Simulation Write Up

Amy Solman

June 12, 2020

1 Aims

To test the hypothesis that the predictions made by Chisholm's model are significantly statistically similar to the results of the island simulation.

To test the hypothesis that, as island area increases, it transitions from a niche-structured regime, to a colonisation-extinction balance regime.

2 Methods

2.1 Simulation Design

2.1.1 Overview

This simulation is designed to mimic the process of island colonisation from a metacommunity. The colonisation process is constrained by a fixed speciation rate within the local community, as the metacommunity is considered to be static throughout the simulation. The colonisation process is also constrained by migration rate . Each island is characterised by number of niches and the size of each niche. The output of the simulation is the final community of each island as well as a timeseries of species richness.

2.1.2 Metacommunity

A metacommunity is generated at the beginning of the simulation using *coalescence_test* function, provided by Dr James Rosindell. The function takes the input parameters: metacommunity size($J\_meta = 10\ 000$) and speciation rate ($nu = 0.001$). It initialises a vector (*lineages*) of length = *J_meta* with 1 as every value. An empty vector (*abundances*) is initialised. The value of *J_meta* is given to *N*. *Theta* is calculated as *nu*(J_meta-1)/(1-nu)*. Then, while *N > 1*, a vector (*linvect*) is created with values 1:length(*J_meta*). A random sample of *linvect* is made (j). A random decimal number is selected between 0 and 1 (*randnum*). If *rannum* is less than *theta/(theta+N-1)*, then the value at *lineages[j]* is appended to *abundances*. Else, another random number (i) is sampled from *linvect*, excluding the last number selected. The values at *lineages[i]* and *lineages[j]* are summed and take the position of *lineages[i]*. *lineages[j]* is then removed from *lineages*, so the vector is one value shorter. The value of N is also decreased by 1. This repeats until $N = 1$. The remaining value in the *lineages* vector is added to *abundances* and the function outputs a vector of simulated species abundances.

The *coalescence_test* function is incorporated into a second function (*metacommunity*), that generates a vector of individuals from the abundance vector. For example, *abundances*(5,4,2,2,1,1) would generate a community *meta*(1,1,1,1,1,2,2,2,2,3,3,4,4,5,6) where each unique number value represents a unique species.

2.1.3 Parameters

The variable parameters of the simulation are: migration rate (range: 0.003-0.06), number of niches (range: 1-20), size of niches (range:1-20). Each unit of space was assumed to host one individual, therefore, number of niches x size of niches = area = size of island population. There are a total of 8000 condition combinations applied during the simulation (20 migration rates, 20 number of niches, 20 size of niches).

2.1.4 Simulation Logic

At timestep i an island is selected. The first niche on that island becomes the focal niche. An individual within that niche is chosen to die. With probability nu (speciation rate), the dead individual is replaced with a new, unique species. With probability m (migration rate), the dead individual is replaced with a randomly chosen propagule from the metacommunity. With probability 1 - nu + m, the dead individual is replaced with a local propagule. The simulation then moves to the next niche on that island. When all niches have been simulated for timestep i, the simulation moves to the next island. When all islands have been simulated for timestep i, the simulation moves to the next timestep i + *1*, and returns to the first island (Figure 1). The total species richness across all niches for each island is calculated and stored at every 10 000 timesteps.

