# Trajectory Patterns in Forest Microbiomes: Broader Reanalysis and Location Effects on Prediction Capabilities

Stanislaw Golebiewski

**Abstract**

Bacterial communities exist in diverse environments and change over time in response to external conditions. Understanding these changes is important in ecology, agriculture, and medicine. In a previously published experiment, bacterial samples collected from beech root puddles showed clear temporal dynamics: five distinct initial clusters converged into two final groups after 7 days of incubation. In this study, we reanalysed that dataset using broader taxonomic units (OTUs) and tested alternative clustering methods. We also incorporated spatial metadata to explore potential location-specific effects. Our results confirmed the original structure and revealed a correlation between cluster composition and latitude. This suggests that environmental factors may influence microbial succession even at small spatial scales. Such analyses may support future efforts to model and predict microbiome trajectories in other real world settings.

## 1 Introduction

Bacteria are everywhere and different estimations have been suggested as to how many bacterial species there are – from millions to trillions of species [1]. In many different environments – human gut, hydrothermal chimneys, soil. These bacteria vary greatly both in taxonomic composition and metabolic activity. As time passes and conditions change, bacteria need to adapt, and so does the coexistence of multiple species. These bacterial communities evolve over time. Understanding or maybe even predicting those changes could have both industrial and healthcare applications. We could enhance the treatment of microbiome-related diseases, e.g. Small Intestinal Bacterial Overgrowth (SIBO), or study bacterial communities in agriculture to achieve better yield and reduce fertilizer use.

One such analysis was carried out by researchers studying microbial communities living in small puddles formed between exposed beech tree roots [2]. They isolated 753 samples, identified the bacterial groups present, split each sample into 4 replicates, and after 7 days performed sequencing again to observe what had changed. This enabled them to study the dynamics and trajectories of the bacterial colonies. What they found was that the initial composition of the samples was not random and could be clustered clearly into 5 classes, each showing similar bacterial taxa. After 7 days of incubation, these communities changed and converged into 2 distinguishable final classes. Moreover, upon assessing each sample, it was found that the initial classes differed in their transition probabilities: some converged into one of the final classes in over

90% of cases, while others had a 50/50 like distribution.

In this study, we aimed to replicate the core clustering-based analysis of bacterial community trajectories, with several modifications to extend its interpretability. First, instead of amplicon sequence variants (ASVs), we used operational taxonomic units (OTUs), providing a broader resolution of taxonomic groups. Second, we integrated the bacterial profiles with spatial metadata to examine possible location-specific trends and transitions, which may serve as a proxy for identifying batch effects or ecological correlations. We also explored alternative clustering strategies and performed a systematic evaluation of different numbers of initial and final clusters to identify parameter combinations that yield the most coherent patterns. Overall, our goal was to refine, validate, and expand upon the original observations through a complementary, data-driven approach.

## 2 Methods and Materials

All analyses in this study were performed on publicly available data and scripts from the original authors' repository [2]. The dataset included ASV abundance profiles, taxonomic annotations, time point metadata, and sample pairs in the 0D (initial) and 7D (final) experimental stages. To gain broader taxonomic resolution, we clustered ASVs into OTUs using vsearch [3] with a 97% similarity threshold. Each OTU was annotated by assigning the most frequent taxonomic label among its composing ASVs.

The analitical pipeline was implemented primarily in R [4] with a Jupyter notebook, with some helper scripts written in Python [5]. Each sample's taxonomic composition was paired with metadata on its isolation location to enable exploration of batch-effects and different community behavior in various places.

Two clustering methods were evaluated: partitioning around medoids with Jensen-Shannon distance (PAM+JSD), and k-means applied to PCA-transformed data (K-means+PCA) [6, 7]. Both methods were applied separately to the initial (0D) and final (7D) time points, using a grid search over all combinations of 2 to 6 clusters per time point. This meant 50 combinations to compare. The quality of each clustering result was assessed using internal validation metrics: silhouette width, Calinski-Harabasz index [8], and Davies-Bouldin index [9]. The functions filtered out low-variance OTUs. The silhouette score was computed directly on the Jensen-Shannon or Euclidean distance matrix; Calinski-Harabasz and Davies-Bouldin were computed using the clusterCrit package [10]. Calinski-Harabasz was also employed in the original study.

The best performing configuration was PAM clustering on Jensen-Shannon distance with 5 clusters at 0D and 2 clusters at 7D, identical to the outcome reported by the original authors [2]. This result was used for downstream analysis.

