Here’s an update to our project. To recap, the challenge was to show that what our method for estimating the C matrix from a community map, based on finding clusters of similar cells in a square gridded landscape, is valid and returns sensible results. I tried doing this by calculating the likelihood of the data given the estimated C and the inferred landscape. Results follow below. (*Note: Results were similar whether I used hierarchical clustering or PAM, compare Figs 1-3 with 5-7.*)

Here’s a general caption to the next few figures.

Top row: **A**: landscape; **B**: C matrix; **C**: Community map

Middle row: **D**: likelihood of the community map given the estimated C matrix and the estimated gridded landscape. Colors correspond to different number of soil types used in the clustering analysis (i.e. the number of clusters being sought); **E**: RV index between the true C matrix (which would be unknown in real life) and the estimated C matrix. This index ranges from 0 (no correlation) to 1 (identical matrices); **F**: Plotting the max likelihood (i.e. top value of each curve in D) against the number of soil types (NST\_ used in the clustering analysis. Red line indicates where the point where the curve has a shoulder, formally the point with the highest magnitude of the second derivative. Notice that regardless of the true NST, the max likelihood seems to always increase NST. I think this suggests overfitting. Hence my idea to look for the shoulder, i.e. the value at which adding more NST doesn’t help as much.

Bottom row: **G**: Same as F, but plotting RV between true and estimated C matrix. Notice how the RV isn’t very sensitive to the number of soil types used unless you use a very bad number. **H**: Linear cell size where the analysis gave its max log likelihood, as a function of the number of soil types used. This index is also not sensitive the NST used; **I**: estimated C matrix using the NST at the shoulder of the curve in F, and the corresponding best cell size from H.

# Hierarchical clustering results

Figure . S = 20 species, true NST = 5 soil types; landscape autocrrelation parameter = 25. The method works great. It identifies the correct NST, and gives us a C matrix that looks very similar to the true C (which is a case of specialists with added randomness).

