Original Article

SAGE Open Medicine

SAGE Open Medicine
Volume 12: 1–11
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20503121241263032
journals.sagepub.com/home/smo
Sage

Evolution of waterborne diseases: A case study of Khyber Pakhtunkhwa, Pakistan

Muhammad Atif[1], Gohar Ayub[2], Fazal Shakoor[1], Muhammad Farooq[1], Muhammad Iqbal[1], Qamruz Zaman[1] and Muhammad Ilyas[3]

Abstract

Objectives: In Pakistan, the degradation of drinking water quality is exacerbated by the increasing population size and rapid industrialization. Contaminated water serves as the predominant source of numerous diseases, including diarrhea, gastroenteritis, and typhoid. This article explores the evolution of waterborne diseases across 21 districts of the Khyber Pakhtunkhwa province in Pakistan by monitoring changes in the clustering solutions.

Methods: The data employed in this study were sourced from 21 districts of KP by the Director-General Health Services. Cluster analysis was utilized to uncover patterns in waterborne disease incidence, while principal component analysis was employed to reveal underlying patterns and reduce dimensionality. Additionally, the MONItoring Clusters (MONIC) framework was applied for change detection, facilitating the identification of significant shifts in disease patterns over time and aiding in the understanding of temporal dynamics.

Results: Our analysis indicates that two clusters survived consistently over time, while other clusters exhibited inconsistency. Profiling of the surviving clusters ($C_{12} \rightarrow C_{24} \rightarrow C_{32} \rightarrow C_{43}$) suggests a gradual increase in cases of bloody diarrhea in the Swat Valley, Hangu, Karak, and Lakki Marwat regions. Similarly, profiling of the surviving clusters ($\odot \rightarrow C_{22} \rightarrow C_{34} \rightarrow C_{44}$) suggests an increase in the acute watery diarrhea (non-cholera) and typhoid fever in the regions of Peshawar, Nowshera, and Swabi.

Conclusion: The findings of this study hold significant importance as they pinpoint the most vulnerable regions for various waterborne diseases. These insights offer valuable guidance to policymakers and health officials, empowering them to implement effective measures for controlling waterborne diseases in the respective regions of Khyber Pakhtunkhwa, Pakistan.

Keywords

Change detection, clustering, windowing approach, waterborne diseases

Date received: 11 May 2024; accepted: 30 May 2024

Introduction

Water is essential for life, and access to clean drinking water is one of human's fundamental rights. Among many other factors, the contamination of food and water is a major source of disease transmission. Every year thousands of children lose their lives due to contaminated sources of water. These sources cause acute diarrhea diseases, typhoid, cholera, and so on, which are termed as waterborne diseases.[1,2] Waterborne diseases pose significant public health challenges worldwide, particularly in regions with inadequate sanitation infrastructure and limited access to clean water. According to UNICEF,[3] water-related infections extinguish 1.8 million lives each year, the leading cause of death, across the world. Among these regions, Khyber Pakhtunkhwa (KP), Pakistan, stands out as a hotspot for waterborne diseases due to rapid population growth, industrialization, and limited infrastructure development.[4] Contaminated water sources serve as breeding grounds for pathogens, leading to the transmission of diarrhea, gastroenteritis, typhoid fever, and

[1]Department of Statistics, University of Peshawar, Peshawar, Pakistan
[2]Department of Mathematics and Statistics, University of Swat, Swat, Pakistan
[3]Department of Statistics, University of Malakand, Chakdara, Khyber Pakhtunkhwa, Pakistan

Corresponding author:
Muhammad Atif, Department of Statistics, University of Peshawar, Peshawar 25120, Pakistan.
Email: m.atif@uop.edu.pk

hepatitis. These diseases not only cause immense suffering but also impose a substantial economic burden on healthcare systems and society at large.[5,6]

There has been a significant increase in waterborne diseases in Pakistan as a result of people being compelled to drink stagnant, dirty water after the floods. More than 660,120 cases of acute, watery diarrhea, skin infections, typhoid, malaria, and dengue fever have been reported.[7,8] The frequent floods in Pakistan destroy the infrastructure, resulting in a lack of toilet facilities, water sanitation, and hygiene. Zahid[5] showed that sources of drinking water and toilet facility are the most common environmental household-level indicators for a high prevalence rate of waterborne diseases. Waterborne infections such as dysentery, cholera, giardiasis, and hepatitides A and E have grown more common as a result of poor sanitation and water quality.

The World Health Organization (WHO) estimates that between 25% and 30% of diseases are gastrointestinal disorders. Almost 46% of the KP population is dependent on the polluted water sources that cause a high risk of waterborne infections. Ahmad et al.[9] conducted a research on drinking water quality in the District Peshawar, KP, Pakistan that contained bacteriological studies and evaluation of antibiotic-resistant bacteria from several drinking water sources.

Pakistan has abundant freshwater resources, but with rising population, urbanization, industry, and inadequate sanitation, the water quality is deteriorating, resulting in a high prevalence rate of waterborne diseases.[10] The in-depth exploration of literature reveals that water consumption or recreational water susceptibility is the root cause of waterborne infection. According to UNICEF,[3] water-related illnesses kill 1.8 million people each year and cause 4 billion cases, making them a major cause of death and morbidity across the world.

In order to prevent the spread of diseases that are transmitted through water, it is essential to effectively monitor water sources. However, monitoring these sources and identifying vulnerable regions with waterborne diseases poses significant challenges. These difficulties are caused by a number of factors, such as the dynamics of waterborne diseases, insufficient infrastructure, and scarce resources.[11,12] Due to the inadequate surveillance infrastructure, they are able to spread unnoticed until they reach dangerous levels. The primary challenge contributing to this phenomenon is the limited availability of water samples over time and the difficulty in detecting certain organisms.[13,14] Consequently, we have designed a methodology for monitoring the evolution of most prevalent waterborne diseases and identify the vulnerable regions. This can be achieved by monitoring and tracing the changes in cluster solutions of data stream over time.

In this study, we implement the clustering algorithm for the segmentation of waterborne diseases dataset. Subsequently, the segment profile diagram was used for detailed profiling of each cluster. This approach enables us to identify regions vulnerable to specific waterborne diseases that are predominant within each cluster. Furthermore, the data stream was discretized by using the landmark window model and the evolution of clusters were traced over time.

