**20XDAG - SOCIAL NETWORK DATA ANALYTICS**

**PROJECT SUMMARY**

**Graph Theory Applications to Comprehend Epidemics Spread of a Disease**

* **GOWTHAM S (22PD13)**
* **SHIVNARAN S (22PD32)**

**ABSTRACT**

This study uses graph theory and statistical modeling to detect rabies outbreak clusters, specifically addressing issues of underreporting. By combining temporal, spatial, and genetic data, the research forms a comprehensive framework for analyzing transmission dynamics. Each type of data is represented as a weighted graph where nodes represent individual cases, and edges capture distances or similarities in time, space, or genetic information. Applying quantile-based cutoffs refines these graphs, retaining only likely transmission links to reveal clearer clusters of related cases.

The analysis involves calculating reproduction numbers (R values) and estimating unobserved cases to gauge the outbreak's full extent, even where underreporting may conceal some cases. Using profile likelihood and maximum likelihood estimation, the study derives R values and confidence intervals, providing insights into the likelihood of sustained transmission within each cluster.

Implementation with the R package `vimes` allows for visualizing clusters under various reporting probabilities and quantile cutoffs. Additionally, the approach assesses importation rates, estimating how many cases may be undetected, thus improving the understanding of disease spread beyond reported cases.

Overall, this graph-based methodology enhances early detection and response capabilities in infectious disease control, making it a versatile tool for managing rabies and other epidemics. It informs public health strategies by identifying significant transmission pathways and potential hotspots, thereby supporting effective, timely interventions.

**Overview**

The primary focus of the uploaded documents is on utilizing graph theory and statistical modeling to detect outbreak clusters in infectious diseases, with a specific emphasis on rabies and zoonotic infections. These methods leverage spatiotemporal and genetic data to reveal patterns of disease transmission that might otherwise be obscured due to underreporting or limited case tracking.

### Methodology

The approach employs graph-based models that represent infected cases as nodes, with edges weighted by temporal, spatial, and genetic distances between cases. In this multi-dimensional framework, each data stream (e.g., date of symptom onset, geographic location, and genetic sequence similarities) is processed through a unique distance function. By calculating distance quantiles, the model applies cutoffs that help filter out unrelated cases, thus isolating probable transmission links. This reduces noise in the dataset, allowing for a refined view of clusters that likely represent actual transmission chains.

#### **Tools and Implementation**

The implementation in R makes use of the vimes package, specifically designed to detect clusters by synthesizing information from diverse data types. The code evaluates the impact of various cutoff thresholds on the resulting cluster structures, exploring different quantile levels to determine optimal separation between clusters.

The model also estimates reproduction numbers (R values) for each cluster, accounting for potential unreported cases by incorporating reporting probabilities. This is achieved through profile likelihood methods, where likelihood profiles are generated to calculate maximum likelihood estimates for R values, along with confidence intervals. These R estimates reveal how likely a detected cluster is to sustain or expand, thus guiding response strategies for controlling transmission.

### Key Findings

1. **Cluster Detection**:

* The vimes implementation in R successfully detected clusters within the dataset by analyzing temporal, spatial, and genetic data of rabies cases. The quantile-based cutoff values helped retain meaningful edges between nodes, resulting in clearer clusters that likely indicate direct transmission links.
* By adjusting quantiles and reporting probabilities, the method demonstrated flexibility in detecting clusters across different scenarios, accounting for variations in data quality and potential underreporting.

1. **Reproduction Number (R) Estimation**:

* The reproduction number (R) for each detected cluster was estimated, with confidence intervals generated via profile likelihood methods. These R values highlighted which clusters had higher potential for sustained transmission.
* Reporting rates were integrated to simulate different levels of underreporting, providing adjusted R estimates that factored in unreported cases. This approach revealed R values within confidence bounds that reflect real-world transmission dynamics, helping to identify clusters with significant transmission potential.

1. **Importation Rates and Unobserved Cases**:

* The model calculated estimated numbers of unobserved cases within each cluster, representing potential importations that might contribute to ongoing transmission. This estimation process used thousands of simulations to derive a range of likely unreported cases, showing how many cases might be overlooked due to reporting limitations.
* Importation rates were converted to annual estimates, allowing for a realistic measure of potential case influx per year.