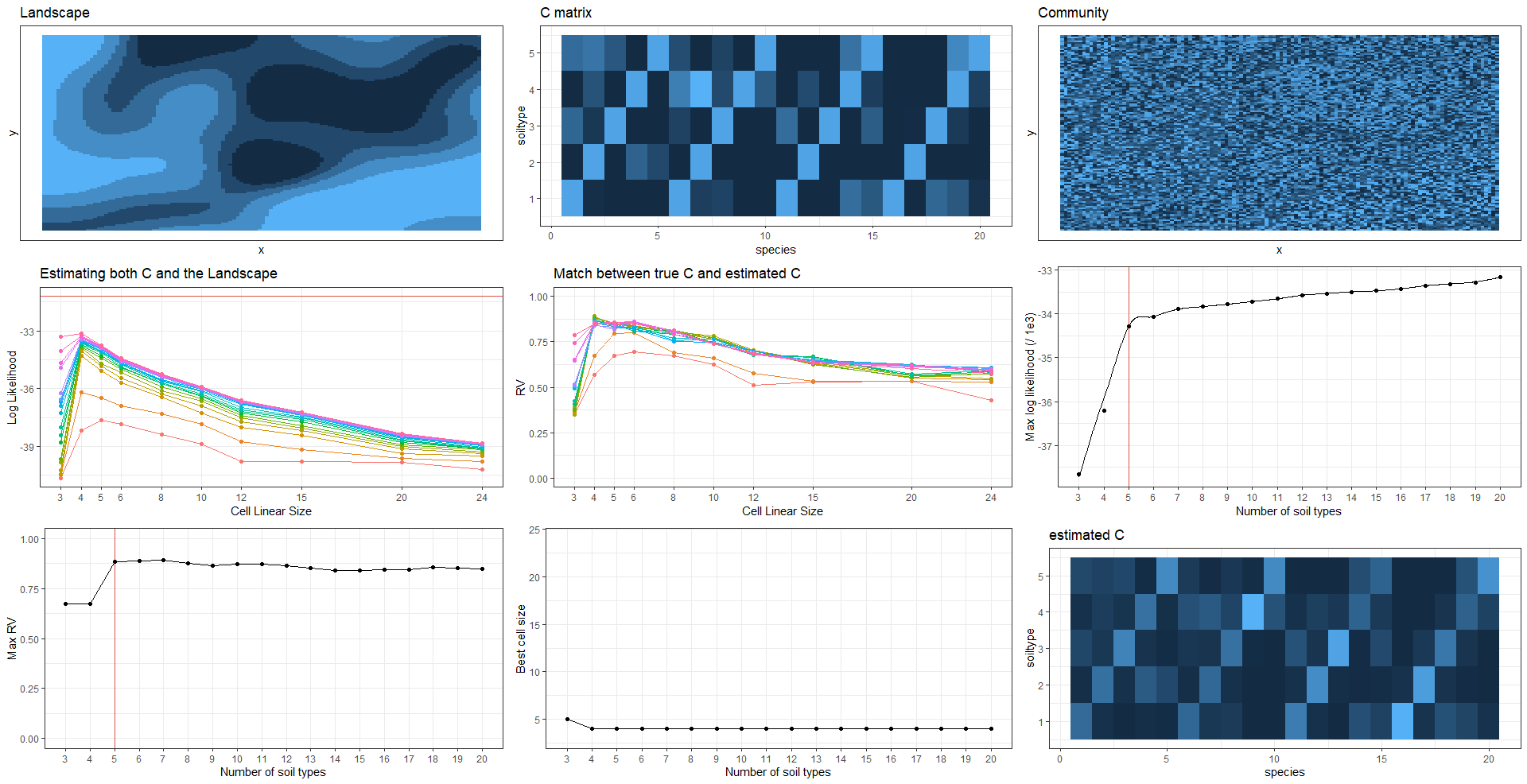


Figure . S = 50 species, true NST = 5 soil types; landscape autocrrelation parameter = 25. When we have many more species (50 instead of 20), we get the true NST wrong but close to reality, and the estimated C matrix maintains the niche structure of the true C (i.e. specialist species occur in groups)

