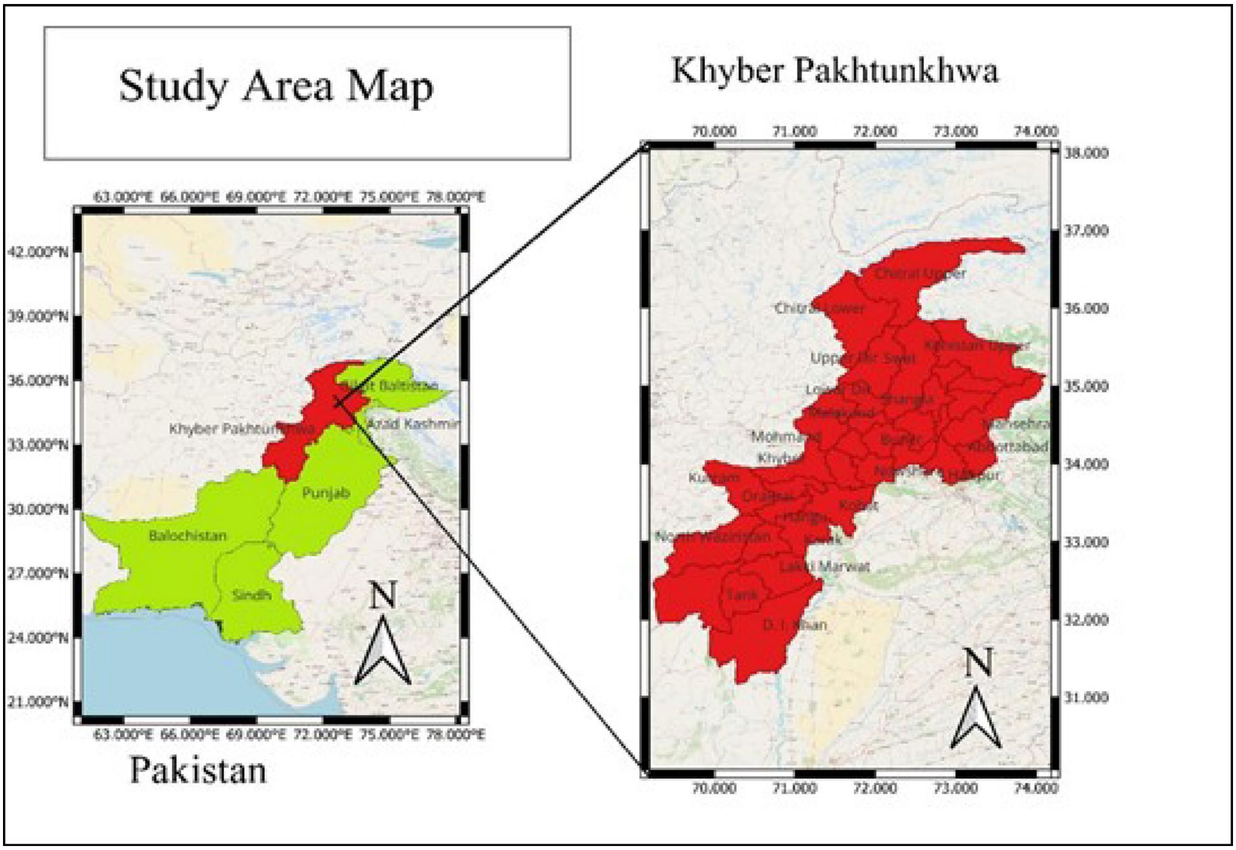


Fig. 1. Study area map of Khyber Pakhtunkhwa, generated using QGIS v3.34 Firenze (https://qgis.org) with administrative boundary data from the HDX[30].

- **Visualization**: A heatmap is used to display the final Relative Risk (RR) matrix, highlighting potential cluster regions through color intensity variations.

```
        P: Population at risk matrix
        K: Typhoid Cases matrix
        α: level of significance (0.10)
Input: P, K and α
Output: Heatmap
    1.  Compute E and R matrices from P and K matrices
    2.  Compute singular vectors:
        a.  Apply truncated SVD to K and E.
    3.  $[\overrightarrow{S_k}, \overrightarrow{T_k}] = 1 - rank\ SVDs(K)$, Cases
    4.  $[\overrightarrow{S_E}, \overrightarrow{T_E}] = 1 - rank\ SVDs(E)\ baseline,$
    5.  Form difference vectors: $[\overrightarrow{D_S}, \overrightarrow{D_T}]$
    6.  for I = 1: m do
    7.  $\mathrm{D}S_i = \mathrm{S}K_i - \mathrm{S}E_i$
    8.  end for
    9.  for i = 1: n do
    10. $\mathrm{D}T_i = \mathrm{T}K_i - \mathrm{T}E_i$
    11. end for
    12. Apply control chart:
        a.  Standardize $[\overrightarrow{D_S}, \overrightarrow{D_T}]$ using robust z-scores.
        b.  Identify abnormal elements exceeding threshold α.
        c.  Determine combined spatiotemporal abnormal components.
    13. Update matrices:
          If abnormal components are found:
            a.  Replace corresponding entries in K by expected values from E.
            b.  Replace corresponding entries in R by the median.
        Repeat Steps 2–5 until no combine abnormal component is found.
        Finalize relative risk matrix:
    14. for i=1:m
    15. for j=1:n
    16. if $R_{i,j} = A_{i,:}$, $A_{i,:}$is median value in the matrix $R$
    17. $R_{i,j} = 1$
    18. else
    19. $R_{i,j} = R_{i,j}$
    20. end
    21. end
In the final updated R, replace all entries other than median by 1.
```

Algorithm. Novel Multi-EigenSpot.

Novel multi-eigenspot

The proposed method targets scenarios where disease case data are aggregated across defined spatial units and temporal intervals. The Population at risk and observed case counts are structured into an $m \times n$ spatiotemporal matrices P and K, where m and n represent the number of spatial regions (Districts) and time points (Months), respectively. Auxiliary matrices, E (expected cases) matrix is computed from spatiotemporal matrices P and K and R (relative risk) is computed from spatiotemporal matrices K and E. The relative risk (RR), a standard epidemiological metric, is computed as the ratio of observed to expected counts.

To extract central spatiotemporal SVD is applied to matrices K and E, yielding LSV and RSV that capture spatial and temporal patterns, respectively. Formally, the decomposition of K is expressed as $K = UDV^t$, where U and V contain the LSV and RSV, and D is a diagonal matrix of singular values. The principal singular vectors of K are denoted as $SK = (sk_1, sk_2, \ldots, sk_m)$ and $TK = (tk_1, tk_2, \ldots, tk_m)$ for the spatial and temporal dimensions, respectively. Correspondingly, the principal vectors for E are $SE = (se_1, se_2, \ldots, se_m)$, and $TE = (te_1, te_2, \ldots, te_m)$. Abnormal spatiotemporal components are identified by computing the differences vectors: $\mathrm{DS} = \mathrm{SK} - \mathrm{SE}$ and $\mathrm{DT} = \mathrm{TK} - \mathrm{TE}$.

Figure 2 presents a systematic workflow of the novel Multi-EigenSpot algorithm, detailing the integration of SVDs decomposition, robust Z-control charts for anomaly detection, and iterative matrix updating to identify spatiotemporal clusters with reduced computational complexity. The proposed algorithms employ a robust Z-control chart to detect joint spatiotemporal abnormal component by analysing both differences vectors, DS and DT. Upon identification of simultaneous joint spatiotemporal abnormal component in these vectors, the observed case matrix K is updated by replacing the corresponding entries with their expected case matrix E values. Concurrently, the relative risk matrix R is updated by replacing the associated abnormal component with the median value. This iterative process continues until no further abnormal component are detected in either

