Detecting Dengue Hotspots in Singapore Using Unsupervised Learning

Jamie Too ZhuoEn

***Abstract*—Dengue continues to be a pressing public health issue in Singapore, with 2020 showing one of the worst outbreaks in recent memory. This study explores how unsupervised learning can be used to detect dengue hotspots, offering valuable insights for data-driven disease tracking and potential strategies for intervention and prevention. For this project, we analysed a dataset released by the National Environment Agency (NEA) that documents the dengue clusters reported between 15 February and 9 July 2020. We applied Expectation-Maximisation (EM) and K-Means clustering, alongside Principal Component Analysis (PCA), Independent Component Analyses (ICA), and Gaussian Random Projection (GRP), to identify spatial-temporal dengue outbreak patterns across subzones. Both models divided the dataset into clusters reflecting varying levels of dengue case intensity, but the EM model consistently delivered stronger overall performance when applied to the entire dataset. The EM clustering model demonstrated stronger statistical and spatial performance compared to the K-Means model, validating the hypothesis to a higher degree that dengue spacial and temporal clusters correspond to higher cases. The results of EM and ICA showed partial support, while EM displayed strong empirical validation. In conclusion, this project has provided fundamental insights into the use of supervised learning as a means of identifying and preventing outbreaks in Singapore.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Dengue remains a persistent public health concern in Singapore, with 2020 marking one of the country's most severe outbreaks. For this project, we analysed a dataset released by the National Environment Agency (NEA) that documents the dengue clusters reported between 15 February and 9 July 2020. This was a period that overlapped with Singapore's COVID-19 nationwide lockdown. Each reported case was tagged to a probable infection location, such as residential block or workplace, and mapped to fine-grained spatial units averaging $1.35km^2$, enabling a detailed study of transmission patterns. This dataset provides an opportunity to examine how dengue clusters evolve under restricted human mobility. Our hypothesis is that certain spatial-temporal clusters correspond to higher case intensities, potentially revealing outbreak hotpots that can inform targeted vector-control strategies. To investigate this, we applied clustering and dimensionality-reduction methods, including Expectation Maximisation (EM), K-Means clustering, Principal Component Analysis(PCA), Independent Component Analysis (ICA), and Gaussian Random Projection (GRP), to uncover meaningful patterns in the spread of dengue during this critical period.

II. Methodology

For this project, we will be exploring 2 different models: Expectation-Maximisation (EM) and K-Means Clustering. EM will be using the following major techniques: ICA, EM and K-Means, while K-Means Clustering will be using, K-Means, PCA, GRP, and EM.

A. Data Organisation

The dataset comprised approximately 160,000 dengue case records, defined by "record_ID", "latitude", "longitude", "date", "case_number", "cluster_label", and "subzone_ID". In each model, the data was organised according to its needs.

For the expectation-maximisation (EM) model, the "date" column of the dataset is converted to datetime. Relevant features that are time-related and geographically related were cleaned up and extracted. The data are then normalised to create a standardised format that improves model performance [1].

For the next K-Means model, the data was aggregated into 214 subzones. Each subzone was assigned a mean latitude and longitude, along with total, average, and maximum case counts to reflect local intensity. This supports analysing broader spatial-temporal patterns and identifying higher-intensity clusters that may indicate potential outbreak hotspots at a meaningful geographical scale.

B. Expectation Maximisation (EM)

The EM algorithm estimates parameters using Maximum Likelihood when data are incomplete or partially hidden [2]. EM also helps uncover overlooked information. It operates through two alternating steps: the Expectation (E) step computes expected values of missing data based on current parameter estimates, and the Maximisation (M) step updates those estimates by maximising the likelihood of the observed data [3].

These steps repeat until convergence, producing stable final estimates. Through this iterative refinement, EM gradually improves parameter accuracy and adapts to the underlying structure [4]. Its flexibility allows it to handle overlapping clusters and varying data shapes effectively, making it valuable for complex datasets.

C. K-Means clustering

Clustering determines whether distinct spatial and epidemiological groupings exist across subzones. K-Means helps to efficiently organizes similar data points into groups by minimising the distance between data points within a cluster, using their K-Means value [5].

K-Means identifies patterns without assumptions about cluster structure, and standardising the features ensures equal influence of spatial and case variables. The optimal cluster number was chosen using the Elbow Method, and the Silhouette Score [6], and then applied and merged back into the dataset for mapping and interpretation. This provides an initial baseline for dengue intensity and spatial proximity.

D. Gaussian Mixture Models GMM

GMM is a probabilistic model incorporating uncertainty, generating data from multiple Gaussian distributions representing clusters [4]. This allows flexible cluster shapes and soft clustering, where data points can belong to multiple clusters simultaneously with varying degrees of membership.