Several heatmaps were created to study the results and find new correlations. First, we generated a global heatmap showing the transition counts between 0D and 7D clusters, similar to those of original authors. Those heatmaps could show both counts and normalized values. A second heatmap splitted these transitions by forest location which enabled direct insights into the different location community dynamics.

To better understand the taxonomic structure of each cluster, we computed the mean relative abundance of each OTU across all clusters and visualized the 25 most important (variable) OTUs in a heatmap using the pheatmap package [11]. OTUs were labeled using genus or species names if they were succesfully found in the previous step. A similar approach was used to display OTUs for each cluster trajectory, showing which taxa dominated particular paths.

A map of sampling locations was prepared to study the sites around London using geopandas [12].

In addition to taxonomic comparisons, spatial metadata was explored by calculating the correlation between coordinates (latitude and longitude) and cluster assignments. Since the number of samples collected from each forest varied a lot, it could fake the results. A bootstrap was used to sample the data multiple times. For each of 100 iterations, an equal number of samples (8, because that was the minimal amount of samples from one of the forests) was randomly drawn from all locations. A Kruskal-Wallis test was applied to assess whether geographic position differed significantly between clusters. The analysis was performed separately for the initial (0D) and final (7D) time points, and separately for latitude and longitude. Summary statistics from the bootstrapped tests included the mean test statistic, mean p-value, and the percentage of runs with p-value below 0.05. We also visualized the distribution of geographic coordinates within clusters using boxplots.

The whole analysis was supported by the package documentations, stack overflow and chatgpt to solve the indexing problems and bugs - as a last resort [13].

## 3 Results

The transitions of each of the 5 initial cluster have been shown on Figure 1 heatmap. We can see, how almost 90% of samples from cluster 1 and 4 transfer to final class 1 while initial classes 3 and 5 follow a more random route.

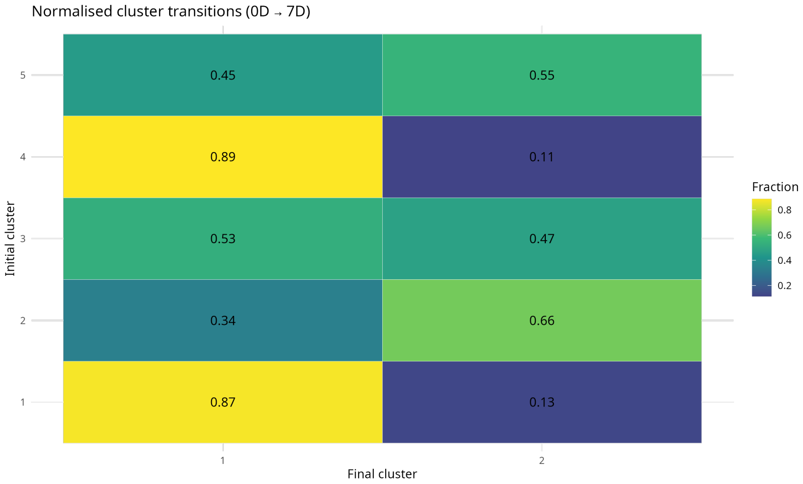

Figure 1: Normalised transition matrix between initial and final clusters. Each cell represents the proportion of samples from one initial cluster that transitioned to a given final cluster.

A new approach is to look at the initial and final classes in each of the forests separately. Figure 2 and 3 show how different the transition percentages are among locations that are not so far from each other. We can clearly see that different locations provided varying number of samples. The most samples were collected from Wytham Woods, Burham Beeches and Greenbroom Cover. Places like Windosr Great Park only provided 4 samples (that were later split into 16 to check reproducibility). Different forests had different percentages of starting classes and their trajectories.