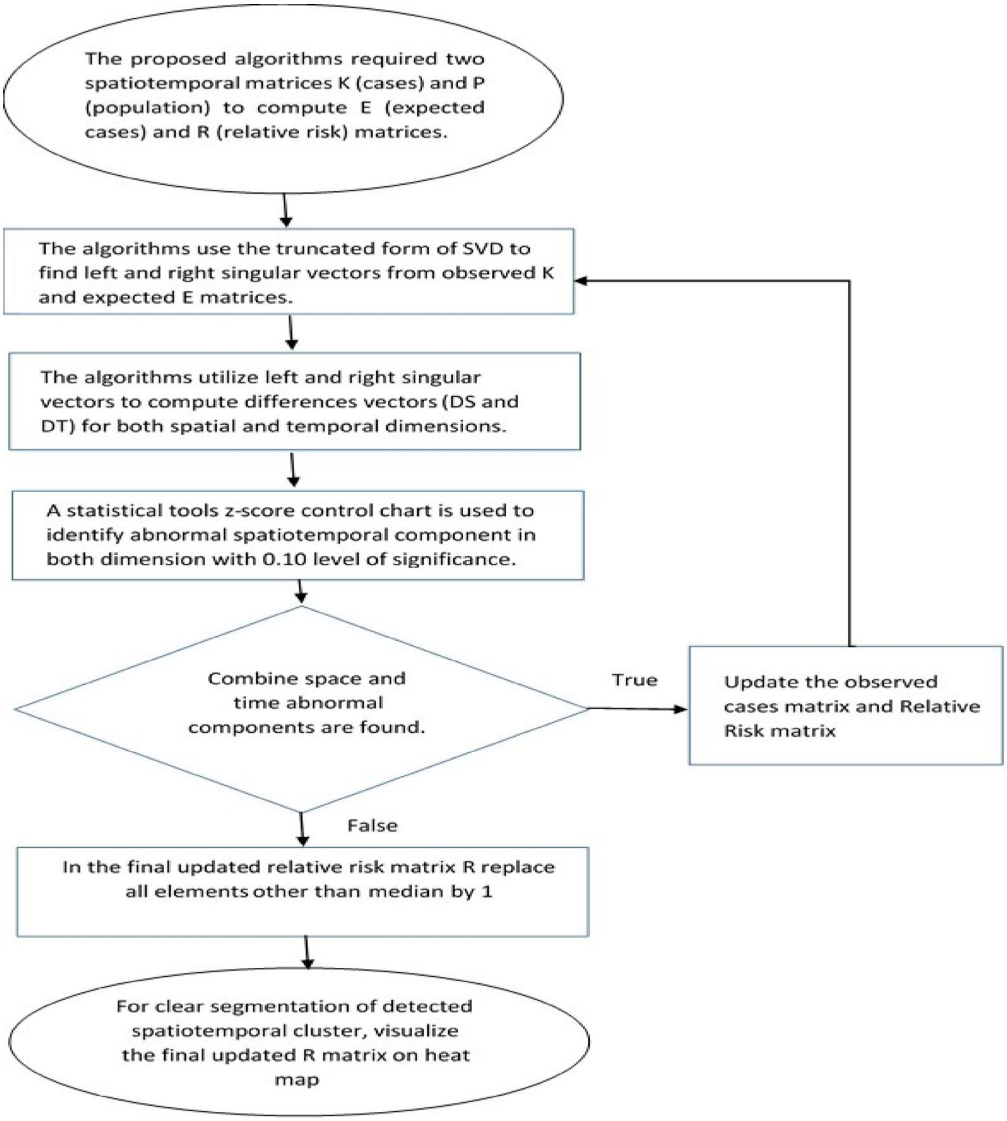

The proposed algorithms required two spatiotemporal matrices K (cases) and P (population) to compute E (expected cases) and R (relative risk) matrices.
The algorithms use the truncated form of SVD to find left and right singular vectors from observed K and expected E matrices.
The algorithms utilize left and right singular vectors to compute differences vectors (DS and DT) for both spatial and temporal dimensions.
A statistical tools z-score control chart is used to identify abnormal spatiotemporal component in both dimension with 0.10 level of significance.
Combine space and time abnormal components are found.
True
Update the observed cases matrix and Relative Risk matrix
False
In the final updated relative risk matrix R replace all elements other than median by 1
For clear segmentation of detected spatiotemporal cluster, visualize the final updated R matrix on heat map

Fig. 2. Flow chart of the proposed algorithms.

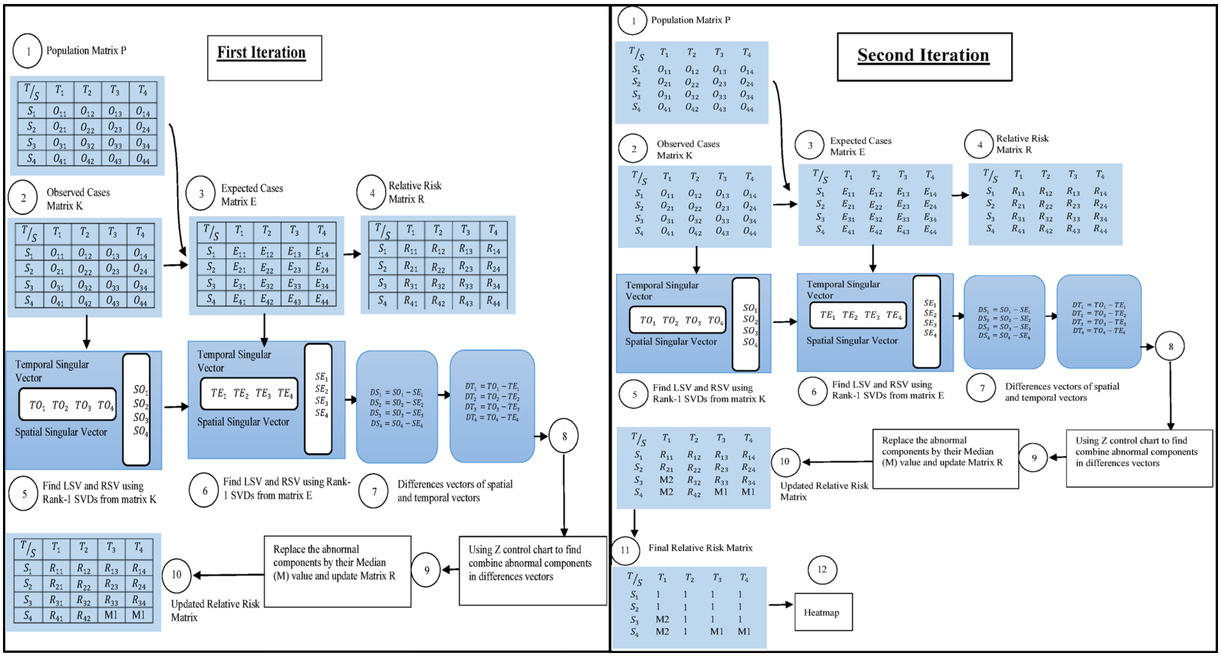

First Iteration
1 Population Matrix P
2 Observed Cases Matrix K
3 Expected Cases Matrix E
4 Relative Risk Matrix R
Temporal Singular Vector
TO_1 TO_2 TO_3 TO_4
Spatial Singular Vector
SO_1 SO_2 SO_3 SO_4
TE_1 TE_2 TE_3 TE_4
SE_1 SE_2 SE_3 SE_4
$DS_1 = SO_1 - SE_1$
$DS_2 = SO_2 - SE_2$
$DS_3 = SO_3 - SE_3$
$DS_4 = SO_4 - SE_4$
$DT_1 = TO_1 - TE_1$
$DT_2 = TO_2 - TE_2$
$DT_3 = TO_3 - TE_3$
$DT_4 = TO_4 - TE_4$
5 Find LSV and RSV using Rank-1 SVDs from matrix K
6 Find LSV and RSV using Rank-1 SVDs from matrix E
7 Differences vectors of spatial and temporal vectors
8
9 Using Z control chart to find combine abnormal components in differences vectors
Replace the abnormal components by their Median (M) value and update Matrix R
10 Updated Relative Risk Matrix
Second Iteration
11 Final Relative Risk Matrix
12 Heatmap
T/S T_1 T_2 T_3 T_4
S_1 S_2 S_3 S_4
M1 M2

Fig. 3. Scheme Illustration of the proposed Algorithm.

the spatial or temporal dimensions. For hotspot visualization, the final updated matrix R is utilized. Elements in R differing from the median are replaced with 1 to mark potential clusters. A heatmap is then generated to provide a clear graphical representation and segmentation of the detected disease clusters. Figure 3 visually demonstrates the iterative cluster detection and removal mechanism of the proposed algorithm, emphasizing how joint spatiotemporal anomalies are identified and suppressed in the relative risk matrix through recursive update.

A comprehensive step-by-step process of how these techniques are integrated within the algorithm is given below.