GMM with EM estimates the likelihood of each data point belonging to each Gaussian component and normalises these probabilities. Its adaptive covariance allows clusters to vary in shape and orientation, enabling effective separation of complex, overlapping data. This helps clarify overlapping regions, accurately identifying cluster centroids and revealing the true underlying structure.

E. Principal Component Analysis (PCA)

PCA was used to simplify the dengue dataset by reducing standardised features into orthogonal components ranked by explained variance [7]. The first components captured most information, and feature loadings revealed whether geographic or epidemiological variables drove variability. This clarified key data patterns and provided insight into how spatial structure relates to dengue intensity.

K-Means clustering was applied to PCA-transformed data to assess whether cluster separation remained consistent after dimensionality reduction and to improve interpretability. Clustering quality was further validated using EM and metrics such as the Silhouette Score, Bayesian Information Criterion (BIC), and Akaike Information Criterion (AIC), ensuring model stability and cluster validity.

F. Independent Component Analysis (ICA)

ICA finds a linear transformation that maximises statistical independence among the components [8]. ICA helps to improve the results of K-Means and EM by separating mixed signals into independent, non-Gaussian components that are more distinct, which allows for easier clustering to identify underlying clusters.

G. Gaussian Random Projection (GRP)

GRP was used as an additional dimensionality-reduction method to test clustering robustness. Unlike PCA, it applies random Gaussian projections while preserving approximate distances. If similar clusters persist under this randomness, it indicates the dataset's spatial-temporal patterns are inherent and not dependent on PCA's structure [9]. The standardised five-dimensional data were reduced to two GRP components and clustered using K-Means with the previously selected k. A GMM was also fitted for comparison. EM and metrics such as silhouette Score, BIC, and AIC were applied to determine clustering stability and overall parsimony.

III. Results

This section presents the outcomes of our exploration of applying clustering and dimensionality reduction algorithms to the dataset. A total of 11 unique demonstrations have been conducted, enabling us to uncover patterns relevant to our initial hypothesis.

A. EM results

```
================================================================================
FINAL EM CLUSTERING MODEL EVALUATION
================================================================================

Model Configuration:
  • Number of clusters (k):     5
  • Covariance type:            full
  • Converged:                  True
  • Iterations:                 96

Performance Metrics:
  • BIC Score:                  -67,829.36 (lower is better)
  • AIC Score:                  -69,205.44 (lower is better)
  • Log-Likelihood:             34,781.72 (higher is better)
  • Silhouette Score:           0.0425 (range: -1 to 1, higher is better)

Cluster Quality:
  • Overall Quality:            GOOD
  • Detail:                     reliable clustering despite low Silhouette
  normal for spatial-temporal disease clustering
  • Avg Assignment Confidence:  98.29%

  Epidemiological Context:
    Low Silhouette (0.04) + High Confidence (98%) indicates:
     Outbreak zones overlap spatially (realistic for disease transmission)
     Clusters distinguished by temporal patterns and severity
     Results are scientifically valid for dengue outbreak analysis
```

Fig. 1. Final EM Clustering Model Evaluation.

The following metrics were used to determine the optimal number of clusters for our model:

- Bayes Information Criterion (BIC): A measure of the goodness of fit.
- Akaike Information Criterion (AIC): An estimate of the relative quality of the statistical model for a given dataset.
- Silhouette Score: A measure of how well an object lies within its cluster.

The optimal number of clusters for our model is the one that minimises the BIC and AIC values while maximising the silhouette score. The best balance between the metrics is achieved at 5 clusters, which yields a BIC score of -67,829.36, an AIC score of -69,205.44, and a silhouette score of 0.0425, as shown in Figure 1. In addition, the overall cluster quality is supported by the high assignment confidence.

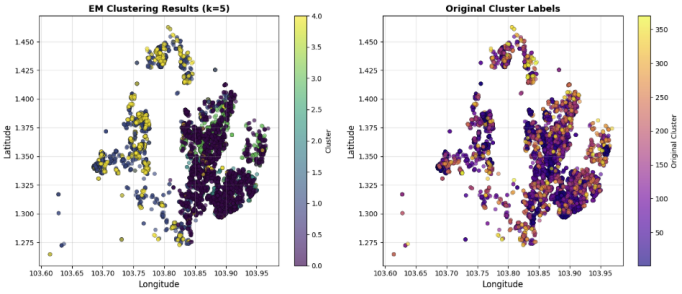

Fig. 2. Geographic scatter plot with EM algorithm applied

Figure 2 presents the geographic scatter plot to visualise dengue cases in Singapore. The right plot shows the original cluster labels. The left plot shows that applying the EM algorithm groups the dengue cases into five distinct groups, suggesting reliable clustering.

