Multivariate and Unsupervised Analysis of Bovine Tuberculosis Breakdowns

Data Cleaning, Clustering, and Pattern Discovery

Victory Chukwudi
Supervisor: Prof. Conor McAloon & Prof. Lennon Onaraigh

November 12, 2025

Data Description and Preparation

Background and Objective

The original dataset, titled `master_tb_victory_16_Oct_2025_encrypted.csv`, contained extensive records of bovine tuberculosis (bTB) testing data. Each record corresponds to an individual herd-level test, covering test results, herd type, county, trading status, and date information. This dataset originally comprised over 2.8 million observations and 107 variables.

However, the data were highly granular, with multiple test entries recorded for the same herd across different time periods within the same breakdown. A *breakdown* (denoted by `bd_no`) represents a continuous episode of infection and testing within a herd. Therefore, the first analytical objective was to transform the raw test-level data into a structured, aggregated dataset where each row represents one breakdown event for one herd.

This transformation reduced the dataset to approximately 85,210 rows and 40 variables, summarized into a new file: `breakdown_summary_advanced.csv`. The goal of this section was to clean, structure, and summarize the data for further statistical and clustering analyses.

Overview of the Processing Steps

The data processing and summarization were performed in R using several key libraries: `dplyr` for data manipulation, and `lubridate` for date parsing and time calculations. The key steps taken are outlined below.

Step 1: Loading the Data

```
master <- read.csv(
  "master_tb_victory_16_Oct_2025_encrypted.csv",
  stringsAsFactors = FALSE,
  colClasses = c(herd_no = "character")
```

```
)
```

The dataset was read into R with `herd_no` stored as a character variable to preserve leading zeros and avoid incorrect numeric conversion. This ensured herd identifiers remained consistent throughout the analysis.

Converting Date Columns

```
master <- master %>%
  mutate(
    starts = suppressWarnings(ymd(starts)),
    ends   = suppressWarnings(ymd(ends))
  )
```

Dates were originally provided in multiple formats (e.g., "01/02/2023", "2023-02-01"). To ensure consistency, all date fields were standardized using `lubridate::ymd()`, which correctly converts these into R date objects. This step allowed accurate computation of breakdown durations later in the process.

Creating a Combined Positive Count

```
pos_cols <- c(
  "total_reactor_skin","total_standard_reactor",
  "total_reactor_slaughter","total_reactor_slaughter_after_skin_test",
  "total_reactor_permit_lesions","severe_reactor",
  "cows_positive","bulls_positive","calves_positive",
  "heifers_positive","steers_positive"
)
master <- master %>%
  mutate(positives = rowSums(across(all_of(present)), na.rm = TRUE))
```

Several columns in the raw dataset recorded the number of animals that tested positive (reactors) under different test types. To simplify analysis, these were summed into one unified column called `positives`. For example, if a herd had:

- 2 reactors in the skin test,
- 1 in the slaughter test,
- 0 in other categories,

then the total for that record would be `positives = 3`.

This aggregation step condensed multiple positive indicators into a single interpretable variable.

Counting the Number of Tested Animals

```
master <- master %>%
  mutate(n_tested = ifelse(!is.na(total_animals), total_animals, NA_integer_))
```

The total number of animals tested during each testing event was extracted from the variable `total_animals`. If this value was missing, it was recorded as `NA`. This step helped quantify the test scale per herd.

Filtering for Breakdown-Related Tests

```
tests <- master %>%
  filter(!is.na(bd_no))
```

Not all test records in the dataset corresponded to active breakdowns. To ensure consistency, only records with a valid breakdown number (`bd_no`) were retained. This reduced the dataset to breakdown-relevant tests only.

Grouping Test Types into Families

```
tests <- tests %>%
  mutate(
    test_group = case_when(
      test_type %in% c("1") ~ "test1",
      test_type %in% c("5","5a","5b","5c","5d","5e","5f","5g","5h") ~ "test5_family",
      test_type %in% c("7","7a","7b") ~ "test7_family",
      ...
      TRUE ~ "other_tests"
    )
  )
```

The dataset contained multiple variations of similar test types (e.g., "5a", "5b"). To make the data more consistent, these were categorized into test families such as `test5_family` or `test7_family`. This grouping allowed easier aggregation of results by test category.

Summarizing Data at the Breakdown Level

```
breakdown_level <- tests %>%
  group_by(herd_no, bd_no) %>%
  summarise(
    herd_type = first(na.omit(herd_type)),
    county = first(na.omit(county)),
    start_date = min(starts, na.rm = TRUE),
    end_date   = max(ends, na.rm = TRUE),
    total_positive = sum(positives, na.rm = TRUE),
    total_tests = n(),
    duration_days = as.numeric(difftime(max(ends), min(starts), units = "days")),
    ...
  )
```

This step grouped all tests belonging to the same herd and breakdown number into one summarized record. Within each group, several calculations were made:

- **Start and End Dates:** Earliest and latest test dates within the breakdown.
- **Total Positives:** Sum of all positive reactors across all tests.
- **Total Tests:** Number of test events conducted.

- **Duration (Days):** Number of days between the first and last test.

Additionally, positive and tested counts were calculated per test type (`pos_test1`, `tested_test1`, etc.), providing a complete view of test outcomes by category.

Creating Herd-Level Summary Statistics

```
herd_summary <- breakdown_level %>%
  group_by(herd_no) %>%
  summarise(
    total_breakdowns = n(),
    completed_breakdowns = sum(!is.na(end_date)),
    ongoing_breakdowns = sum(is.na(end_date)),
    mean_days = mean(duration_days, na.rm = TRUE),
    median_days = median(duration_days, na.rm = TRUE),
    sd_days = sd(duration_days, na.rm = TRUE)
  )
```

Once breakdown-level summaries were computed, an additional layer of aggregation was added at the herd level. This allowed evaluation of:

- How many breakdowns each herd experienced in total.
- How many of those breakdowns were completed or ongoing.
- The average, median, and standard deviation of breakdown durations per herd.

These statistics later supported comparative analysis between herds and clusters.

Merging and Exporting the Final Dataset

Finally, the herd-level summaries were merged with the breakdown-level data, and the resulting dataset was exported:

```
write.csv(final_df, "breakdown_summary_advanced.csv", row.names = FALSE)
```

The final dataset contained 85,210 breakdowns and 40 descriptive variables, making it suitable for subsequent analytical stages such as PCA, FAMD, and clustering.

Principal Component Analysis (PCA) and K-Means Clustering

Data Sampling

The next analytical phase applied **Principal Component Analysis (PCA)** combined with **K-Means clustering** to identify natural groupings among herd breakdowns.

Due to hardware limitations, it was computationally infeasible to process all 85,000 records at once. To ensure model feasibility and efficient computation, a random sample of **5,000 breakdowns** was selected. This subset still preserved the statistical structure of the full dataset while making the PCA and clustering procedures manageable on a standard personal computer.

Feature Preparation

From the breakdown summary dataset, key numeric variables were selected — including:

- `total_positive`, `total_tests`, and `duration_days`,
- Positive test counts by test type (`pos_test1`–`pos_test10_family`),
- Positive rates per test (computed as pos_testX/tested_testX).