1. The total observed cases matrix is denoted by K, and the population at risk matrix is denoted by P.

$$K = \begin{bmatrix} k_{11} & \cdots & k_{1n} \\ ? & \ddots & ? \\ k_{m1} & \cdots & k_{mn} \end{bmatrix}, P = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ ? & \ddots & ? \\ p_{m1} & \cdots & p_{mn} \end{bmatrix}$$

Where k_{11} is the total disease in the first region (district), first time point (Month), p_{11} is the total population at risk in the first region, first time point,m is the total spatial dimensions, and$n$total time points.

2. Compute the expected disease cases E and relative risks R matrices for K and P matrices.

$$E = \begin{bmatrix} E_{11} & \cdots & E_{1n} \\ ? & \ddots & ? \\ E_{m1} & \cdots & E_{mn} \end{bmatrix} \text{and} R = \begin{bmatrix} R_{11} & \cdots & R_{1n} \\ ? & \ddots & ? \\ R_{m1} & \cdots & R_{mn} \end{bmatrix}$$

The primary objective of computing the relative risk matrix R is to enable effective visualization of disease clusters through a heatmap representation.

3. The one-rank SVDs are used to obtain the principal left and right singular vectors for matrices K and E. Our approach only requires the principal singular vector corresponding to the highest eigenvalue, as the first principal singular vector explains the majority of variance in the data. While full-rank SVDs decompose a matrix into a combination of orthogonal vectors, one-rank SVDs capture the most significant singular value and corresponding singular vectors, effectively representing the matrix with a single dominant direction. For matrix K, the principal left singular vector is denoted as $SK = (sk_1, sk_2, \ldots, sk_m)$ and the principal right singular vector is denoted as $TK = (tk_1, tk_2, \ldots, tk_m)$. Similarly, for matrix E, the principal left singular vector is denoted as $SE = (se_1, se_2, \ldots, se_m)$, and the principal right singular vector is denoted as $TE = (te_1, te_2, \ldots, te_m)$. The elements in the principal left singular vectors correspond to the components in the spatial dimension, while the elements in the principal right singular vectors correspond to the components in the temporal dimension.

4. Abnormal components are identified by computing the difference vectors between the corresponding singular vector pairs: the spatial differences vector as DS = SK – SE and the temporal differences vector as DT = TK – TE.
5. Standardized z-score vectors are computed from the differences vectors DS and DT. A robust z-score control chart is then applied to both vectors at a significance level α = 0.10. Elements yielding left-tailed p-values less than α are considered out of control, indicating abnormal components within the spatial and temporal dimensions, respectively.
6. If simultaneous abnormal component/components is/are detected in both DS and DT, the observed case matrix K is updated by replacing the elements corresponding to the joint abnormal spatial and temporal component with their respective expected values. Likewise, the relative risk matrix R is updated by substituting the affected entries with the median value.
7. Identify any additional abnormal components in the spatial and temporal dimensions. Repeat Steps (01–06) until no abnormal components are found in either dimension.
8. The elements in the recently updated matrix R, corresponding to the components (spatial/temporal) not classified as abnormal, are substituted with the value of 1.
9. Visualize the final updated relative risk matrix R on a heatmap for clear segmentation and interpretation of detected disease cluster.

Results and discussions

Typhoid, caused by Salmonella enterica serotype Typhi, is classified as a WBD. Typhoid is a major public health concern in several countries with limited resources, including Pakistan. Annually, 9 to 12 million individuals are affected by typhoid globally. Typhoid is a major health concern in Pakistan, with thousands of cases reported annually, particularly in regions characterized by poor sanitation and restricted access to clean water[33]. The World Health Organization (WHO) said that Pakistan is at higher risk for typhoid fever, especially in Khyber Pakhtunkhwa, where environmental and infrastructure conditions increase the likelihood of outbreaks.

This study examined typhoid data in KP, identifying many spatiotemporal hotspots with significantly higher case counts. Identifying these clusters is crucial for public health, enabling the early detection of high-risk areas, which allows for timely action, resource allocation, and awareness campaigns. Spatiotemporal cluster identification provides governments with evidence-based insights to improve infrastructure, execute immunization programs, and monitor disease transmission, therefore reducing morbidity and mortality associated with typhoid. Figure 4 displays the population distribution across KP 35 districts, revealing demographic disparities that underpin the normalization of disease incidence rates and the calculation of expected case thresholds in spatiotemporal analysis.

From Figs. 5 and 6, it is clear that maximum typhoid disease cases are recorded in district Bannu and Peshawar, and in the months of May, July, and October.

The alpha threshold was set at 0.10 because, in common diseases, observed cases typically exceed expected cases in most regions. Setting the alpha threshold at 0.05 or 0.01 could limit the identification of multiple high-risk locations or lead to undetected hotspots. The results were verified by displaying the observed and expected typhoid cases for each of the 35 districts every month (January 2024 to December 2024) in the graphs illustrated in Fig. 7.

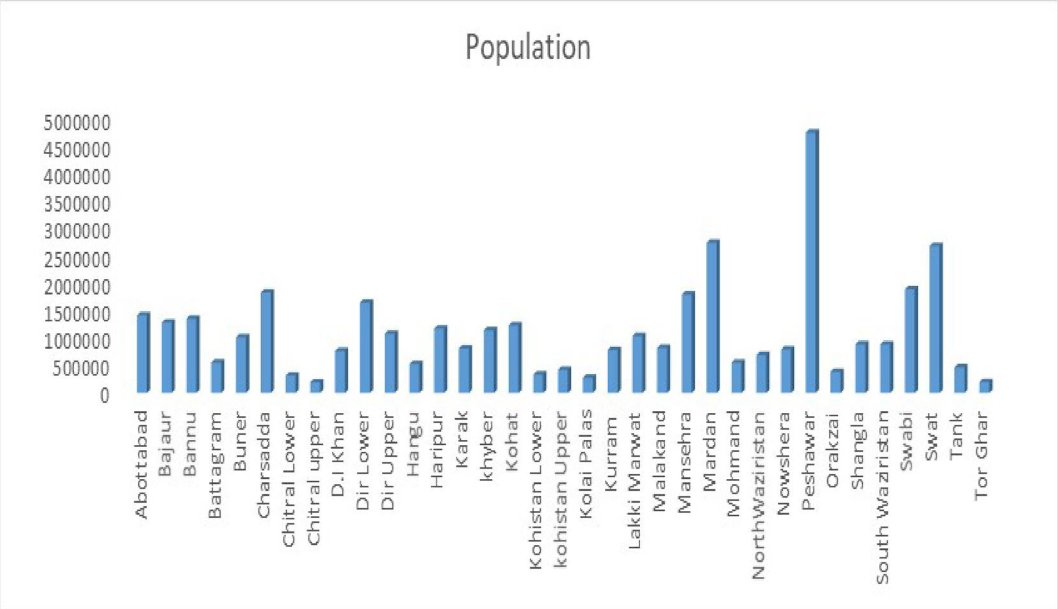


Fig. 4. Total KP district wise population.

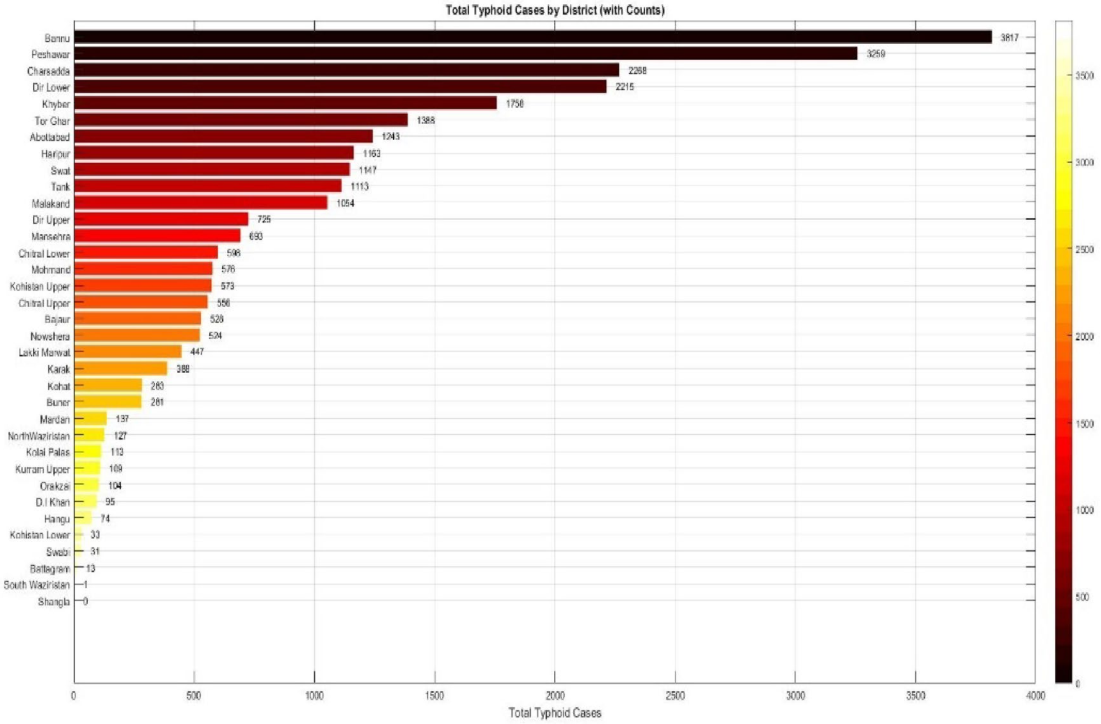


Fig. 5. District-wise total recorded typhoid disease cases of KP 2024.

