

# BG4104 Group Assignment 2

Group 1

## 1 Introduction

Maternal mortality refers to the death of a woman during pregnancy, childbirth, or within 42 days after the pregnancy due to related complications [1]. It remains a major challenge in Southeast Asia (SEA), where large inequalities persist across countries [2]. The Maternal Mortality Ratio (MMR) is the standard indicator for tracking these outcomes [1]. Globally, an estimated 260,000 women died from pregnancy-related causes in 2023, many of which were preventable with quality care [2, 3]. This report hypothesises that some SEA countries are systematically under-capturing maternal deaths due to weak vital registration systems and limited measurement capacity. As a result, officially reported MMRs may appear lower than the true burden, requiring organisations such as the UN to rely on statistical modelling to produce more accurate estimates. Under-reporting has serious consequences: maternal deaths lead to long-term social and economic impacts such as increased newborn mortality, higher household health costs, and reduced human capital [7]. At the national level, inaccurate MMR data impede effective policymaking and progress toward SDG 3.1 [2]. Therefore, this report examines the gap between observed maternal mortality figures and modelled estimates, arguing that weaknesses in measurement and registration capacity could contribute to this discrepancy.

## 2 Methodology

### 2.1 Data Set Description

The key variables in this dataset are `env_mat` (recorded maternal deaths), `truemat_vr` (verified maternal deaths from vital registration), `obs_matdeaths` (observed maternal death counts), `final_pm` (proportion of maternal deaths) and `final_mmr` (MMR per 100,000 live births). Supplementary fields such as `citation_short` and `citation_long` provide bibliographic details of data sources. Indicators like `complete_vr` and `incomplete_vr_multiplier` reflect data completeness and adjustment factors for underreporting. Additional variables such as `icd_utilized`, `study_period_coverage` and `check_outside_of_vr` capture classification methods and the temporal scope of each dataset entry.

Dataset indicators of quality assurance, such as `include`, `usability_percentage` and `include_reason`, determine whether data points were incorporated into the final analyses. Within the dataset, core identifiers such as country code and year range are fully populated; however, some derived fields contain missing values that may be due to differences in data reporting and health system capabilities across regions and years.

Abbreviations used include ISO (International Organisation for Standardisation), VR (Vital Registration), PM (Proportion of Maternal deaths), ENV (Envelope, referring to total deaths used as model inputs) and ICD (International Classification of Diseases). Model identifiers such as BMI, BMAT and BMIS refer to baseline maternal mortality estimation models, while RHO denotes correlation coefficients used to assess reliability and data calibration.

### 2.2 Data Preprocessing

#### 2.2.1 Data Cleaning and Preparation

The dataset was first filtered to isolate only the countries in SEA using their ISO3 codes. From the full dataset, rows where the `iso_alpha_3_code` matched SEA country codes (e.g. `["MYS", "SGP", "THA"]`) were retained, ensuring regional focus and enabling country-level analyses relevant to local health system challenges. After selection, columns with large amounts of missing data were removed to prevent skewed or unreliable imputation from affecting later analyses.

#### 2.2.2 Preparing the Data for Modelling

Columns with more than 150 missing values were identified using `.isnull().sum()` and removed with `.drop(columns=...)`. These columns included redundant or non-informative features, such as duplicate year columns, full country names and citation information. This enabled the preparation of the final data matrix for clustering. Row counts before and after each filtering step were checked to track the impact of these removals.

The dataset was then filtered to retain only numeric columns using `select_dtypes(include='number')`, as clustering, Principal Component Analysis (PCA) or Independent Component Analysis (ICA) and correlation methods required numeric inputs. Remaining rows containing missing numeric values were subsequently dropped using `dropna()`, resulting in a fully complete numeric matrix. Following this, all numeric features were standardised using the `StandardScaler` from `scikit-learn`, transforming them into a zero mean and unit variance. This scaling step was essential as each variable differs substantially in magnitude and unscaled features could otherwise dominate and distort clustering outcomes.

Subsequently, exploratory checks were performed before the final clustering. Heatmaps and scatter-plots were generated to visually inspect the correlations between the features, identify outliers and any potential correlations. Lastly, to better visualise the patterns and test the robustness of the clusters, dimensionality reduction techniques PCA, ICA and Randomised Projections (RP) were applied. These methods helped compress the data and identify the

clearest, most distinct signals, allowing for a more confident identification of the trends in the health system reporting and quality.

### 3 Exploratory Data Analysis (EDA)

EDA of the SEA maternal dataset revealed correlations impacting MMR. Countries with more observed maternal deaths compared to total deaths among women of reproductive age 15–49 reflected positive correlations. This demonstrates the possibility of countries having higher levels of risk and mortality for women due to systemic health and societal vulnerabilities.