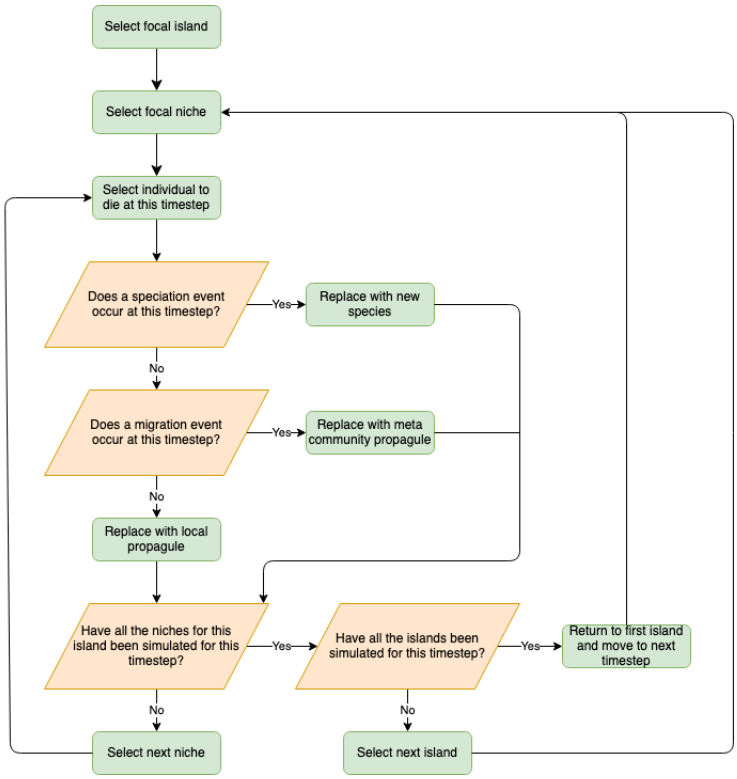


Figure 1: Flowchart of simulation design

2.1.5 High Performance Computing

The metacommunity and simulation code are contained in *ClusterSim.R*. *ClusterSim.R* functions are sourced by *ClusterCode.R* and given the input parameters: *J_meta* = 10 000, *nu* = 0.001, *num_m_rates* = 20, *max_k_num* = 20, *max_k_size* = 20, *wall_time* = 1380, *output_file_name* = output_file_name (where each simulation is given a unique file name "simulation_timeseries_*i*". *ClusterRun.sh* is used to run on the cluster, with a time limit of 24:00:00. *wall_time* is given as 1380 minutes (23 hours) within the function, to ensure

all simulations are completed before the cluster run ends. 100 parallel simulation were run on the Imperial College London High Performance Computing service. This generated a total of 800 000 islands, simulated for $>$ 50 000 timesteps.

2.2 Data Preparation and Timeseries Plots

The 100 simulation results were imported into *DataPrep.R*. The data from each island was isolated and configured into a data frame with simulation number, migration rate, area, number of niches and number of species (*SimModelFitData.csv*). A second data frame was generated for timeseries plotting, with simulation number, island number, migration rate, timestep and species richness timeseries for each island (*SimTimeseriesPlotData.csv*).

To ensure the simulation had run long enough for each island to reach dynamic equilibrium the species richness timeseries of simulations 25, 50 and 75 were plotted(*TimeseriesPlot.R*)(Figure 2).

2.3 Analysis

```
chisholm_model <- function(area, theta, m0, K) {
  rho = 1
  K = K
  Js = area*rho
  J_stars = Js/K
  ms = m0/sqrt(area)
  gamma_stars = J_stars*ms/(1-ms)
  return(theta*(digamma(theta/K+gamma_stars*
  (digamma(gamma_stars+J_stars)-digamma(gamma_stars)))-digamma(theta/K)))
}
```

An analysis script (*Analysis.R*) imported the prepared data (*SimModelFitData.csv*) for each island across all simulations (800 000). Model estimated species richnesses were generated by giving the island parameters (*m0 = m*sqrt(area)*) and an estimated *theta* (*theta = 2*J*nu*, where J = area) for each island to the *chisholm_function* (above). The results of the simulation and those estimated by the *chisholm_function* were bound together in a data frame. Mean species richness results for each combination of island area, migration rate and number of niches across all 100 simulations was calculated and stored.

2.3.1 Within Simulation Analysis

Repeatability The repeatability of the simulation was assessed by running a linear model of species richness, with simulation number as independent variable. The repeatability statistic was calculated by running an ANOVA on the model and finding the among simulation and within simulation variabilities.

Normality The normality of the simulation data was assess by generating histograms of island species richnesses. It was necessary to check the normality of my data as linear regression assumes normality.

Collinearity The collinearity of the simulation independent variables was assessed by using the R base function pairs() to produce a scatterplot of matrices. This allowed for visual inspection of the relationships between migration rate, island area and number of niches. It is important to assess the collinearity of covariates because it can inflate variation. Large amounts of collinearity cause covariates to have larger standard errors and it becomes less likely to detect a significant result. The Variance Inflation Factor (VIF) was calculated to ensure the collinearity between number of niches and island area was within acceptable limits for this analysis.

Linear Regression Linear regression was carried out to investigate whether there was a statistically significant relationship between the dependent variable (species richness) and independent variables (area and niches). Migration rate was excluded from the linear regression analysis as previous collinearity statistics indicated it was not significant in predicting species richness (this is later confirmed in multivariate analysis). The mean species richness for each unique set of parameters across all 100 simulations was calculated. Area and niche variables were then z-transformed so that they could better meet the assumptions of a linear model and the intercept statistic would then give the mean of the dependent sample. Histograms of the residuals of each linear model were generated to check for homogeneity of the variances, to validate the assumptions of the regression analyses. The models were validated by visual inspection of the distribution of the residuals.

Multivariate Analysis Multivariate analysis of the mean simulation data, including number of niches, island area and migration rate was performed to investigate to what extent the three independent variables predicted species richness. A second multivariate analysis was carried out, excluding migration rate. Model results were plotted to check the distribution of the residuals.

2.3.2 Simulation/Model Comparison

Outliers Both the simulation dataset and model estimated species richnesses were checked for outliers using boxplots. This ensured analysis was not skewed by anomalous results.

Mean, range, variance, standard deviation and standard error The mean, range, variance, standard deviation and standard error of the simulation and the *chisholm_function* estimated species richnesses were calculated.

Paired-sample t-Test A paired-sample t-test was used to assess whether the mean difference between the simulation results and *chisholm_function* estimates was zero. A paired-sample t-test was used because both sets of species richness data were calculated from the same set of parameters (migration rate, number of niches, island area) and were directly comparable across pairs.