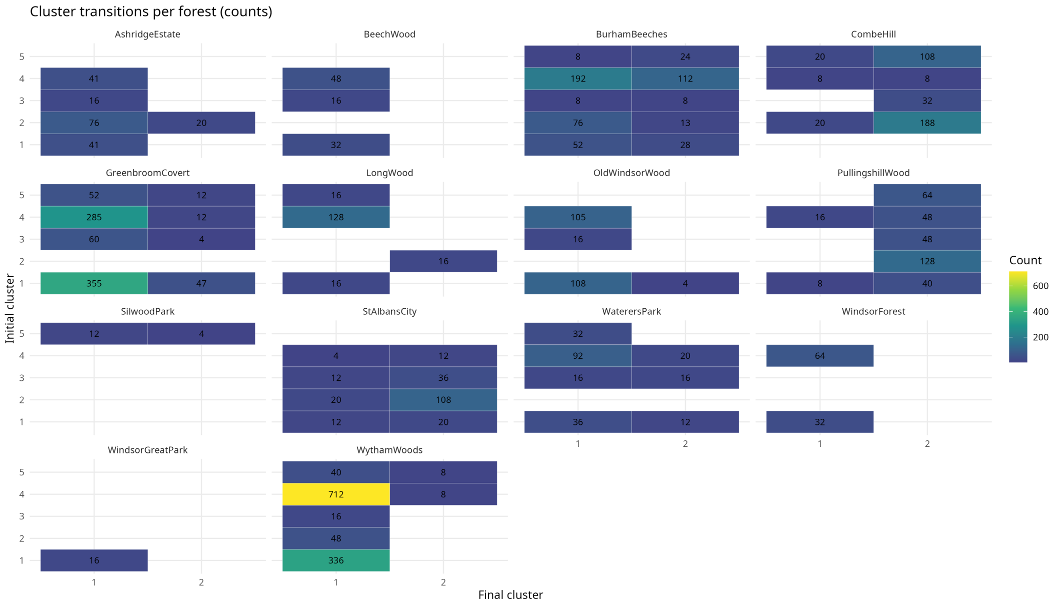

Figure 2: Cluster transitions per forest. Raw counts of transitions from initial to final clusters, shown separately for each forest location.

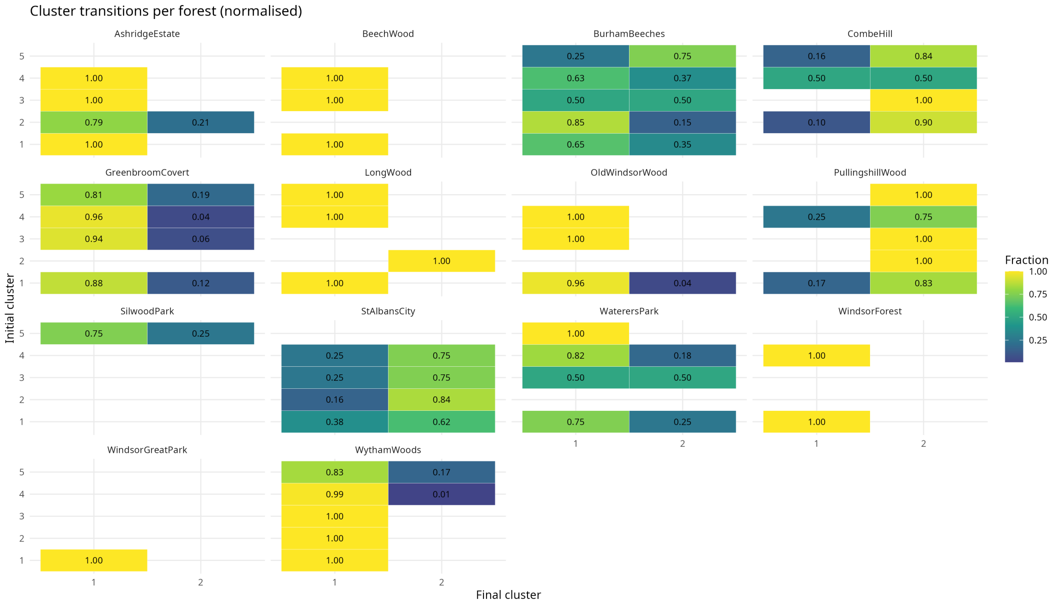

Figure 3: Normalised cluster transitions per forest. Percentage of transitions from initial to final clusters, shown separately for each forest location.

The internal clustering evaluation indicated that the PAM-JSD method consistently outperformed k-means on PCA-transformed data (Table 1). The highest silhouette score (0.067) and the most favourable Calinski-Harabasz and Davies-Bouldin values were achieved for the configuration with 5 initial and 2 final clusters. This configuration was selected for further analysis. It is the same configuration, as the authors have used in their original paper.