In the last couple of decades, researchers have increasingly focused on investigating changes in the patterns of underlying populations. The literature has proposed numerous models and algorithms for monitoring and tracing cluster solutions in temporal streams. These approaches offer valuable tools for analyzing dynamic data and capturing evolving patterns over time.[15–20] These algorithms are widely used in various domains for tracking changes in cluster solutions, allowing for the detection of evolving patterns and anomalies in data streams. The study conducted by Atif et al.[20] showcases the practical applications and significance of tracing cluster evolution across a range of real-life datasets. The research highlights the process of segmenting data streams and emphasizes the importance of monitoring changes in clustering solutions. Through their findings, the study illustrates how change detection can offer valuable insights for policymakers, enabling them to effectively address clusters that evolve over time.

Research objectives

1. To investigate the evolution of waterborne diseases in Khyber Pakhtunkhwa and identify the regions vulnerable to exposure.
2. To monitor and trace changes in cluster solutions of waterborne disease dataset.

Methods

Dataset

This study is an observational investigation utilizing secondary data collected by the Director-General Health Services (DGHS) from 21 districts of KP, Pakistan. Among various health concerns, waterborne diseases present significant threats to public health. These diseases often stemming from contaminated water sources, inadequate sanitation infrastructure, and limited access to clean water. To evaluate the existing healthcare and safety conditions, the DGHS conducted data collection on waterborne diseases across 21 districts of KP. Over a span of 28 weeks, this study examined six variables related to waterborne diseases. These variables included acute watery diarrhea (AWD) (cholera), AWD (non-cholera), bloody diarrhea, acute viral hepatitis (AIS), typhoid fever, and extensively drug-resistant (XDR) typhoid. A total of 478 cases of waterborne diseases were observed during the study span. In this study, we utilize this dataset to monitor the evolution of waterborne diseases in KP, aiming to gain insights into their dynamics and identify regions requiring targeted intervention. In the initial step, we cleaned the dataset by removing cases with missing values and outliers.

Clustering

Clustering, a cornerstone technique in machine learning and data analysis, involves grouping similar data points together based on their similarities. The goal of cluster analysis is to partition a dataset into groups such that objects within the same group are more similar to each other than to those in other groups. Overall, cluster analysis is a powerful tool for exploring and summarizing complex datasets, aiding in data interpretation, and decision-making processes. Clustering has numerous applications across various domains, including customer segmentation, image segmentation, outliers detection, and document clustering.[21,22] The process of clustering typically involves the following steps.

Choosing a proximity measure: In cluster analysis, proximity measure, is used to quantify the similarity or dissimilarity between data points. This measure defines the distance between pairs of observations in the feature space and forms the basis for clustering algorithms to group similar data points together. Common proximity measures include Euclidean, Manhattan, Minkowski, Cosine Similarity, and Correlation. The choice of proximity measure depends on the nature of the data and the characteristics of the clustering problem.

Choosing a clustering algorithm: Select an appropriate clustering algorithm that fits the characteristics of the dataset and the objectives of the analysis. Some common clustering algorithms include *k*-means, hierarchical clustering, density-based clustering, and Gaussian mixture models.

Determining the number of clusters: Identifying the optimal number of clusters is a critical aspect of cluster analysis, ensuring that the resulting clusters accurately represent the underlying structure of the data. The process involves selecting the number of clusters that best represent the underlying structure of the data while avoiding over-fitting. Several techniques can be employed for this purpose, each offering insights into the most suitable clustering solution.

Interpreting and visualizing results: Examine the resulting clusters to understand their characteristics and interpret the patterns present in the data. This may involve visualizing the clusters using techniques such as scatter plots, dendrograms, heatmaps, and *t*-distributed Stochastic Neighbor Embedding (*t*-SNE).

In this article, we employed the standard *k*-means algorithm to cluster the dataset. The variables used for clustering included AWD (non-cholera), AWD (cholera), bloody diarrhea, AIS, typhoid fever, and XDR typhoid. To determine the optimal number of clusters, we utilized the elbow method, silhouette score, and Gap statistics. These techniques allowed us to identify the most suitable number of clusters that best captured the inherent structure of the data. The Euclidean distance function was utilized as a dissimilarity measure index.

Windowing approach

Over the recent past, a number of applications in real life have been generating data streams, where data items are continually produced by different sources over time. Unlike traditional datasets that are static and stored in databases, data streams are dynamic in nature and continuously updated with new observations. As a result, the underlying structure is nonstationary and undergoes evolution over time. Due to their high volume, velocity, and variability, data streams present unique challenges for processing, storage, and analysis. In order to achieve this, the continuous data stream must be discretized into subsets according to some ordered phenomena. The term "windowing approach" refers to this discretization of the stream into smaller groups. The windows represent segments of the data stream and are defined based on specific criteria, such as time intervals or the occurrence of certain events. Data within each window is then analyzed or processed independently, allowing for the detection of patterns, trends, or anomalies within that time frame. In this research article we implement the landmark window models to discretize the stream and accumulate the data items at successive time points. The landmark window model is a specific approach to windowing in data stream processing. In this model, the data stream is discretized into subsets or windows based on specific landmark points in time. In the landmark window model, data points are accumulated from a specific landmark time t_1 up to the current time point t_i. This means that the window includes all data points that have been observed between the landmark time t_1 and the current time t_i, that is,

$$\mathrm{D_i} = \bigcup_{\mathrm{i=1}}^{\mathrm{t}} \mathrm{d_i},\ \mathrm{t} = 2,\ldots,\mathrm{n}$$

where *D* represents the window, *d* represents the data items accumulated within the window, and *n* represents the total number of time points observed up to the current time.

Tracing changes in cluster solutions

Monitoring changes in cluster solutions of streaming data involves tracking how clusters evolve over time. This process is crucial for understanding how patterns within the stream change over different time periods and identifying any shifts in cluster memberships. By systematically monitoring changes in cluster solutions of temporal data, researchers can gain valuable insights into the temporal dynamics of the data and make decisions accordingly. Some of the famous algorithms for monitoring changes in cluster solutions are given in Table 1.

To monitor and trace the evolution of clusters extracted from the re-clustering of cumulative datasets, a framework known as the MONIC algorithm was introduced.[23] In this paper, we implement the *clusTransition* package in R-software for change detection in the waterborne disease dataset.[24,25] This helps in understanding the evolution of waterborne diseases in KP.

Table 1. Models and algorithms for monitoring changes in cluster solutions.