All missing values were replaced with zeros, and each variable was standardized (zero mean, unit variance) using the `scale()` function in R. This ensures that all variables contribute equally to the PCA and clustering steps.

Principal Component Analysis (PCA)

PCA was applied to reduce the dimensionality of the dataset and identify the most important underlying patterns. Figure 1 presents the scree plot of explained variance by principal component.

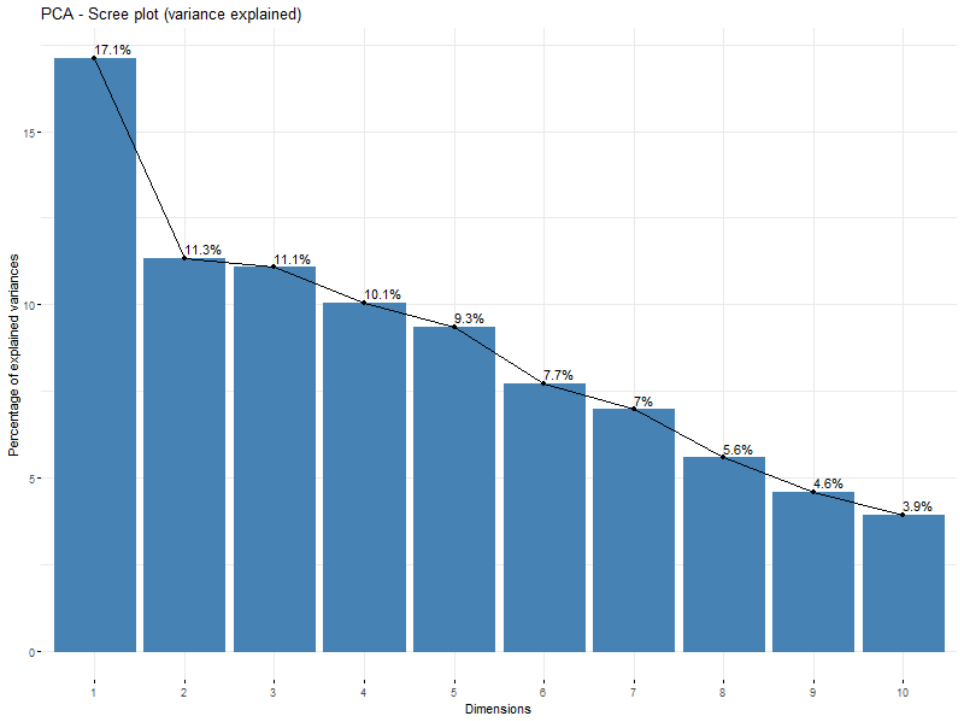

Figure 1: PCA Scree Plot showing proportion of variance explained by each principal component.

The first principal component (PC1) explained approximately **17.1%** of the variance, while PC2 explained around **11.3%**. Together, the first few components captured over 60% of total variation. This indicates that most of the information in the dataset can be represented in a smaller number of uncorrelated dimensions, improving interpretability.

Determining the Optimal Number of Clusters

K-Means clustering was applied to the PCA-transformed data to group herds with similar breakdown characteristics. Two diagnostic plots were used to determine the best number of clusters (k):

- The **Elbow Method**, which assesses the reduction in within-cluster variance as k increases.
- The **Silhouette Score**, which evaluates how well-separated the clusters are.

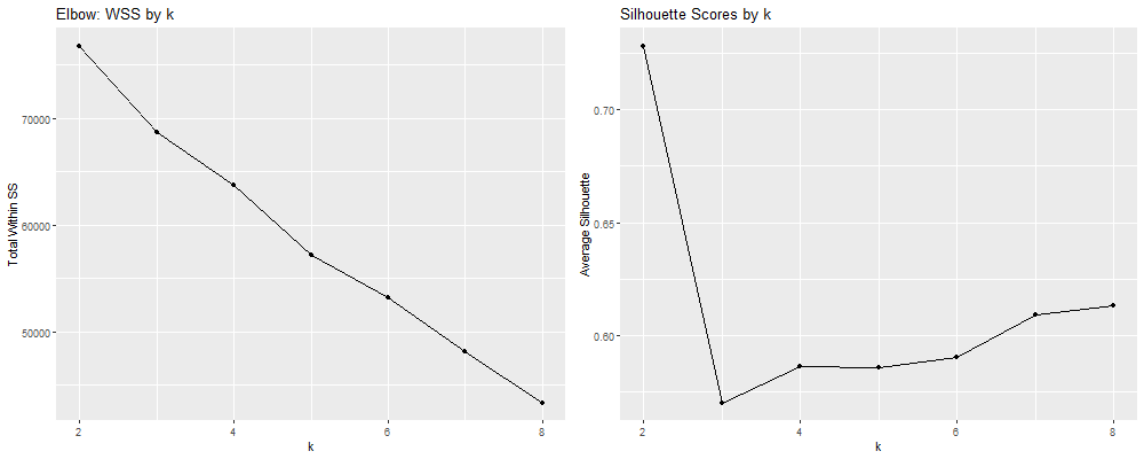

Figure 2: Cluster number diagnostics using the Elbow and Silhouette methods.

From Figure 2, the silhouette score peaked at $k = 2$, indicating that two broad, well-separated clusters best describe the sampled breakdowns.

Cluster Visualization

The PCA biplot (Figure 3) visualizes the herds along the first two principal components, colored by their assigned K-Means cluster.

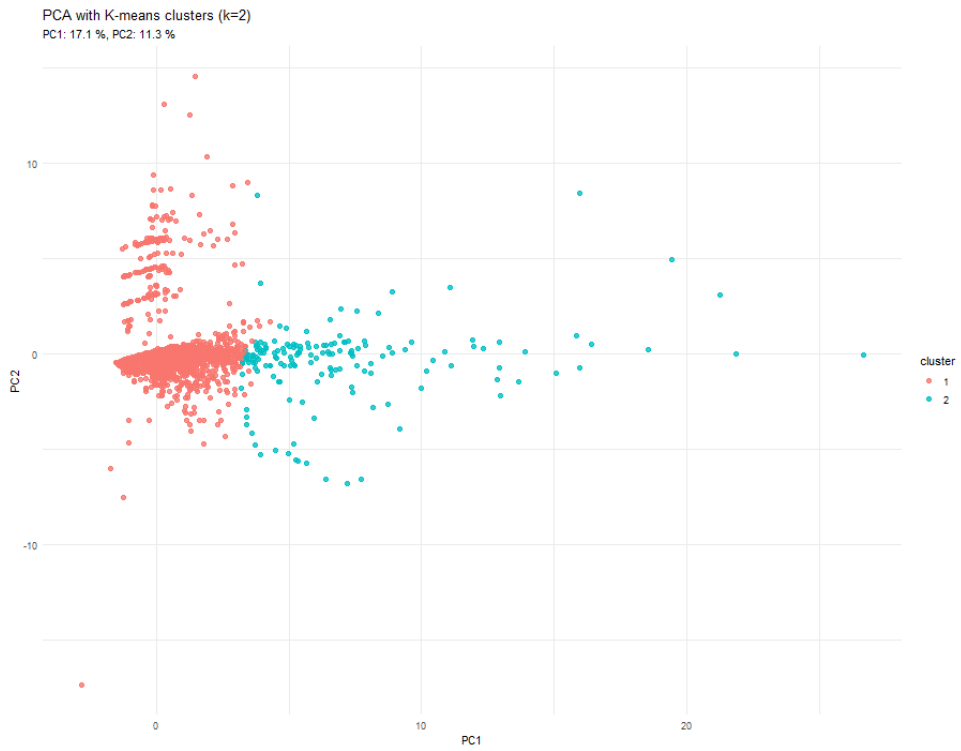

Figure 3: PCA scatter plot of breakdowns colored by K-Means cluster assignment ($k = 2$).