1. **Temporal and Spatial Dimensions**:

* By extending analysis across a 3-year temporal window, the study captured longer-term trends in transmission. Similarly, spatial cutoffs enabled detection of regional transmission dynamics, identifying geographic hotspots likely linked by animal or human movement patterns.

### Graph Theory in Epidemic Modeling

The second file broadens the perspective by exploring graph theory applications in epidemic spread modeling. It provides a comprehensive examination of how different graph structures, such as regular trees and network hubs, can represent patterns of disease spread across human and animal populations. The document discusses:

* **Contact Networks**: Nodes (representing individuals or animals) and edges (representing interactions or movements) are mapped out to simulate how diseases propagate through populations.
* **Cluster-based Control**: The study highlights how strategically disrupting high-degree nodes (such as trading hubs or densely connected individuals) can potentially mitigate outbreak sizes.
* **Modeling Zoonotic Spread**: Graphs are constructed to map animal movement patterns and potential transmission routes for zoonotic pathogens. The paper provides context with examples such as avian influenza, Ebola, and SARS, illustrating how graph theory can predict the spread and help design targeted interventions.

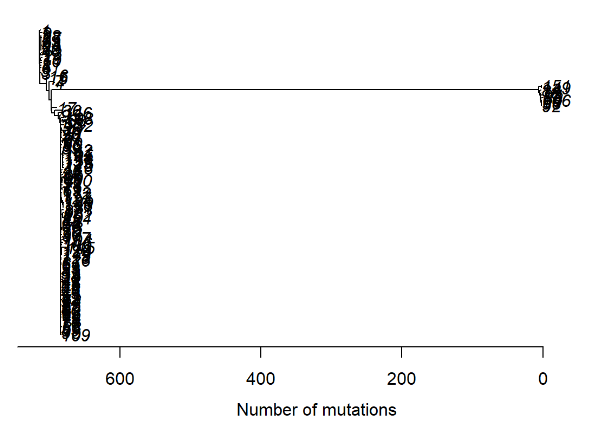
### Insights Gained

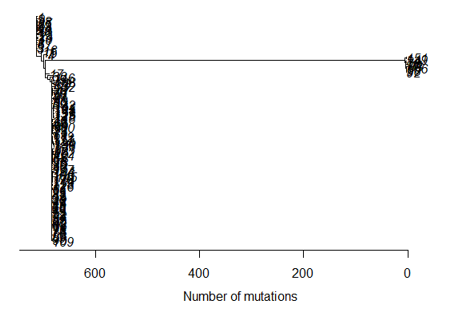
1. **Enhanced Cluster Detection Through Multi-Dimensional Analysis**:
   * Combining temporal, spatial, and genetic data significantly improves cluster detection accuracy, allowing for a comprehensive view of transmission patterns. This multi-dimensional approach helped distinguish actual transmission links from random associations, providing a more realistic depiction of outbreak spread.
2. **Effective Use of Graph Theory in Epidemic Control**:
   * Graph-based models offer a powerful tool for identifying clusters within infectious disease outbreaks, especially for zoonotic infections like rabies. By modeling the disease as a network of interconnected cases, it is possible to identify transmission patterns, high-risk clusters, and potential pathways for disease spread.
   * The insights into graph properties, such as node centrality and edge weighting, underscore how certain nodes (individuals or locations) could act as transmission hubs. Targeting these hubs for intervention may prove effective in outbreak control.
3. **Impact of Reporting Rates on Outbreak Interpretation**:
   * The inclusion of underreporting factors highlighted how outbreak dynamics change with reporting limitations. By adjusting for different reporting rates, the model provides a range of R estimates and unobserved cases, offering a nuanced view that accounts for incomplete surveillance data.
   * This adjustment is particularly valuable for rabies and similar diseases where full case reporting is unlikely, allowing for more accurate estimates of outbreak size and spread potential.
4. **Quantile-Based Cutoffs as a Filtering Mechanism**:
   * The use of quantile-based cutoffs in temporal, spatial, and genetic data streams proved effective in reducing noise while retaining meaningful connections. This method ensures that only relevant links between cases are included in cluster analysis, enhancing the accuracy of the final clusters.
   * Different cutoff choices affect the size and composition of clusters, demonstrating that flexibility in filtering criteria can tailor the analysis to the specific characteristics of each outbreak or dataset.
5. **Broad Applicability Across Infectious Diseases**:
   * While the study focused on rabies, the framework and methods used have broad applicability to other infectious diseases. The ability to combine spatiotemporal and genetic data makes it adaptable to diseases where multiple data sources can enhance understanding of transmission pathways.
   * This approach is particularly relevant for zoonotic diseases, where animal movement and environmental factors play a significant role, as seen in previous outbreaks of avian flu, SARS, and MERS.