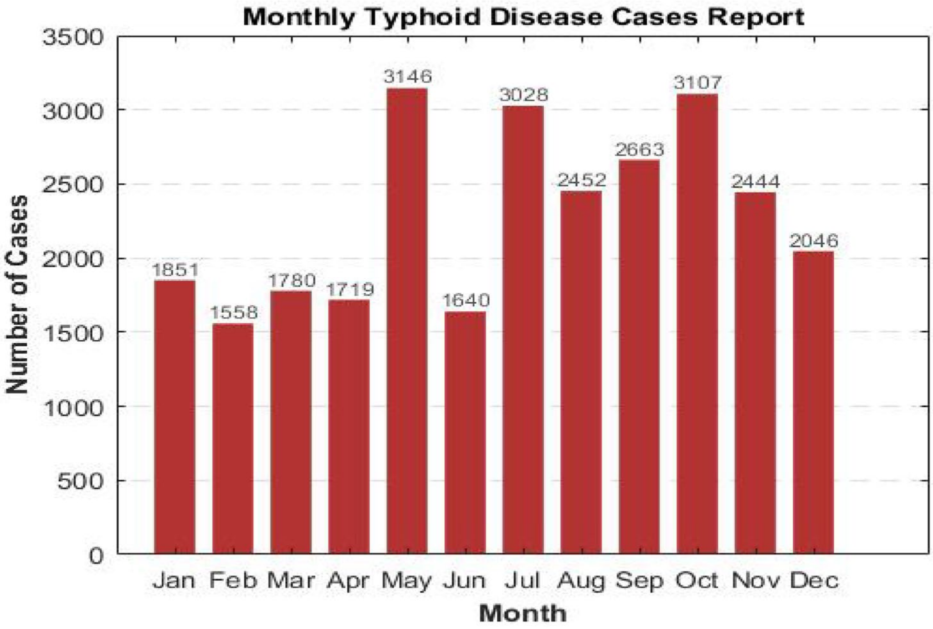


Fig. 6. Monthly observed typhoid disease cases of KP 2024.

The heatmap illustrates the spatiotemporal distribution of the relative risk matrix for typhoid fever over 35 districts in Khyber Pakhtunkhwa, Pakistan. Figs. 8 and 9 show that the first likely and most significant cluster was detected in the districts of Bannu and Tor Ghar throughout May and October, with an average relative risk (RR) of 1.767, as denoted by a deep red colour. The second likely cluster identified in the districts of Bannu, Dir Lower, Khyber, Tank, and Tor Ghar throughout July, with an average RR of 1.663, highlighted in red. The third cluster emerged in August inside the districts of Bannu, Khyber, Tank, and Tor Ghat, with a relative risk (RR) of 1.587. The fourth likely cluster was detected in the districts of Bannu (August and November), Charsadda (May, October, and November), Dir Lower (May, October, and November), Khyber (August), Tank (August),

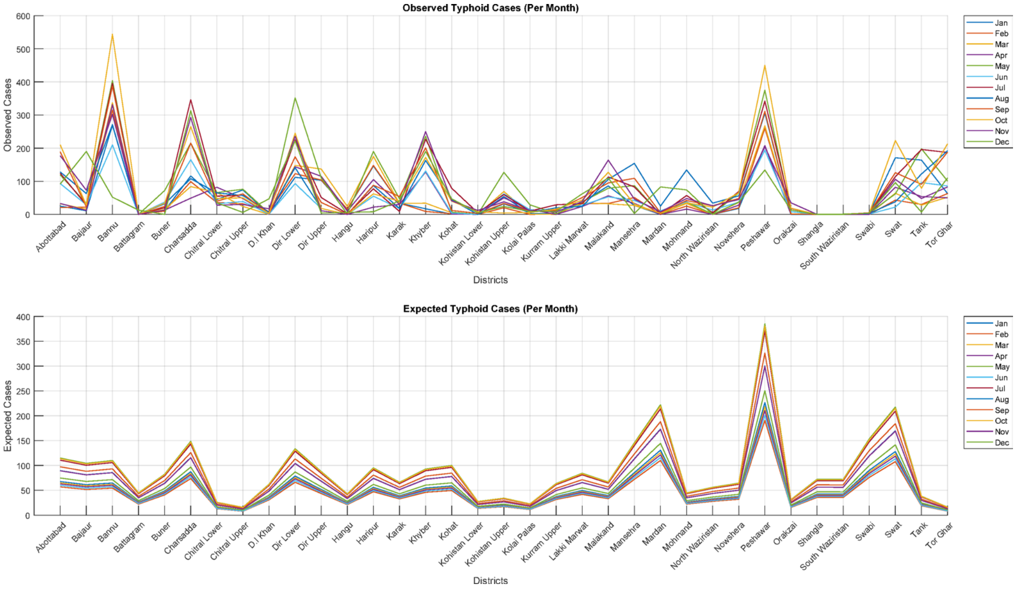
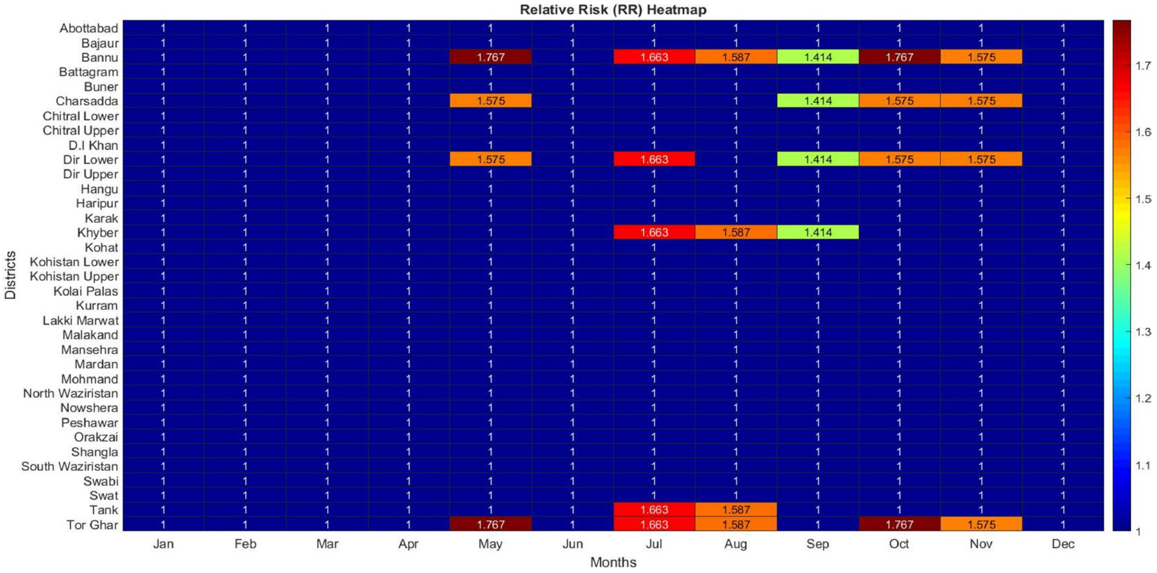


Fig. 7. Observed and expected typhoid cases of KP 2024.