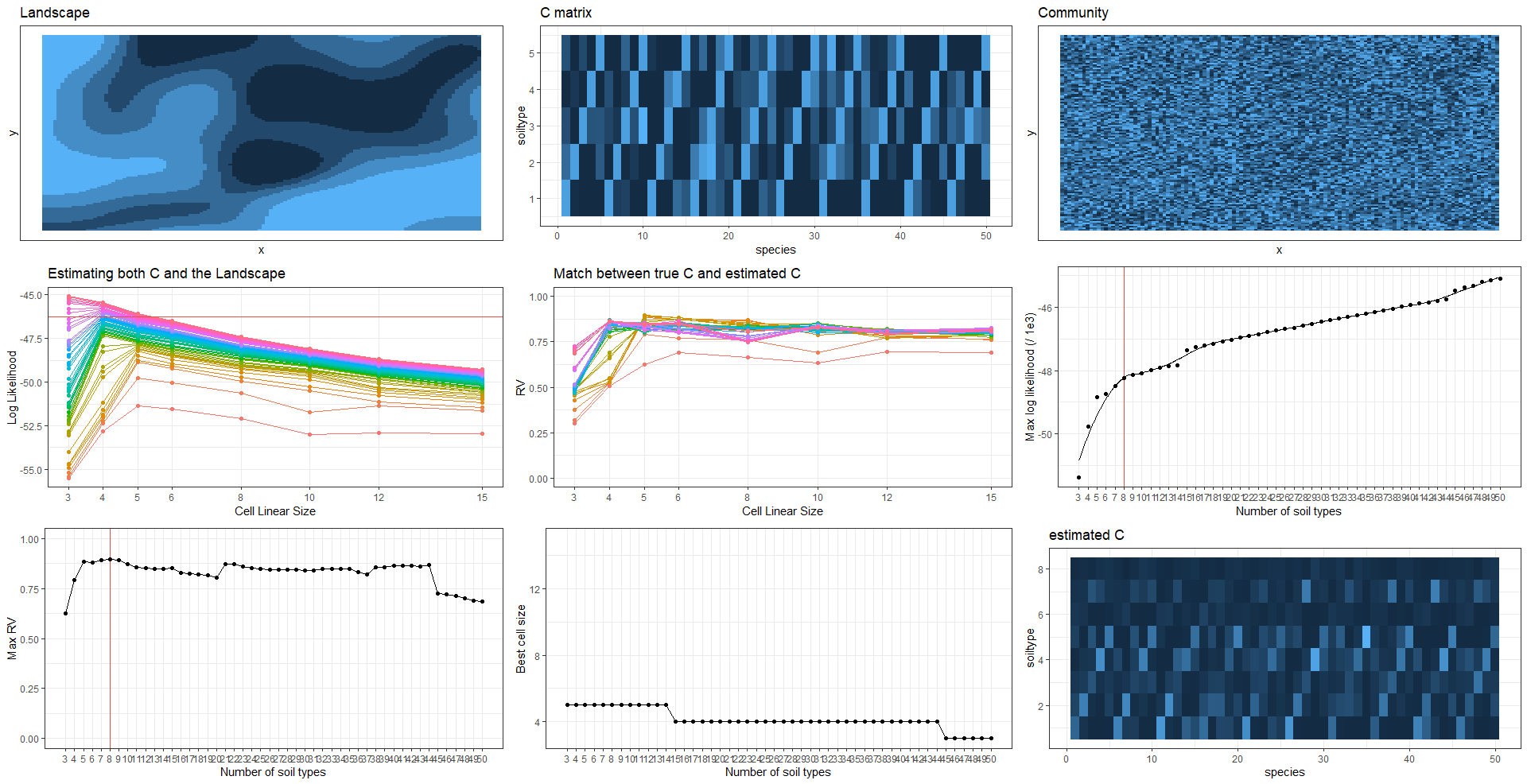


Figure . S = 20 species, true NST = 20 soil types; landscape autocrrelation parameter = 50. Here our method gets lost. It wrongly identifies 13 as the best NST, and gives us a C matrix that isn’t particlualry similar to the true one. Interestingly, even if we force the right NST, the estimated C matrix doesn’t look much better than this one (see next figure).