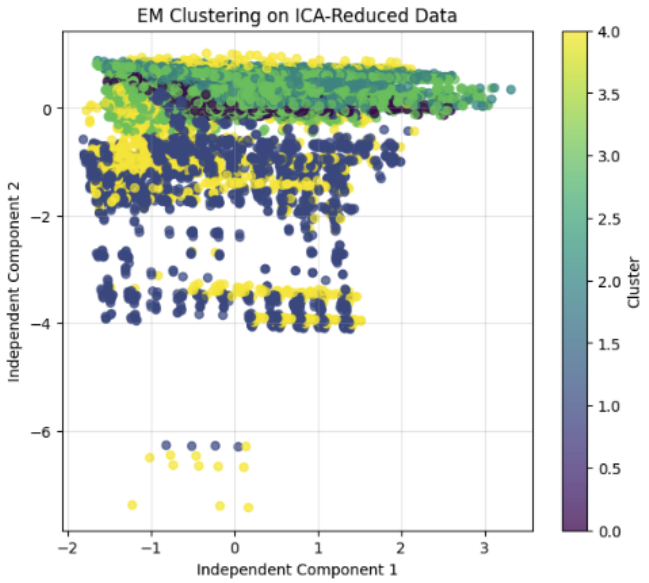

Fig. 3. EM clustering on ICA-Reduced Data

1) EM ICA: Figure 3 shows the results of the EM algorithm in the ICA-reduced space in two independent components. The blue clusters appear clear and independent, with fairly little overlap between other clusters, which demonstrates that the data features can distinguish the groups clearly. However, the boundaries of the green and purple clusters are less distinct, indicating a possible limitation in how well the clusters are defined after dimensionality reduction.

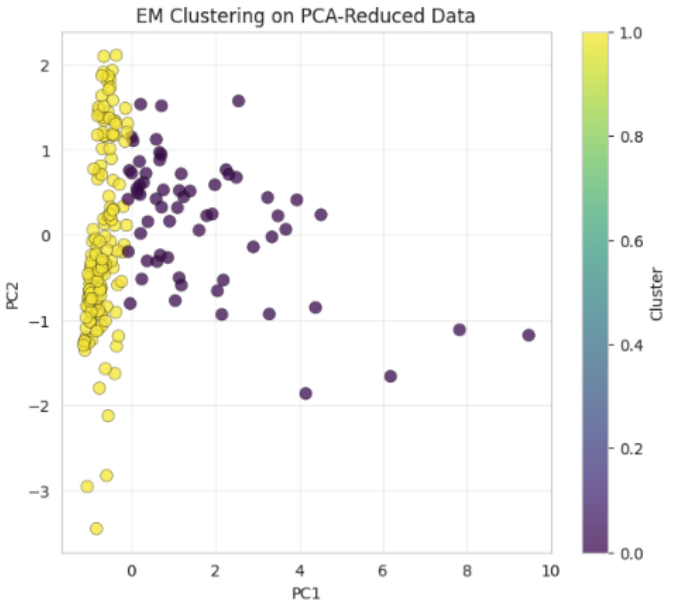

Fig. 4. EM Clustering on PCA-Reduced Data

2) EM PCA: Figure 4 shows the results of the EM algorithm in the PCA-reduced space in two principal components. A clear boundary with minimal overlap is formed at PC1 = 0, separating the data into two distinct clusters. Compared to the yellow cluster, the purple cluster appears more spread out, which suggests that the data retains significant variability. The reduction in clustering quality is also supported by a reduction in the silhouette score when comparing K-Means to EM: K-Means achieves a silhouette score of 0.618, whereas EM achieves a silhouette score of 0.410.

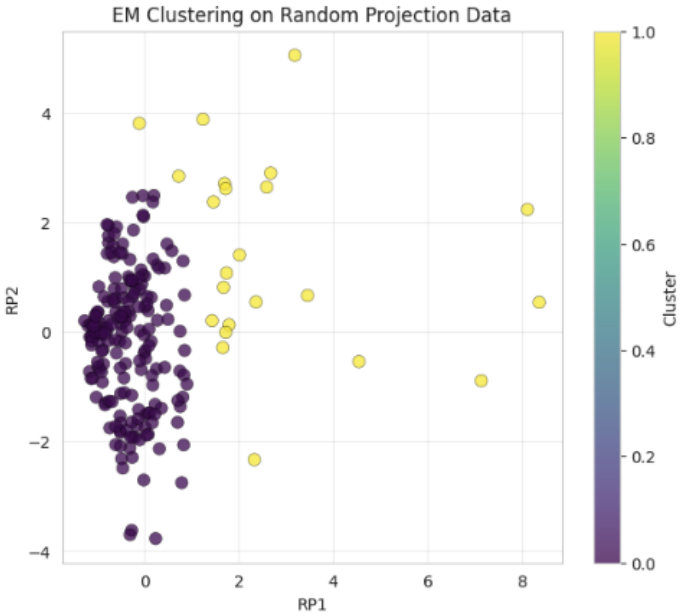

Fig. 5. EM Clustering on Random Projection Data

3) EM RP: Figure 5 shows the results of the EM algorithm in the Random Projection space that has been reduced to two dimensions. A clear boundary with minimal overlap is formed at RP1 = 1, which separates the data into two distinct clusters. The dense purple cluster suggests high feature similarity, whereas the sparse yellow cluster contains more outliers.