Non-Linear Least Squares Fitting The model was also fit to the data using non-linear least squares (NLLS) fitting. The 800 000 simulated islands were subset by migration rate and number of niches (400 combinations). For each of the 400 datasets, the mean species richness for each island area was calculated. The *chisholm_function* was then fit to the mean species richness dataset for each unique combination of parameters using NLLS fitting.

2.4 Repeat Analysis with Reducing Data Set

All analysis was repeated 20 times whilst reducing the dataset by limiting maximum island area (Table 1). This allowed for comparative multivariate analysis, to ascertain if the relative significance of the independent parameters changed as island area decreased. This was also useful in assess whether the model was better at predicting species richness on smaller or larger islands.

All analysis and plotting were carried out using R 3.6.0.

3 Results

3.1 Within Simulation Results

3.1.1 Timeseries

3 of 100 simulation were selected and plotted as a timeseries (Figure 2). From this plot it was determined that the islands reached dynamic equilibrium at approximately 25 000 timesteps. The simulation had run for enough time to reach dynamic equilibrium.

Table 1: Reducing dataset to smaller islands

Max Island Size	Num. Islands
400	800000
200	660000
133	534000
100	452000
80	392000
66	340000
57	306000
50	274000
44	248000
40	236000
36	216000
33	194000
30	182000
28	170000
26	158000
25	154000
22	140000
22	140000
21	136000
20	132000

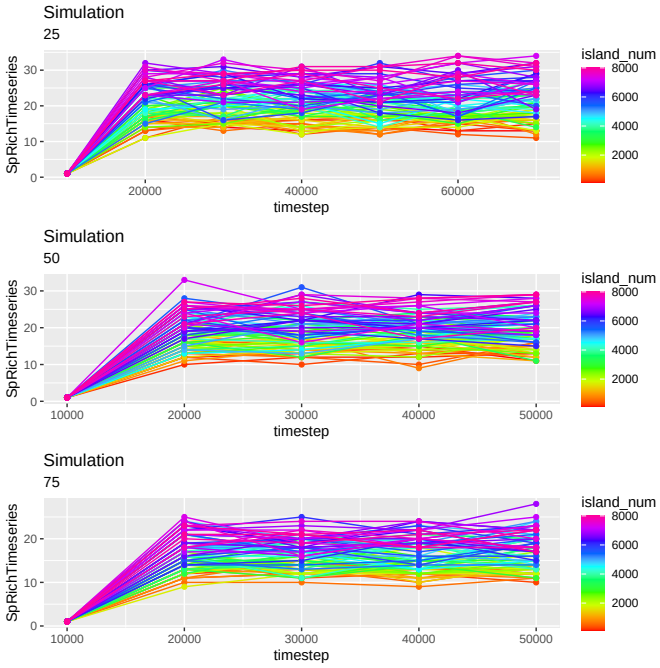


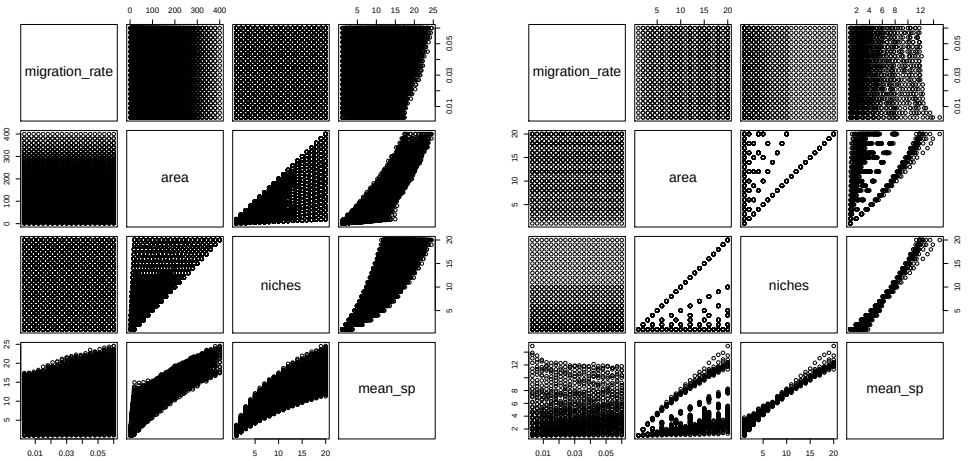
Figure 2: Timeseries plot of three simulations

3.1.2 Repeatability

The mean sum of squared variance among simulations was 8648.9 (Vg). The mean sum of squared variance within simulations was 30.3 (Vr). The repeatability of the simulations was calculated as Vg / (Vg + Vr) = 0.997. $>$ 99% of the variation in species richness is determined by between-simulation differences. The

simulations are giving consistent values of species richness, indicating the method is robust.