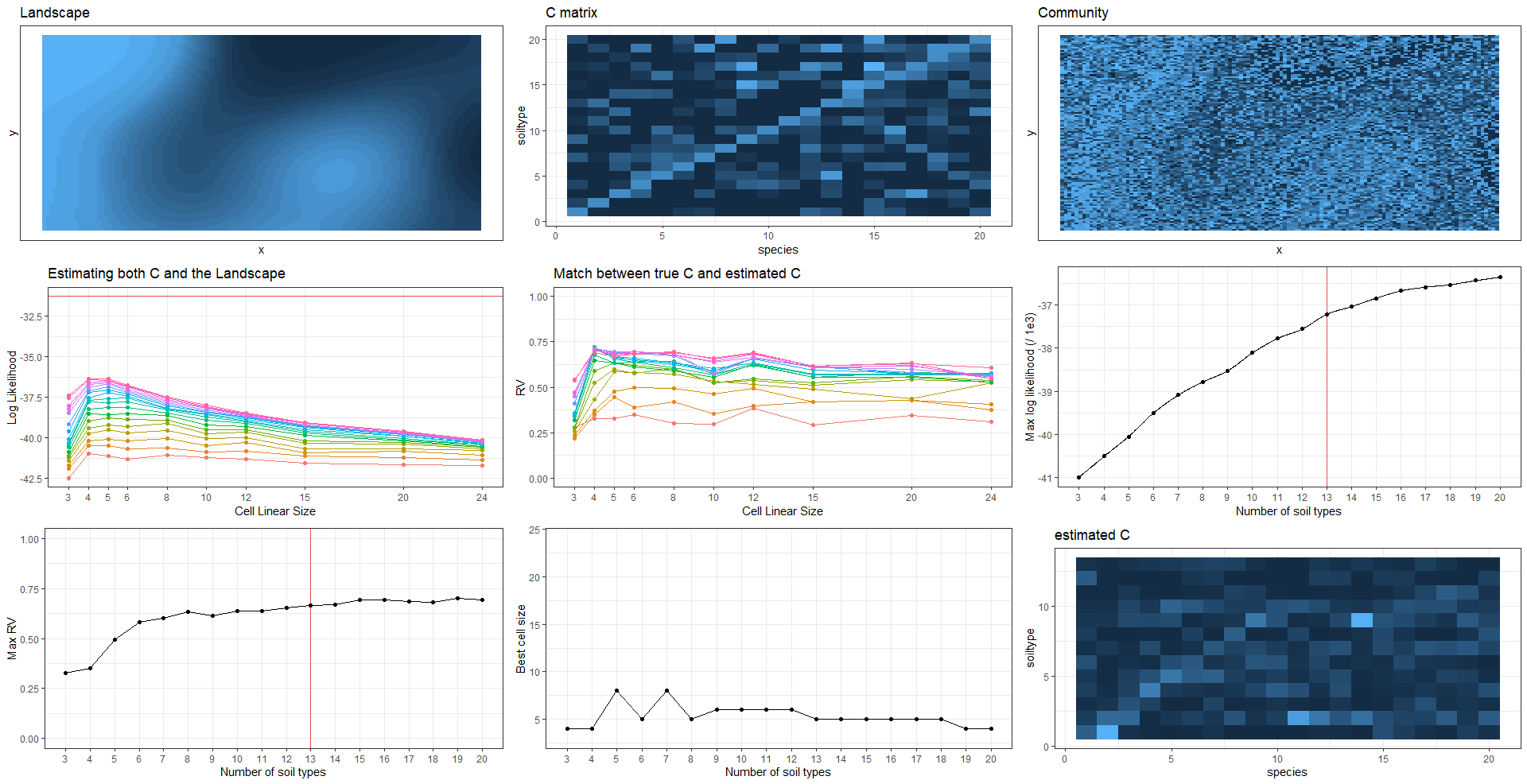
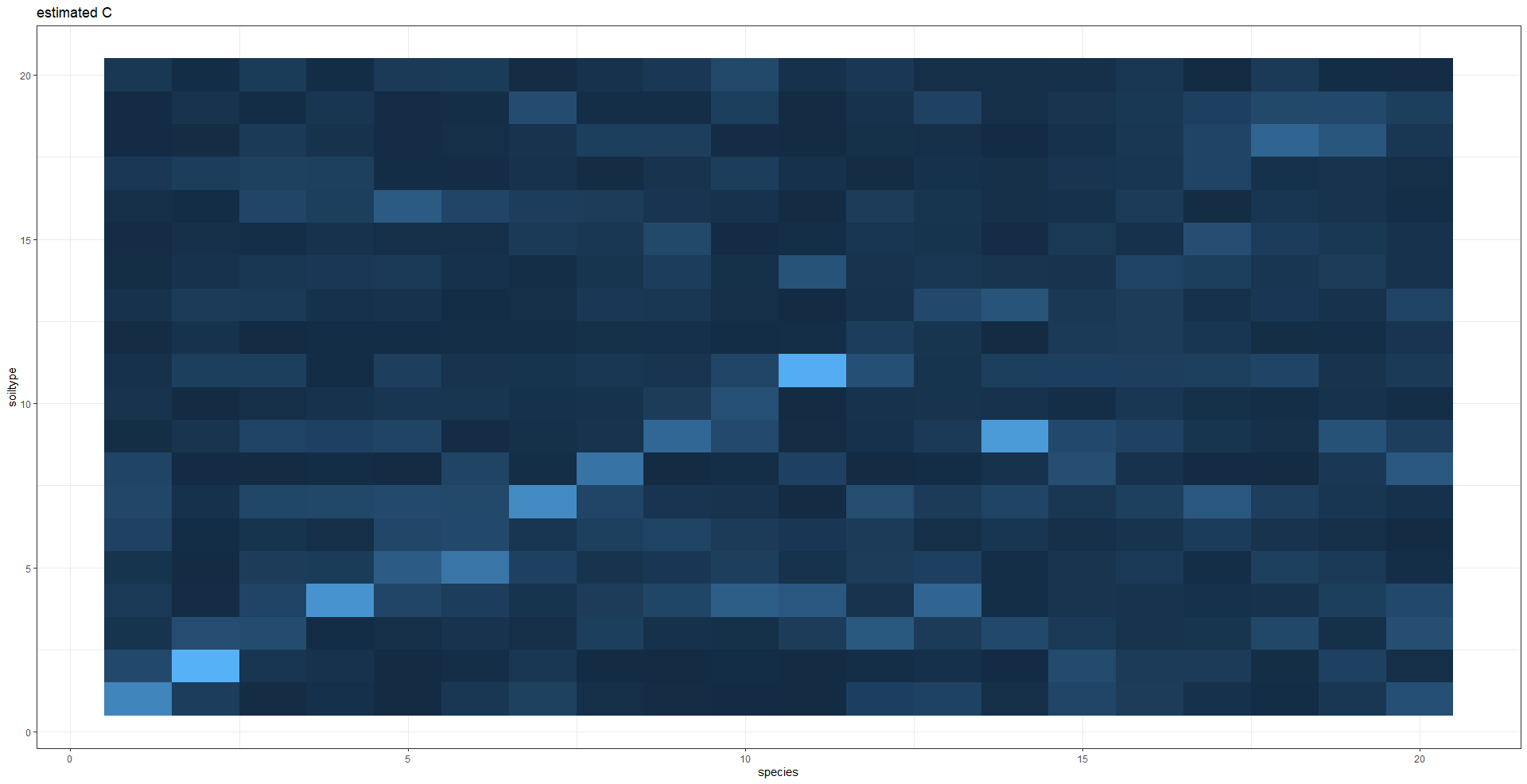


Figure . Same parameters as above, but I ignore the elbow and estimate C using NST = 20 (the true value). It looks closer to the true C but not by a whole lot more than the est C above.