**RESULTS AND INFERENCES**

* This plot visualizes a phylogenetic tree based on DNA sequence data, arranged with branches in a ladderized structure for clarity. The x-axis, labeled "Number of mutations", represents genetic distances or mutations along branches from the root to each tip. This helps in understanding evolutionary relationships and mutation counts across the tree's branches.

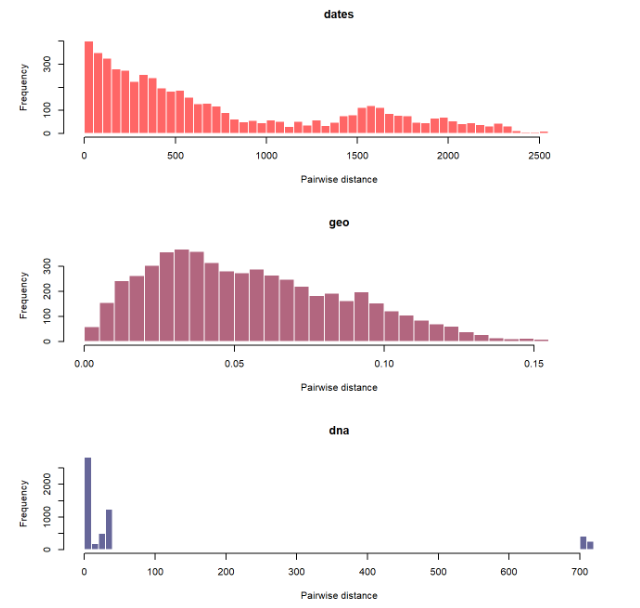
(fig-1 considers underreporting and fig-2 does not)



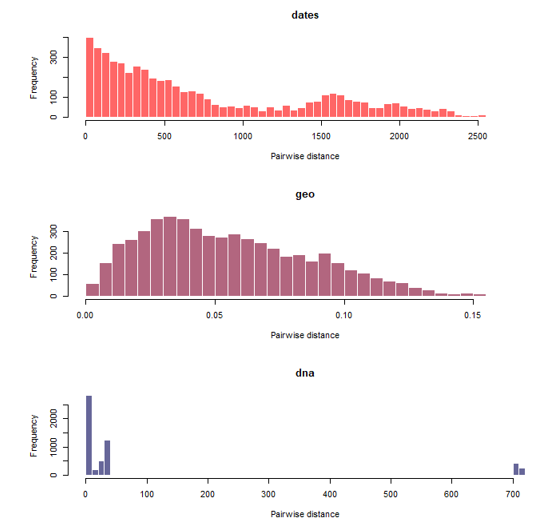


* The plot generated by the vimes package visualizes spatial and temporal distribution of genetic data (e.g., pathogen sequences) across locations. Each class in nclass = 60 represents a subdivision of the data for detailed insight into clustering patterns or trends over time and space.

Considering Underreporting :

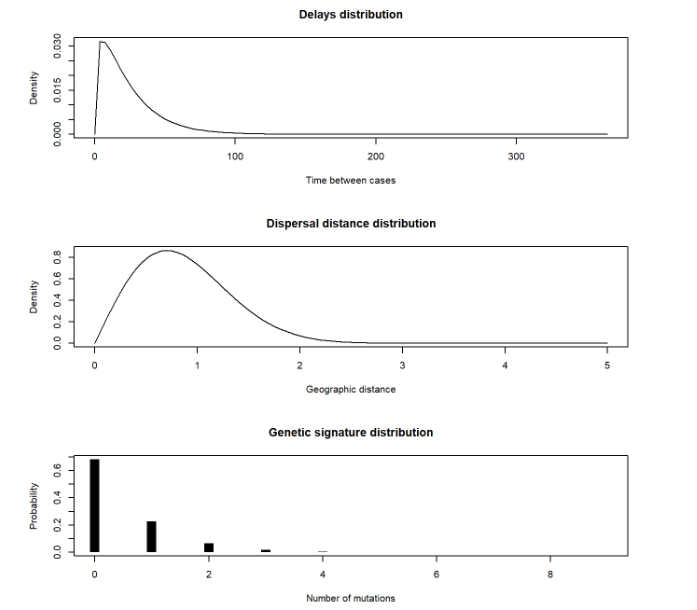


Without considering Underreporting :