3.1.3 Collinearity



(a) maxArea = 400 units (b) maxArea = 20 units

Figure 3: Collinearity plots for full data set (a) and reduced data set (b)

The collinearity statistic for migration rate and area, as well as migration rate and niches was 0. A lack of collinearity between migration rate and the other covariates is to be expected as they are not logically linked in the simulation, nor would they be in a real-world scenario. Area and niches had a collinearity score of 0.659. The VIF score for collinearity between niches and area was 1.77. The standard errors of number of niches are therefore inflated by $\sqrt{1.77} = 1.33$, which is within an acceptable range.

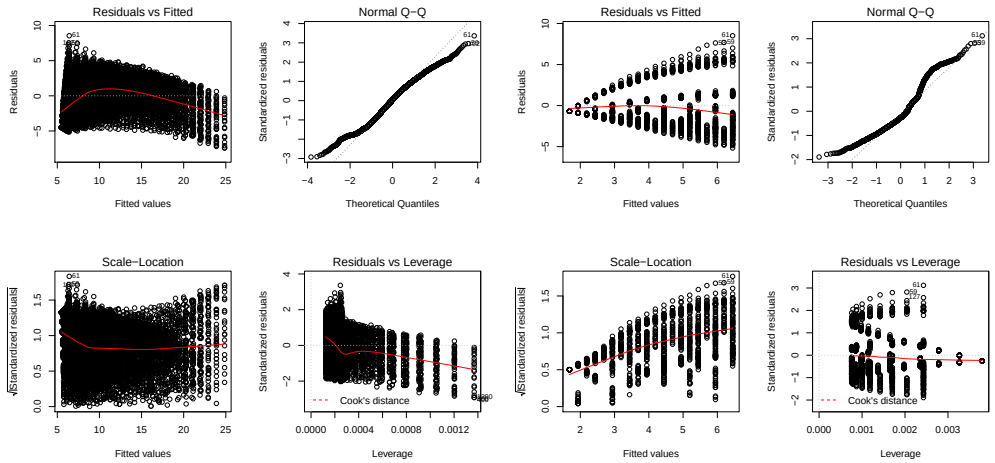
The collinearity statistic for migration rate and area, as well as migration rate and niches in the most reduced dataset was 0. Area and niches had a collinearity score of 0.355 and VIF score of 1.14. The standard errors of number of niches are therefore 1.07. This is less than that of the whole dataset and therefore in an acceptable range. Visual inspection of pairs plots (Figure 3) supported the collinearity statistics.

3.1.4 Linear Regression Analysis

Table 2: Linear regression results for z-transformed area and niches

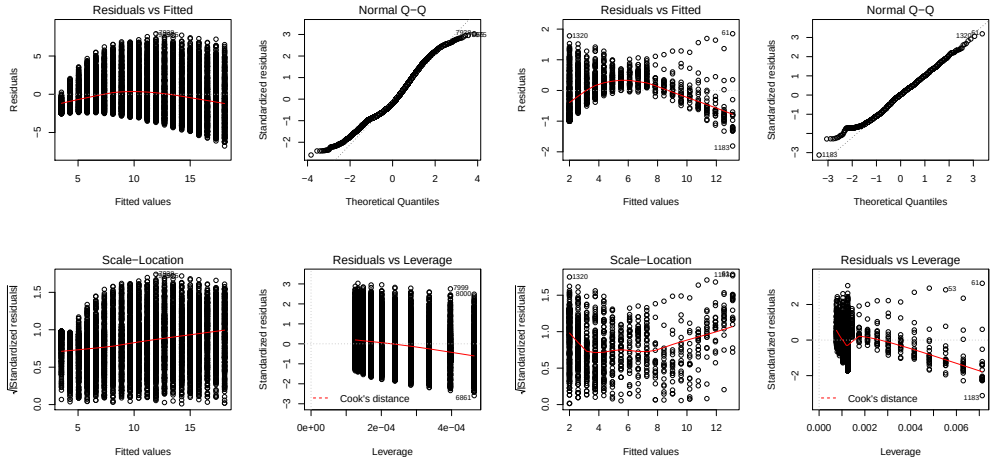
maxArea	Coefficients	Estimate	Standard.Error	t.value	p.value	R.Squared	Variate
400	(Intercept)	10.793	0.028	379.953	< 0.001	0.755	Area
400	z.area	4.458	0.028	156.928	< 0.001	0.755	Area
400	(Intercept)	10.793	0.029	369.481	< 0.001	0.741	Niches
400	z.Knum	4.416	0.029	151.172	< 0.001	0.741	Niches
20	(Intercept)	4.419	0.075	58.742	< 0.001	0.2	Area
20	z.area	1.365	0.075	18.146	< 0.001	0.2	Area
20	(Intercept)	4.419	0.016	275.93	< 0.001	0.964	Niches
20	z.Knum	2.998	0.016	187.161	< 0.001	0.964	Niches

The linear model results of species richness and z-transformed area gave an r-squared value of 0.755 with a p-value < 0.001 (Table 2). This indicated that there was a positive, statistically significant relationship between area and species richness for the entire dataset.