# PAM Results

Figure . S = 20 species, true NST = 5 soil types; landscape autocrrelation parameter = 25. The method works great, and very similar to hierarchical clustering (Fig 1).

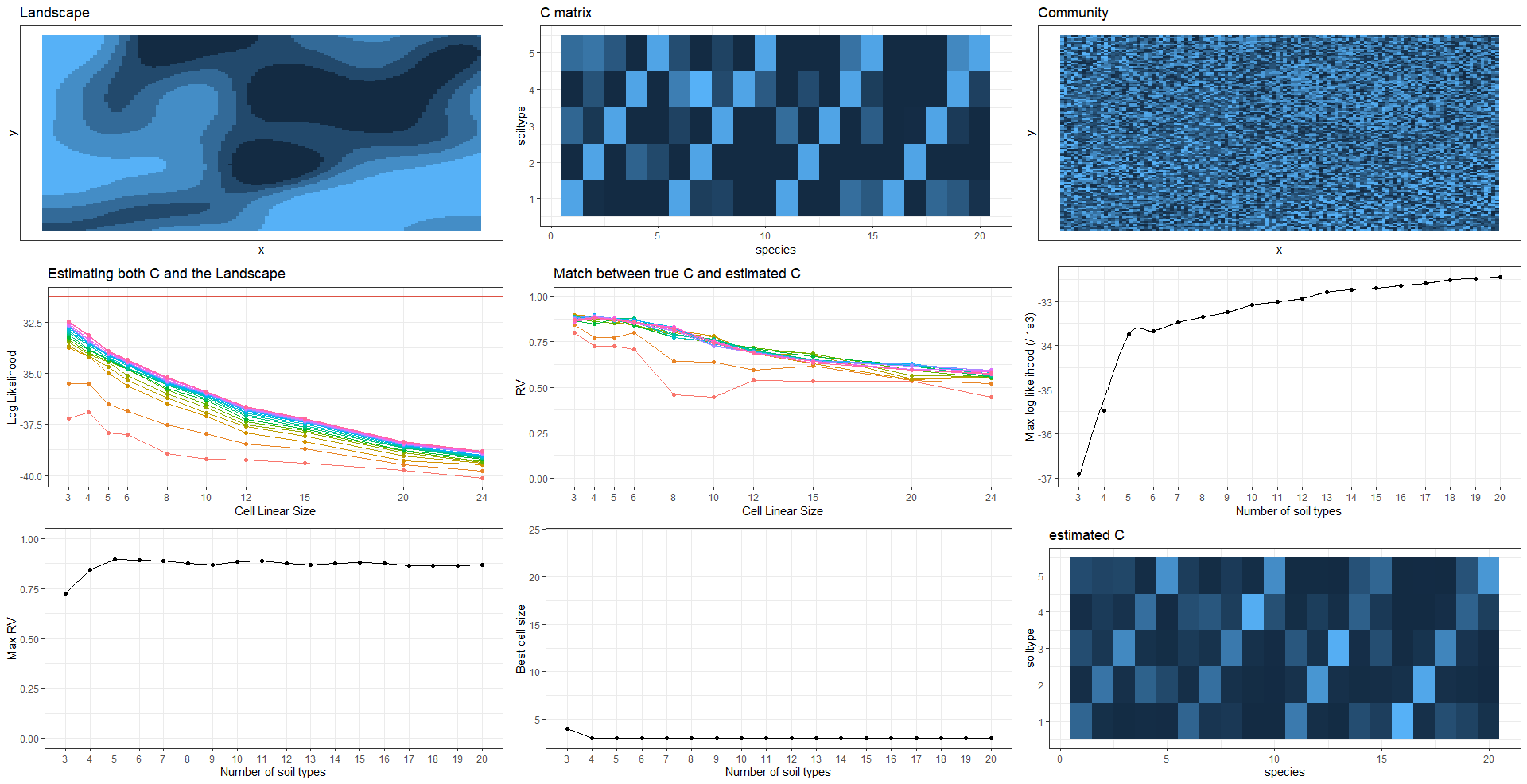


Figure . S = 50 species, true NST = 5 soil types; landscape autocrrelation parameter = 25. Results are similar to Fig. 2.