Algorithm Name	Description
MONIC	A heuristic framework for monitoring changes in cluster solutions of temporal datasets
DStream	A clustering algorithm designed for streaming data that adapts to concept drift and changes in data distribution
BIRCH	An algorithm for clustering large datasets incrementally while maintaining low memory usage
OPTICS	An algorithm for density-based clustering that provides a hierarchical view of clusters and can adapt to changes in density
CluStream	A stream clustering algorithm that continuously updates cluster models and monitors changes in cluster structure over time
FOCUS	An algorithm based on the self-organizing maps to monitor changes in temporal datasets
CEDAS	Online approach for clustering evolving data streams into arbitrary-shaped clusters in real time

Source: Atif et al.[19]

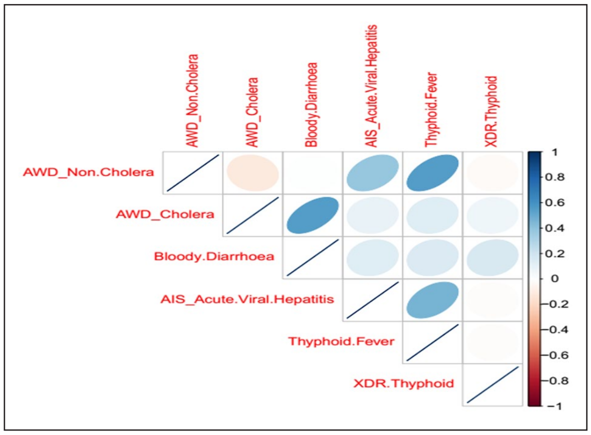


Figure 1. Correlation plot for variables.

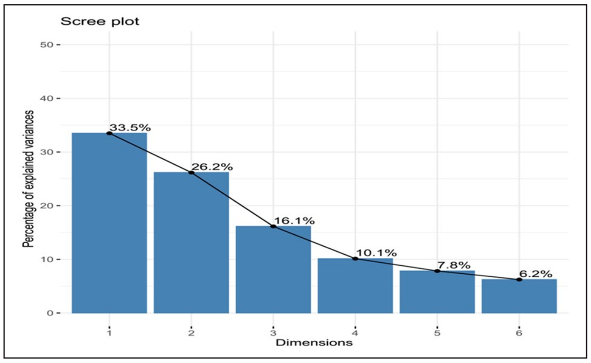


Figure 2. Explained variation by PCs.

Results

Figure 1 demonstrates the correlation plot between variables, where the darker shade represents strong correlation and lighter shade represents weak or no correlation. The plot suggests a mild correlation of AWD (non-cholera) with AIS and typhoid fever. Similarly, AWD (cholera) is correlated with bloody diarrhea, while AIS is correlated with typhoid fever.

Since some of the variables in our dataset are correlated, Principal Component Analysis (PCA) allows us to gain insights into the relationships and patterns among the diseases represented by the variables. Based on the scree plot analysis presented in Figure 2, we decide to retain only two dimensions, as the first two principal components collectively explain approximately 60% of the total variation in the dataset.

Figure 3 demonstrates the contribution of each variable to the corresponding PCs. Subplot A illustrates that typhoid fever, AIS, and AWD (non-cholera) significantly contribute to the variability explained by the first PC. Similarly, in subplot B, AWD (cholera), and bloody diarrhea are shown to make substantial contributions to the variability explained by the second PC. In subplot C, it is observed that AWD (cholera), bloody diarrhea, and XDR typhoid constitute one dimension, whereas, AWD (non-cholera), typhoid fever, and AIS constitute the second dimension, suggesting a separate source of variability.

Figure 4 illustrates the determination of the optimal number of clusters in the dataset using three different methods: the gap statistic, silhouette statistic, and elbow method. The elbow method indicates that additional clusters beyond the fourth add only a small value in minimizing the within-cluster variation, suggesting that four clusters may be optimal. Similarly, according to the gap statistic, the optimal number of clusters is $k=4$, with $k=5$ being a potential contender. The silhouette statistic also supports the conclusion that $k=4$ clusters is optimal based on the dataset. These analyses provide consistent evidence that $k=4$ clusters is the most suitable choice for partitioning the data.

Figure 5 provides a profiling of each individual cluster in the dataset generated by the k-means algorithm. The largest cluster, Cluster 3, represents 37% of the data items and is characterized by a relatively high proportion of bloody diarrhea patients and an extremely low proportion of AWD (non-cholera) and typhoid fever patients. The second largest cluster, Cluster 1, represents 28% patients and is distinguished by a very high proportion of AWD (cholera) and typhoid fever patients. The Cluster 2, comprising 28% of the data items, exhibits a high proportion of AIS patients. The Cluster 4 is characterized by an extremely high proportion of

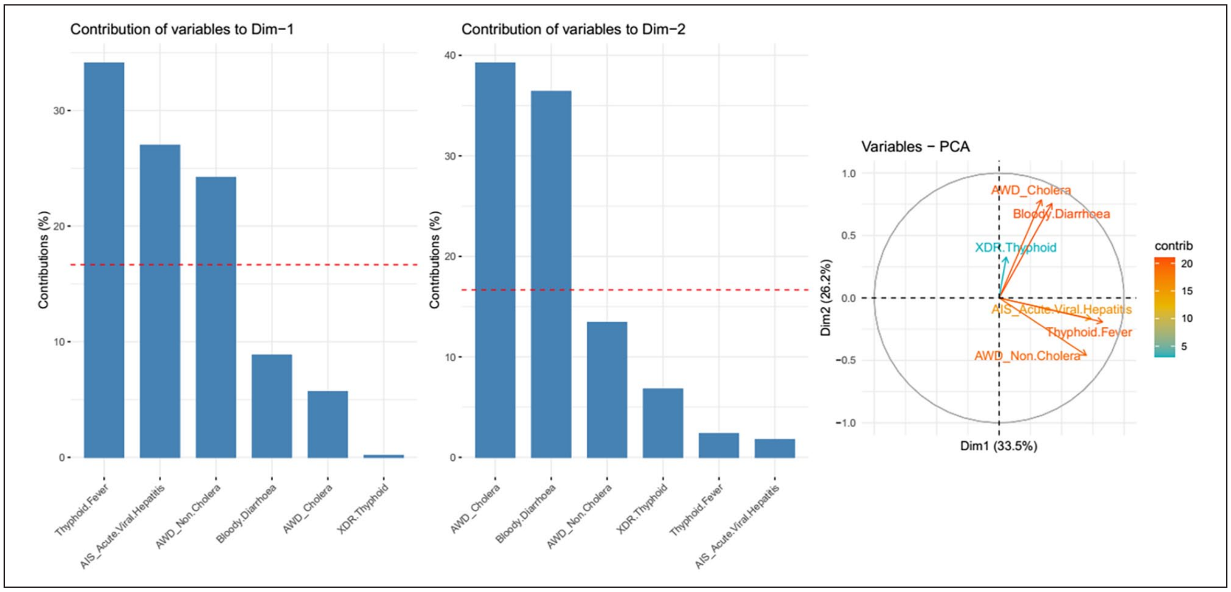


Figure 3. Contribution of PCs to the dimensions.