More intriguingly, the relationship between usability percentage and MMR was strongly negative, supporting the idea that system quality is closely tied with reduced maternal risk. Here, the usability percentage served as a proxy for the maturity and transparency of a country's health reporting. Countries scoring higher on usability showed lower MMRs, suggesting that vital registration systems not only reflect better data quality but also link to actual improvements in health outcomes.

Finally, the total number of deaths in relation to live births indicated positive correlations, consistent with population size effects. Comparing these axes, we can identify whether a country's maternal death burden scales simply with population or whether particular systems are either outperforming or underperforming given their demographic context. Unexpectedly, some populous countries, despite being burdened by sheer scale, still demonstrate relatively strong performance once registry quality and intervention factors are considered.

## 4 Machine Learning Algorithms

### 4.1 Expectation Maximisation (EM)

EM is a common algorithm used to fit a GMM. EM is an iterative optimisation method used to estimate the parameters in probabilistic models involving hidden or missing variables. EM aims to maximise the likelihood of the observed data by alternating between 2 key steps until convergence. The Expectation Step, which estimates the probability of each data point belonging to each Gaussian component based on current parameters and the Maximisation Step, which updates the model parameters by maximising the expected complete-data log-likelihood [8]. This will produce a new parameter estimate which is guaranteed to increase or maintain the data likelihood.

The GMM employs a probabilistic model to represent data as a combination of several Gaussian distributions, each with its own mean and variance values. It is also weighted by mixing coefficients that indicate the influence of each distribution within the entire dataset. The GMM is val-

able in clustering and density estimation as it models complex multimodal datasets where data points group around different centers rather than just one [9].

### 4.2 K-Means Clustering

K-Means clustering, an unsupervised learning algorithm, partitions a dataset into  $k$  clusters by minimising the within-cluster sum of squared distances to each centroid. The algorithm iteratively assigns each data point to the nearest centroid and updates centroids as the mean of their respective clusters until assignments stabilise. This method is efficiently used during segmentation and compression; however, it is highly susceptible to centroid initialisation, often assuming roughly spherical and similarly sized clusters under the square Euclidean distance and converges to local optima. However, running the algorithm multiple times with different initialisations can help improve results [10].

### 4.3 Randomised Projections (RP)

RP is a dimensionality reduction technique that maps high-dimensional data into a lower-dimensional subspace using a random matrix. This method effectively preserves pairwise distances between data points while reducing computational costs. Unlike PCA, it does not rely on complex matrix decompositions or prior knowledge of data distribution. RP are particularly useful for simplifying large datasets and speeding up machine learning workflows while maintaining the essential geometric structure [11].

### 4.4 Principal Component Analysis (PCA)

PCA is an unsupervised dimensionality reduction technique that transforms possibly correlated variables into a smaller set of uncorrelated principal components (PC). PC captures the largest variance in the data, allowing complex datasets to be represented in fewer dimensions with minimal information loss. PCA starts by mean-centring and standardising the data, then computes PCs as linear combinations of the original variables based on eigenvectors and eigenvalues from the covariance or correlation matrix. Projecting the data onto these top components simplifies analysis and preserves most of the original structure [12].

### 4.5 Independent Component Analysis (ICA)

ICA is an unsupervised technique that separates multivariate data into statistically independent and non-Gaussian components. It assumes observed data are linear mixtures of unknown source signals and aims to recover these sources by maximising their independence. Unlike PCA, which finds

uncorrelated directions of maximum variance, ICA identifies independent factors. This makes ICA effective for blind source separation, where the goal is to extract independent signals from mixed observations, such as distinguishing individual voices in audio or isolating noise from biomedical data [13].

## 5 Results and Discussion

### 5.1 Clustering Hypothesis

K-Means is better suited than the Gaussian Mixture Model (GMM) for clustering MMR data in SEA because of its simpler and more direct methodological analysis. It partitions data into hard-assigned, spherical clusters purely based on distance to centroids, resulting in clear boundaries that are easy to interpret. K-Means excels in situations where distance-based clusterings correspond to real differences, producing easily interpretable, sharply bounded groups that reflect discrete system levels, such as countries with strong, transitional, or developing maternal health systems.

In contrast, while GMM offers probabilistic assignments and can capture uncertainty or overlap, it often merges moderate and low-risk countries, blurring important boundaries and complicating interpretation. With its efficiency, stability, and superior alignment with the natural grouping of this MMR dataset, K-Means provides more practical and meaningful clusters for regional health analysis.

### 5.2 K-Means Clustering

The algorithm was implemented using the `KMeans` class from `sklearn.cluster`. It critically involves determining an appropriate number of clusters,  $K$ . The Elbow Method was employed for this purpose.