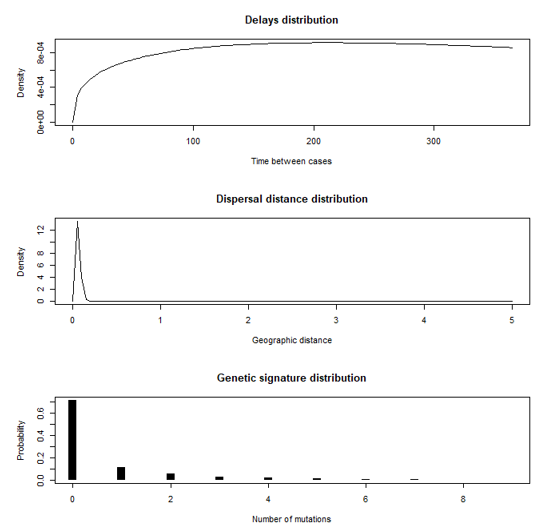


* Three separate plots for temporal, spatial, and genetic pairwise case functions, each with specified parameters. The plots illustrate the distributions over time, distance, and genetic divergence within given ranges.

Considering Underreporting :

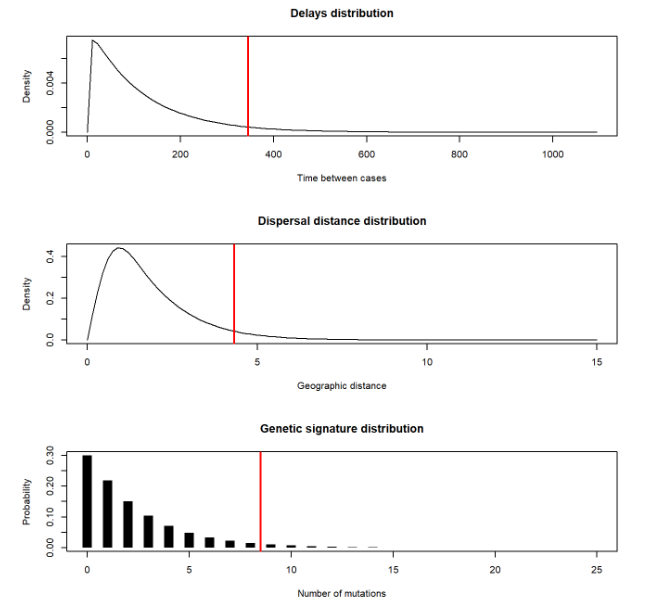


Without considering Underreporting :

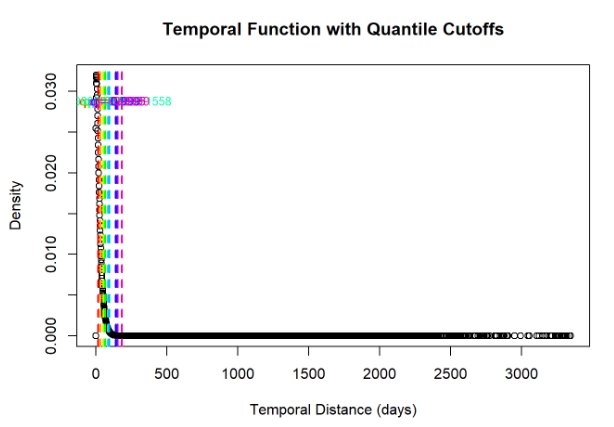


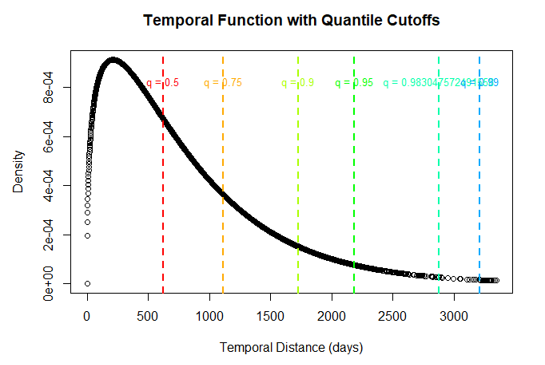
* This plots gives the temporal, spatial, and genetic distance functions with a 95% quantile cutoff, incorporating a reporting probability (pi) and quantiles. Colored overlays represent various quantiles, highlighting thresholds and distributions across each dimension.

