Landscape Genetics

James O'Reilly
Student Number: `r0773125`

Study the relationships between AR and F_{ST} on the one hand (i.e. the genetic parameters), and the geographical information on the other hand.

1 Statistical Analyses of Barrier Impact

There are a number of statistical analyses which could be used to investigate the impact of barriers on F_{ST}. The natural way to investigate this impact would be to first calculate the pairwise F_{ST} or allelic richness (AR) for each population and then investigate how these affected by both the number and type of barriers between these populations. In order to isolate and determine the effect of these barriers, we will need to account for other geographical variables in our analyses (e.g. distance), which might also effect pairwise F_{ST} and allelic richness.

In order to quickly get a rough idea of the patterns in our dataset, we can use descriptive statistics and plots such as correlation and scatter plots. These analyses aren't statistically robust, and don't prove any causal relationship, but they can point us in the right direction or highlight interesting features. For example, the mean allelic richness of each population is shown in Figure 1. This plot tells us that the upstream populations have similar AR, while the downstream populations are dissimilar, with some populations having higher AR than upstream populations. We can use this to help interpret the results when we model AR as a dependent on a number of other geographical features.

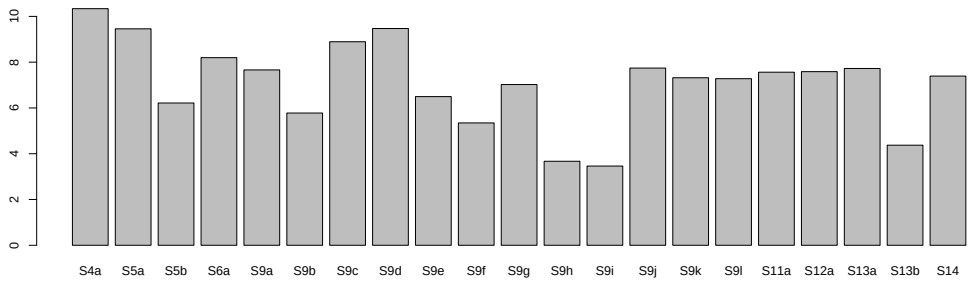

Figure 1: Mean allelic richness for each population.

Before building models for F_{ST} or AR, we should also look at the independent variables and see how they relate to each other. This allows us to detect potential multicollinearity. Table 1 shows the Pearson correlation values and significance values for the `barriers`, `habitat_width`, and `upstream_distance` variables. In each case, the correlations are significant, but not particularly high, which suggests there is not a big problem with multicollinearity.

	Barriers	Habitat Width	Upstream Distance
Barriers	–	0.007	0.0006
Habitat Width	-0.57	–	0.0084
Upstream Distance	-0.69	-0.56	–

Table 1: Correlation and significance values for barriers, habitat width, and upstream distance.

A more rigourous test for multicollinearity is the variance inflation factor (VIF), which takes into account how the independent variables covary with the dependent variable. A high VIF indicates that the

associated independent variable is highly collinear with the other variables in the model. Different papers across statistics, economics, and biology give different limits for an 'acceptable' VIF threshold, with the most common being being $VIF < 5$ or $VIF < 10$. For a simple linear model of AR against `barriers`, `habitat_width`, and `upstream_distance`, the VIF values were 2.1, 1.6, and 2.1, respectively. This further indicates that multicollinearity is not an issue.

We can then investigate the relationship between the dependent variable (allelic richness or F_{ST}) and the independent variables. This can be done visually with a scatterplot matrix (though I don't have space to include it). A more rigourous way to investigate this relationship is using simple and partial Mantel tests. The simple Mantel test checks for correlation between the dependent variable and a single independent variable. A partial Mantel test checks for correlation between the dependent variable and independent variable, after correcting for the effect of a third independent variable. The simple Mantel tests of F_{ST} against `distance` and `barriers` gives significant Mantel scores (using Pearson) of 0.46 and 0.69, respectively. Partial Mantel tests for F_{ST} against `distance` and `barriers` showed that the effect of `barriers` on F_{ST} is still substantial ($r = 0.6023$), even when accounting for the effect of distance. This suggests that barriers, and not distance control the balance of gene flow and drift.

1.1 Spatial Linear Models for Allelic Richness and FST

Another statistical analysis we can perform is to create and compare different linear regression models for AR and F_{ST}. The models are compared using the corrected Akaike Information Criterion (AICc), which weighs the accuracy of the model against the complexity. Note that the 'best' model really depends on our goals. If we are interested in interpretability and determining the relative importance of independent variables, then simple models are better. If we only care about prediction accuracy, then we should use more complex models. Table 2 shows the models considered for allelic richness, along with their AICc values.

Model	AICc	Index
$AR \sim BARRIERS + WIDTH + UDIST$	1.09	1
$AR \sim BARRIERS + WIDTH$	1.60	2
$AR \sim BARRIERS + UDIST$	-1.30	3
$AR \sim WIDTH + UDIST$	11.40	4
$AR \sim BARRIERS$	0.42	5
$AR \sim WIDTH$	19.29	6
$AR \sim UDIST$	11.12	7

Table 2: Comparison of candidate models for AR using AICc. The best performing model is highlighted in blue.