In this visualization, two distinct clusters can be observed:

- **Cluster 1 (red)** – Represents breakdowns with relatively lower positive counts and shorter durations.
- **Cluster 2 (blue)** – Characterized by herds with higher total positives, higher testing frequency, or extended breakdown durations.

Interpretation and Insights

The PCA and K-Means combination successfully reduced data complexity and revealed underlying herd-level structures:

- Breakdown events differ primarily in positivity rates and duration.
- Some herds experience more persistent or complex breakdowns.
- The two-cluster structure indicates possible differentiation between *short-term* and *long-term* breakdowns.

Principal Component Analysis (PCA) and K-Means Clustering on Sampled Data (10,000 Rows)

Due to computational limitations, the full dataset (85,210 rows) was too large to process efficiently on a standard workstation. To maintain analytical quality while avoiding memory overload, a random sample of 10,000 breakdown-level records was extracted. This approach provided a representative subset for pattern detection and model testing.

Dimensionality Reduction using PCA

Principal Component Analysis (PCA) was performed on scaled numeric variables (e.g., total positives, total tests, duration, and test-wise positive counts). The aim was to reduce data complexity while retaining most of the variation in herd-level testing behavior.

Figure 4 shows the Scree Plot, illustrating how much variance each principal component (PC) explains. The first two components accounted for approximately 17% and 11.3% of total variance, respectively, while the first ten PCs cumulatively explained nearly 80% of the total data variability.

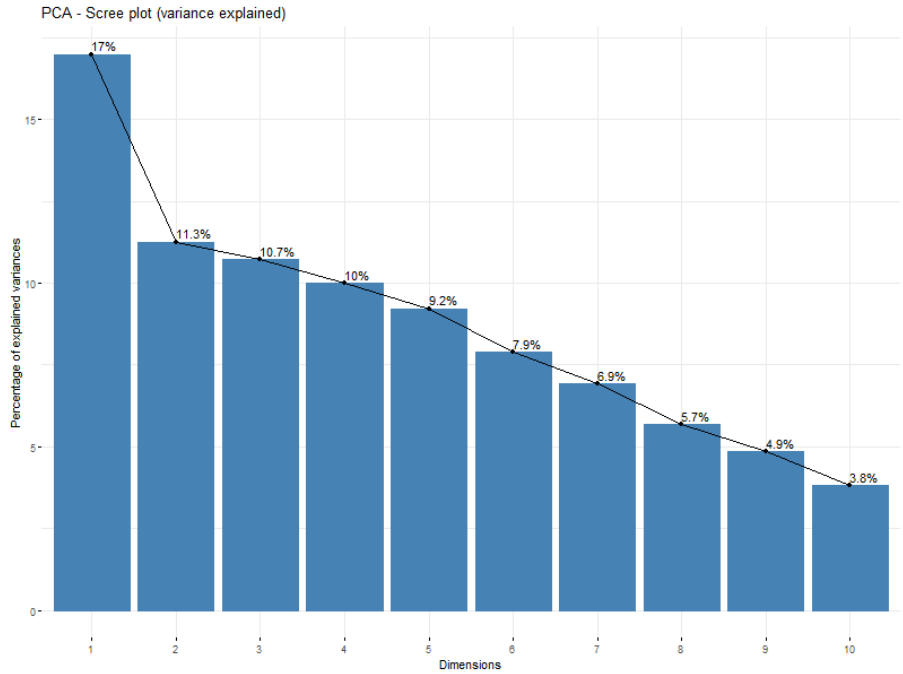

Figure 4: PCA Scree plot (variance explained) for 10,000-row sample.

This means that much of the information within the original 40-dimensional dataset can be represented effectively within the first few principal components, simplifying the subsequent clustering process.

Determining Optimal Number of Clusters (k)

The K-Means algorithm was applied to the PCA-transformed dataset. To identify the optimal number of clusters, both the Elbow Method and Silhouette Analysis were used (Figure 5).

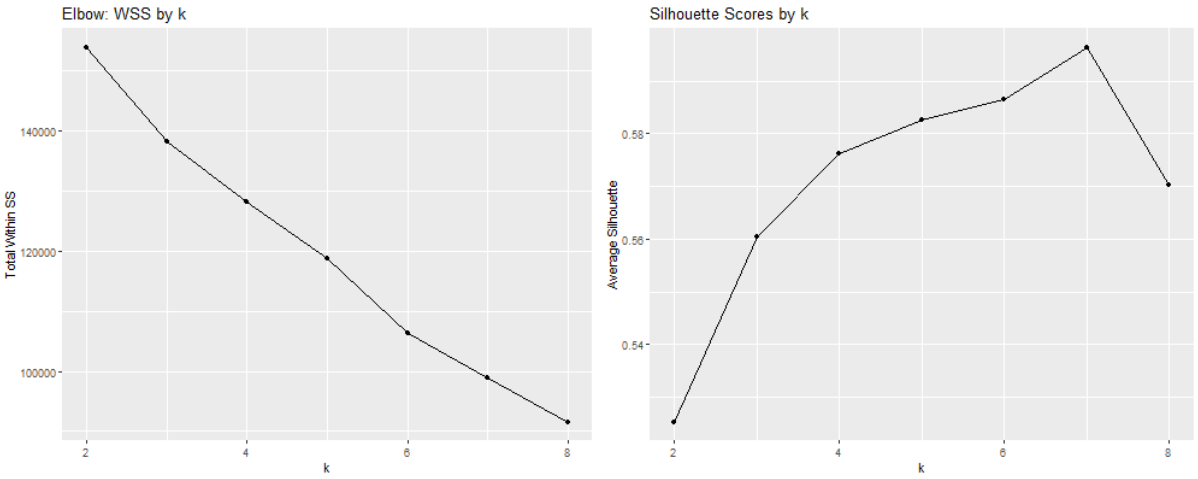

Figure 5: Elbow (WSS) and Silhouette plots used to determine optimal k.

- The **Elbow curve** (left panel) shows the total within-cluster sum of squares (WSS) gradually flattening around $k = 7$, indicating diminishing returns in cluster compactness beyond that point.
- The **Silhouette curve** (right panel) measures how well-separated the clusters are. The highest average silhouette score was also found near $k = 7$.

Therefore, $k = 7$ clusters was selected as the optimal configuration for the 10,000-row dataset.