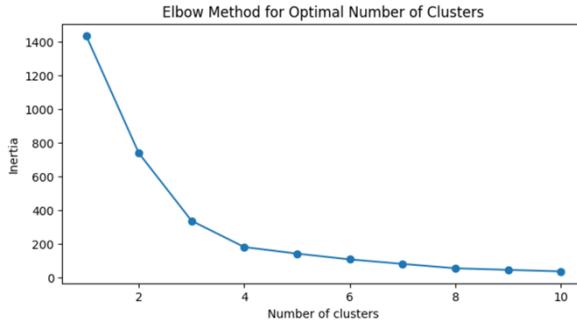


Figure 1: Elbow Method for Optimal Number of Clusters in K-Means Cluster Evaluation

The Elbow Method helps determine the optimal number of clusters for K-Means by plotting inertia, the sum of squared distances between data points and their nearest cluster centroid, against different values of  $K$ . Inertia naturally decreases as the number of clusters increases, but slows beyond a

certain point. Based on Figure ??,  $K = 3$  clusters were identified as the most suitable for this dataset.

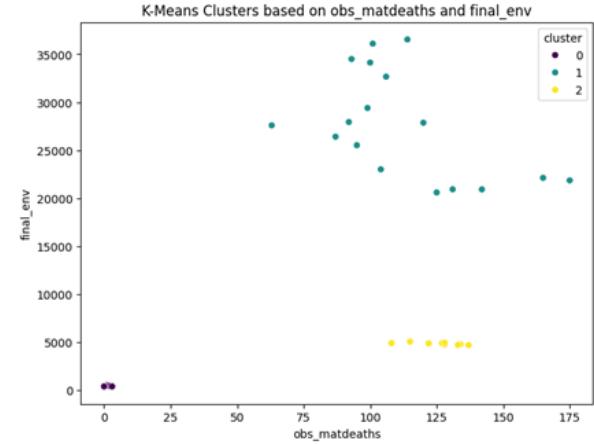


Figure 2: K-Means Cluster of `obs_matdeaths` vs `final_env`

Figure 2 compares `obs_matdeaths` with the estimated total deaths of women of reproductive age (`final_env`), revealing three distinct clusters representing countries with varying mortality profiles. Cluster 0 includes countries with very low observed maternal deaths and low overall female mortality, reflecting strong vital registration systems and effective maternal healthcare. Cluster 2 represents countries with moderate to high observed maternal deaths but lower total female deaths, indicating maternal mortality constitutes a significant share of reproductive-age mortality. Cluster 1 captures countries exhibiting both high maternal deaths and high total mortality, indicating broader systemic health challenges.

### 5.3 Gaussian Mixture Models (GMM)

SEA countries are represented in respective datasets as probabilities, allowing smooth or overlapping boundaries while capturing uncertainty in real system transitions. They are further assigned to probable cluster memberships (E-step), where their cluster characteristics are updated to fit the data better (M-step).

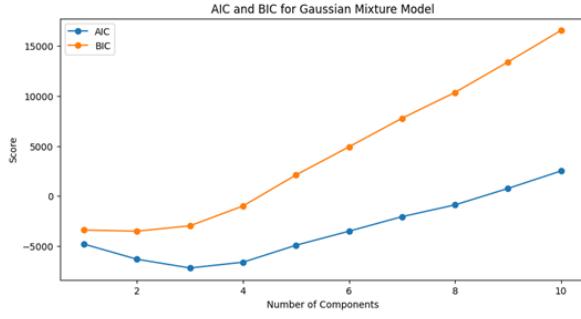


Figure 3: AIC and BIC for GMM to Determine the Number of Components

In Figure 3, the fit of GMM models is evaluated using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) with different numbers of clusters and components. This compares the balance between the model that fits the data and the model's complexity to avoid overfitting. Both AIC and BIC reached their minima at two clusters, with BIC being more conservative at 2 components or clusters.

This indicates the strongest structure in the dataset is split between high-performing health systems and those with significant challenges. The binary split indicates that SEA countries have either developed registries or face major challenges, with few truly transitional states.

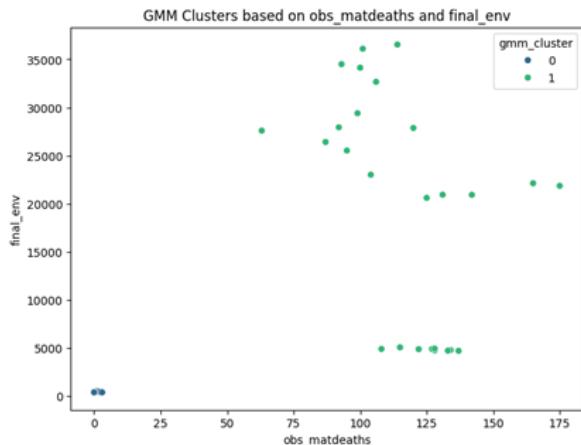


Figure 4: GMM Cluster of `obs_matdeaths` vs `final_env`

Comparing `obs_matdeaths` and total estimated deaths worldwide (`final_env`), Cluster 0 comprises countries whose reporting aligns with global models of an effective healthcare system, while Cluster 1 indicates higher observed and estimated deaths or larger discrepancies. GMM's soft approach allows it to capture the distinction of countries that excel in one dimension but lag in another, reflecting the reality that improvements in either health or data systems often proceed at different rates and

flexible cluster membership better captures these transitional states.