Table 1: Evaluation metrics for clustering (Silhouette, Calinski-Harabasz and Davies-Bouldin) for selected methods and parameters.

| Silhouette | Calinski-Harabasz | Davies-Bouldin | Method | Init | Final |
|---|---|---|---|---|---|
| 0.0674 | 298.36 | 2.08 | PAM-JSD | 5 | 2 |
| 0.0591 | 261.41 | 2.28 | PAM-JSD | 6 | 2 |
| 0.0566 | 140.88 | 3.50 | KMEANS-PCA | 2 | 2 |
| 0.0480 | 53.60 | 6.28 | KMEANS-PCA | 3 | 2 |
| 0.0404 | 170.89 | 5.02 | PAM-JSD | 5 | 4 |

Another step in our analysis was to repeat the analysis with OTUs instead of ASVs. Figure 4 shows the 25 most variable OTUs across all clusters, both initial and final. Most of them were successfully identified with taxonomic names. What is visible from the heatmap, is that the most important OTUs for initial class categorization were not the same as the most important ones for final classes. This suggest a dynamic change in the bacteria community composition during the experiment and emergence of new ”leaders” in each group.

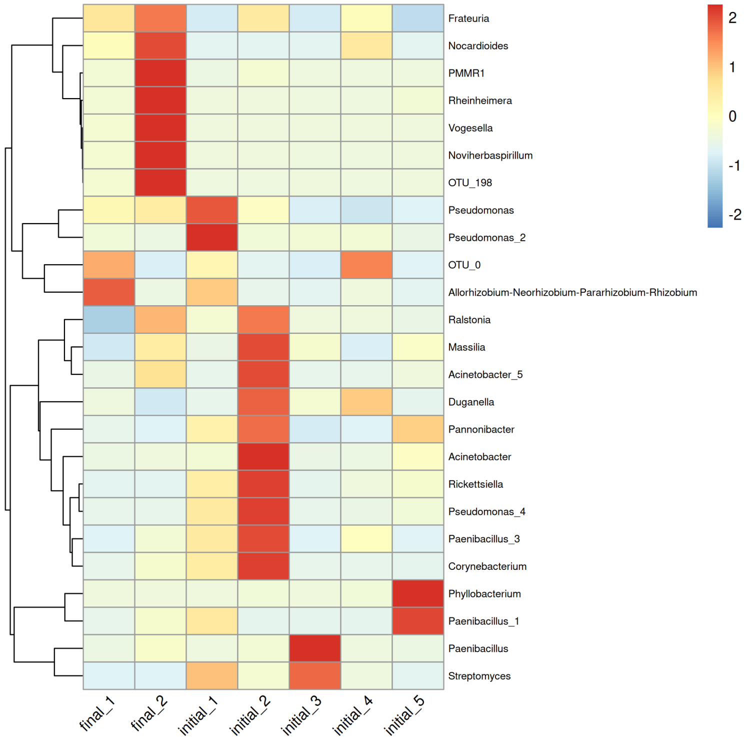

Figure 4: Heatmap showing the 25 most variable OTUs across all clusters. Clustering was performed on z-score normalised abundances. Taxonomic labels are used as row names.

A similar approach was used to check the trajectories and their most variable OTUs. We can see that some of the OTUs (the top 7 rows in Figure 5) were characteristic of communities converging to final class 2 no matter the initial class. Identifying such organisms could help us predict the trajectories of new samples.

Other bacteria groups were strongly connected to singular trajectories which are marked with dark red on the heatmap. What is also worth noting, is that the final class 1 has less characteristic OTUs than class 2. This would suggest that class 1 is the default trajectory and class 2 is a more dependent one.