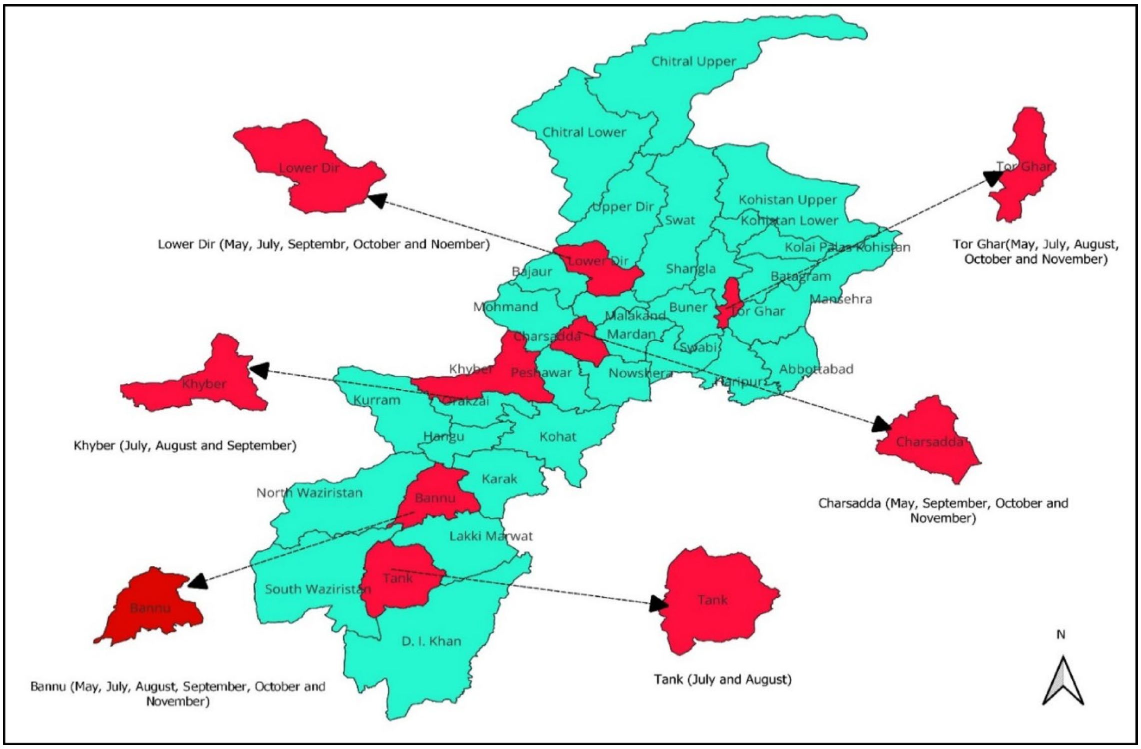
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Abottabad	1	1	1	1	1	1	1	1	1	1	1	1
Bajaur	1	1	1	1	1	1	1	1	1	1	1	1
Bannu	1	1	1	1	1.767	1	1.663	1.587	1.414	1.767	1.575	1
Battagram	1	1	1	1	1	1	1	1	1	1	1	1
Buner	1	1	1	1	1	1	1	1	1	1	1	1
Charsadda	1	1	1	1	1.575	1	1	1	1.414	1.575	1.575	1
Chitral Lower	1	1	1	1	1	1	1	1	1	1	1	1
Chitral Upper	1	1	1	1	1	1	1	1	1	1	1	1
D.I Khan	1	1	1	1	1	1	1	1	1	1	1	1
Dir Lower	1	1	1	1	1.575	1	1.663	1	1.414	1.575	1.575	1
Dir Upper	1	1	1	1	1	1	1	1	1	1	1	1
Hangu	1	1	1	1	1	1	1	1	1	1	1	1
Haripur	1	1	1	1	1	1	1	1	1	1	1	1
Karak	1	1	1	1	1	1	1	1	1	1	1	1
Khyber	1	1	1	1	1	1	1.663	1.587	1.414	1	1	1
Kohat	1	1	1	1	1	1	1	1	1	1	1	1
Kohistan Lower	1	1	1	1	1	1	1	1	1	1	1	1
Kohistan Upper	1	1	1	1	1	1	1	1	1	1	1	1
Kolai Palas	1	1	1	1	1	1	1	1	1	1	1	1
Kurram	1	1	1	1	1	1	1	1	1	1	1	1
Lakki Marwat	1	1	1	1	1	1	1	1	1	1	1	1
Malakand	1	1	1	1	1	1	1	1	1	1	1	1
Mansehra	1	1	1	1	1	1	1	1	1	1	1	1
Mardan	1	1	1	1	1	1	1	1	1	1	1	1
Mohmand	1	1	1	1	1	1	1	1	1	1	1	1
North Waziristan	1	1	1	1	1	1	1	1	1	1	1	1
Nowshera	1	1	1	1	1	1	1	1	1	1	1	1
Peshawar	1	1	1	1	1	1	1	1	1	1	1	1
Orakzai	1	1	1	1	1	1	1	1	1	1	1	1
Shangla	1	1	1	1	1	1	1	1	1	1	1	1
South Waziristan	1	1	1	1	1	1	1	1	1	1	1	1
Swabi	1	1	1	1	1	1	1	1	1	1	1	1
Swat	1	1	1	1	1	1	1	1	1	1	1	1
Tank	1	1	1	1	1	1	1.663	1.587	1	1	1	1
Tor Ghar	1	1	1	1	1.767	1	1.663	1.587	1	1.767	1.575	1

Fig. 8. Heatmap of typhoid disease.

and Tor Ghar (August and November), with an average relative risk of 1.587, denoted by a dark yellow colour. A fifth likely cluster occurred in September over numerous regions, including Bannu, Charsadda, Dir Lower, and Kohat, with a relative risk (RR) of 1.414. The RR 1 shows that no abnormal case was found in these districts. From both Figures, it is clear that Bannue, Charsadda, Dir Lower, Khyber, Tank, and Tor Ghar districts are highly affected by typhoid disease during the year 2024 for various months, suggesting them as alarming typhoid disease hotspots.

Performance evaluation

Figure 10 illustrates the efficiency of the proposed approach, through which we conducted a comparison study against the Eigen Spot and Multi-Eigen Space algorithms for spatiotemporal disease clusters identification. The map illustrates that the Eigen Spot algorithms identify only a single cluster, Bannu (October), while missing other broad disease clusters. Multi-Eigen space approaches inaccurately identify a disease cluster in the temporal domain, detecting numerous spatiotemporal disease clusters including Bannu (March, April, May, October, and December), Kohat (March and April), D.I. Khan (February), and Swat (June and October). No reported cases



Lower Dir (May, July, Septembr, October and Noember)

Tor Ghar(May, July, August, October and November)

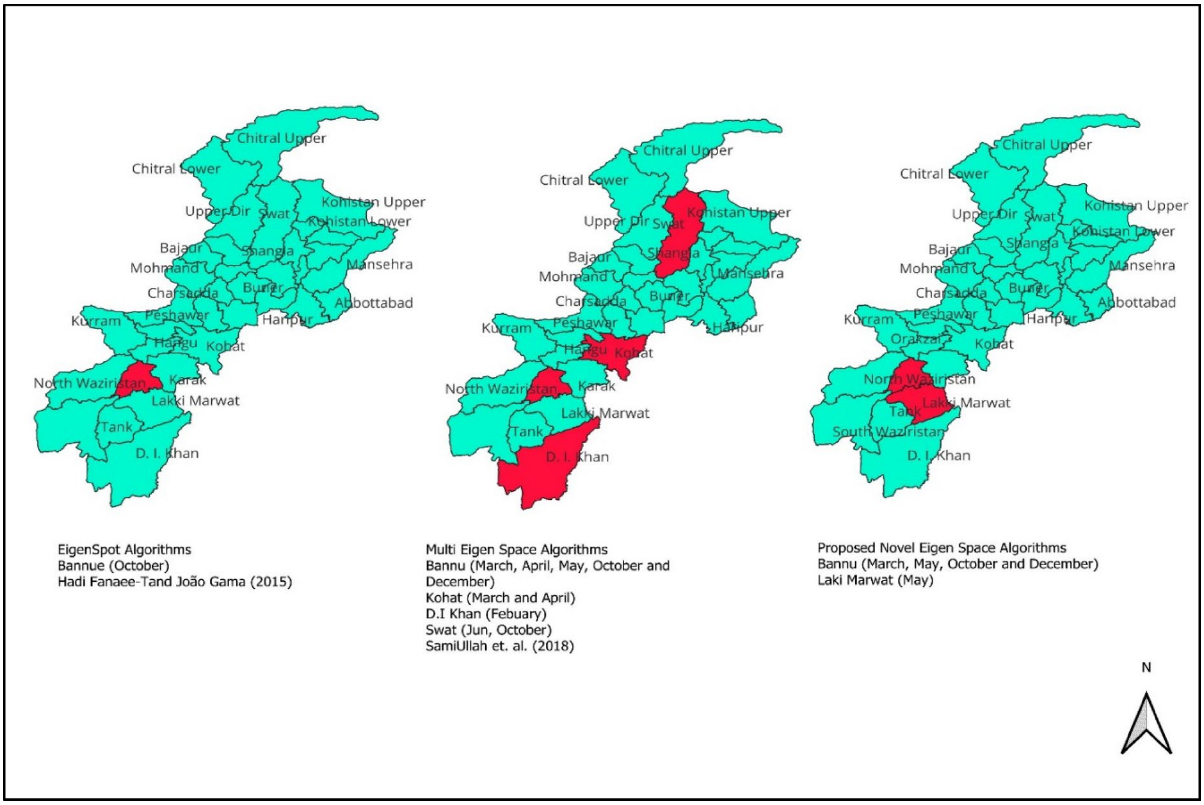
Khyber (July, August and September)