#### 5.4 K-Means vs GMM

Across all three plots, K-Means produced sharper and more defined cluster boundaries, while GMM consistently merged low and moderate mortality cases into broader groups. In the cluster visualisations, GMM's two-cluster structure reflects a simpler split between well-functioning systems and challenged systems, showing gradual transitions and overlapping regions. K-Means, by contrast, offers more granular separation and captures countries in intermediate or transitional stages, revealing clearer stepwise progressions in health system development, such as distinctions among small, medium and large population structures.

The data confirms the hypothesis that K-Means would be better suited than GMM for clustering MMR data, as it produced sharper and more interpretable system tiers. K-Means aligns closely with the three-level structure that will be used throughout the report and is taken as the primary clustering model. However, GMM also plays a supporting role in illustrating uncertainty and overlaps between systems.

#### 5.5 Dimensionality Reduction Hypothesis

Among linear dimensionality-reduction (LDR) algorithms, ICA is hypothesised to be the most effective in analysing maternal mortality data as it identifies statistically independent components, allowing for effective separation of hidden factors that variance-based methods, such as PCA, or distance-preserving techniques like RP might blend. While PCA is valuable for identifying major axes of variation, such as MMR or data usability, its reliance on variance suggests a likelihood of missing or blending subtler, independent sources of variation that may influence complex outcomes. RP is fast and preserves relative distances for computational efficiency, but it does not offer interpretability or insight into true drivers of variation. Comparatively, ICA's capacity to recover independent signals enables clearer, more meaningful clustering of SEA countries with mixed or outlier profiles, revealing influences masked by other methods, hence allowing ICA to have an edge over PCA and RP's analysis.

##### 5.5.1 PCA

After performing PCA, a new dataframe containing PC1 and PC2 was created and used as input for both the K-Means and GMM clustering algorithms. Based on the Elbow Method, K-Means performed best with  $K = 3$  clusters, while the AIC and BIC evaluations for the GMM model indicated that 8 components were optimal.

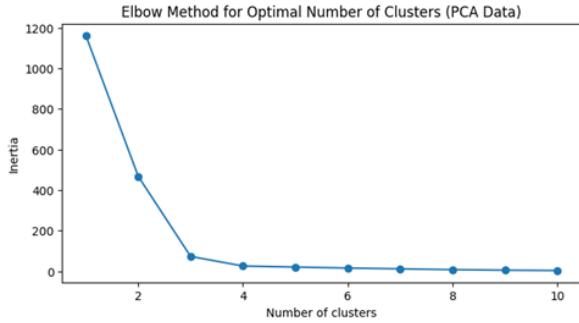


Figure 5: Elbow Method for Optimal Number of Clusters after PCA

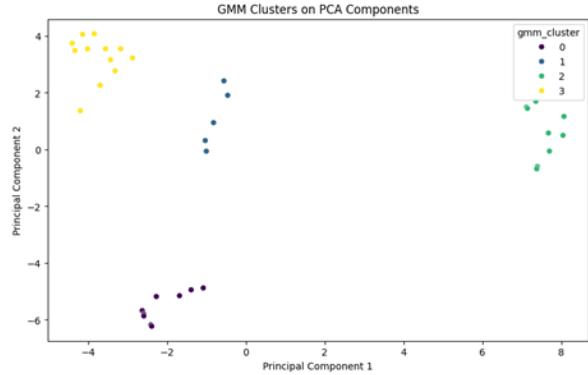


Figure 8: GMM Cluster on PCA Components

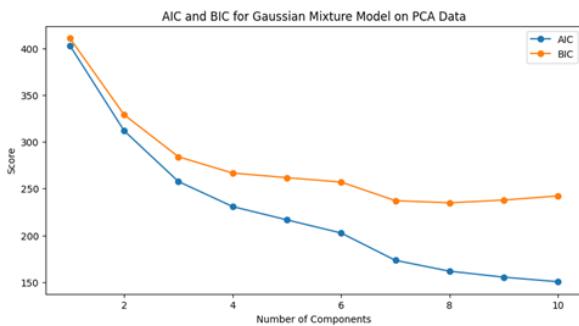


Figure 6: AIC and BIC for GMM after PCA

These components preserved the key structural patterns of the dataset, with PC1 capturing overall system quality and PC2 revealing additional underlying trends. The explained variance values further show that PCA successfully reduced dimensionality while maintaining the dataset's essential information, making it effective for clustering and preserving its global structure.