(a) maxArea = 400 units (b) maxArea = 20 units

Figure 4: Model validation: Diagnostic plots of area and species richness linear models with full data set (a) and reduced data set (b)



(a) maxArea = 400 units (b) maxArea = 20 units

Figure 5: Model validation: Diagnostic plots of niche and species richness linear models with full data set (a) and reduced data set (b)

Visual inspection of residual vs fitted results for the entire dataset shows somewhat linear and constant variance (Figure 4 (a)). The normal Q-Q plot showing standardised residuals plotted against theoretical quantiles falls along a relatively straight line, with some deviation at the highest and lowest values. Diagnostics indicate the model works fairly well for the data.

The r-squared value for the area/species richness linear model reduced from 0.755 to 0.2 as maximum island area was reduce from 400 units to 20 units. This indicated that area had a weaker positive relationship with species richness as island area decreased (Figure 6).

Visual inspection of the residual vs fitted plot for the most reduced dataset showed fairly linear but non-constant variance (Figure 4 (b)). Higher fitted values had a greater residual range than lower fitted values.

The normal Q-Q plot followed a relatively straight line, with some deviation at the lowest values. Diagnostics indicate the model works somewhat well for the data at lower island areas.

The linear model result of species richness and z-transformed niches gave an r-squared value of 0.741 with a p-value < 0.001 (Table 2). This indicated that niches had a positive, statistically significant relationship with species richness.

Visual inspection of the residual vs fitted plot for the entire dataset showed fairly linear but non-constant variance (Figure 5 (a)). Higher fitted values had a greater residual range than lower fitted values. The normal Q-Q plot followed a relatively straight line, with some deviation at the highest and lowest points. Diagnostics indicate the model works fairly well for the data.

The r-squared value for niches/species richness linear model increased from 0.741 to 0.964 as maximum island area was reduced from 400 units to 20 units (Table 2). This indicated that number of niches had a stronger positive relationship with species richness as island area decreased (Figure 6).

Visual inspection of the residual vs fitted plot for the most reduced dataset (Figure 5 (b)) showed weakly linear variance. The normal Q-Q plot followed a relatively straight line, with some deviation at lower values. Diagnostics indicate the model worked somewhat well for the reduced dataset.

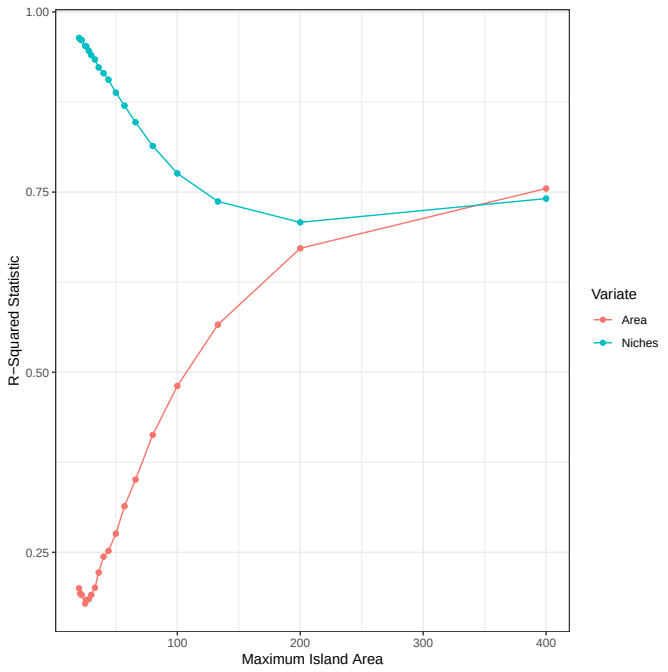


Figure 6: Change of R-squared statistic for number of niches and island area, with increasing island size

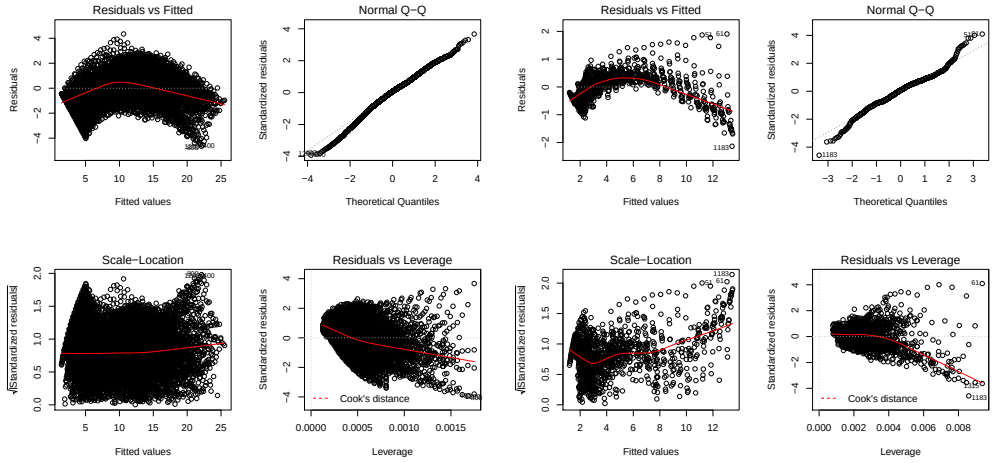
Linear regression analysis for the reducing datasets indicated that at 350 units the islands transitioned from a niche dominated regime to an area dominated regime (Figure 6).