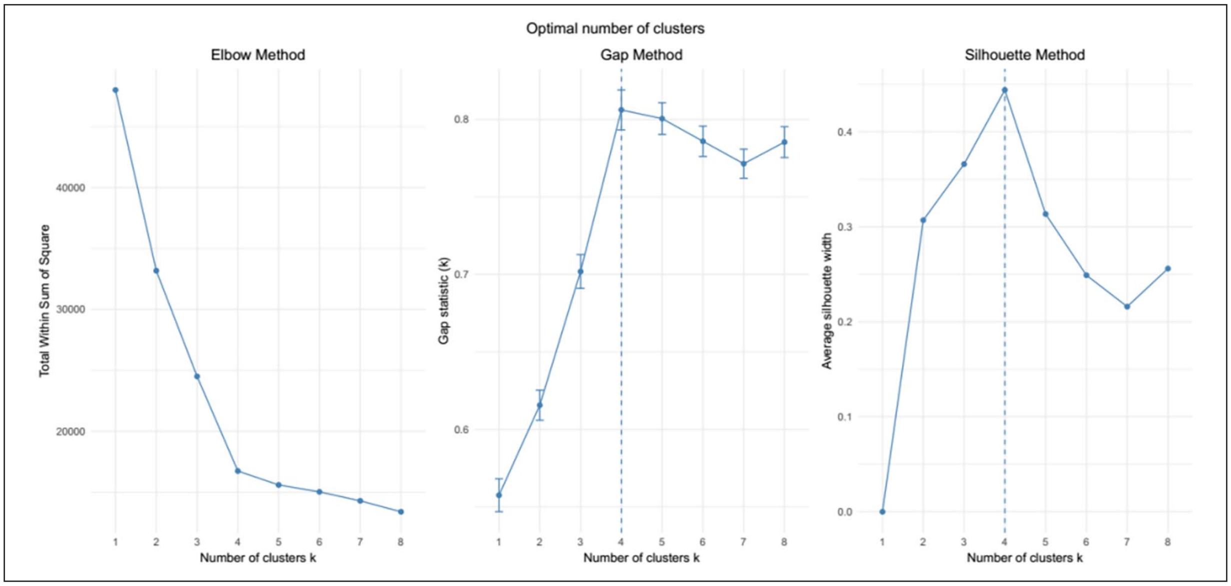


Figure 4. Optimal number of clusters using elbow, gap, and silhouette methods.

AWD (non-cholera) and typhoid fever patients. The disease such as XDR typhoid is homogeneously distributed among all segments.

Table 2 (provided in Annexure I) presents the district-wise summary of all clusters in the dataset. Utilizing the segment profiling plot, we identified the most vulnerable districts to various waterborne diseases on the map of KP, as depicted in Figure 6. The largest cluster (Cluster 3) comprises of Swat Valley (Dir Upper, Dir Lower, Malakand, Swat, Shangla, and Buner), Hangu, Karak, and Lakki Marwat. These regions are identified as highly susceptible to bloody diarrhea. However, this cluster exhibits relatively lower risk levels for AWD (non-cholera) and typhoid fever. The second largest cluster (Cluster 1) includes districts such as Charsadda, Mardan, Swabi, Haripur, Abbottabad, and Dera Ismail Khan. These regions are identified as being exposed to both AWD (cholera) and Typhoid fever patients. Cluster 2 encompasses Battagram, Kohat, Bannu, and Tank districts. These regions are identified as being extremely vulnerable to AIS disease. However, the districts belonging to Cluster 2 are considered safe in terms of AWD (non-cholera) disease. Similarly, Peshawar, Nowshera, and Swabi belong to Cluster 4 and are identified as being extremely vulnerable to

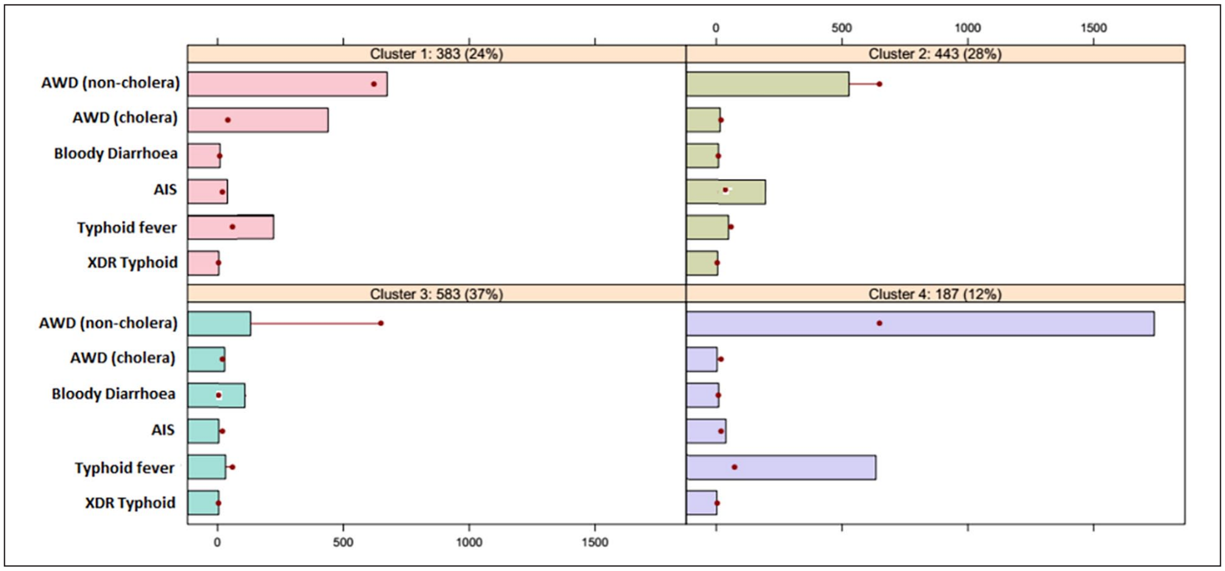


Figure 5. Segment profiling plot.

Table 2. Optimal number of clusters in each windowpane.