Charsadda (May, September, October and November)

Bannu (May, July, August, September, October and November)

Tank (July and August)

Fig. 9. Detected typhoid disease clusters in Khyber Pakhtunkhwa, mapped in QGIS v3.34 Firenze (https://qgis.org) using administrative boundaries from HDX[30].



EigenSpot Algorithms
Bannue (October)
Hadi Fanaee-Tand João Gama (2015)

Multi Eigen Space Algorithms
Bannu (March, April, May, October and December)
Kohat (March and April)
D.I Khan (Febuary)
Swat (Jun, October)
SamiUllah et. al. (2018)

Proposed Novel Eigen Space Algorithms
Bannu (March, May, October and December)
Laki Marwat (May)

Fig. 10. Comparative maps of disease clusters identified by EigenSpot, Multi-EigenSpace, and the proposed method. Maps created in QGIS v3.34 Firenze (https://qgis.org) with shapefile data from HDX[30].

Features	Multi EigenSpace Algorithm	Novel Multi-EigenSpace Algorithm
Data (Matrix Size)	35 by 12	35 by 12
Method	SVD	SVDs (Truncated form of SVD)
Loop	Nested Loops	Vectorized
Allocation	Grow in Loop	Pre-Allocated
Relative Efficiency	Baseline	5 to 10 times faster
Estimated Computation time	1 to 3 s	0.1 to 0.5 s

Table 1. Computational time of the Multi-EigenSpace algorithm vs. the NovelMulti-EigenSpace algorithm.

Approaches	Identified Spatiotemporal cluster/detected Spatiotemporal hotspots
EigneSpot	Tank (Jul)
Novel Multi-EigenSpot (Proposed)	Bannu (May, Jul, Aug, Sep, Oct and Nov), Charsadda (May, Sep and Nov), Dir Lowe (May, July, Sep, Oct and Nov), Khyber (July, Aug and Sep), Tor Ghar (May, Jul, Aug, Oct and Nov) and Tank (Jul and Aug)
SatScan	Bannu (Feb and Mar) and Charsadda (Nov and Dec)
DBSCAN	Bannu (May, Jul, Sep, Oct and Nov), Charsadda (May, Jul, Nov and Dec), Khyber (July, Aug and Sep), Tor Ghar (May, Jul, Oct and Nov) and Tank (May and Jul)

Table 2. Comparison of disease clusters identified by EigenSpot. Novel Multi-EigenSpot (proposed), SaTScan, and DBSCAN.

were noted in April; however, a temporal cluster was found in Bannu during that month. The suggested Novel Eigen Space algorithms effectively resolve these two limitations. The suggested techniques identify multiple disease clusters while minimizing false positives in temporal clusters with no observed cases. The map clearly illustrates that the suggested techniques identify real clusters with higher precision. This demonstrates its superiority in identifying significant hotspots in sparse data, making it a more accurate approach for cluster discovery. Table 1 quantifies the 5–10× speed advantage of the novel algorithm over Multi-EigenSpace, achieved through SVDs truncation and vectorization (0.1–0.5 s vs. 1–3 s for 35 × 12 matrices.

As shown in Table 2, EigenSpot is limited to detecting a single hotspot (Tank, July), while SaTScan identifies only a few clusters due to its circular window constraint. DBSCAN performs better, capturing multiple clusters, but its results are sensitive to parameter selection, sometimes leading to fragmented or spurious detections. In contrast, the proposed Novel Multi-EigenSpot method consistently identifies a broader set of epidemiologically reasonable clusters across districts and months, capturing both temporal recurrence and spatial irregularity. This demonstrates its superior robustness and practical applicability in real-world surveillance settings where outbreak signals are rare and irregular.

Figure 11 provides a visual comparison of spatiotemporal clusters detected by the proposed Novel Multi-EigenSpot, DBSCAN, and SaTScan. SaTScan's results are constrained by its circular scan windows, leading to under-detection of irregular cluster shapes. DBSCAN identifies several clusters, but its outputs vary depending on parameter choices, sometimes overestimating cluster boundaries. The proposed method, however, delineates multiple realistic clusters with higher spatial precision and temporal consistency, closely matching the epidemiological distribution of typhoid in KP. This visualization reinforces the interpretability and robustness of the proposed algorithm over existing approaches.

Figure 12 shows the comparative performance of EigenSpot family, SaTScan, and DBSCAN across multiple evaluation metrics, including Precision, Recall, F1-score, Robustness Index, and computational efficiency. The results show that the proposed Novel Multi-EigenSpot constantly achieves the maximum accuracy (Precision, Recall, and F1-score above 80%) and robustness while maintaining the computational time, even on a logarithmic scale. In contrast, SaTScan and DBSCAN demonstrate relatively lower detection accuracy and robustness, joined with higher computational costs. These results highlight the superior stability of efficiency, sensitivity, and robustness achieved by the Novel Multi-EigenSpot method.

From Table 3, it is clear that the proposed method consistently outperforms existing approaches. Unlike SaTScan, which is limited to circular or elliptical clusters, and DBSCAN, which is highly sensitive to parameter tuning, the proposed framework efficiently identifies multiple irregularly shaped clusters with minimal parameter dependence. Compared to EigenSpot and Multi-EigenSpot, it demonstrates stronger performance on sparse data, faster computation through truncated SVDs, and improved interpretability via clear heatmap visualizations.

From Table 4, it is evident that the proposed method maintains the highest robustness to missingness across all situations. While the performance of all methods drops as missingness increases, Novel Multi-EigenSpot consistently achieves superior F1-scores with lower variability, demonstrating resilience to both MCAR and MNAR patterns. In contrast, SaTScan shows the weakest robustness, and DBSCAN and EigenSpot exhibit