3.1.5 Multivariate Regression Analysis

Multivariate analysis results for the whole dataset showed that the combined effects of number of niches, island area and migration rate gave an r-squared value of 0.947 (Table 3). The combined effects of niche and area gave an r-squared value of 0.902. This indicates that migration rate had a weak positive effect on species richness across the entire dataset.

Table 3: Multiple regression analysis for z-transformed migration, area and niches

maxArea	Coefficients	Estimate	Standard.Error	t.value	p.value	R.Squared	Variate
400	(Intercept)	10.793	0.013	814.755	< 0.001	0.947	NicheAreaMigration
400	z.Knum	2.613	0.018	148.326	< 0.001	0.947	NicheAreaMigration
400	z.area	2.736	0.018	155.289	< 0.001	0.947	NicheAreaMigration
400	z.migration	1.091	0.013	82.343	< 0.001	0.947	NicheAreaMigration
400	(Intercept)	10.793	0.018	599.386	< 0.001	0.902	NicheArea
400	z.Knum	2.613	0.024	109.118	< 0.001	0.902	NicheArea
400	z.area	2.736	0.024	114.241	< 0.001	0.902	NicheArea
20	(Intercept)	4.419	0.013	342.973	< 0.001	0.977	NicheAreaMigration
20	z.Knum	2.876	0.014	208.599	< 0.001	0.977	NicheAreaMigration
20	z.area	0.343	0.014	24.881	< 0.001	0.977	NicheAreaMigration
20	z.migration	0.13	0.013	10.06	< 0.001	0.977	NicheAreaMigration
20	(Intercept)	4.419	0.013	330.626	< 0.001	0.975	NicheArea
20	z.Knum	2.876	0.014	201.089	< 0.001	0.975	NicheArea
20	z.area	0.343	0.014	23.986	< 0.001	0.975	NicheArea



(a) maxArea = 400 units (b) maxArea = 20 units

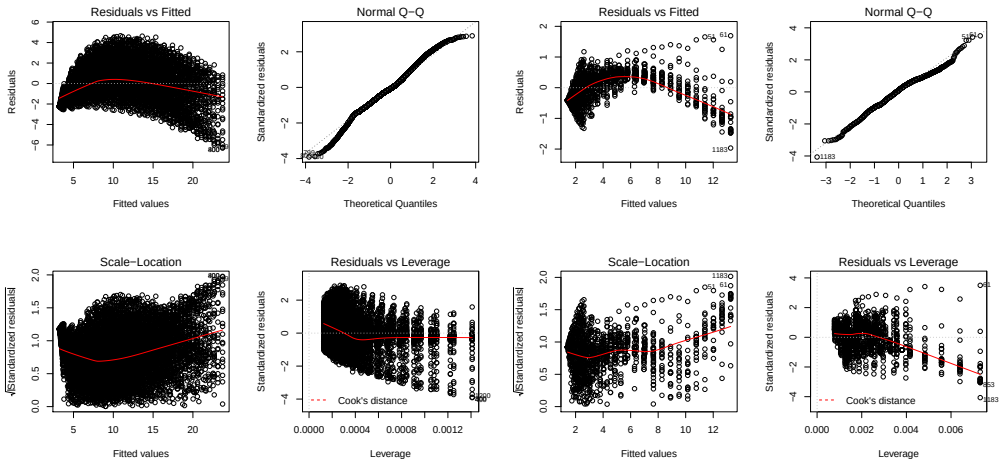
Figure 7: Model validation: Residuals plots of multivariate analysis of niche, area, migration and species richness with full data set (a) and reduced data set (b)

Visual inspection of the residual results of the full dataset for migration/area/niche multivariate analysis showed non-linear but fairly evenly distributed variance (Figure 7 (a)). The Q-Q plot follows a relatively straight line.

Visual inspection of the residual results of the full dataset for area/niche multivariate analysis showed fairly linear, fairly constant variance (Figure 8 (a)). The Q-Q plot follows a relatively straight line.

Multivariate analysis results for the most reduced dataset showed that the combined effects of number of niches, island area and migration rate, gave an r-squared value of 0.977 (Table 3). The combined effects of niche and area gave an r-squared value of 0.975, indicating that migration rate had almost no effect on species richness at smaller island area.

Visual inspection of the residual results of the reduced dataset for migration/area/niche multivariate analysis showed non-linear, non-constant variance (Figure 7 (b)). The Q-Q plot follows a relatively straight line with some deviation at the lowest and highest points.