IV. K-Means

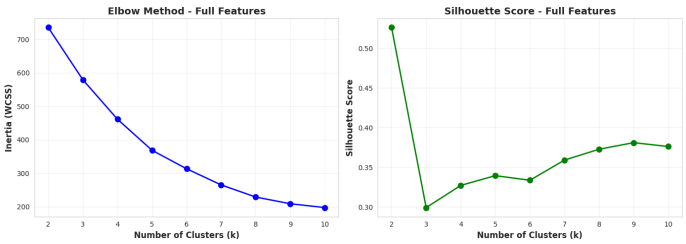

Fig. 6. K-Means Elbow Method and Silhouette Scores

1) K-Means Elbow Method and Silhouette Scores: Figure 6 shows two plots: Inertia against Number of Clusters and Silhouette Score against Number of Clusters.

The inertia plot showed a gradual decrease as k increased. Using the elbow method, we aim to look for a sharp bend in this graph, where the rate of decline in inertia slows drastically. However, this is not present in this graph. This suggests that our dataset may not contain naturally distinct, separate clusters.

The silhouette score was also calculated and plotted. Scores closer to 1 indicate data points are well separated. However, the plot shows a maximum score of 0.526, with the majority of silhouette scores around 0.30 to 0.39. This suggests that the clusters formed are only moderately separated and that some overlap or ambiguity exists.

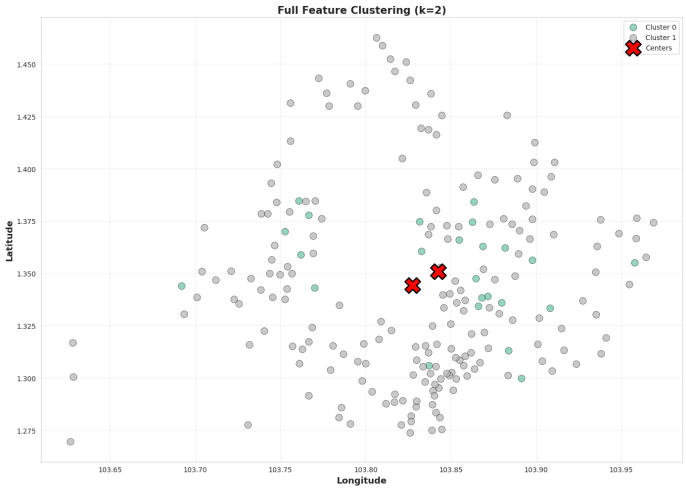

Fig. 7. K-Means Full Feature Clustering

2) K-Means Geographical Visualisation: Figure 7 shows the Full Feature Clustering, using Latitude and Longitude for its axes. It provides a geographical visualisation of clusters, showing that some clusters are geographically dispersed and not strictly spatially compact. The number of clusters used was 2, as it was established to have the highest, but imperfect, separation indicated by the peak silhouette score.

The map indicates that the clusters are not grouped but scattered, making it difficult to clearly determine the hotspot zones. Cluster 0 and cluster 1 have plot points that are spatially indistinguishable from each other. Furthermore, the cluster centres marked should provide a reference for the relative centre location for each group; however, the clusters are widely scattered, and these cluster centres cannot provide an accurate centre location for each cluster. We are unable to determine clear spatial hotspots.

```
Cluster Profiles (Full Features):
              n_subzones  avg_total_cases  sum_cases  max_total  \
cluster_full
0                     24          1214.50      29148       3473
1                    190            99.87      18976        876

              avg_cases_per_report  avg_max_cases  avg_lat  avg_lon
cluster_full
0                             6.41          35.75     1.35   103.84
1                             1.74           5.06     1.34   103.83
```

Fig. 8. K-Means Cluster Summary Table

3) K-Means Cluster Summary Table: Figure 8 shows the cluster summary table, generated after applying K-Means clustering. The algorithm has identified a distinct separation between subzones with high and low case intensities. Cluster 0, although smaller in number (24 subzones), exhibits dramatically higher average and total case counts compared to Cluster 1, suggesting that these areas likely represent significant outbreak hotspots. However, the fact that both clusters share similar average geographic coordinates implies that these high-intensity zones are not isolated to a particular region and are instead interspersed with regions of lower intensity. This means that while K-Means successfully differentiates subzones based on case load, it does not uncover sharply distinct geographic clusters. Thus, the table suggests that case intensity is the primary driver of cluster separation, rather than clear-cut spatial boundaries, highlighting a limitation in using standard K-Means for spatial hotspot detection in this context.