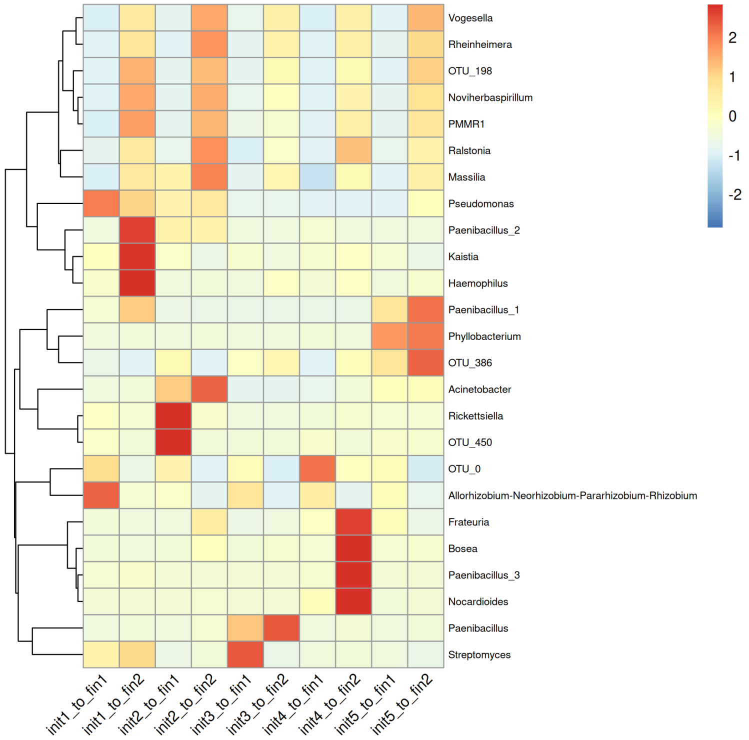

Figure 5: Heatmap of 25 most variable OTUs across all cluster trajectories (from initial to final). This visualises how OTU abundances shift along different cluster paths.

Figure 6 shows a simple map of the sampling locations. They are all west of London, near the Oxfor University. A lot of the locations are concentrated in one spot which could also introduce some bias. Those locations consisted mainly of forests, natural reserves but also big and small city parks.

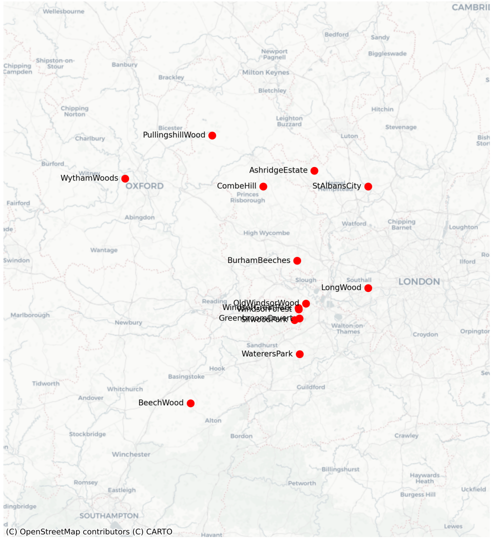

Figure 6: Sampling locations used in the study (manually curated).

A further analysis of location metadata revealed some correlation between the coordinates and initial and final classes (Figure 7). The latitude comparison among the locations, showed that sites up north we more likely to fall into the final class 2. Also, different initial classes were more or less dependent on the latitude. For example the initial cluster 3 showed in both southern and northern locations while initial cluster 2 only apperaed in the most northern regions. Latitude did not show significant results.

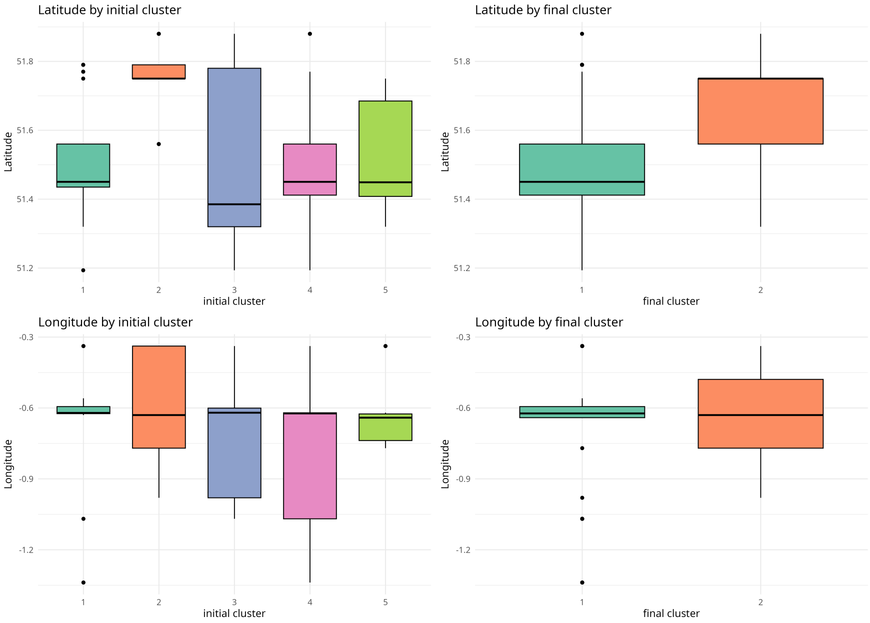

Figure 7: Boxplots showing the distribution of geographic coordinates per cluster. Separate plots are shown for both timepoints (initial and final) and for both coordinates (latitude and longitude).