Considering Underreporting :



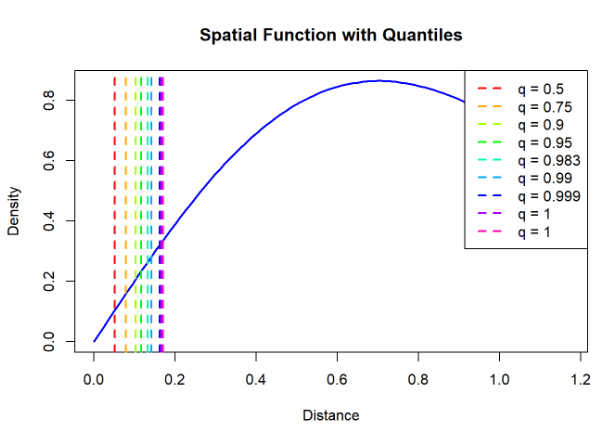
* This plots gives the temporal function with quantile cutoffs as vertical lines, adding labels for each quantile along the x-axis for clarity. Each quantile is color-coded, providing a visual representation of density distribution across different temporal distances. (fig-1 considers underreporting and fig-2 does not)

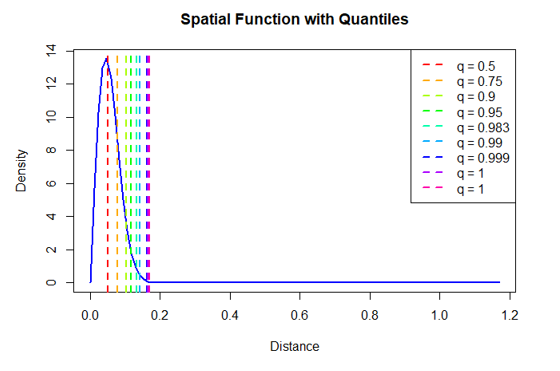




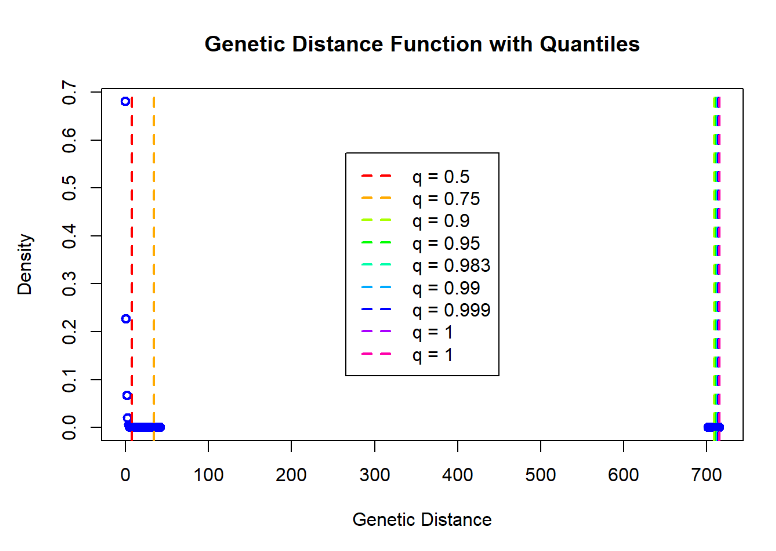
* Plots the spatial function with quantile cutoff lines and labels at each quantile, with color-coded lines marking distances based on quantiles of `D\_geo`. A legend provides quantile values, aiding in visual interpretation of the density distribution by distance.