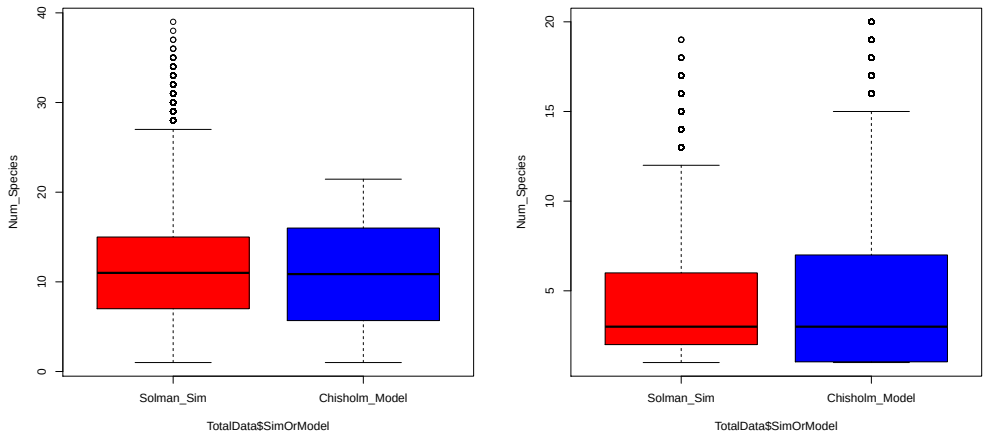
(a) maxArea = 400 units (b) maxArea = 20 units

Figure 8: Model validation: Diagnostic plots of multivariate analysis of niche, area and species richness with full data set (a) and reduced data set (b)

Visual inspection of the residual results of the reduced dataset for area/niche multivariate analysis showed non-linear, non-constant variance (Figure 8 (b). The Q-Q plot follows a relatively straight line.

3.2 Simulation/Model Comparison

3.2.1 Outliers



(a) maxArea = 400 units (b) maxArea = 20 units

Figure 9: Boxplot of simulation and model estimation data for the full dataset (a) and area reduced dataset (b)

Comparison of boxplots of simulated and model estimated species richnesses for the full dataset show outlying simulation data points (Figure 9 (a)). On examining the data, the plotted points represented 1054 islands that had species richness values >= 29. As the number of islands present in this outlier group was so high, it was inappropriate to remove them from the data set as they represent a significant contribution to the

overall mean.

Boxplots of the reduced data indicate outliers in both the simulation and model estimate datasets (Figure 9 (b)). 3370 islands in the simulation data set had species richness >= 13 and 10 000 islands in the model estimation dataset had species richness >=16. The outliers do not appear to be random deviates from the data, but represent a considerable proportion of the overall dataset and are not removed from this analysis.

3.2.2 Mean, range, variance, standard deviation and standard error

Table 4: Mean, range, variance, SD and SE for simulation and model estimates

Mean	Range	Variance	Standard.Deviation	Standard.Error	maxArea	ModelOrSim
10.793	1 - 39	31.322	5.597	0.006	400	SolmanSim
10.71	1 - 21.45	34.647	5.886	0.007	400	ChisholmMod
4.419	1 - 19	10.456	3.234	0.009	20	SolmanSim
5.146	1 - 20	26.206	5.119	0.014	20	ChisholmMod

For the full dataset, the mean number of species on simulated islands was 10.79 (SD 5.60, SE 0.006, range: 1-39) (Table 4). The mean number of species estimated by the model was 10.71 (SD 5.89 , SE 0.007, range: 1-21.45). For the reduced island area dataset, the mean number of species on simulated islands was 4.42 (SD 3.23, SE 0.009, range: 1-19). The mean number of species estimated by the model was 5.15 (SD 5.12, SE 0.014, range: 1-20).

3.2.3 Paired-sample t-Test

Table 5: Paired-sample t-test for simulation data and model estimations

Mean.Difference	t.value	p.value	DF	conf..low	conf..high	Method	Alternative	maxArea
0.08	20.64	< 0.001	799999	0.08	0.09	Paired t-test	two.sided	400
-0.73	-108.12	< 0.001	131999	-0.74	-0.71	Paired t-test	two.sided	20

The mean difference between simulation results and model estimations for the full dataset was 0.08, with a t-value of 20.64 and a p-value of < 0.001 (DF 799999, CI low 0.08, CI high 0.09). The mean difference between simulation results and model estimations for the most reduced dataset were -0.73, with a t-value of -108.12 and a p-value of < 0.001 (DF 131999, CI low -0.74, CI high -0.71). In both cases, $p < 0.001$, therefore we reject the null hypothesis that there is no difference in the simulated and model estimated datasets. However, the mean differences are extremely small and in real-world terms represent < species, a logical impossibility.