Windowpane	D_1 (week 1–week 8)	D_2 (week 9–week 16)	D_3 (week 17–week 24)	D_4 (week 25–week 29)
k_i	3	4	4	4

AWD (non-cholera) and typhoid fever. This observation underscores the importance of implementing targeted public health interventions in these areas to address the high prevalence of these waterborne diseases. By focusing resources and efforts on these vulnerable districts, authorities can effectively mitigate the spread of diseases and improve overall public health outcomes in the region.

Change detection

The stream of data was discretized by accumulating it over the weeks. This was done by implementing the landmark window to the stream, accumulating it at successive time points. A window of 8 weeks was used. The implementation of the landmark window model generates four windowpanes, which comprise data evolving during $[t_1, t_i]$. Table 3 summarizes the optimal number of clusters in each windowpane of cumulative dataset estimated using elbow, gap, and silhouette methods. The details of estimating optimal number of clusters in each windowpane are provided in Figure 12 in Supplemental files. These methods provide consistent evidence that the optimal number of clusters in windowpanes D_1, D_2, D_3, and D_4 is 3, 4, 4, and 4, respectively.

Clustering cumulative datasets at successive time points results in a series of cluster solutions, with each solution corresponding to a specific windowpane of the dataset. The survival thresholds of $\tau=0.6$, 0.7, 0.8, and 0.9 were used to monitor any changes in these cluster solutions over time. Figure 7 demonstrates the survival ratio of clusters for different values of survival threshold. It is evident that $\tau \leqslant 0.7$ produces very stable cluster solutions at successive time points. As the survival threshold τ exceeds 0.7, it signifies a strong temporal dependency in the data. In this scenario, only a few clusters from previous time points survived, while new clusters emerge in the dataset.

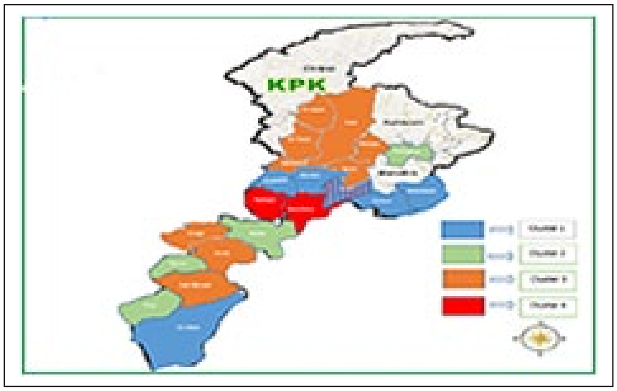


Figure 6. Vulnerable regions with waterborne diseases.

Figure 8 in the Annexure demonstrates the changes in cluster solution for $\tau=0.7$ and $\tau=8$. For $\tau=0.7$, hardly any changes are detected, while for $\tau=0.9$, only one cluster survived. For small τ, we observed that the resulting clusters

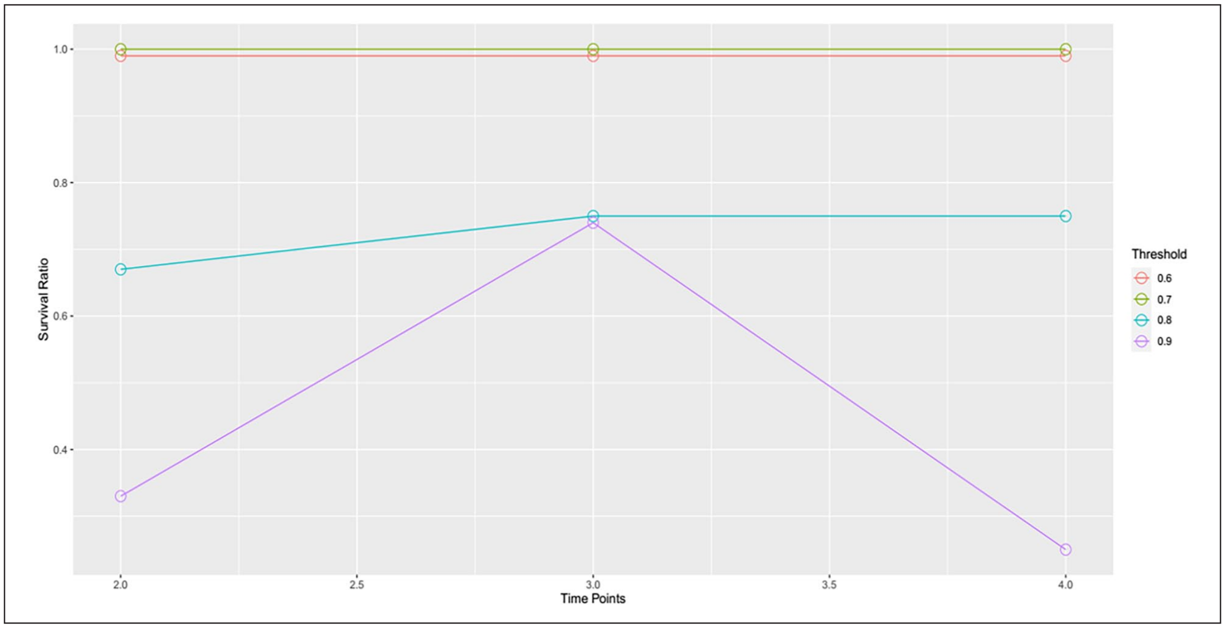
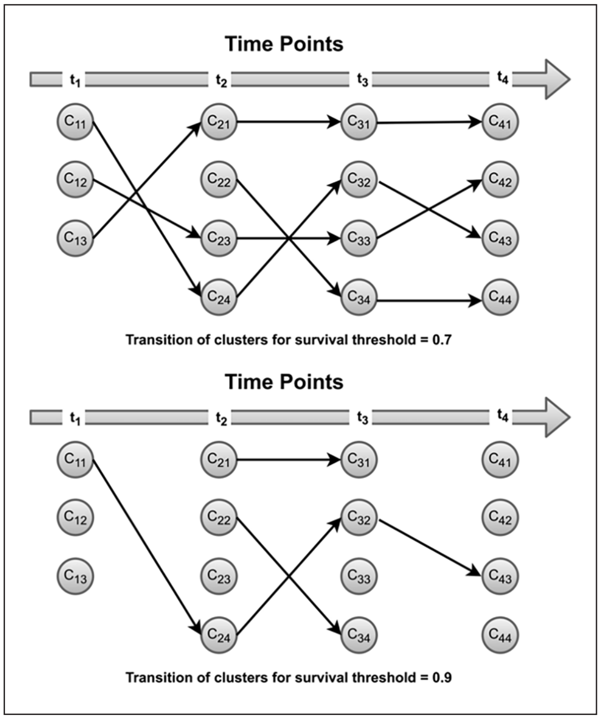


Figure 7. Effect of survival threshold on survival ratio.