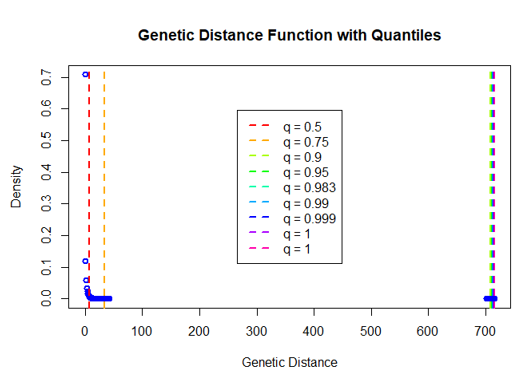
(fig-1 considers underreporting and fig-2 does not)





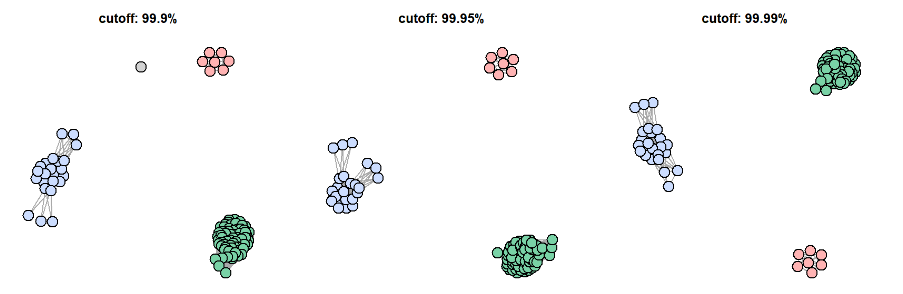
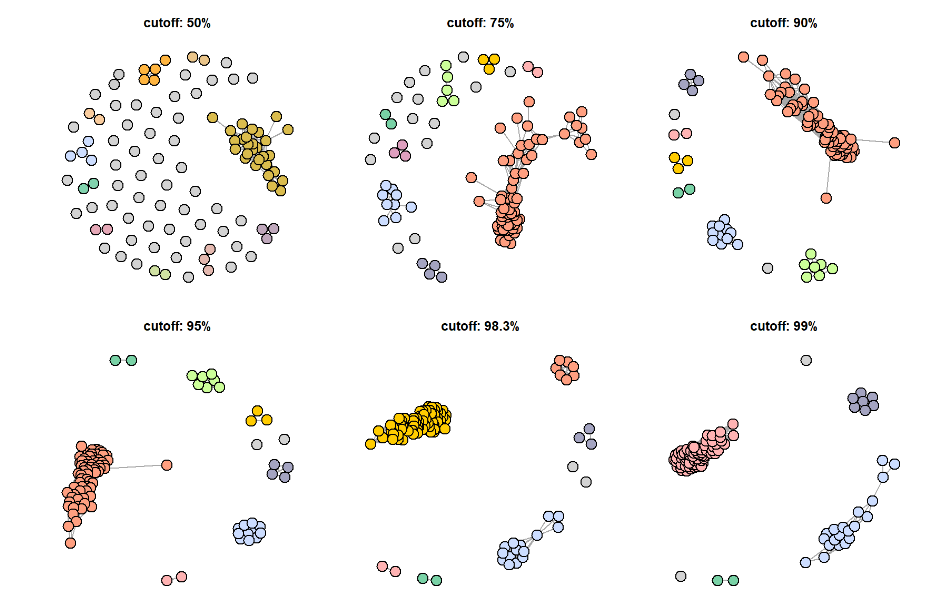
* This plot gives the genetic distance function with vertical dashed lines indicating quantile cutoffs for `D\_dna`, each labeled and color-coded to show different quantile levels. A centered legend details the quantile values, enhancing interpretability of genetic distance distribution. (fig-1 considers underreporting and fig-2 does not)