Cluster Visualization and Interpretation

The final clustering results are shown in Figure 6, where each color represents one cluster in the 2D PCA space (PC1 vs. PC2).

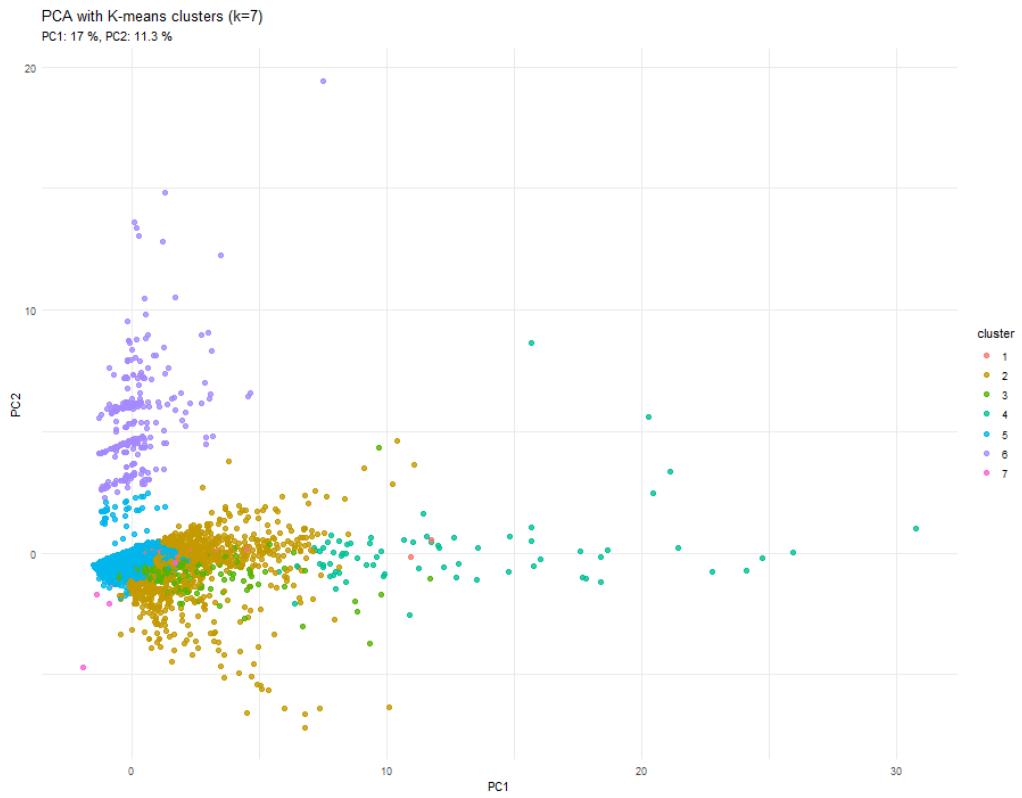

Figure 6: PCA scatter plot showing K-Means clusters ($k = 7$) for 10,000-row sample.

The visual separation between colored groups suggests that the clustering successfully identified several distinct behavioral patterns in the TB breakdown data:

- Some clusters (e.g., yellow and cyan) are densely concentrated near the origin, representing herds with moderate test activity and typical breakdown durations.
- A few clusters (e.g., purple and green) extend along PC1, indicating herds with significantly higher total positives or longer breakdown durations.
- Smaller, isolated clusters (e.g., pink) correspond to rare breakdown profiles — possibly herds with extreme test results, unusually persistent infections, or unique testing patterns.

Overall, this 10,000-row sample confirmed that herd-level TB breakdowns naturally fall into multiple distinct categories, consistent with complex underlying epidemiological and operational behaviors.

Interpretation Summary

The PCA and K-Means results demonstrate the following:

1. The dataset has an intrinsic multidimensional structure, but most of its variation can be summarized using a few principal components.
2. The seven-cluster configuration ($k = 7$) provides a meaningful balance between detail and interpretability.

3. Different clusters likely represent contrasting epidemiological patterns — such as short-duration vs. long-duration breakdowns, low-positivity vs. high-positivity herds, or variations in testing frequency.

These findings laid the foundation for further unsupervised learning analyses using mixed-data clustering (FAMD + HCPC) and nonlinear approaches (SOM) in later sections.

Factor Analysis of Mixed Data (FAMD) and Hierarchical Clustering on Principal Components (HCPC)

Motivation

After applying PCA and K-means clustering on numeric variables, the next goal was to explore both **numeric and categorical features** (e.g., herd type and county) together. Principal Component Analysis (PCA) cannot handle categorical variables directly, so we used **Factor Analysis of Mixed Data (FAMD)** — a method that integrates both numeric and categorical variables into a common low-dimensional representation.

Following FAMD, we performed **Hierarchical Clustering on Principal Components (HCPC)** to identify natural groups of breakdowns (clusters) based on these combined features.

Due to computational limitations and the large size of the dataset (over 85,000 rows), a representative random sample of 5,000 rows was used for this stage. This sampling ensured manageable processing without losing the main data patterns.

Data Preparation and Steps

1. Selected both numeric variables (e.g., duration, test counts, positives) and categorical ones (herd type, county).
2. Cleaned the data by removing empty or zero-only columns.
3. Converted appropriate columns to factors for mixed-data handling.
4. Imputed missing values using `imputeFAMD()` from the `missMDA` package.
5. Ran the FAMD to extract the main dimensions summarizing variability in the dataset.
6. Applied HCPC to these dimensions to form meaningful clusters.

Key Results

FAMD Scree Plot Interpretation The scree plot (Figure 8) shows the percentage of total variance explained by each principal dimension of the Factor Analysis of Mixed Data (FAMD). The first few dimensions capture the largest share of variability in the dataset and thus represent the most informative features.

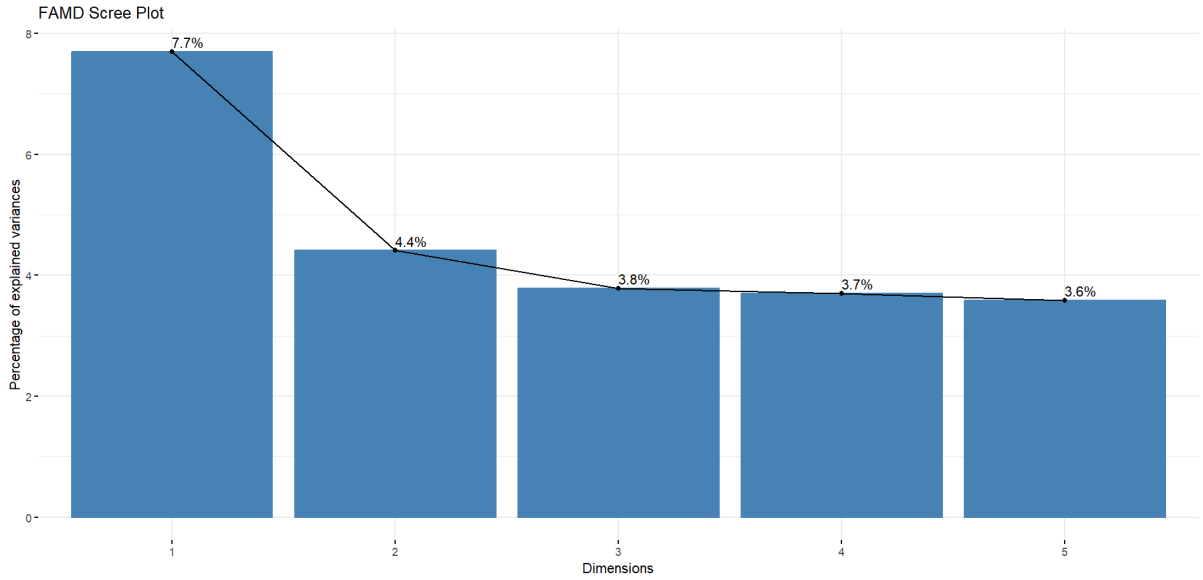

Figure 7: FAMD Scree Plot showing the proportion of explained variance for the first five dimensions.