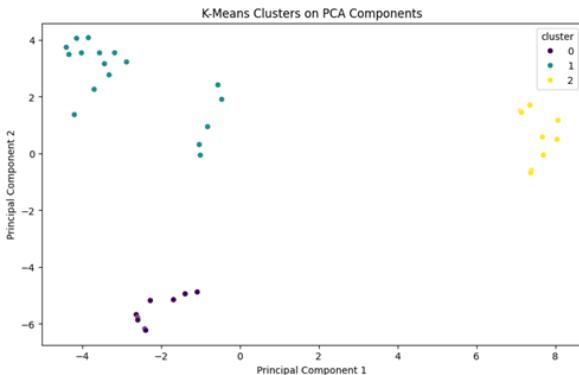


Figure 7: K-Means Cluster on PCA Components

### 5.5.2 ICA

Using ICA alongside both clustering models, the Elbow curve after ICA indicates a sharp drop at 3 clusters for K-Means, suggesting that additional clusters offer little improvement in explaining variance. The AIC and BIC curves for GMM showed that both decrease as the number of components increases, with the BIC minimum value at 6 clusters and AIC further decreasing, suggesting slightly more granularity in the data's structure after ICA.

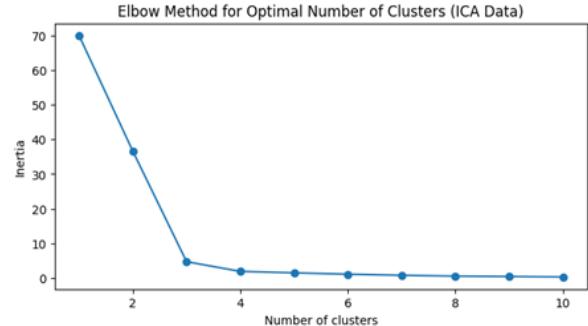


Figure 9: Elbow Method for Optimal Number of Clusters after ICA

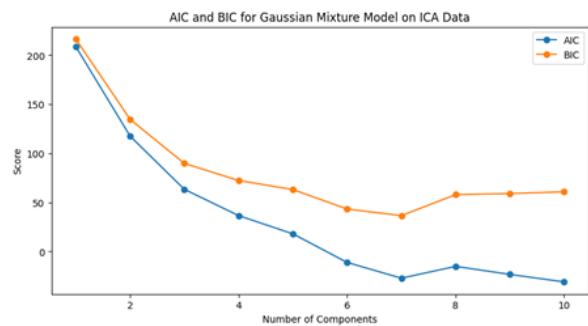


Figure 10: AIC and BIC for GMM after ICA

The cluster assignment plots using ICA components make these divisions visually distinct. The three compact and well-separated clusters in the

K-Means plot after ICA indicate the data's structure as highly distinguishable along the independent axes extracted by ICA. In contrast, GMM clustering after ICA yields four clusters, capturing subtle groupings or transitional subpopulations that K-Means might miscategorise.