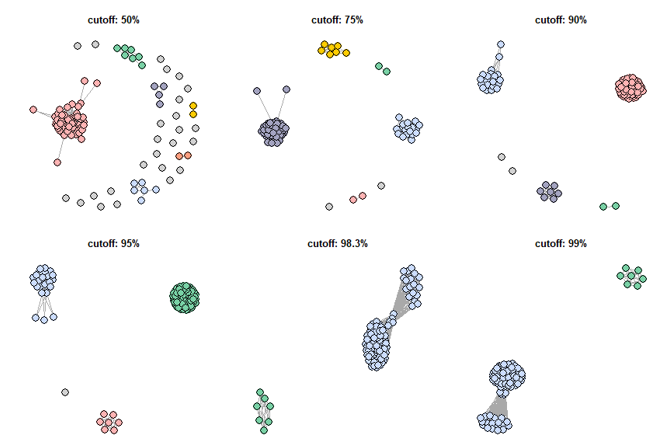


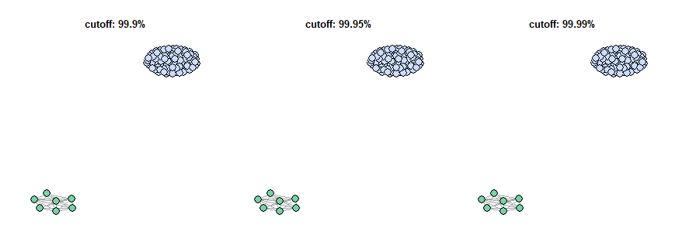
* This plot generates results for different cutoff and reporting rate combinations, visualizing each in a 3x3 grid with no vertex labels. Each plot shows the network graph corresponding to a specific quantile cutoff, as labeled by percentage.

Considering Underreporting :

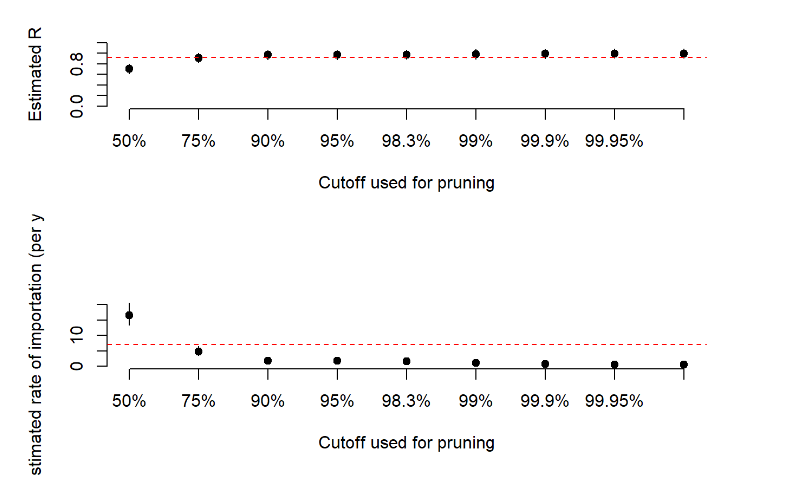


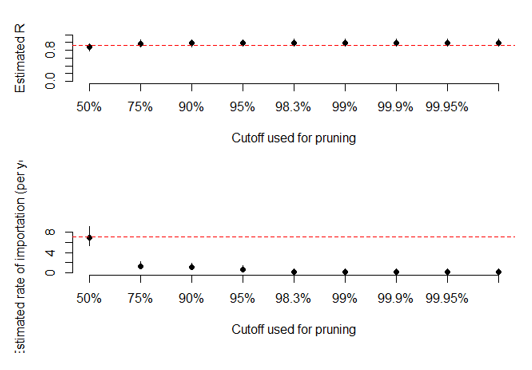
Without considering Underreporting :





* This estimates the reproduction number R and importation rates using cluster sizes and reporting rates, generating confidence intervals for each. It then plots estimated R values with confidence bounds and importation rates per year for different quantiles, marking central, lower, and upper estimates, and overlaying the baseline reproduction number.





**CONCLUSION**

This study demonstrates the effectiveness of graph theory and statistical modeling in detecting outbreak clusters, particularly for diseases like rabies where underreporting is common. By integrating temporal, spatial, and genetic data, the approach enhances cluster accuracy and provides essential insights into transmission dynamics, allowing for better-informed public health responses. The flexible use of quantile cutoffs and reporting rates further refines cluster detection, making it adaptable to different disease scenarios. Overall, this framework supports early detection, intervention planning, and improved outbreak control, with potential applications across a range of infectious diseases.

**DATASETS :**

* rabies\_data\_dna\_2003.csv

<https://drive.google.com/file/d/1bK6iqgVjQVzUMqBM9p-lhfJ6LIpcVSOp/view?usp=drive_link>

* rabies\_data\_linelist\_2003.csv

<https://drive.google.com/file/d/16qnAVxCe2mhKBv5zTQ25cDR-8J8pTleB/view?usp=drive_link>

**RESEARCH PAPERS :**

* Base Research Paper :

[https://doi.org/10.18662/brain/12.2/198](https://doi.org/10.18662/brain/12.2/198%20)