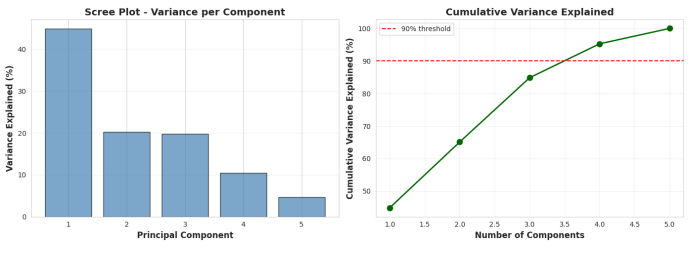

Fig. 9. K-Means Scree and Cumulative Variance Explained Plots

4) K-Means Scree Plot and Cumulative Variance Explained: Figure 9 shows the Scree Plot and Cumulative Variance Explained. In the scree plot, there is a notable drop in variance explained in the first few Principal Components. This

was corroborated as well in the cumulative variance explained plot, where a threshold of 90 per cent was obtained after the number of PCs reached at least 3.

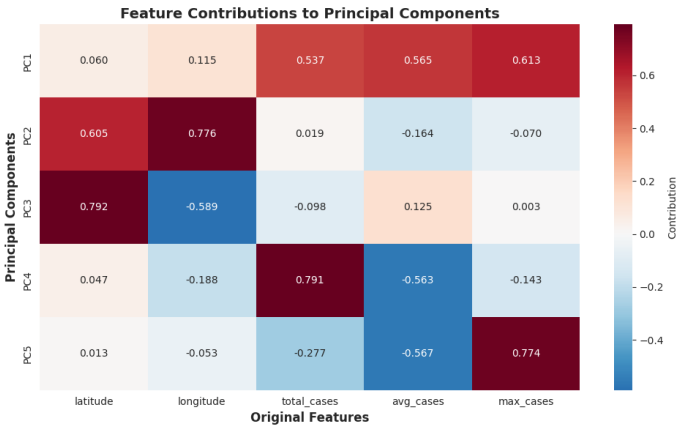

Fig. 10. K-Means Feature Contribution Heatmap

5) K-Means Feature Contributions to Principal Components: Figure 10 shows the results of the EM algorithm in the Random Projection space that has been reduced to two dimensions. A clear boundary with minimal overlap is formed at RP1 = 1, which separates the data into two distinct clusters. The dense purple cluster suggests high feature similarity, whereas the sparse yellow cluster contains more outliers.

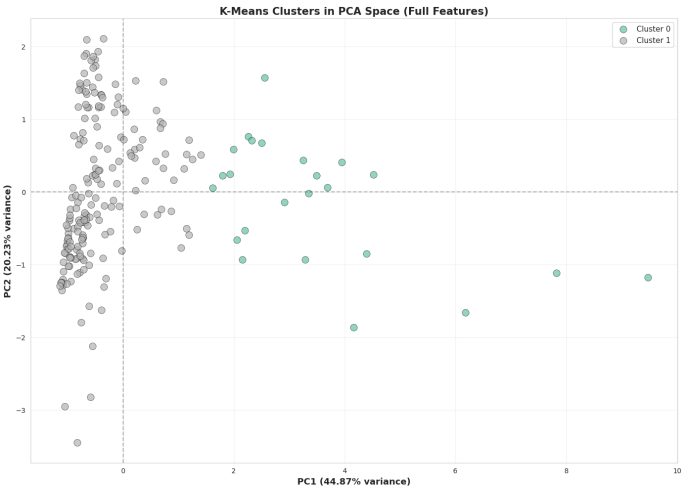

Fig. 11. K-Means Clusters in PCA Space

6) K-Means PCA: Figure 11 shows the plot using K-Means Clusters in PCA space. PC1 and PC2 were selected as previously discussed. This PCA plot reveals a small group of outlier subzones with much higher case intensity (Cluster 0), clearly separated from the large majority of subzones (Cluster 1) that are tightly packed near the origin. The strong divide along PC1 highlights that case numbers drive the clustering more than location, signalling a few distinct hotspots within otherwise similar regions.

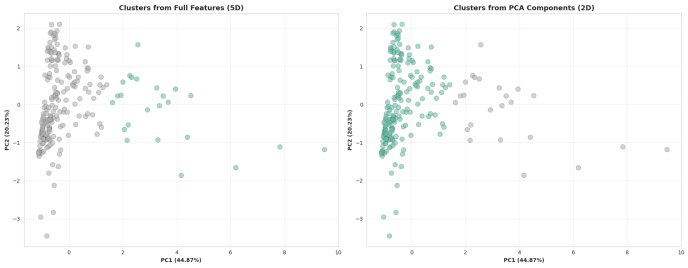

Fig. 12. K-Means Clusters in Full Features vs PCA Space

7) K-Means Full Features vs PCA: Figure 12 shows the plots for clusters using all original features and for clusters using features identified using PCA. Both methods show a small set of outliers with high PC1 values and a large group near the origin, but the cluster boundaries are even fuzzier after PCA reduction. Overall, neither approach finds sharply defined clusters, suggesting that boundaries between outbreak intensity groups are gradual rather than distinct. This further emphasises that K-Means may not be sufficient in determining hotspots.