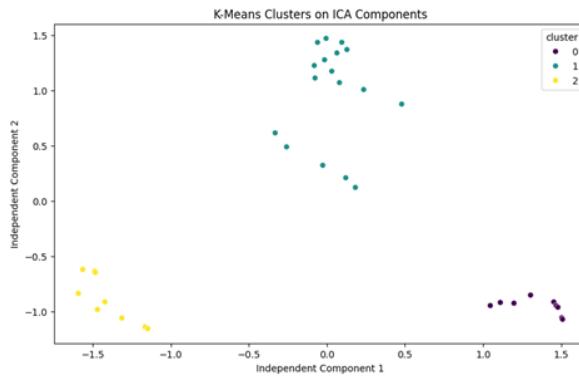


Figure 11: K-Means Cluster on ICA Components

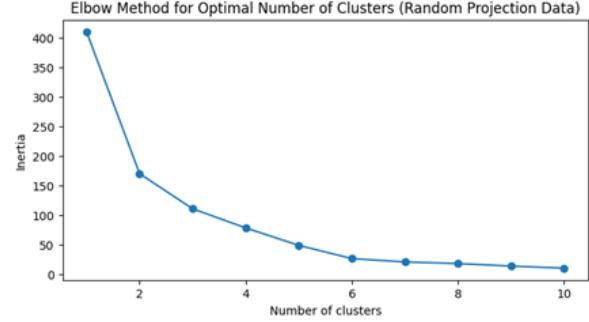


Figure 13: Elbow Method for K-Means after RP

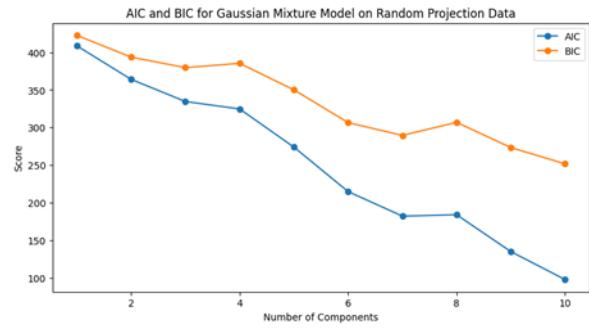


Figure 14: AIC and BIC for GMM after RP

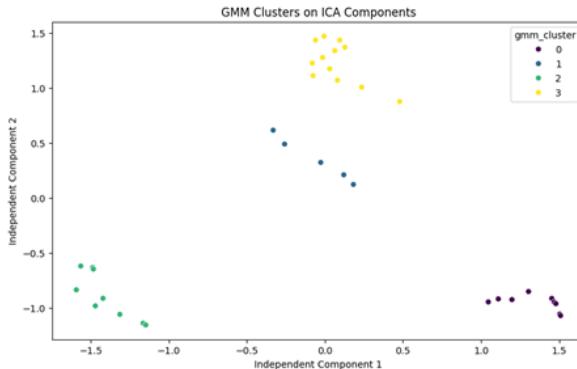


Figure 12: GMM Cluster on ICA Components

K-Means clustering on the RP components groups the countries into three broad clusters that align with high-performing, transitional and developing systems (Figure 15), showing that the projected data space still reflects the main structure of the maternal mortality data.

GMM clustering on the same RP components instead yields four clusters, but the separation is not pure. Clusters 1, 2 and 3 sit close together with a small degree of overlap, suggesting that the extra GMM clusters mostly capture small variations within the intermediate group rather than clear distinctions.

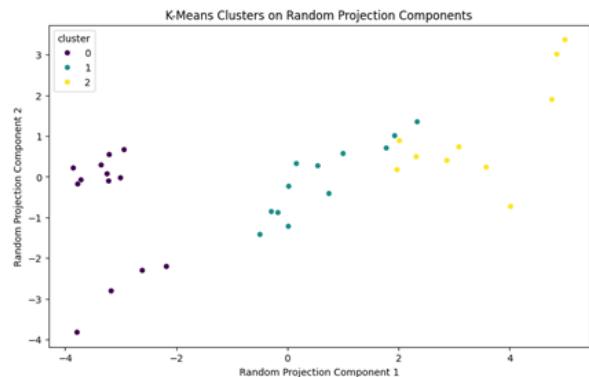


Figure 15: K-Means Cluster on RP Components

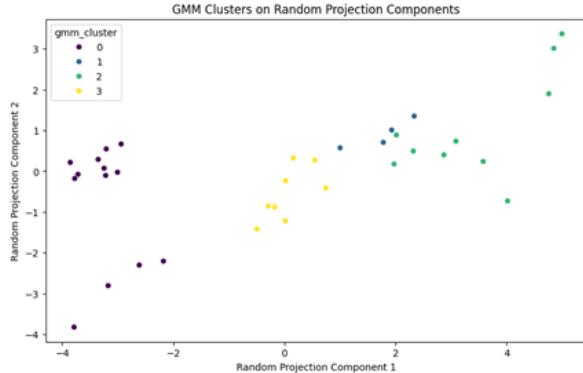


Figure 16: GMM Clusters on RP Components

## 5.6 PCA vs ICA vs RP

Across the dimensionality-reduction approaches applied, PCA, ICA, and Gaussian Random Projections (GRP) each offered complementary insights into the maternal mortality dataset.

PCA captures the dominant axes of variation, which are primarily usability and MMR. PCA produced clear and interpretable clusters when broad, correlated differences existed across countries. ICA, in contrast, identified statistically independent components, exposing hidden factors that influenced health system performance and reporting quality. This allowed ICA to separate countries with mixed or atypical profiles more effectively, revealing structure that PCA's variance-driven approach could overlook.

GRP added a further layer of validation by compressing the dataset from (35, 41) to (35, 3) while still preserving the overall cluster structure. The consistency of cluster patterns before and after projection showed that the core relationships in the data were strong enough to withstand dimensionality reduction.

The data confirms the hypothesis that ICA would be the most effective dimensionality reduction method for this dataset, as independent components better capture the distinct drivers of maternal mortality and reporting quality in SEA. ICA not only preserved the broad trends highlighted by PCA and GRP, but also exposed additional separation between transitional and outlier systems that the other methods blurred, providing the most informative and discriminative representation for downstream clustering and interpretation.

