Supplement of Biogeosciences, 19, 2507–2522, 2022
https://doi.org/10.5194/bg-19-2507-2022-supplement

Biogeosciences
Open Access

*Supplement of*

# Gaps in network infrastructure limit our understanding of biogenic methane emissions for the United States

**Sparkle L. Malone et al.**

*Correspondence to:* Sparkle L. Malone (smalone@fiu.edu)

## Supplement

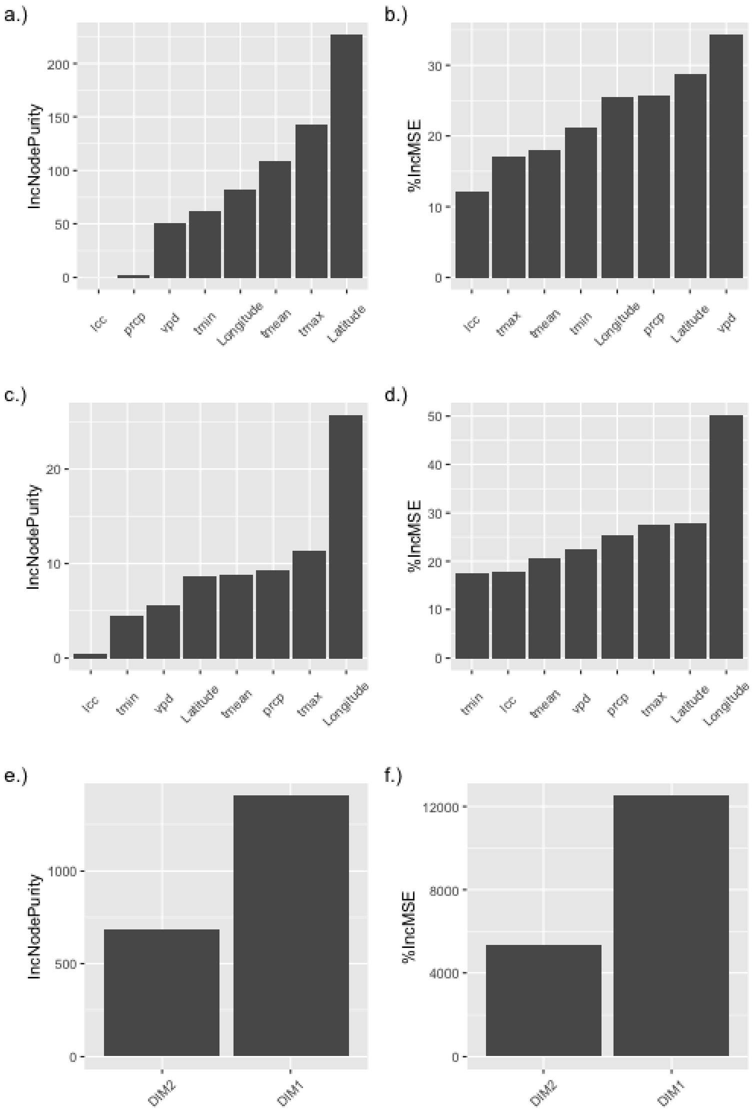

**Figure S1: Variable importance plots for randomforest models of the first (a - b) and second (c-d) dimensions of dissimilarity in the US. Land cover (lcc), climate (prcp, tmean, tmax, tmin, and vpd), and location explained 99% of the variance in each dimension. The random forest model for the clusters used the first and second dimensions of dissimilarity, which had an out-of-bag estimate of the error rate of 3.06%.**

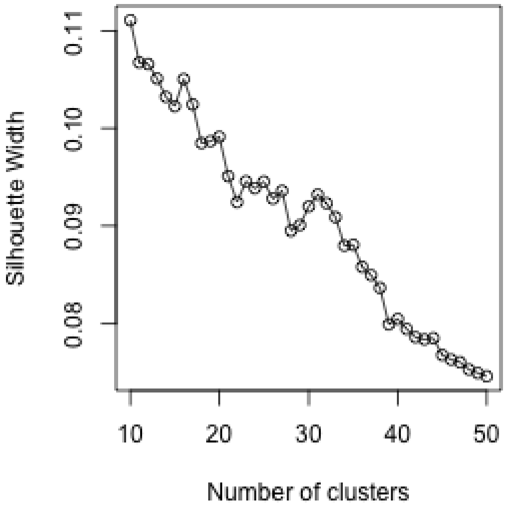

**Figure S2: Silhouette plot showing the number of clusters versus the average silhouette width. We explored clusters from 2 to 50 and selected 10 clusters for the highest silhouette width when considering clusters from 10 to 50.**

### Cluster Analysis

We evaluated the stability of the cluster solution presented (Figure 2) by comparing it to the cluster solutions of 6 example cases (Table S1). The example cases (Samples 1-5) use the same analysis approach as the presented solution (Figure 2). While Sample 1 is from a systematically random sample, samples 2-5 are from a randomly sampled dataset. To understand the impacts of subsampling, we derived a cluster solution for each sample and compared it to the standard solution (Figure 2). The package *bigmds* uses the divide-and-conquer MDS approach (Delicado and Pachon-Garcia 2020). This algorithm partitions the data into subsamples (n=60,000), where classical methods can work. In order to align all the solutions, the Procrustes formula is used (Borg and Groenen 2005). Following the MDS we used the kmeans function to cluster the bigmds (Forgy 1965; Hartigan and Wong 1979; Lloyd 1982).