As shown, the first dimension accounts for approximately **7.7%** of the total variance, followed by **4.4%** for Dimension 2, and around **3.8%–3.7%** for subsequent dimensions. Although the variance explained by each individual axis is modest, this pattern is typical for datasets containing many mixed-type variables (both numeric and categorical). Together, the first three dimensions capture a sufficient proportion of variability to describe the main trends and are therefore used for further interpretation and clustering in the HCPC stage.

Scree Plot of FAMD Dimensions The scree plot (Figure 8) shows how much variance is explained by each FAMD dimension. The first three dimensions capture the majority of variability in the dataset and were retained for further interpretation.

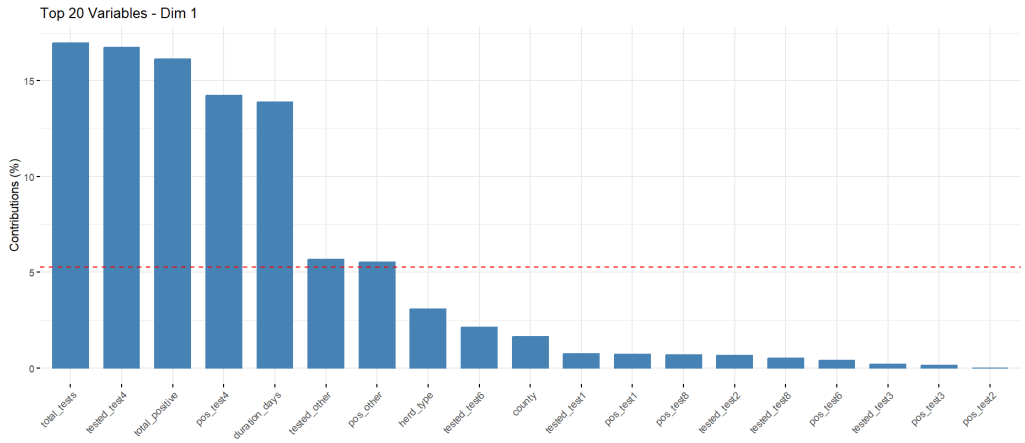

Figure 8: FAMD Scree and Contribution Plot for Dimension 1

Variable Contributions per Dimension Each FAMD dimension represents a major underlying trend in the data. The following figures show which variables contributed most to each dimension.

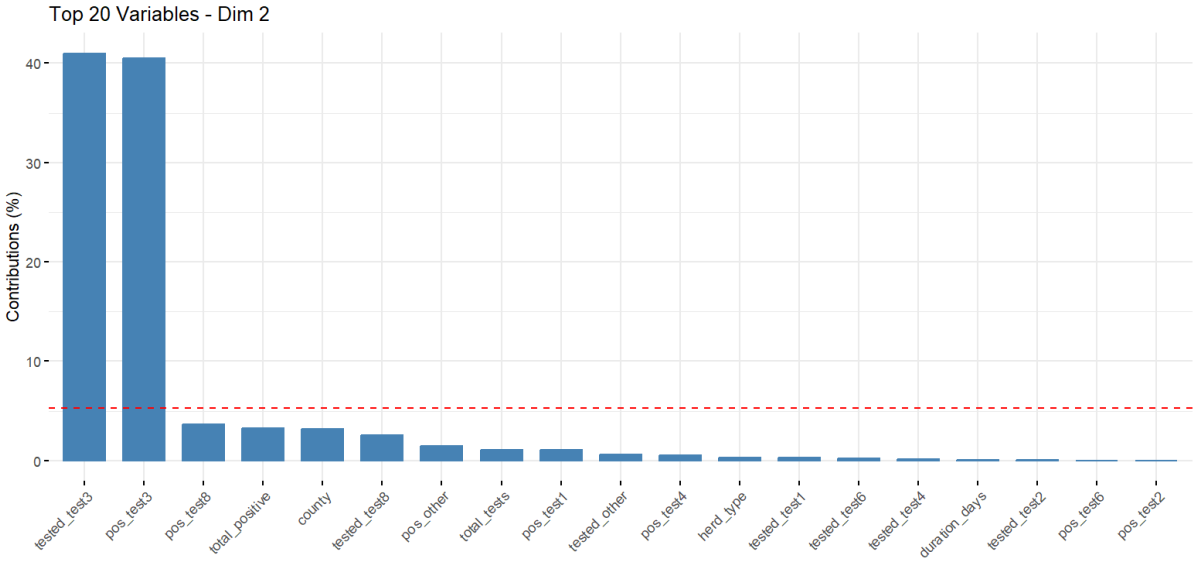

Figure 9: Top 20 Variable Contributions for Dimension 2

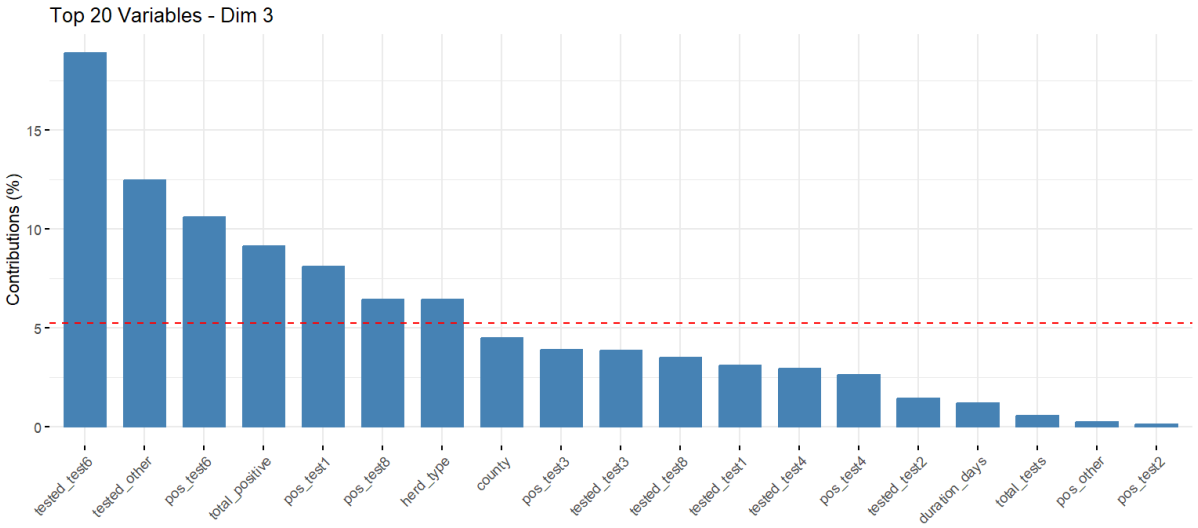

Figure 10: Top 20 Variable Contributions for Dimension 3

Interpretation of the FAMD Dimensions

Dimension 1 – Testing Intensity and Total Positives This first axis captures overall testing activity and infection level. Variables such as `total_tests`, `tested_test4`, `total_positive`, and `pos_test4` contribute most strongly. Breakdowns with higher values on this axis correspond to herds that experienced **longer durations**, conducted **more tests**, and recorded **higher numbers of positives**. In simple terms, this dimension distinguishes "quiet or short-lived" breakdowns from "busy, prolonged, or severe" ones.