3.2.4 Non-Linear Least Squares Fitting

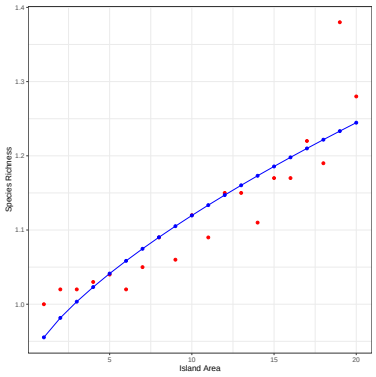
NLLS fitting was used to fit the *chisholm_function* to 400 subsets of island data (unique migration rate + unique niche number). This was to test whether, through NLLS fitting, the model could produce robust parameter estimates of the known migration and niche values for each dataset. 396 out of 400 model fits had a r-squared value of > 0.9. The remaining four fits had r-squared values of between 0.789 and 0.875. This indicates that for all datasets the model fit the data well.

Table 6 shows parameter estimates and statistical values for NLLS fitting 1, 190 and 400. All three show high r-squared values with significant niche parameters ($p < 0.001$). Migration rate is significant for fits 190 and 400 ($p < 0.001$). *Theta* estimates for all three fits are much higher than realistically possible. *m0* estimations are within the range of true values where ($m = m0/sqrt(area)$). Niche (K) estimations are more acurate for islands with lower number of niches (< 10), but fail to achieve accurate estimates for islands

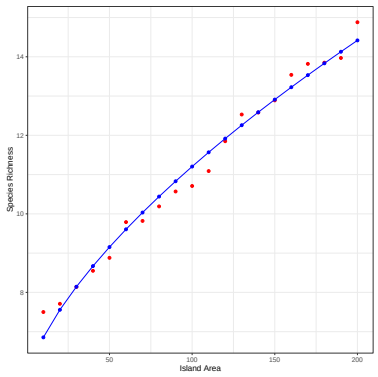
Table 6: NLLS fitting of Chisholm model to simulation data, parameter statistics for 3 out of 400 fits

fit.num	term	Estimate	std.error	Statistic	p.value	r.squared	migration.rates	Niches
1	theta	830.814	448295.261	0.002	0.999	0.789	0.003	1
1	m0	0.019	0.007	2.521	0.022	0.789	0.003	1
1	K	0.954	0.027	35.316	0	0.789	0.003	1
190	theta	39104.546	1019087.267	0.038	0.97	0.983	0.03	10
190	m0	0.155	0.005	30.586	0	0.983	0.03	10
190	K	6.553	0.191	34.355	0	0.983	0.03	10
400	theta	24571.929	1454366.305	0.017	0.987	0.997	0.06	20
400	m0	0.183	0.006	29.073	0	0.997	0.06	20
400	K	10.731	0.151	70.938	0	0.997	0.06	20

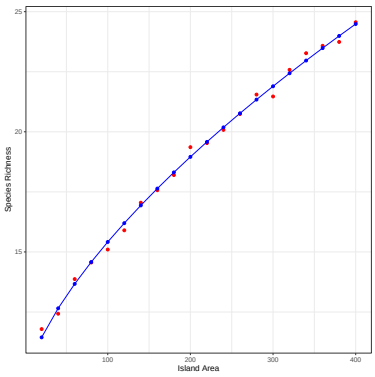
with >= 10 niches. Visual inspection of the direct model estimations and NLLS fitting reveal both were successful in predicting species richness on the simulated islands (Figure 10).



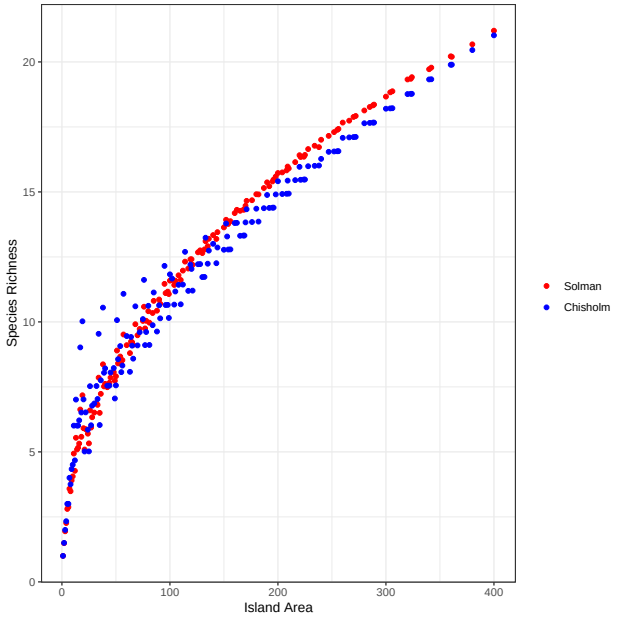
(a) NLLS Plot 1



(b) NLLS Plot 190



(c) NLLS Plot 400



(d) Direct estimations of species richness

Figure 10: Comparison of model fitting directly using the *chisholm_function* on mean dataset (d), and using NLLS fitting for subsets of data (a, b, c)