**Table S1: Example cases to test the stability of the cluster solution.**

| **Example** | **Description** | **Sample Size** | **R Packages** |
|---|---|---|---|
| Sample 1 | Systematic random sample across all climate and land cover classes. | 20,000 | *kmed. MASS, PAM* |
| Sample 2 | Random Sample | 20,000 | *kmed. MASS, PAM* |
| Sample 2 | Random Sample | 20,000 | *kmed. MASS, PAM* |
| Sample 3 | Random Sample | 20,000 | *kmed. MASS, PAM* |
| Sample 4 | Random Sample | 20,000 | *kmed. MASS, PAM* |
| Sample 5 | Random Sample | 20,000 | *kmed. MASS, PAM* |
| BigMDS | divide_conquer_mds | 8,268,498 | *bigmds, stats* |

To compare cluster solutions we measured the stability of a cluster (Figure S3 and Figure S4a.). Stability was measured as the consistency of clustering within the presented solution/standard (Figure 2). This approach is not influenced by the cluster label changing, but by the presence of multiple clusters occurring within the standard cluster. A stability of 100% indicates that all samples within the standard cluster belong to the same cluster in the sample solution.

Regardless of the sampling, similar clustering patterns were obtained (Figure S4) and the mean stability was 82% across the six examples explored (Appendix. 2). The mean stability of the bigmds example was 65%. The cluster stability was low between neighboring clusters NEb and NEa, and between Eb and the SE cluster. Altogether these results show that failing to subsample the data properly leads to declines in cluster stability and changes in the clustering method can lead to differences in the final cluster solution (Figure S4).