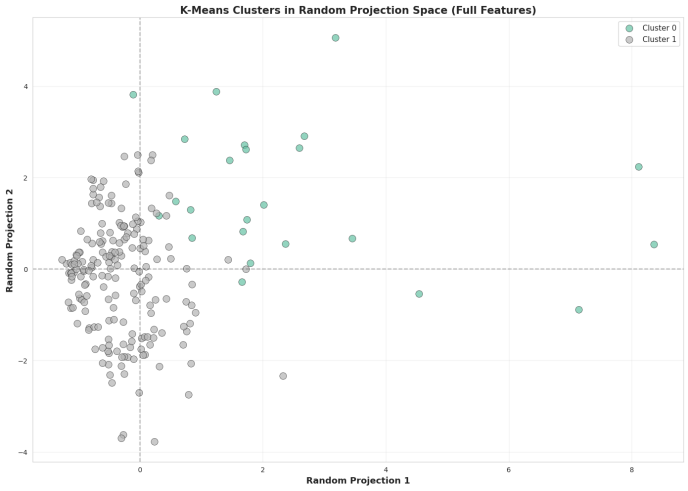

Fig. 13. K-Means Clusters in Random Projection Space

8) K-Means RP: Figure 13 shows the Random Projection Cluster Plot. In this plot, the clusters overlap substantially. There is minimal difference between the cluster segregation in the random projection space compared to the original fifth dimension plot. Typically, cluster overlap would indicate that random projection has obscured some of the group structure; however, as previous plots have shown, the original full feature separation was weak. This instead shows that our clustering solution is not sufficient in determining any clusters within the dataset.

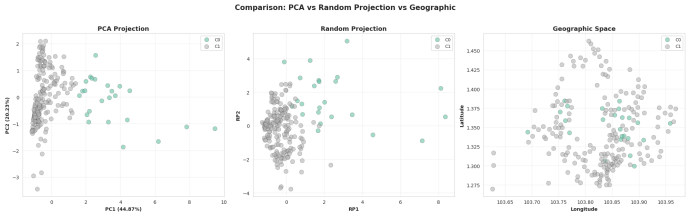

Fig. 14. K-Means Clustering Method Comparison

9) K-Means PCA vs RP vs Geographic: Figure 14 shows the three different clustering methods: PCA projection (left), random projection (middle), and K-Means (right). In all three plots, there are no clear, distinct clusters that can be formed. The two clusters (C0 and C1) are heavily intermixed, with no sharp visual separation between them. While the PCA plot shows that a few C0 points are slightly apart along the PC1 axis, there remains a substantial overlap between the two clusters. The random projection also similarly fails to uncover clear boundaries, with clusters distributed throughout the space. The geographic plot further reinforces this finding, with both clusters scattered throughout the physical landscape with no dominant spatial grouping or contiguous hotspot region for either cluster. The consistent lack of separation in all projections strongly suggests that the chosen features and K-Means clustering do not reveal natural or distinct groupings within the data. There is little evidence for well-defined outbreak hotspots, and any apparent clusters may be artefacts of the algorithm rather than meaningful patterns.

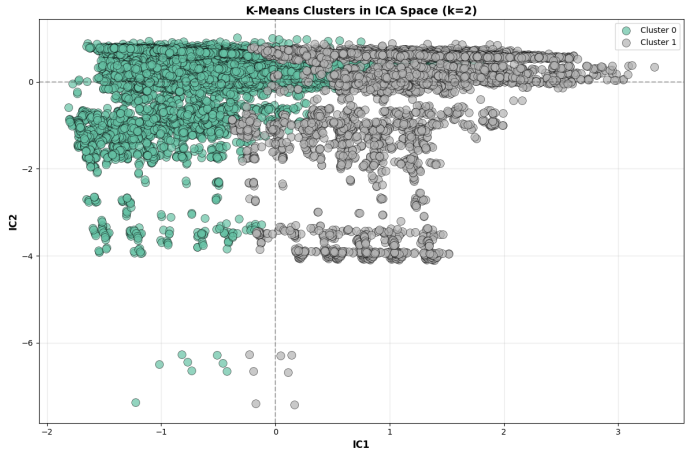

Fig. 15. K-Means Clusters in ICA Space

10) K-Means ICA: Figure 15 shows K-Means Clusters in ICA space. This plot shows how K-Means clustering divides the data using independent component axes (IC1 and IC2) instead of principal components. The two clusters are primarily separated vertically along IC2, with green (Cluster 0) dominating lower IC2 values and grey (Cluster 1) higher ones. Here, the split suggests that certain independently varying patterns in the data strongly drive the cluster assignments. Still, the clusters overlap horizontally and across layers, meaning the division is clearer along specific features but not absolute, indicating moderate separability with some mixing between groups.