Transition of clusters for survival threshold = 0.7

Transition of clusters for survival threshold = 0.9

Figure 8. Transition of clusters for different survival threshold.

were highly inconsistent. Conversely, for very large τ, hardly any changes were detected in the clusters over time. Therefore, we decided to select a survival ratio of 0.8 for further analysis of the evolution of disease. This threshold strikes a balance between capturing meaningful changes in cluster composition while minimizing inconsistency in the clustering results.

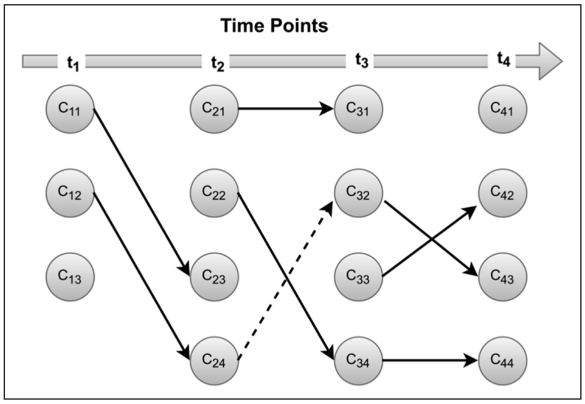


Figure 9. Evolution of clusters over time. The nodes represent the clusters at respective.

Figure 9 demonstrates the evolution of clusters over time for $\tau = 0.8$. It is evident that one cluster survives from time point t_1 till time point t_4 $(C_{12} \rightarrow C_{24} \rightarrow C_{32} \rightarrow C_{43})$, growing more diffuse than its ancestor clusters. Similarly, one cluster that emerged at time point t_2 survived till time point t_4 $(\odot \rightarrow C_{22} \rightarrow C_{34} \rightarrow C_{44})$. One cluster disappears at each successive time point $(C_{13} \rightarrow \odot C_{23} \rightarrow \odot\, C_{31} \rightarrow \odot)$. Two clusters t_4 $(C_{12} \rightarrow C_{24} \rightarrow C_{32} \rightarrow C_{43})$ and $(\odot \rightarrow C_{22} \rightarrow C_{34} \rightarrow C_{44})$ are important clusters, which require detailed analysis and special attention. The other clusters are inconsistent and is constantly disappearing at $\tau = 0.8$ threshold.

Figure 10 demonstrates the profiling plot of survived clusters $(C_{12} \rightarrow C_{24} \rightarrow C_{32} \rightarrow C_{43})$. The plot clearly indicates that the proportion of bloody diarrhea patients is increasing gradually with the passage of time. This suggests that the water quality in Swat Valley, Hangu, Karak, and Lakki

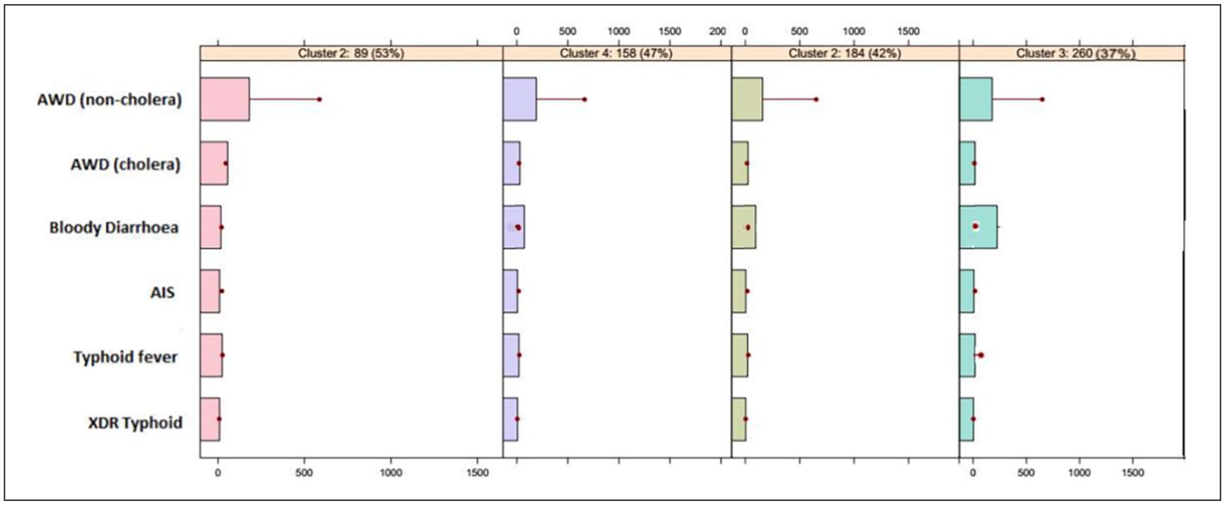


Figure 10. Profiling plot of survived cluster ($C_{12} \rightarrow C_{24} \rightarrow C_{32} \rightarrow C_{43}$).

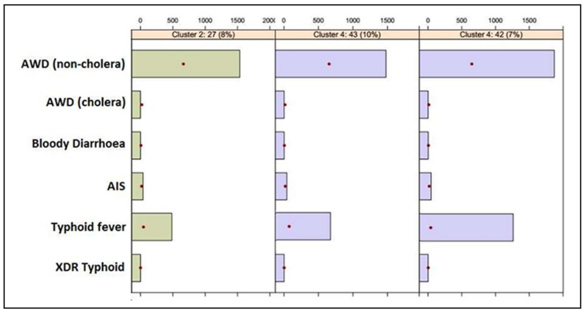


Figure 11. Profiling plot of survived cluster ($\odot \rightarrow C_{22} \rightarrow C_{34} \rightarrow C_{44}$).

Marwat is contaminating day by day causing an increase in the bloody diarrhea in the region.

Similarly, Figure 11 suggests an increase in the AWD (non-cholera) and typhoid fever in the regions of Peshawar, Nowshera, and Swabi.

