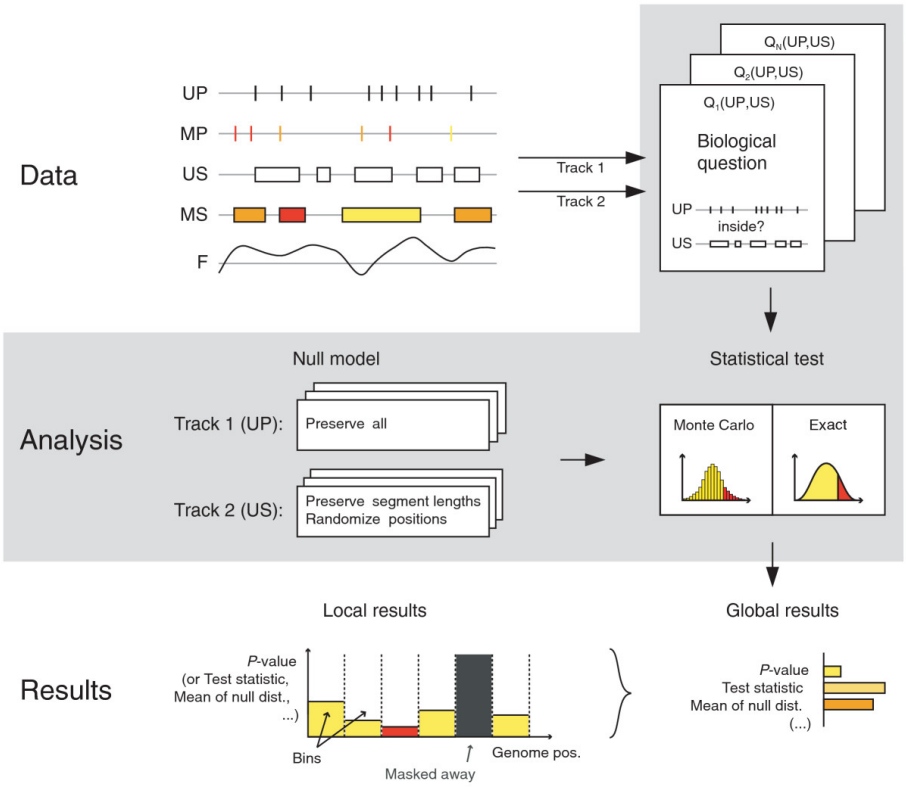
Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Monte Carlo null models in ecology



Note no SAMBA/20/13

Authors Egil Ferkingstad
Lars Holden
Geir Kjetil Sandve

Date 5th September 2013

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Monte Carlo null models in ecology**
Authors	**Egil Ferkingstad** <egil.ferkingstad@nr.no> **Lars Holden** <lars.holden@nr.no> **Geir Kjetil Sandve** <geirksa@ifi.uio.no>
Date	5th September 2013
Publication number	SAMBA/20/13

Abstract

We present a discussion on null models for Monte Carlo randomization tests in ecology, particularly for species competition problems. This report is a companion paper (supplementary material) to the article "Monte Carlo null models in genomics" (Ferkingstad et al., 2013).

Keywords	Monte Carlo methods, hypothesis testing, ecology
Target group	Researchers
Availability	Open
Project	SFI 20 GenomeBrowser
Project number	220372
Research field	Statistics
Number of pages	6

An ecological presence–absence matrix has a row for each species and a column for each site, where the value of an element in the matrix is $a_{ij} = 1$ if the species i is present at site j, and 0 otherwise. According to MacArthur competition theory, co-occurrence of different species at the same location is rarer than independence, and testing this hypothesis is a topic of great interest in ecology (Gotelli, 2000). The Monte Carlo testing approach is indispensable for this type of problem, but for more than two decades the choice of randomization null model has been fraught with controversy (Gotelli and Graves, 1996; Manly, 2007). At least nine different null models may be defined by keeping the row/column sums fixed, or proportional with the observed values, or equiprobable. In addition, it is an open question whether one should remove elements in the state space where row or column sums are equal to zero.

A general observation is that when increasing the randomness of the null model, the p-values decrease. However, we will show that for extreme values of the test statistics we can get the opposite ordering. We will illustrate ordering of null models using one synthetic and one real example. Similar types of data and randomization techniques are also found in the data mining literature, see Gionis et al. (2007).

We test whether the species avoid competition using the checkerboard score

$$C = \sum_{i,j} 2(S_i - Q_{ij})(S_j - Q_{ij})/R(R-1),$$

where the sum is over all combinations of rows i and j, S_i is the number of 1's in the row, Q_{ij} is the number of sites with a 1 in both row i and j, and R is the number of rows (Stone and Roberts, 1990). This score measures the tendency for "checkerboard" patterns to appear in the presence–absence matrix, giving a measure of competition avoidance.