## 6 Conclusion

The clustering analysis revealed three main groups in the K-Means results as follows: a high-performance cluster (SGP, MYS) which is characterised by the lowest MMR, high usability percentages and minimal discrepancies between observed and estimated deaths; a transitional cluster

(THA, VNM) with moderate mortality, medium usability and moderate reporting discrepancies; and a developing-systems cluster consisting of the remaining SEA nations with the highest mortality, lowest usability and largest discrepancies.

The GMM model produced two broader clusters as follows: the advanced systems group, combining both high and transitional performers (SGP, MYS, THA) with stronger data quality and more consistent reporting; and the developing systems group, comprising all other SEA countries, marked by higher mortality and weaker data systems. Key findings revealed strong links between system quality and reporting accuracy, with usability percentage emerging as a strong predictor of data reliability.

K-Means was more effective at identifying distinct stages of development, whereas GMM better captured gradual transitions, though both methods agreed on the strongest and weakest performers.

Dimensionality-reduction methods showed that these cluster patterns were robust to changes in the feature space. ICA provided the clearest separation of transitional and outlier systems, while PCA and GRP confirmed that the main structure of the data persisted even in low-dimensional representations.

Overall, the analysis highlights a clear need for targeted improvements in developing systems, staged progression toward advanced systems, and a stronger emphasis on data quality. Policy priorities include reducing ill-defined deaths, increasing usability, and strengthening vital registration. Future work should combine hard and soft clustering with dimensionality reduction to monitor progress over time and to support practical, data-driven strategies for reducing maternal mortality across the region.

## 7 References

- [1] “NVSS - Maternal mortality - FAQ,” Centers for Disease Control and Prevention, <https://www.cdc.gov/nchs/maternal-mortality/faq.htm> (accessed Nov. 13, 2025).
- [2] “Maternal mortality,” World Health Organization, <https://www.who.int/news-room/fact-sheets/detail/maternal-mortality> (accessed Nov. 16, 2025).
- [3] “Trends in maternal mortality 2000 to 2023: Estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division,” World Health Organization, <https://www.who.int/publications/i/item/9789240108462> (accessed Nov. 13, 2025).
- [4] “Trends in maternal mortality 2000 to 2023: Estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division,” World Health Organization, <https://www.who.int/publications/i/item/9789240108462> (accessed Nov. 16, 2025).
- [5] “[maternal and newborn] - mortality/causes of death,” World Health Organization, <https://www.who.int/data/gho/data/themes/topics/topic-details/mca/maternal-and-newborn—mortality-causes-of-death> (accessed Nov. 11, 2025).
- [6] MSD for mothers in Asia, [https://www.msdfornmothers.com/docs/MSD\\_for\\_Mothers\\_in\\_Asia\\_Pacific.pdf](https://www.msdfornmothers.com/docs/MSD_for_Mothers_in_Asia_Pacific.pdf) (accessed Nov. 8, 2025).
- [7] Trends in maternal mortality estimates 2000 to 2023, <https://www.unfpa.org/sites/default/files/pdf/9789240108462-eng.pdf> (accessed Nov. 8, 2025).
- [8] “Expectation-maximization algorithm - ml,” GeeksforGeeks, <https://www.geeksforgeeks.org/machine-learning/ml-expectation-maximization-algorithm/> (accessed Nov. 8, 2025).
- [9] “Expectation-maximization algorithm - ml,” GeeksforGeeks, <https://www.geeksforgeeks.org/machine-learning/ml-expectation-maximization-algorithm/> (accessed Nov. 8, 2025).
- [10] E. Kavakoglu and V. Winland, “What is K-means clustering?,” IBM, <https://www.ibm.com/think/topics/k-means-clustering> (accessed Nov. 15, 2025).
- [11] E. Bingham and H. Mannila, Random projection in dimensionality reduction: Applications to image and text data, <https://cs-people.bu.edu/evimaria/cs565/kdd-rp.pdf> (accessed Nov. 15, 2025).
- [12] “What is Principal Component Analysis (PCA)?,” IBM, <https://www.ibm.com/think/topics/principal-component-analysis> (accessed Nov. 8, 2025).
- [13] “Independent component analysis - ML,” GeeksforGeeks, <https://www.geeksforgeeks.org/machine-learning/ml-independent-component-analysis/> (accessed Nov. 9, 2025).