Discussions

In Pakistan, the drinking water quality is being eroded due to the threatening growth of population size and rapid industrialization. Contaminated water is the primary source of several diseases such as diarrhea, gastroenteritis, and typhoid.[26] Findings of this study shed light on the complex relationship between waterborne diseases, their prevalence, and dynamics in KP, Pakistan. For instance, typhoid fever and AIS are often linked to fecal contamination of water sources, with poor sanitation and inadequate hygiene practices contributing to their common transmission route.[23] The correlation between these diseases suggests common risk factors and environmental conditions conducive to their spread, emphasizing the need for comprehensive sanitation and hygiene interventions to prevent their transmission. Similarly, study by O'Reilly et al.[27] have highlighted the role of bacterial pathogens such as *Vibrio cholerae* and *Escherichia coli* in the etiology of both AWD (cholera) and bloody diarrhea cases. Interestingly, the contribution of XDR typhoid to the principal components was found to be minimal, indicating that this disease is spread homogeneously across KP. This suggests that XDR typhoid may pose a consistent and widespread threat to public health in the region, warranting further attention and targeted intervention efforts.[28]

Waterborne diseases are among the foremost causes of death across the globe. It is suspected that contaminated water is the primary source of spreading these diseases. Inadequate access to clean water and poor sanitation facilities significantly contribute to the spread of these diseases. For instance, cholera, caused by the bacterium *Vibrio cholerae*, is often linked to drinking water contaminated with fecal matter. Similarly, typhoid fever, which is caused by *Salmonella typhi*, is frequently transmitted through ingestion of water or food that has been contaminated by the feces of an infected person.[29] Unfortunately, the identification of risk factors related to waterborne diseases is a difficult task. Because, on one hand, the time to time water samples is not available for all regions. Similarly, on the other hand, some pathogens are difficult to detect. Effective prevention strategies include ensuring access to safe drinking water, improving sanitation and hygiene practices, and implementing robust water quality monitoring systems. These measures are essential to reduce the burden of waterborne diseases and improve public health outcomes worldwide.[30] For the identification of the vulnerable regions to different waterborne diseases, we cluster the dataset and detailed profiling of each segment was studied. The diseases were used to generate clusters, and then summaries of districts were used to identify the vulnerable regions. The study's finding reveals that bloody diarrhea is extremely prevalent in Swat Valley (which includes Dir Upper, Dir

Lower, Malakand, Swat, Shangla, and Buner), Hangu, Karak, and Lakki Marwat. However, this is extremely safe zone for AWD (non-cholera) and typhoid fever. Rahman et al.[31] suggested that renovating existing drinking water sources and implementing new safe drinking water schemes could significantly reduce the prevalence of waterborne diseases and lower household healthcare costs in the region. Similarly, Charsadda, Mardan, Swabi, Haripur, Abbottabad, and Dera Ismail Khan are extremely vulnerable to bloody diarrhea and typhoid fever, but are considered safe for AWD (non-cholera). The AIS is more prevalent in the regions of Battagram, Kohat, Bannu, and Tank. Similarly, the AWD (non-cholera) and typhoid fever is extremely prevalent in Peshawar, Charsadda, and Swabi.

Monitoring changes in cluster solutions identifies regions facing persistent water quality challenges and increasing prevalence of waterborne diseases. Specifically, analysis reveals concerning trends in the water quality of Peshawar, Nowshera, and Swabi, where pollution levels are consistently rising, leading to a significant increase in cases of AWD (non-cholera) and typhoid fever.[32,33] These regions require urgent attention and targeted interventions to address the underlying causes of water pollution and mitigate the associated health risks. Furthermore, the Swat Valley, Hangu, Karak, and Lakki Marwat are experiencing a rapid increase in cases of bloody diarrhea. This alarming trend underscores the need for immediate action to improve sanitation infrastructure, enhance water treatment processes, and implement effective public health measures to prevent the spread of these diseases in these regions.

While this study contributes to understanding the dynamics of waterborne diseases in KP, it is essential to acknowledge its limitations to provide a foundation for future research and public health interventions.

This study does not account for all potential confounding factors that could influence the relationship between water quality and disease prevalence. Factors such as socioeconomic status, access to healthcare, and environmental variables could have significant impacts on disease transmission but were not explicitly considered in the analysis. Additionally, as an observational study, the findings shed light on the dynamics of waterborne diseases and identify vulnerable regions affected by them. However, it's crucial to acknowledge that causality cannot be established solely from observational data. While we observed correlations between certain variables, further research, including longitudinal studies or randomized controlled trials, is needed to explore the causal relationships.

Conclusion

The findings of this study underscore the relationships among different waterborne diseases in KP, Pakistan, emphasizing the need for comprehensive strategies to address public health challenges. Identifying vulnerable regions to different waterborne diseases through clustering and detailed profiling enables targeted interventions and resource allocation for effective disease prevention and control. The findings highlight alarming trends, such as the rapid increase in cases of bloody diarrhea in Swat Valley, Hangu, Karak, and Lakki Marwat, and rising pollution levels in Peshawar, Nowshera, and Swabi leading to an increase in cases of AWD (non-cholera) and typhoid fever. Urgent action is needed to improve sanitation infrastructure, enhance water treatment processes, and implement effective public health measures to prevent the spread of these diseases.

Based on the findings of this study, several recommendations can be made to address the challenges posed by waterborne diseases and improve public health outcomes. Firstly, it is essential to enhance water quality monitoring by implementing robust systems across all districts to identify and mitigate sources of contamination promptly. Additionally, investing in infrastructure projects to improve access to clean water and sanitation facilities, particularly in vulnerable communities, is crucial. Public health campaigns should be launched to raise awareness about the importance of clean water, proper hygiene practices, and the prevention of waterborne diseases. Furthermore, developing targeted interventions for regions identified as high risk based on the clustering analysis is necessary, focusing on disease prevention and control measures. Finally, further research, including longitudinal studies or randomized controlled trials, is necessary to explore causal relationships and inform more targeted public health interventions.

Acknowledgements

Not applicable.

Author contributions

Study conception and design: Muhammad Atif and Fazal Shakoor. Data collection: Gohar Ayub and Muhammad Ilyas. Interpretation of results: Muhammad Farooq and Qamruz Zaman. Draft manuscript Preparation: Muhammad Atif and Muhammad Farooq. Supervision: Muhammad Iqbal and Qamruz Zaman. All authors reviewed the results and approved the final version of the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Ethical approval

Our study did not require an ethical board approval because there was no human interaction in this study.

Informed consent

Not applicable.

Trial registration

Not applicable.

ORCID iD

Muhammad Atif https://orcid.org/0000-0002-4139-8292

Supplemental material

Supplemental material for this article is available online.