Dimension 2 – Test-Specific Behavior (Test 3 Dominance) Dimension 2 is primarily driven by `tested_test3` and `pos_test3`, which together account for over 80% of the variation on this axis. This indicates that some herds are characterized by repeated

or heavy use of a specific test (Test 3), possibly reflecting a particular diagnostic approach or regional testing policy. Hence, Dimension 2 represents the **testing strategy effect**.

Dimension 3 – Alternative Testing Patterns (Test 6 and Other Tests) The third axis highlights less common test types such as `tested_test6`, `tested_other`, and `pos_test6`. These breakdowns are likely cases with alternative or confirmatory tests being used. This axis separates herds with these specialized testing patterns from those using the standard test protocols.

FAMD Variable Factor Map

The Factor Analysis of Mixed Data (FAMD) variable map (Figure 11) visualizes how both numerical and categorical variables contribute to the first two principal dimensions. Dimension 1 (7.7% of total variance) primarily separates herds based on overall testing intensity and breakdown size, while Dimension 2 (4.4%) captures differences related to specific test types.

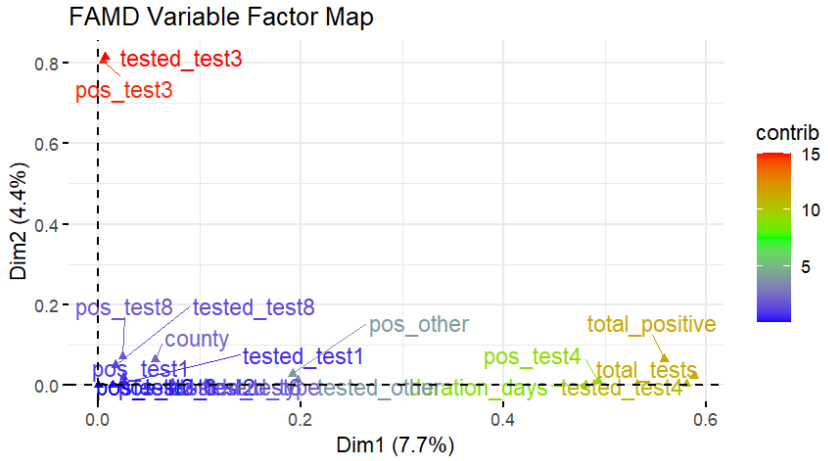

Figure 11: FAMD Variable Factor Map showing variable contributions on Dimensions 1 and 2.

Variables such as `total_tests`, `tested_test4`, and `total_positive` load strongly on Dimension 1, reflecting herds with longer breakdown durations and more extensive testing. On the other hand, `pos_test3` and `tested_test3` dominate Dimension 2, representing variation in specific test outcomes and the proportion of animals tested under certain schemes. Categorical factors like `county` and `herd_type` also contribute but to a lesser degree, indicating that regional and herd management factors influence but do not fully explain the breakdown clustering.

Overall, the map shows that FAMD successfully integrates mixed data, allowing interpretation of both test-related and categorical effects in a unified, low-dimensional space.

Hierarchical Clustering on Principal Components (HCPC)

After performing the FAMD analysis, Hierarchical Clustering on Principal Components (HCPC) was conducted to identify groups of herds with similar tuberculosis (TB) breakdown profiles. HCPC combines the dimensional reduction of FAMD with hierarchical clustering to create interpretable, data-driven groups.

- Each cluster represents herds sharing similar breakdown characteristics—such as total number of tests, duration of infection, positivity rate, and test-type distribution.
- The axes correspond to the first three FAMD dimensions, which capture the majority of the variance in herd-level features.
- Clusters were automatically determined based on Euclidean distances in the FAMD-reduced space, leading to four distinct herd clusters.

Cluster Visualization: Dimensions 1 and 2

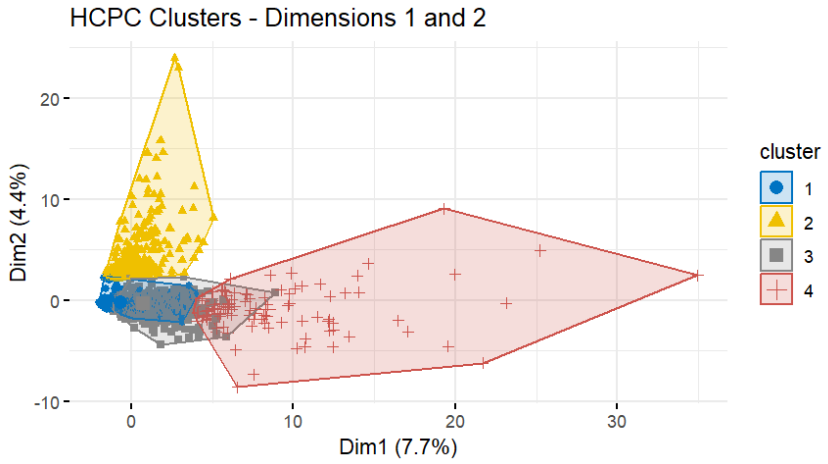

Figure 12: HCPC Clusters visualized on Dimensions 1 (7.7%) and 2 (4.4%).

In the first projection, Cluster 1 (blue) and Cluster 3 (grey) overlap considerably near the origin, indicating herds with moderate breakdown characteristics—average number of tests, positives, and duration. Cluster 2 (yellow) extends vertically along Dimension 2, representing herds distinguished by higher testing activity or variation in specific test types. Cluster 4 (red) spreads horizontally along Dimension 1, containing herds with the longest durations and highest total positives, likely reflecting more severe or persistent breakdowns.