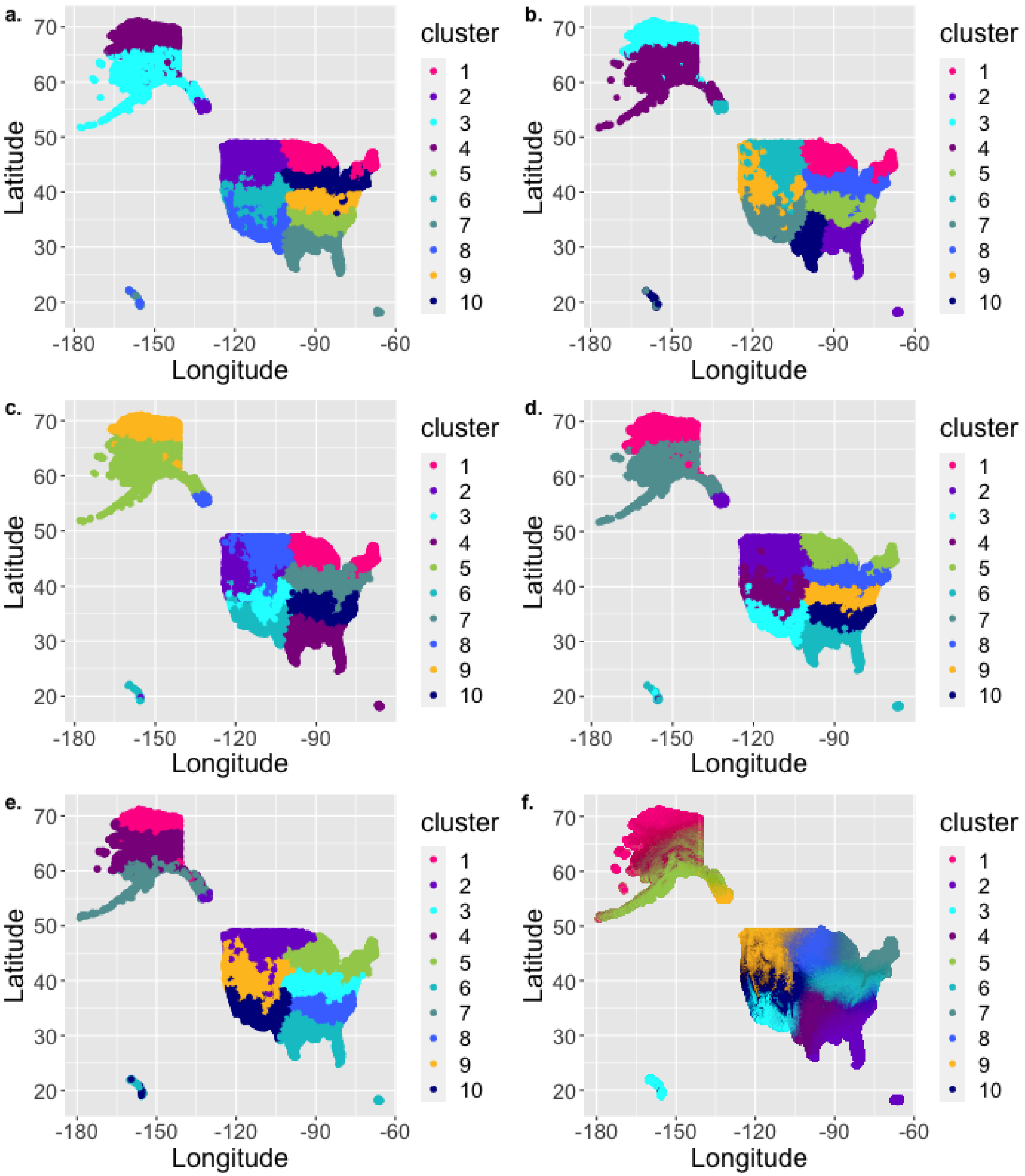

**Figure S3. The cluster solution for a. sample 1, b. sample 2, c. sample 3,  d. sample 4, e. sample 5, and d. the Bigmds (Table S1).**

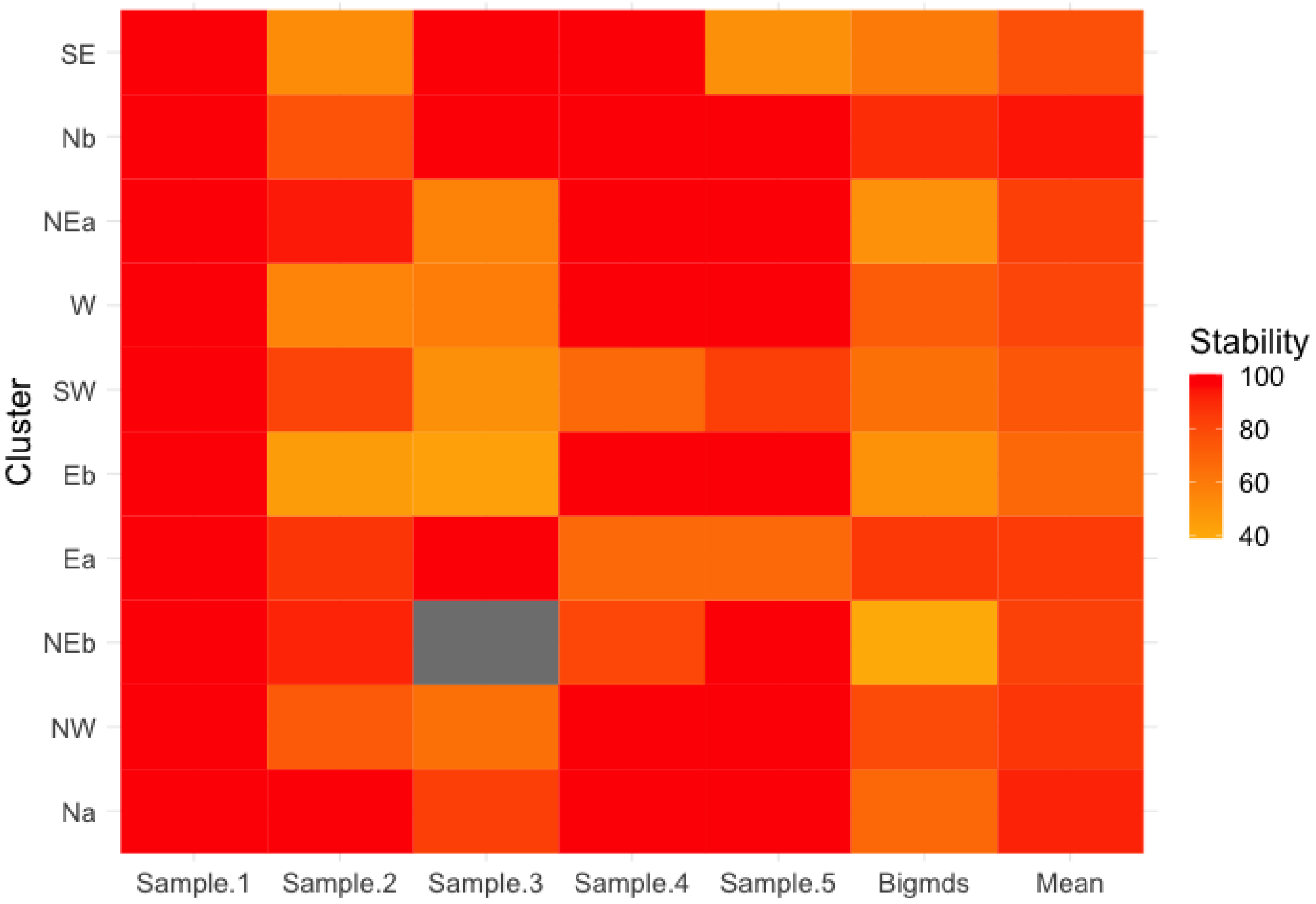

**Figure S4. The stability of the example cases compared to the standard cluster solution (Figure 2). The mean stability is 82% across all cluster solutions. The gray region was not sampled by sample 3.**