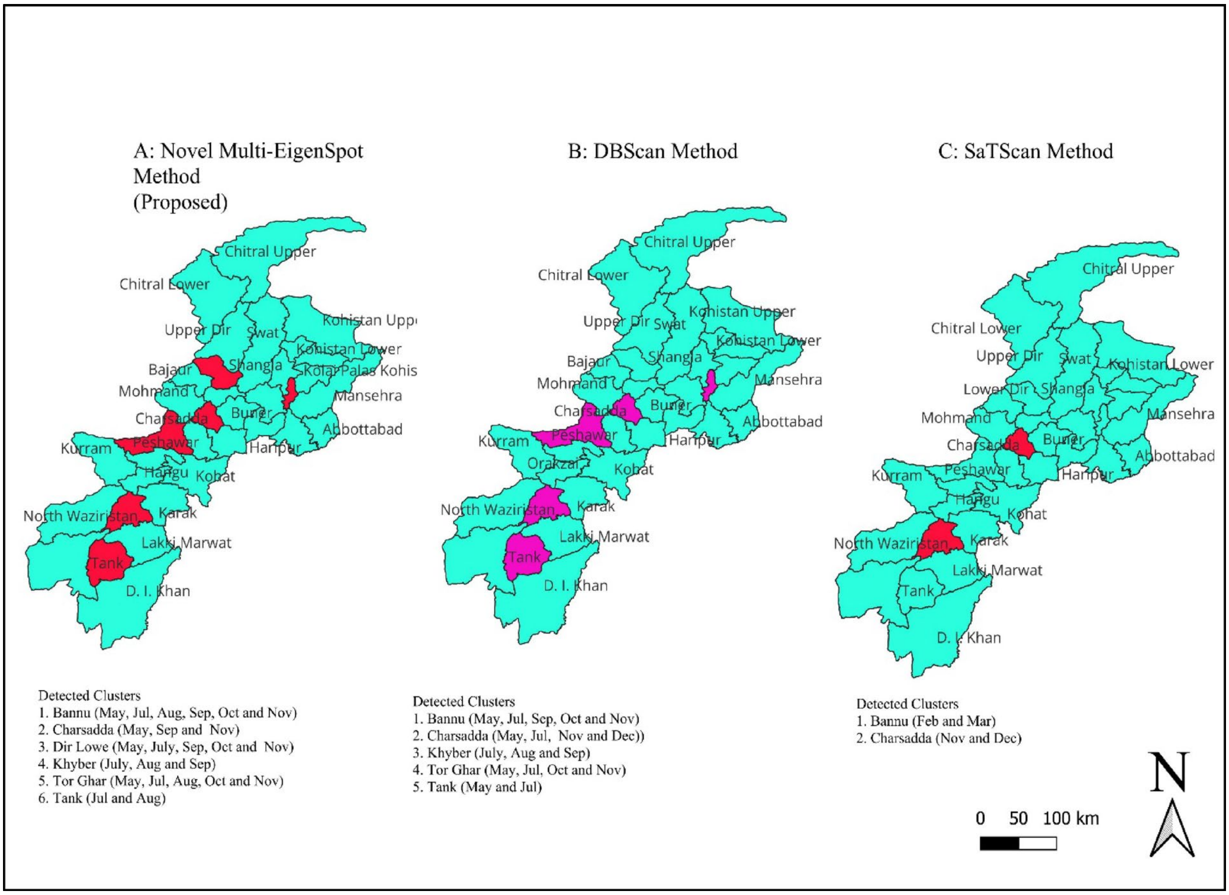


Fig. 11. Spatiotemporal clusters detected by the proposed method, DBSCAN, and SaTScan, generated in QGIS v3.34 Firenze (https://qgis.org) using shapefile data from HDX[30].

moderate but less stable performance. These results highlight the ability of the proposed framework to handle data imperfections such as missing values and zeros without relying on distributional assumptions.

Discussion and conclusion

The proposed Novel EigenSpace method markedly advances spatiotemporal cluster detection by addressing two critical shortcomings of the original EigenSpot and Multi-EigenSpot algorithms: the inability to identify multiple, rare-disease hotspots and prohibitive computational demands. When applied to 2024 typhoid case data from Khyber Pakhtunkhwa, our approach consistently revealed distinct clusters, both spatially across districts such as Bannu, Charsadda, Dir Lower, Khyber, Tank and Tor Ghar, and temporally during high-incidence months, where previous methods either missed secondary hotspots or generated false positives in periods with zero observed cases. By integrating truncated SVDs with a robust Z-control chart in a recursive update scheme, we realize a five-to-ten-fold acceleration in computation while preserving the sensitivity of cluster detection in inherently sparse rare disease datasets. This advancement not only optimizes the efficiency of real time epidemiological monitoring but also strengthens the robustness of public-health conclusions derived from erratic, irregularly shaped outbreak signals.

In summary, this study introduces a computationally efficient, distribution-free algorithm that successfully detects multiple spatiotemporal clusters of rare diseases, overcoming limitations of shape assumption, data sparsity, and processing time. Although the heatmap representation of the finalized relative-risk matrix offers an intuitive means of pinpointing hotspots, the current framework does not elucidate the directional propagation or transmission dynamics that give rise to these clusters. Furthermore, by aggregating data into discrete sub-regional and monthly slices, the method may obscure phenomena that span administrative borders or persist across overlapping temporal intervals.

Future research will focus on incorporating spatiotemporal network models to infer propagation routes, adopting rolling-window analyses for continuous-time detection, and integrating auxiliary covariates (e.g., environmental or mobility data) to contextualize outbreak drivers. By broadening the Novel EigenSpace paradigm along these lines, we aspire to furnish a comprehensive platform for adaptive epidemiological monitoring and precision targeted intervention design.

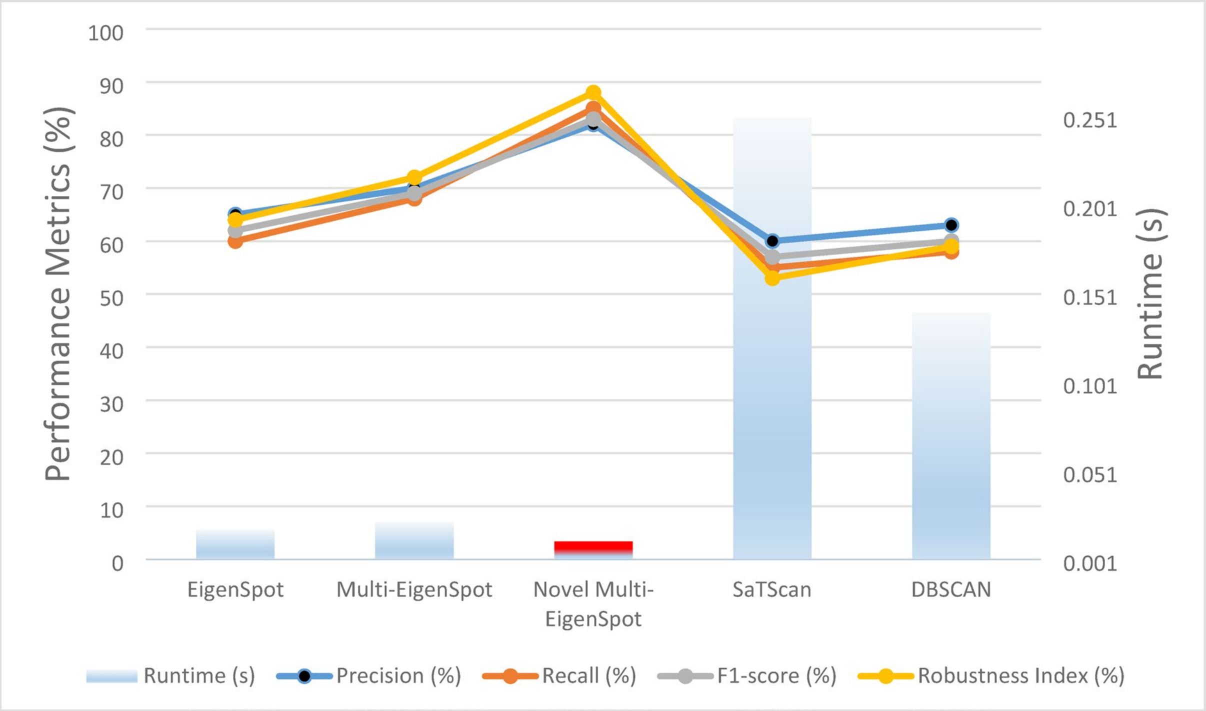


Fig. 12. Benchmarking EigenSpot Family Against SaTScan and DBSCAN: Precision, Recall, F1-score, Robustness, and Runtime.

Criterion	SaTScan (Scan Statistic)	DBSCAN (Density-Based)	EigenSpot	Multi-EigenSpot	Novel Multi-EigenSpot
Cluster Shape Handling	Circular / Elliptical only	Arbitrary / Irregular	Single hotspot only	Multiple hotspots, irregular shapes	Multiple irregularly shaped clusters
Distribution Assumptions	Parametric	None (non-parametric)	None (distribution-free)	None (distribution-free)	None (distribution-free)
Sensitivity to Parameters	Low (defaults work well)	High (eps, minPts tuning critical)	Low	Low	Low (only α threshold)
Handling Sparse / Rare Data	Weak (low power for rare events)	Moderate (clusters may fragment)	Weak	Moderate (but false positives in temporal dimension)	Strong (robust to zeros, rare counts, and outliers)
Computational Efficiency	High cost	Efficient	Faster (1–2 s)	Slower (1–3 s due to iterative SVDs)	Fastest (0.1–0.5 s via truncated SVDs)
Interpretability	Moderate (map outputs, but rigid clusters)	Moderate (depends on parameter tuning)	Low (single hotspot only)	Moderate (multiple clusters but false positives)	High (clean heatmap visualization, interpretable clusters)

Table 3. Comparative characteristics of baseline and eigenspace-based methods for Spatiotemporal cluster detection.