The Kruskal–Wallis tests revealed a strong correlation between the final classes and geographic latitude at both timepoints. However, longitude showed no significant differentiation between clusters, with p-values remaining high and low percentages of runs below the 0.05 threshold. This suggests that microbiome community structure may follow a latitude gradient, potentially showing environmental or biogeographic variation between forests.

Table 2: Summary statistics of Kruskal-Wallis bootstrap tests between clusters and geographical coordinates.

| Timepoint | Coordinate | Mean KW | Mean p-value | p $< 0.05$ (%) |
|---|---|---|---|---|
| Final | Latitude | 17.7 | 0.00026 | 100 |
| Final | Longitude | 0.79 | 0.49 | 1 |
| Initial | Latitude | 20.8 | 0.0029 | 99 |
| Initial | Longitude | 5.86 | 0.29 | 15 |

## 4 Discussion

The original study focused on the compositional dynamics of bacterial communities over time but did not include any analyses involving metadata such as sampling location. Our results show that spatial metadata, particularly latitude, may correlate with community structure, even within a relatively narrow geographic area. This suggests that subtle environmental gradients could influence the observed clustering patterns.

It is also worth noting that the 7-day incubation period used in the experiment represents a very short ecological timescale. In fact, the act of sample collection and incubation itself likely introduced strong selective pressures independent of the original substrate. In situ studies, with continuous sampling and minimal perturbation, would be more appropriate for learning natural community trajectories.

While the observed latitude correlation is statistically supported, it does not necessarily reflect major climatic differences. Rather, it may be a connected to other local variables, such as canopy coverage, ambient temperature (e.g., urban vs. forest microclimates), wind exposure, or altitude. Future work could benefit from recording and integrating such parameters into multivariate models. It would also be really heplfull to find locations that are further apart to support the actual climate-latitude changes. In this case the coordinates might randomly align with the types of locations. Fore example sites up north could be more remote than those closer to London or Oxford.

Finally, this type of analysis serves as a template for broader applications. All of the above-mentioned comments only focus on this single study but actually the whole idea and statistical analysis might be really useful. Understanding the evolution of microbial communities is crucial in many contexts from agriculture to medicine. If we can reliably cluster and predict microbial transitions, personalised interventions for microbiome-related diseases, such as SIBO, could become significantly more targeted and effective. Also it is really important to note, that different communities trajectories were found to be predictable only in some cases. We already noticed that some of the initial classes had a 50/50 chance of landing in any final class. It is possible, that other studies not connected to puddles under the trees could show a really different result.

## References

[1] Curtis, T. P., Sloan, W. T. & Scannell, J. W. Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences* **99**, 10494–10499 (2002).

[2] Pascual-García, A., Rivett, D. W., Jones, M. L. & Bell, T. Replicating community dynamics reveals how initial composition shapes the functional outcomes of bacterial communities. *Nature Communications* **16**, 3002 (2025).

[3] Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. Vsearch: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2024). URL `https://www.R-project.org/`.

[5] Van Rossum, G. & Drake Jr, F. L. *Python 3 Reference Manual*. Python Software Foundation (2009). URL `https://www.python.org`. Version 3.x.

[6] Jin, X. & Han, J. *K-Means Clustering*, 563–564 (Springer US, Boston, MA, 2010).

[7] Jin, X. & Han, J. *K-Medoids Clustering*, 564–565 (Springer US, Boston, MA, 2010).

[8] Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics* **3**, 1–27 (1974).

[9] Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-1**, 224–227 (1979).

[10] de Amorim, R. & Ghouila-Houri, M. *clusterCrit: Clustering Indices* (2022). URL `https://cran.r-project.org/package=clusterCrit`. R package version 1.2.8.

[11] Kolde, R. pheatmap: Pretty heatmaps (2019). URL `https://cran.r-project.org/package=pheatmap`. R package version 1.0.12.

[12] Jordahl, K. *et al.* geopandas/geopandas: v0.8.1 (2020).

[13] OpenAI. Chatgpt (june 2025 version) (2025). Accessed June 2025. URL: `https://chat.openai.com`.