Cluster Visualization: Dimensions 1 and 3

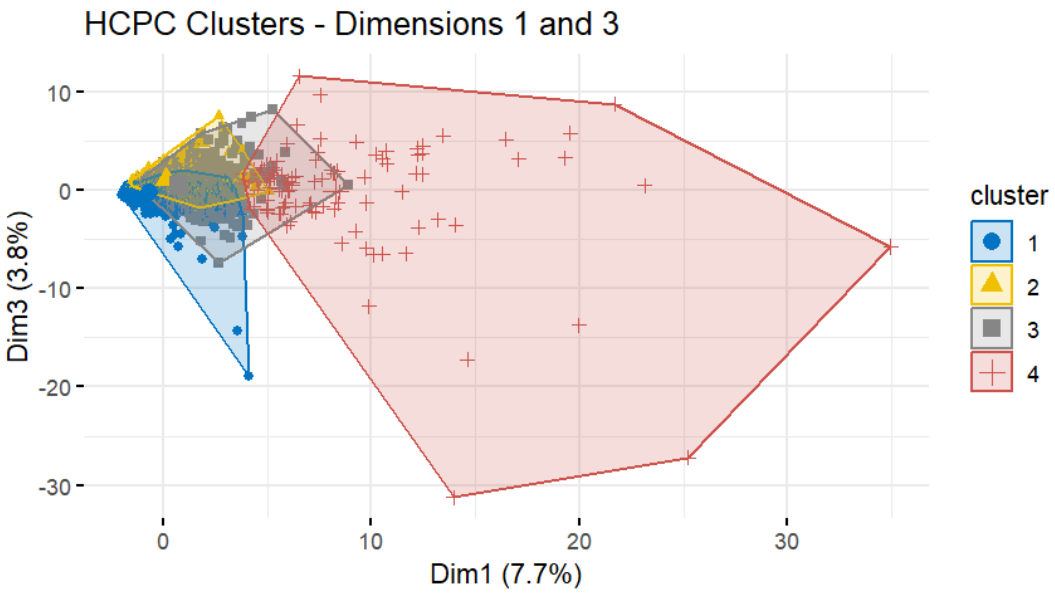

Figure 13: HCPC Clusters visualized on Dimensions 1 (7.7%) and 3 (3.8%).

The second projection (Dim 1 vs. Dim 3) confirms the separation of Cluster 4, emphasizing its distinction along both the duration and positivity axes. Cluster 1 herds remain near the origin with compact distribution, while Clusters 2 and 3 diverge slightly along Dimension 3, possibly reflecting variation in county or herd type composition.

Cluster Visualization: Dimensions 2 and 3

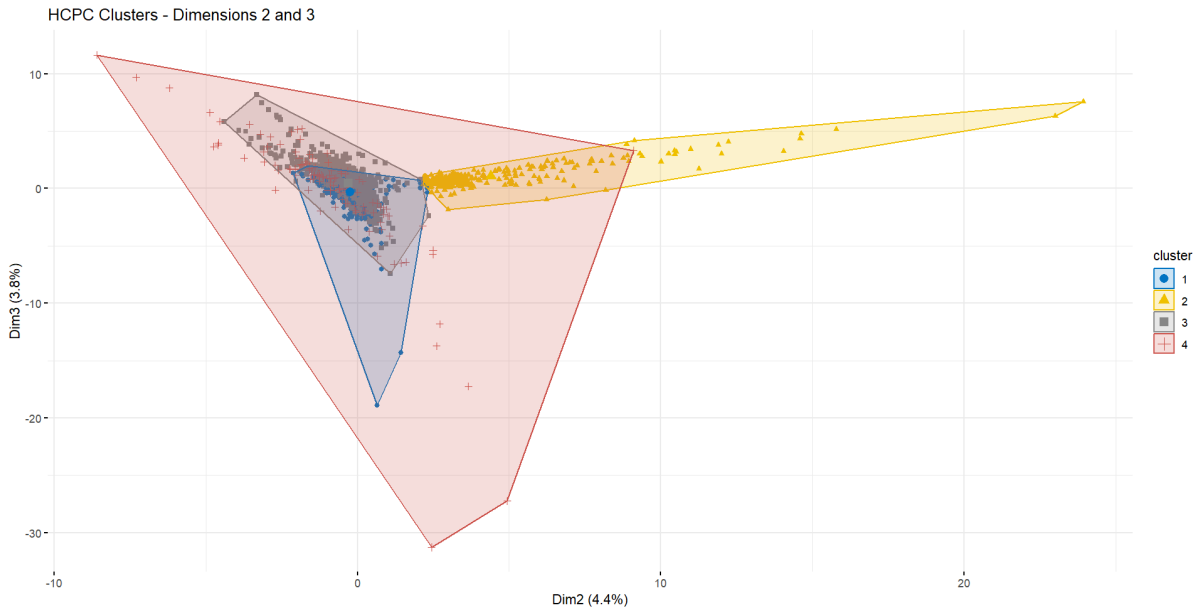

Figure 14: HCPC Clusters visualized on Dimensions 2 (4.4%) and 3 (3.8%).

The projection of Dimensions 2 and 3 highlights clearer boundaries between the clusters. Cluster 2 (yellow) extends strongly along Dimension 2, which is driven by the intensity

and frequency of testing variables (e.g., `tested_test3`, `pos_test3`). Cluster 4 (red) dominates along Dimension 3, associated with high total positives and long breakdown durations. Cluster 1 (blue) and Cluster 3 (grey) remain relatively central, indicating more controlled or short-term breakdown events.

Interpretation and Epidemiological Insights

The HCPC approach identifies four distinct profiles of herd breakdowns:

1. **Cluster 1 (Low-Intensity):** Short-duration breakdowns with few positive cases likely early detection or rapid containment.
2. **Cluster 2 (Testing-Intensive):** Herds undergoing frequent or repeated testing, possibly under surveillance or post-outbreak monitoring.
3. **Cluster 3 (Moderate/Mixed):** Average-duration breakdowns with balanced test outcomes, representing typical TB control responses.
4. **Cluster 4 (High-Intensity / Persistent):** Long-lasting breakdowns with high numbers of positive reactors and extensive testing—suggesting persistent infection or incomplete clearance.

Both `herd_type` and `county` contributed to the clustering structure, implying that geographic and management factors play a significant role in TB dynamics. These findings complement the earlier PCA and K-means results by offering a mixed-data clustering approach that better captures categorical influences alongside numeric testing variables.

Interactive 3D Cluster Visualization

To further explore the spatial structure of the hierarchical clusters, an interactive three-dimensional plot was generated using the `plotly` package in R (Figure 15). This visualization maps the first three FAMD dimensions, with points colored by their assigned HCPC cluster.

The 3D perspective provides a clearer sense of separation among clusters, revealing how herds group along different axes of variation. Clusters occupy distinct regions of the space, suggesting that the chosen features particularly breakdown duration, total tests, and positivity measures effectively differentiate herd-level TB breakdown patterns.