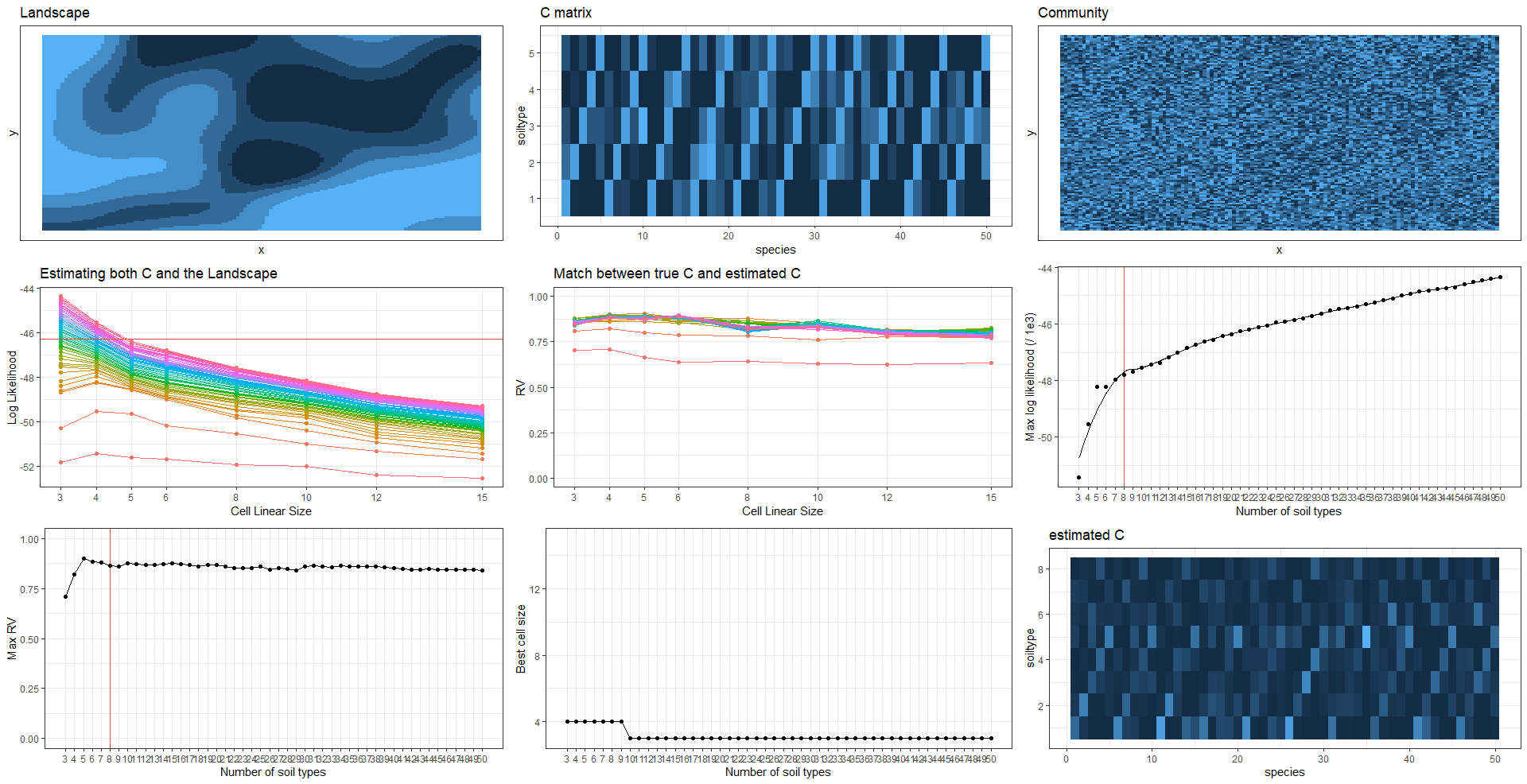


Figure . S = 20 species, true NST = 20 soil types; landscape autocrrelation parameter = 50. Here our method gets lost. Results are similar to Fig 3.