Note that all of the best performing models included the `barriers` variable. Model 3 had the best overall performance. The results of an ANOVA (not included here) on the full model showed that barriers are the only significant predictor of allelic richness. The final model, with coefficients and intercepts, is given by

$$AR \sim 1.076\text{UDIST} + -0.159\text{BARRIERS} + 7.642 \tag{1}$$

More than 15 different linear models were compared for F_{ST}. I don't have space to show every model. Instead I will choose the models which are most interesting and highlight those. Table 3 shows some of the models considered, along with the AICc values.

Looking at Table 3, we see that the best models contain barriers as an independent variable. In fact, the best model contains only barriers as a singular independent variable. Models which do not include barriers score poorly in comparison. The final model, with coefficients and intercepts, is given by

Model	AICc	Index
$F_{ST} \sim BARRIERS + DIST + WIDTH + UDIST$	-57.79	1
$F_{ST} \sim BARRIERS + WIDTH$	-64.09	2
$F_{ST} \sim BARRIERS + UDIST$	-62.07	3
$F_{ST} \sim BARRIERS + DIST$	-62.69	4
$F_{ST} \sim BARRIERS$	-64.81	5
$F_{ST} \sim DIST$	-55.97	6
$F_{ST} \sim UDIST$	-53.91	7

Table 3: Comparison of candidate models for F_{ST} using AICc. The best performing model is highlighted in blue.

$$AR \sim 0.0064\text{BARRIERS} + 0.0451 \tag{2}$$

Performing a multiple regression on the full model and then calculating significance scores via permutation show that barriers are the only significant predictor of F_{ST}. Lastly, we can use the full model to predict pairwise F_{ST} based on geographical features. Plots for the visualising observed and predicted F_{ST} using multi-dimensional scaling are given in Figure 2. It is clear that the prediction is not perfect. The correlation coefficient between observed and predicted F_{ST} is 0.54. This is the proportion of genetic variation which we can explain using the geographical data given in the model.

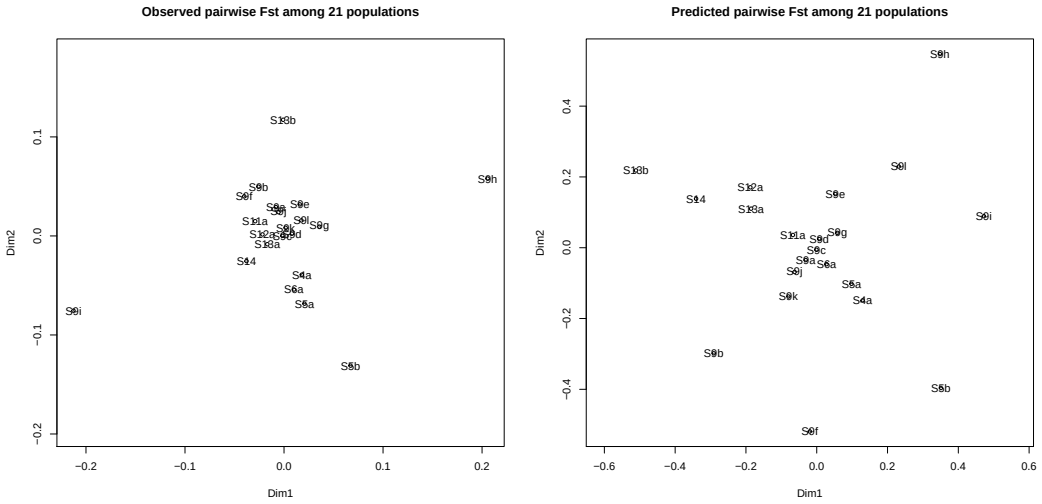

Figure 2: Visualising observed and predicted F_{ST} using multi-dimensional scaling

Do you think there is a straightforward way to measure barrier impact?
Initially, I would have said no. However, it is clear after doing this analysis that it is somewhat straightforward to determine the *relative* impact of barriers on genetic variation between populations, when compared to other factors. However, this only gives the *relative impact*, and there are many other geographic and non-geographic variables which could be included in this analysis. In that sense, it's not so straightforward, simply due to the number of possible variables which also affect genetic variance.

Do you have suggestions for improving the study? In order to improve the study, I would try to determine the effects of specific types of barriers and barrier height. One could also try to include more geographical variables in the study, to try to explain a greater percentage of the genetic variance. It might also be possible to include interactions between geographical variables in the linear models.

References

[1] Raeymaekers, J. A., Maes, G. E., Geldof, S., Hontis, I., Nackaerts, K., and Volckaert, F. A. Modeling genetic connectivity in sticklebacks as a guideline for river restoration. *Evolutionary applications 1*, 3 (2008), 475–488.