## 8 Appendix

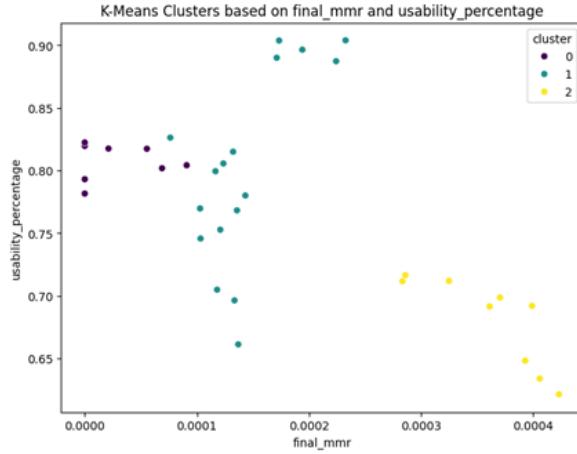


Figure 17: K-Means Cluster of `final_mmr` vs `usability_percentage`

Figure 17 shows a negative correlation between MMR `final_mmr` and data usability `usability_percentage`, indicating that higher maternal mortality may be associated with reduced data quality. Cluster 0 represents countries with low MMR and high usability, while Cluster 1 includes countries with moderate MMR and variable data quality and Cluster 2 consists of countries with high MMR and lower usability. This pattern illustrates that poorer registry quality is associated with maternal mortality.

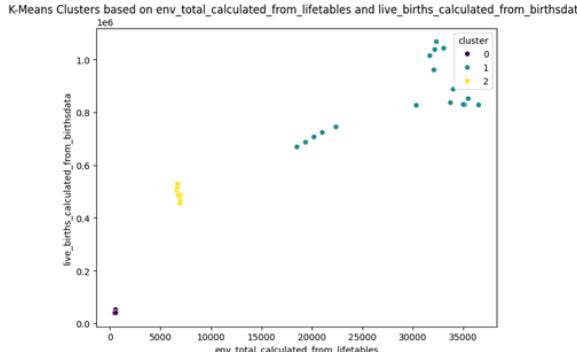


Figure 18: K-Means Cluster of `env_total_calculated_from_lifetable` vs `live_births_calculated_from_birthsdata`

Figure B shows a positive correlation between estimated deaths of women aged 15 to 49 derived from total deaths `env_total_calculated_from_lifetable` and estimated live births `live_births_calculated_from_birthsdata`, reflecting demographic expectations that countries with larger populations experience higher numbers of both live births and deaths among women of

reproductive age.

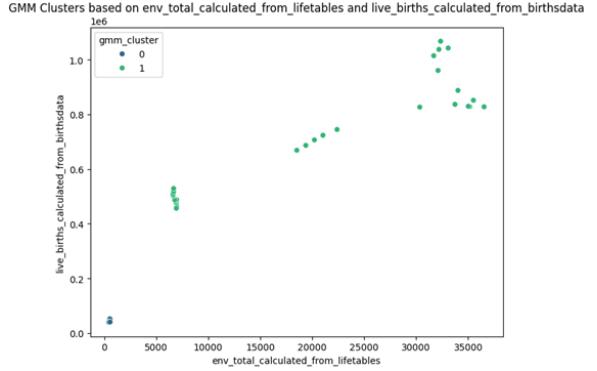


Figure 19: GMM Cluster of `env_total_calculated_from_lifetable` vs `live_births_calculated_from_birthsdata`

Comparing the data in Figure C, we identified that the smaller, high-performing countries group is separate from the larger, more challenged ones. This approach allows countries in transition or with mixed characteristics to share correlations between the two groups. This suggests that demographic scale influences maternal health outcomes in SEA, where the two main clusters found by GMM correspond to plausible substantive differences in health system development and reporting effectiveness.

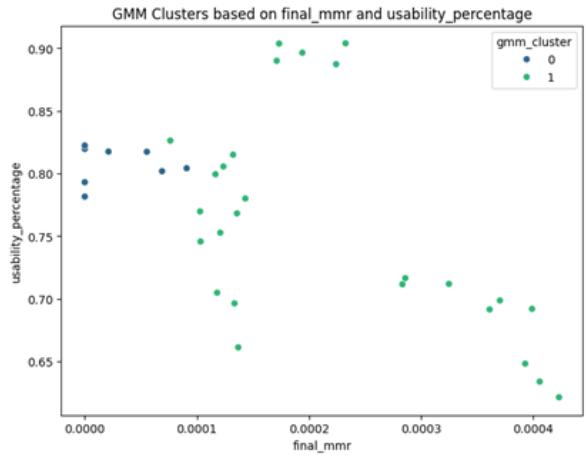


Figure 20: GMM Cluster of `final_mmr` vs `usability_percentage`

From Figure D, Cluster 0 gathers countries with low MMR and high usability, while Cluster 1 groups those with higher mortality and poorer usability. The absence of a middle cluster suggests that healthcare system transitions are rare or short-lived in this context, consistent with the tracking of maternal mortality by international agencies. Effective registration systems can significantly reduce deaths, with the binary split from GMM giving a clean regional overview.