Method	MCAR 0%	MCAR 5%	MCAR 10%	MCAR 20%	MNAR 20% (zeros)
EigenSpot	0.63 ± 0.08	0.59 ± 0.09	0.52 ± 0.10	0.44 ± 0.12	0.39 ± 0.14
Multi-EigenSpot	0.68 ± 0.07	0.63 ± 0.08	0.56 ± 0.09	0.48 ± 0.11	0.42 ± 0.13
Novel Multi-EigenSpot	**0.75 ± 0.06**	**0.71 ± 0.07**	**0.66 ± 0.08**	**0.58 ± 0.10**	**0.51 ± 0.12**
SaTScan	0.57 ± 0.09	0.51 ± 0.10	0.44 ± 0.12	0.37 ± 0.13	0.33 ± 0.14
DBSCAN	0.61 ± 0.08	0.55 ± 0.09	0.49 ± 0.11	0.42 ± 0.12	0.37 ± 0.13

Table 4. Robustness to missingness (RR = 1.6, |S|=6, |T|=2, Irregular)(Mean F1 ± SD over 100 replicates). EigenSpot methods are distribution-free. Missingness patterns (MCAR: Missing Completely at Random, MNAR: Missing Not at Random) are imposed only for robustness stress-testing.

Data availability

The dataset used in this study is publicly available at the National Institute of Health Pakistan website: [https://www.nih.org.pk/phb/weekly-bulletin](https:/www.nih.org.pk/phb/weekly-bulletin) . Researchers may use this dataset freely for replication and validation purposes.

Received: 7 July 2025; Accepted: 23 September 2025
Published online: 29 October 2025

References

1. Jawad, S., Thirthar, A. A. & Nisar, K. S. The impact of climate change on flowering plants-bees-Vespa orientalis model. *Results Control Optim.* **20**, 100583 (2025).
2. Tiwari, S. S. K., Schmidt, W. P., Darby, J., Kariuki, Z. G. & Jenkins, M. W. Intermittent slow sand filtration for preventing diarrhoea among children in Kenyan households using unimproved water sources: randomized controlled trial. *Trop. Med. Int. Health*. **14**, 1374–1382 (2009).
3. Noureen, A., Aziz, R., Ismail, A. & Trzcinski, A. P. The impact of climate change on waterborne diseases in Pakistan. *Sustain. Clim. Chang.* **15**, 138–152 (2022).
4. Prüss, A., Kay, D., Fewtrell, L. & Bartram, J. Estimating the burden of disease from water, sanitation, and hygiene at a global level. *Environ. Health Perspect.* **110**, 537–542 (2002).
5. Stauber, C. E., Ortiz, G. M., Loomis, D. P. & Sobsey, M. A randomized controlled trial of the concrete biosand filter and its impact on diarrheal disease in Bonao. *Dominican Repub.* (2009).
6. Perveen, S. Drinking water quality monitoring, assessment and management in pakistan: A review. *Heliyon* **9**, (2023).
7. Atif, M. et al. Evolution of waterborne diseases: A case study of Khyber Pakhtunkhwa, Pakistan. *SAGE Open. Med.* **12**, 20503121241263032 (2024).
8. Qudsia, S. et al. Prevalence of waterborne diseases in different union councils of Abbottabad district. *World* **6**, 45 (2025).
9. Ahmed, S., Qadir, A., Khan, M. A., Khan, T. & Zafar, M. Assessment of groundwater intrinsic vulnerability using GIS-based DRASTIC method in district Haripur, Khyber Pakhtunkhwa, Pakistan. *Environ. Monit. Assess.* **193**, 487 (2021).
10. Thirthar, A. A., Alaoui, A. L., Roy, S. & Tiwari, P. K. Fractional and stochastic dynamics of predator–prey systems: The role of fear and global warming. *Eur. Phys. J. B*. **98**, 147 (2025).
11. Thirthar, A. A. et al. Climate change impacts on tri-trophic predator–prey model of the interactions between wolves, ungulates, and plants. *Fractals* **2540106** (2025).
12. Wang, H. & Rodríguez, A. Identifying pediatric cancer clusters in Florida using loglinear models and generalized Lasso penalties. *Stat. Public. Policy Phila. Pa.* **1**, 86 (2014).
13. Amin, R., Bohnert, A., Holmes, L., Rajasekaran, A. & Assanasen, C. Epidemiologic mapping of Florida childhood cancer clusters. *Pediatr. Blood Cancer*. **54**, 511–518 (2010).
14. Kulldorff, M. A Spatial scan statistic. *Commun. Stat. Theory Methods*. **26**, 1481–1496 (1997).
15. Kulldorff, M., Athas, W. F., Feurer, E. J., Miller, B. A. & Key, C. R. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, New Mexico. *Am. J. Public. Health*. **88**, 1377–1380 (1998).
16. Neill, D. B. *Detection of Spatial and Spatio-Temporal Clusters* (Carnegie Mellon University, 2006).
17. Iyengar, V. S. Space-time clusters with flexible shapes. *MMWR Suppl.* **54**, 71–76 (2005).
18. Takahashi, K., Kulldorff, M., Tango, T. & Yih, K. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *Int. J. Health Geogr.* **7**, 14 (2008).
19. Tango, T. A Spatial scan statistic with a restricted likelihood ratio. *Jpn. J. Biom*. **29**, 75–95 (2008).
20. Duczmal, L. & Assuncao, R. A simulated annealing strategy for the detection of arbitrarily shaped Spatial clusters. *Comput. Stat. Data Anal.* **45**, 269–286 (2004).
21. Neill, D. B., Moore, A. W., Sabhnani, M. & Daniel, K. Detection of emerging space-time clusters. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* 218–227 (ACM, Chicago Illinois USA, (2005). https://doi.org/10.1145/1081870.1081897
22. Dong, W. et al. Detect irregularly shaped spatio-temporal clusters for decision support. In *Proceedings of IEEE International Conference on Service Operations, Logistics and Informatics* 231–236 (IEEE, 2011).
23. Costa, M. A. & Kulldorff, M. Maximum linkage space-time permutation scan statistics for disease outbreak detection. *Int. J. Health Geogr.* **13**, 20 (2014).
24. Neill, D. B. An empirical comparison of Spatial scan statistics for outbreak detection. *Int. J. Health Geogr.* **8**, 20 (2009).
25. Fanaee-T, H. & Gama, J. Eigenspace method for Spatiotemporal hotspot detection. *Expert Syst.* **32**, 454–464 (2015).
26. Ullah, S., Daud, H., Dass, S. C., Fanaee-T, H. & Khalil, A. An eigenspace approach for detecting multiple space-time disease clusters: Application to measles hotspots detection in Khyber-Pakhtunkhwa, Pakistan. *Plos One*. **13**, e0199176 (2018).
27. Bryan, K. & Leise, T. The $25,000,000,000 eigenvector: The linear algebra behind Google. *SIAM Rev.* **48**, 569–581 (2006).
28. Bell, R. M., Koren, Y. & Volinsky, C. The Bellkor solution to the Netflix prize. *KorBell Team's Rep. Netflix* **2**, (2007).
29. Humanitarian Data Exchange |. Find & Use Crisis Data | HDX. https://data.humdata.org/
30. Pakistan - Subnational. Administrative Boundaries | Humanitarian Dataset | HDX. https://data.humdata.org/dataset/cod-ab-pak
31. Visual Style Guide. QGIS Web Site. https://qgis.org/styleguide/
32. National Institutes of Health. *Islamabad Pakistan* https://www.nih.org.pk/phb/weekly-bulletin
33. Stanaway, J. D. et al. The global burden of typhoid and paratyphoid fevers: A systematic analysis for the global burden of disease study 2017. *Lancet Infect. Dis.* **19**, 369–381 (2019).

Acknowledgements

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FFR-2025-2942-02".

Author contributions

Writing - original draft: Conceptualization: Muhammad Fayyaz Project administration, Supervision: Alamgir Solution Methodology, Software, Formal analysis: Sami UllahWriting–review & editing, Investigation, Validation: Hameed AliProject administration, Solution methodology: Abdulrahman Obaid AlshammariVisualization, Writing - review & editing, Formal analysis: Zeineb Klai Writing - review & editing, Investigation: Bilal Himmat.

Declarations

Competing interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Grammarly in order to readability and avoid grammatical mistakes. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Additional information

Correspondence and requests for materials should be addressed to Z.K. or B.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.