V. Discussion

A. Model Performance

The EM clustering model demonstrated strong overall performance when applied to the spatial-temporal dengue data set. From the results obtained, the model converged smoothly after 96 iterations, indicating that the algorithm has reached a stable solution [3]. The use of the full covariance matrices also allows each cluster to be according to the structure of the data used, which then allows the algorithm to capture the correlation between the latitude, longitude and date from the data sets used. The flexibility of the model allowing each cluster to take its own unique shape and orientation, is suitable for epidemiological data since outbreaks of diseases generally do not form neat and spherical clusters [9]. From the results obtained, it can also be observed that there is a high log-likelihood and low AIC and BIC scores means that there is an effective balance between the fit of the model and its complexity [10]. The high assignment confidence of 98.29 per cent also means that it can determine the cluster's membership with certainty even in areas where there are overlapping spatial patterns.

Overall, the EM model performed well on the spatial-temporal dengue data set. It was effective in identifying meaningful spatial-temporal patterns and potential dengue hotspots.

The K-Means clustering model showcased moderate performances when applied across the dengue data, demonstrating mild separation of clusters in determining spatial-temporal groupings with different dengue case intensities. The application of the elbow method, followed by the analysis of the silhouette score, led to the decision in k = 2 as the optimal number of clusters. The value of clusters had a resulting silhouette score of 0.526, which indicated moderate overlap among the clusters.

Upon the application of PCA and Random Projection to the K-Means model, clusters showed partial separation, indicating the existence of differing intensities of dengue cases. However, visualisation of the geographical plot revealed a lack of partitioning between clusters. The clusters were interspersed instead of being geographically isolated, exposing the limitations of the K-Means model to fully break down the complexity of the dataset's non-linear spatial spread.

This limitation results from the assumption of spherical, isotropic clusters of the same size, rendering this model to be compatible for this dataset, where the distribution of dengue cases are rarely in regular, spherical patterns, but are instead in spatially skewed patterns [11]. Subzones of varying intensities would hence have been placed into the same cluster, caused by the inflexible centroid boundaries.

While the K-Means model could identify general differences of dengue intensity, it displayed difficulty in translating these identifications into clear spatial-temporal boundaries. Overall, the EM model outperformed the K-Means model, suggesting

that in this context of dengue cases, patterns are better analysed through probabilistic distinguishments than distance-based.

B. Relation to Hypothesis

The clustering results obtained by the EM model supported the hypothesis that certain spatial–temporal clusters correspond to higher dengue case intensities and therefore represent potential outbreak hotspots. The model consistently grouped together dense case areas and the dates when the cases were high to form clusters that highlighted areas with a higher risk of dengue transmission. The pattern is epidemiologically meaningful since it is consistent with real dengue outbreaks, which usually spreads gradually from one area to neighbouring areas rather than being isolated as one small cluster. Moreover, the clusters generated by the EM model also align with the behaviour of the spreading as it highlights the areas and time period where the dengue cases are higher and more concentrated. The high assignment confidence increases the credibility of these identified hotspots, suggesting that the clusters are reliably formed.

In contrast, the K-Means model generated more diffuse and overlapping clusters. Despite its ability to distinguish intensity patterns, its inability to form clear spatial boundaries, as well as its moderate silhouette scores showed that this model was only able to partially capture the heterogeneity of dengue transmission, which consequently led to unreliability of results and thus only partial support towards the hypothesis.

The EM model has demonstrated stronger statistical and spatial performance compared to the K-Means model, validating the hypothesis to a higher degree that dengue spacial-temporal clusters correspond to higher case intensities, indicating potential outbreak hotspots.

C. Potential Improvements for EM Model

The overlapping nature of the clusters causes undefined boundaries. This is due to the EM model's nature to assign probabilities rather than hard boundaries, thus making it difficult to identify where one cluster ends and begins. This causes hotspot borders for dengue to be harder to identify due to the lack of distinction of the boundaries [12].

Another limitation would be that the EM model assumes that each cluster follows a Gaussian Shape. However, realistically, dengue outbreaks can show skewed distribution or sudden spikes which do not fit the Gaussian Curve. Hence, the complex outbreak patterns are oversimplified, causing an abrupt increase in case numbers to not get captured accurately [13].

To improve the definition of the boundaries, we can combine the EM model with spatial post-processing [14]. Along with the flexibility of the EM model, which handles the overlapping clusters, the combination of spatial post-processing helps to provide a more interpretable cluster border to better identify outbreaks.

Another improvement would be to use alternative distributions such as the skew-t distribution, which can handle irregular and skewed clustering shapes that do not follow the Gaussian Distribution. This allows for the model to capture sudden spikes or asymmetrical outbreak patterns in realistic dengue data.