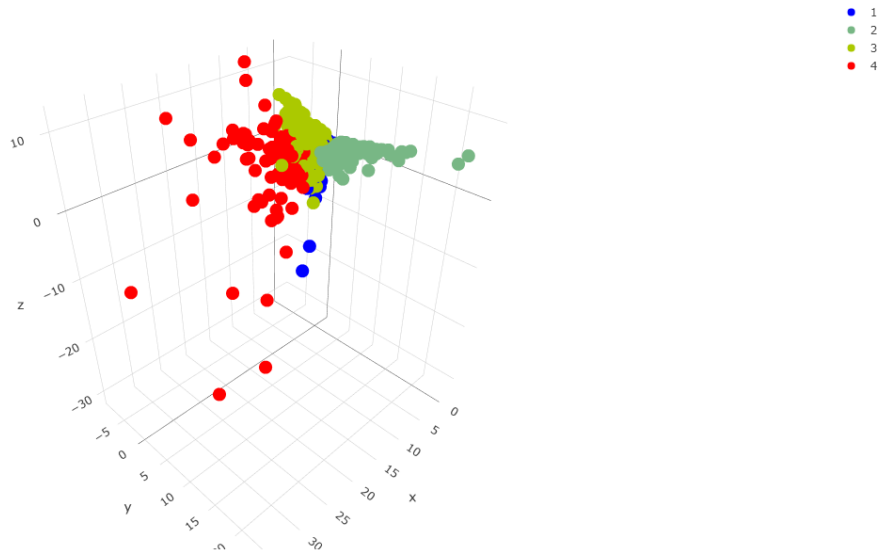

Figure 15: Interactive 3D scatter plot (via `plotly`) showing herd-level clusters across the first three FAMD dimensions.

Self-Organizing Map (SOM) Analysis

To complement the PCA and FAMD approaches, a **Self-Organizing Map (SOM)** was trained to visualize and explore complex, non-linear patterns in the breakdown dataset. SOM is an unsupervised neural network algorithm that projects multidimensional data onto a two-dimensional grid, preserving the topological relationships among the input data.

Model Training and Convergence

Figure 16 shows the *training progress* of the SOM over 100 iterations. The decreasing mean distance between input vectors and their closest map units (best matching units) indicates stable convergence. This means the SOM successfully learned the underlying structure of herd-level breakdown features without overfitting.

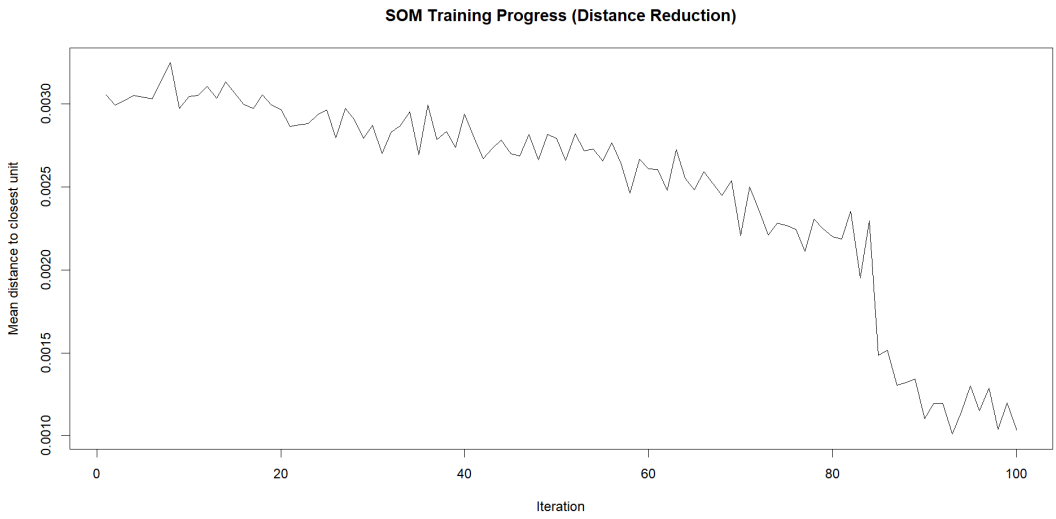

Figure 16: SOM training progress showing reduction in mean distance to closest unit, indicating model convergence.

Node Density and Distribution

The node count map (Figure 17) illustrates how herds were distributed across the SOM grid. Denser nodes (shaded in yellow and orange) indicate regions where many herds share similar breakdown profiles, while sparse or grey nodes represent rarer breakdown patterns. The overall structure suggests that the majority of herds cluster around a few dominant behavioral profiles.

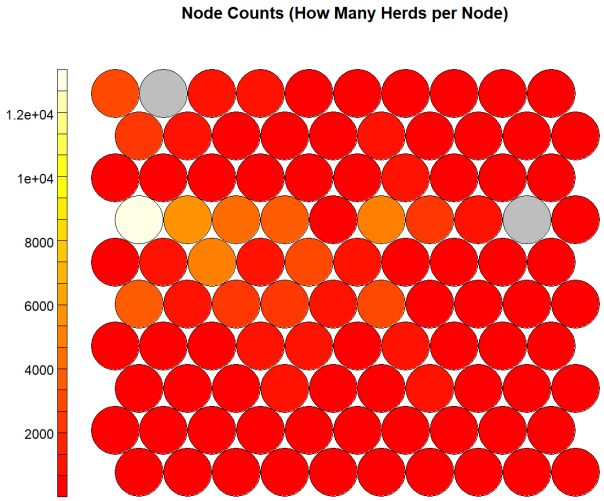

Figure 17: Node count map indicating how many herds are represented by each SOM node. Dense nodes correspond to common breakdown patterns.

Feature Codebook Visualization

The SOM codebook (Figure 18) displays the feature patterns across all nodes. Each small pie-like symbol within a node represents the relative strength of the input variables

(e.g., `duration_days`, `total_positive`, `pos_test1--8`). Green shades correspond to breakdown duration and total positives, while orange and brown shades represent different test-type positivity features. Distinct color patterns across the grid reflect variations in herd breakdown behavior and test response profiles.

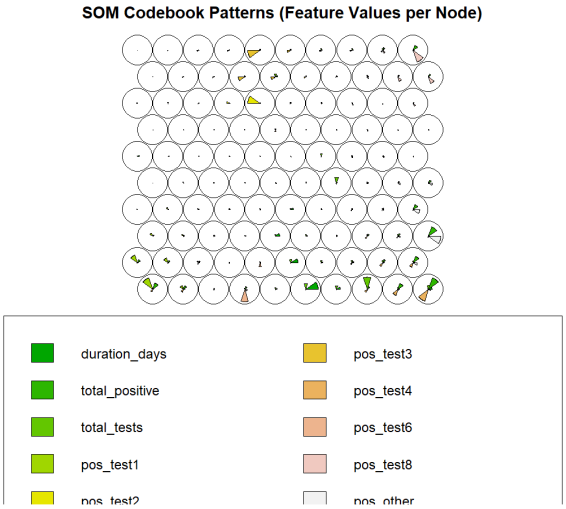

Figure 18: SOM codebook showing feature contribution patterns across nodes. Each node represents a unique combination of breakdown features.

Interpretation

The SOM revealed clear spatial organization of herds:

- Clusters of high `duration_days` and `total_positive` correspond to persistent, high-intensity breakdowns.
- Nodes dominated by low positivity tests (`pos_test1--3`) represent short-lived or limited outbreaks.
- Intermediate nodes capture transitional patterns, suggesting herds with moderate test activity or mixed test-type involvement.

Overall, the SOM provided a **non-linear, topology-preserving view** of the dataset, highlighting subtle gradients in herd breakdown behavior that were not fully captured by linear methods like PCA or FAMD.