To provide a realistic test of species competition, it can be argued that row and/or column sums should be preserved. Some species are much more common than others. This leads to large variation in the row sums. Correspondingly, if some sites have larger number of species than the other sites, then column sums vary. This may be considered a fundamental property of the phenomenon, that must be preserved in the randomization (Gotelli, 2000).

Consider the following three permutation algorithms:

R,C: permute, but preserve row and columns sums

R: permute, but preserve row sums

F: permute freely

and their corresponding null models $P_0^{(R,C)}$, $P_0^{(R)}$ and $P_0^{(F)}$. The three algorithms defined above are null ordered, that is, $P_0^{(R,C)} \leq_0 P_0^{(R)} \leq_0 P_0^{(F)}$. It is easily seen that we obtain the maximum value of the checkerboard score under full permutations for particular patterns of the row and column sums. Correspondingly, we obtain the maximum value of the checkerboard score for fixed row sums, for very particular row sums. If we assume

Table 1. Presence–absence matrix for West Indian finches. Missing entries denote zeroes. Rows represent species, while columns represent islands. The last row and column show the row and column sums, respectively, while the lower right entry shows the total sum. Labels are omitted, as they are unimportant for this discussion. For further details, see Gotelli (2000).

	1																	1
	1																	1
										1								1
										1								1
1													1					2
			1															1
	1	1																2
				1	1	1	1	1	1	1	1			1	1	1	1	12
							1											1
1	1	1	1										1					5
	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	17
1																			1
		1																	1
				1	1	1	1												4
1																			1
	1	1	1																3
	1																		1
4	7	5	4	3	3	3	4	2	2	4	2	1	2	2	2	2	2	1	55

that X does not have these particular row and column sums, then we have

$$\max\{R_0^{(R,C)}\} < \max\{R_0^{(R)}\} < \max\{R_0^{(F)}\}$$

where $R_0^{(A)}$ represents the range for the checkerboard score, starting with the matrix X, and where A represents one of the three algorithms. This means that we could, in some cases, have a p-value ordering where a higher level of null model preservation leads to lower p-values.

Gotelli (2000) studies a presence–absence matrix for West Indian finches on islands in the Pacific. The matrix is displayed in Table 1. A Monte Carlo test using this matrix is illustrated in Table 2. The matrix has 17 rows, 19 sites, 55 non-zero elements and a checkerboard score equal to 3.8. There are large differences among row sums. This indicates that some species are much more common than others, and the probability for a value $a_{ij} = 1$ is much larger for some rows than for other rows. This makes it difficult to argue for random permutations of values between the rows. The column sums are quite equal and typical for a random permutation between the columns, when the row sums are fixed.

The null model using full permutations and the null model keeping row and column sums fixed both gave an estimated p-value equal to zero (based on 10,000 samples). Note that the sampled values are all larger than the observed value in the former case, while

Table 2. Checkerboard score for three different algorithms using data from the finches–island matrix

Algorithm	Min	$q_{0.05}$	Mean	$q_{0.95}$	Max
Full permutation	4.5	6.1	6.8	7.4	7.9
Row sum fixed	1.6	2.4	2.9	3.4	4.0
Row, column sum fixed	2.3	2.5	2.7	2.9	3.6

they are all smaller than the observed value in the latter case. The state space with row sums fixed gave a p-value equal to 0.0004. In our opinion, this last p-value is the most reasonable, since the column sums in our dataset are typical for this state space. Using the state space with both row and column sums fixed gives very little flexibility to make changes— seemingly too little to make a fair test. Using full permutations does not reflect the fact that some species are more common than others.

Table 2 shows checkerboard scores for three different algorithms using data from the finches–island matrix based on 10,000 samples. The finches–island matrix has a checkerboard score equal to 3.8, which is outside the range of simulated values from all three algorithms except for a few runs when only the row sums are fixed. It is not reasonable to use the full permutation algorithm since this does not consider the unequal distribution of species. The column sums are reasonably balanced. Hence, there is no reason not to use the state space where only the row sums are fixed. The state space with both row and column sums fixed is narrower, but also further apart from the data matrix.

References

Ferkingstad, E., Holden, L., and Sandve, G. K. (2013). Monte carlo null models for genomic data. Submitted.

Gionis, A., Mannila, H., Mielikäinen, T., and Tsaparas, P. (2007). Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data*, 1(3):14.

Gotelli, N. (2000). Null model analysis of species co-occurrence patterns. *Ecology*, 81(9):2606–2621.

Gotelli, N. and Graves, G. (1996). *Null models in ecology*. Smithsonian Institution, Washington, D.C.

Manly, B. (2007). *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman & Hall/CRC, Boca Raton, 3rd edition.

Stone, L. and Roberts, A. (1990). The checkerboard score and species distributions. *Oecologia*, 85(1):74–79.