References

1. Cisse G. Food-borne and water-borne diseases under climate change in low- and middle-income countries: Further efforts needed for reducing environmental health exposure risks. *Acta Trop* 2019; 194(12): 281–188.
2. Butt M and Khair SM. Cost of illness of water-borne diseases: a case study of Quetta. *J Appl Emerg Sci* 2016; 5(2): 133–143.
3. UNICEF. The state of the world's children 2008, https://www.unicef.org/ reports/state-worlds-children-2008 (2008, accessed 5 May 2024).
4. Qamar K, Nchasi G, Mirha HT, et al. Water sanitation problem in Pakistan: a review on disease prevalence, strategies for treatment and prevention. *Ann Med Surg* 2022; 82: 104709
5. Zahid J. *Impact of clean drinking water and sanitation on water borne diseases in Pakistan*. Technical report, Sustainable Development Policy Institute, 2018.
6. Howard G. The future of water and sanitation: global challenges and the need for greater ambition. *Water Infrastruct Ecosyst Soc* 2021; 70(4): 438–448.
7. Javed A and Kabeer A. Enhancing waterborne diseases in Pakistan and their possible control. *Am Scient Res J Eng Technol Sci* 2018; 49(1): 248–256.
8. Cann KF, Thomas DR, Salmon RL, et al. Extreme water-related weather events and waterborne disease. *Epidemiol Infect* 2013; 141(4): 671–686.
9. Ahmad B, Liaquat M, Ali J, et al. Microbiology and evaluation of antibiotic resistant bacteria profiles of drinking water in Peshawar, Khyber Pakhtunkhwa. *World Appl Sci J* 2014; 30(11): 1668–1677.
10. Shahzad S, Ali N, Hussain J, et al. Prevalence of water-borne diseases and perception of quality of drinking water in the low socioeconomic area of Islamabad, Pakistan: a cross-sectional study. *South Asian J Emerg Med* 2019; 2(1): 29–36.
11. Tillett HE, de Louvois J and Wall PG. Surveillance of outbreaks of waterborne infectious disease: categorizing levels of evidence. *Epidemiol Infect* 1998; 120(1): 37–42.
12. Morua AR, Halvorsen KE and Mayer AS. Waterborne disease-related risk perceptions in the Sonora river basin, Mexico. *Risk Analy* 2012; 31(5): 866–878.
13. Ruiz-Moreno D, Pascual M, Emch M, et al. Spatial clustering in the spatio-temporal dynamics of endemic cholera. *BMC Infect Dis* 2010; 10: 51.
14. Jabeen A, Huang X and Aamir M. The challenges of water pollution, threat to public health, flaws of water laws and policies in Pakistan. *J Water Res Protect* 2015; 7(17): 151516–151526.
15. Chakrabarti D, Kumar R and Tomkins A. Evolutionary clustering. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining – KDD 06*, New York, NY, USA, 2006.
16. Kim M-S and Han J. A particle-and-density based evolutionary clustering method for dynamic networks. *Proc VLDB Endowm* 2009; 2(1): 622–633.
17. Spiliopoulou M, Ntoutsi E, Theodoridis Y, et al. Monic: modeling and monitoring cluster transitions. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2006, pp. 706–711.
18. Fahy C and Yang S. Finding and tracking multi-density clusters in online dynamic data streams. *IEEE Trans Big Data* 2022; 8: 178–192.
19. Atif M, Shafiq M, Farooq M, et al. Monitoring changes in clustering solutions: a review of models and applications. *J Probab Stat* 2023; 2023: 7493623.
20. Atif M, Shafiq M and Leisch F. Applications of monitoring and tracing the evolution of clustering solutions in dynamic datasets. *J Appl Stat* 2023; 50(4): 1017–1035.
21. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Comput Appl Math* 1987; 20: 53–65.
22. Tibshirani R, Walther G and Hastie T. Estimating the number of data clusters via the gap statistic. *J Royal Stat Soc B* 2001; 63(2): 411–423.
23. Asadi F, Trinugroho JP, Hidayat AA, et al. Data mining for epidemiology: the correlation of typhoid fever occurrence and environmental factors. *Proc Comput Sci* 2023; 216: 284–292.
24. R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. https://www.R-project.org/
25. Atif M and Leisch F. ClusTransition: an R package for monitoring transition in cluster solutions of temporal datasets. *PLoS One* 2022; 17(12): e0278146.
26. Rahmat ZS, Zubair A, Abdi I, et al. The rise of diarrheal illnesses in the children of Pakistan amidst COVID-19: a narrative review. *Health Sci Rep* 2023; 6(1): e1043.
27. O'Reilly CE, Jaron P, Ochieng B, et al. Risk factors for death among children less than 5 years old hospitalized with diarrhea in rural western Kenya, 2005–2007: a cohort study. *PLoS Med* 2012; 9(7): e1001256.
28. Klemm EJ, Shakoor S, Page AJ, et al. Emergence of an extensively drug-resistant *Salmonella enterica* serovar Typhi clone harboring a promiscuous plasmid encoding resistance to fluoroquinolones and third-generation cephalosporins. *mBio* 2018; 9(1): e00105–e00118.
29. Ali M, Nelson AR, Lopez AL, et al. Updated global burden of cholera in endemic countries. *PLoS Neglect Trop Dis* 2015; 9(6): e0003832.
30. Fewtrell L and Bartram J. *Water quality: guidelines, standards and health. Assessment of risk and risk management for water-related infectious disease*. Geneva, Switzerland: World Health Organization, 2001.
31. Rahman M, Ali S and Hayat N. Households health cost from water borne diseases in District Swat, Khyber Pakhtunkhwa, Pakistan. *IRASD J Eco* 2022; 4(4): 633–646.
32. Awan F, Ali MM, Afridi IQ, et al. Drinking water quality of various sources in Peshawar, Mardan, Kohat and Swat districts of Khyber Pakhtunkhwa province, Pakistan. *Braz J Biol* 2022; 84: e255755.
33. Kibria Z, Khan MN, Aleem S, et al. Linkages between poverty and food insecurity in Pakistan: evidence from urban and rural households in Peshawar. *Pak J Med Sci* 2023; 39(2): 479–484.

Annexure

Table 3. District wise summary of each cluster.

District	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Abbottabad	38	27	11	0
Bannu	1	54	21	0
Battagram	0	45	0	0
Buner	0	7	69	0
Charsadda	46	22	8	0
Dera Ismail Khan	56	4	0	16
Dir Lower	57	15	35	0
Dir Upper	0	0	76	0
Hangu	0	0	76	0
Haripur	21	18	23	14
Karak	0	0	76	0
Kohat	18	58	0	0
Lakki Marwat	0	15	61	0
Malakand	12	34	30	0
Mardan	36	32	5	3
Nowshera	14	0	0	62
Peshawar	19	7	2	48
Swabi	41	0	0	35
Shangla	0	33	43	0
Swat	24	11	32	9
Tank	0	61	15	0