Before the implementation of potential improvements, however, the consideration of additional computational costs should be taken into account. The introduction of spatial post-processing or alternative distributions could potentially increase model complexity, negatively impacting training durations and computational costs. Despite the enhanced performance the model could potentially gain, a balance must be maintained between model precision and computational cost.

VI. Conclusion

This project compared EM and K-Means models, alongside PCA, ICA and RP, to identify spatial-temporal dengue outbreak patterns across subzones. Both models segmented the dataset into clusters representing differing dengue intensities, but the application of the EM model was superior in performance. Limitations of the K-Means model were clearly observed through diffuse and overlapping clusters, indicating its inability to analyse the irregular spatial distribution within the dataset. EM's approach through probability was significantly better in modelling the dataset, achieving clearer separations of clusters that were statistically robust and geographically meaningful as well.

While the EM model may have displayed stronger performance in comparison to K-Means, this improvement came with a higher computational cost and training duration, due to the computationally complex nature of EM compared to K-Means [15]. Conversely, K-Means was computationally efficient. While it may not have been suitable for analysing this particular dataset, it remains a strong option when rapid analysis of simpler datasets is required.

Evaluating the hypothesis that spatial-temporal clusters correspond to higher case intensities, indicating potential outbreak hotspots, results of K-Means showed partial support, while EM displayed strong empirical validation. The K-Means model only managed to achieve a general differentiation of intensity levels, while EM more successfully identified hotspot clusters that were consistent with known outbreak patterns.

All these findings have thus underlined the importance of selecting the right models depending on the structure of datasets, while also keeping computational efficiency into consideration, emphasising on EM model's superiority in analysing complex epidemiological datasets where statistical, as well as spatial boundaries, are overlapping. To conclude, this project has provided fundamental insights into the application of unsupervised learning for dengue hotspot detection, yielding benefits to data-driven disease surveillance as well as potential intervention and prevention strategies.

REFERENCES

[1] "Data normalization explained: The complete guide — Splunk," *Splunk*. https://www.splunk.com/en_us/blog/learn/data-normalization.html

[2] M. Haugh, *The EM algorithm*. 2015. [Online]. Available: https://www.columbia.edu/~mh2078/MachineLearningORFE/EM_Algorithm.pdf

[3] GeeksforGeeks, "ExpectationMaximization Algorithm ML," *GeeksforGeeks*, Sep. 08, 2025. https://www.geeksforgeeks.org/machine-learning/ml-expectation-maximization-algorithm/

[4] T. Pawar, "Gaussian Mixture Models Explained: Applying GMM and EM for Effective Data Clustering," Medium, May 07, 2024. https://medium.com/@tejaspawar21/gaussian-mixture-models-explained-applying-gmm-and-em-for-effective-data-clustering-ca24f8911609

[5] E. Kavlakoglu and V. Winland, "K-Means Clustering," What is k-means clustering?, Oct. 21, 2025. https://www.ibm.com/think/topics/k-means-clustering

[6] A. Tomar, "Elbow Method in K-Means Clustering: Definition, Drawbacks, vs. Silhouette Score," Built In, Mar. 13, 2025. https://builtin.com/data-science/elbow-method

[7] IBM, "What is PCA," IBM, Oct. 21, 2025. https://www.ibm.com/think/topics/principal-component-analysis

[8] E. Bingham and H. Mannila, Random Projection in Dimensionality Reduction: Applications to Image and Text Data, pp. 245–250, Aug. 2001, doi: 10.1145/502512.502546.

[9] D. Ruiz-Moreno, M. Pascual, M. Emch, and M. Yunus, "Spatial clustering in the spatio-temporal dynamics of endemic cholera - BMC infectious diseases," BioMed Central, https://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-10-51

[10] (gaussian) mixture models and the expectation maximization algorithm, https://www.cse.chalmers.se/research/lab/mlcourse/lecture_10_em_gmm_2018.pdf

[11] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little, "What to do when K-means clustering fails: A simple yet principled alternative algorithm," PloS one, https://pmc.ncbi.nlm.nih.gov/articles/PMC5036949/

[12] F. L. P. Oliveira, A. L. F. Cançado, G. de Souza, G. J. P. Moreira, and M. Kulldorff, "Border Analysis for spatial clusters," International journal of health geographics, https://pmc.ncbi.nlm.nih.gov/articles/PMC5816564/

[13] S. D. Tomarchio, L. Bagnato, and A. Punzo, "Model-based clustering using a new multivariate skew distribution - advances in data analysis and classification," SpringerLink, https://link.springer.com/article/10.1007/s11634-023-00552-8

[14] "Bulent Sankur: Bulent Sankur," Bulent Sankur — Bulent Sankur, https://academics.boun.edu.tr/bulent.sankur/

[15] P. Rk, "Comparing the EM algorithm and K-means clustering," Medium, https://medium.com/@prithivrk7/comparing-the-em-algorithm-and-k-means-clustering-629b7c80a003