

THE EFFECT OF MIGRATION ON A SIMILARITY INDEX

Jack C. Yue¹ and Murray K. Clayton²

1. Department of Statistics, National Chengchi University, Taipei 11605, Taiwan, R.O.C.

2. Department of Statistics, University of Wisconsin-Madison, Madison, WI. 53706, U.S.A.

csyue@nccu.edu.tw

Key Words: Jaccard's index; Migration; Similarity index; Spatial randomness

ABSTRACT

Fewster and Buckland (2001) defined a similarity index between two communities by allowing changes between sites to reduce the influence of local discrepancies. The similarity index of Fewster and Buckland is calculated to attain the maximum similarity between two communities in the presence of migration. Instead of maximizing similarity, we propose random migration to measure the similarity of two communities with two types of stochastic migration. The similarity values based on the proposed methods can be treated as the expected value of similarity under migration. We use computer simulation and empirical examples to demonstrate our approach.

1. INTRODUCTION

Comparing the similarities of two populations, or evaluating the change of a population over time has been a widely studied topic in ecology, biology, and biogeography. The definition of a similarity measure plays a critical role in the assessment of how close two populations are.

To apply a similarity index, the attributes of interest in each population should be identified first, and then the similarity index defined as a function of these attributes. Jaccard's index is a commonly used similarity measure to compare populations, defined as the ratio of the number of shared species to the number of distinct species in two populations. Although Jaccard's index can only be applied to presence/absence data and possibly underestimates the similarity between two communities (Yue et al., 2001, and Yue & Clayton, 2005), it still receives much attention because of its calculational simplicity.

The interest in evaluating the similarity or overlap of populations based on Jaccard's index can be extended to the spatial realm. Building on notions introduced by Perry and Hewitt (1991), Fewster and Buckland (2001) proposed a similarity index of two populations in which animal counts (or presence/absence) are "swapped" between adjacent sites. This is designed to take into account the fact that observed animal locations do not necessarily correspond to the animal's preferred habitat, and therefore movement of animals might distort spatial similarity. Their similarity index is equal to the sum of a simple matching coefficient (similar to Jaccard's index) and $2/N$ times the number of swaps required to achieve maximum similarity, where N is the number of sites.

In this paper, we are also interested in calculating a similarity index by taking into account the effect of migration. Our approach is also for presence/absence data, and we will use the Jaccard index as the similarity measure. However, our approach is based on the idea

of a stochastic migration, instead of maximizing similarity as in Fewster and Buckland. In particular, we consider two types of stochastic migration. In the next section we first define the notation used in this paper and the assumptions related to migration. In Section 3, we discuss the notion behind the proposed approach and compare it with that of Fewster and Buckland (2001). Empirical examples are given in Section 4, following by the discussion of measuring similarity after migration.

2. MIGRATION ASSUMPTION

Similar to Fewster and Buckland, the proposed similarity index will be based on binary data (i.e., 0 representing absence and 1 representing presence). In order to simplify the discussion, unless specified, we shall assume that 1 at a site represents only one individual at this site. However, unlike the simple matching coefficient (SMC) used in the paper of Fewster and Buckland, we will use Jaccard's index as the measure of similarity. Note that Jaccard's index is defined as $\theta_j = s_0 / (s_1 + s_2 - s_0)$, where s_i is the number of non-empty sites in population i , $i = 1, 2$, and s_0 is the number of matched non-empty sites.

Fewster and Buckland defined the best attained match (BAM) to be the maximum SMC attained by allowing swaps between two mismatched (i.e., 0 and 1) sites. This kind of swapping would lessen the influence of migration given that animals' moves are constrained to their adjacent sites. Swapping mismatched sites is a good choice for eliminating the local

discrepancy causing by migration, but it is likely to over-estimate the similarity between two less similar populations if migration did not happen, or to misreport the similarity if animals' moves are not constrained to their neighborhood. If there is no information regarding how animals migrate, it is natural to consider random swapping (i.e., animals can move to any sites, including matched and mismatched sites) and use the average similarity value as the similarity level for migration. This kind of migration can be treated as a generalization of swapping only the mismatched sites in Fewster and Buckland. In particular, we include two variations of random swapping: unrestricted (UR) and adjacent neighbor (AN). In the unrestricted case, an individual is allowed to move into any site, regardless of whether this site would have two or more than two individuals. In other words, suppose there are m occupied sites in a map of n sites. Then the unrestricted model is equivalent to sampling at random m occupied sites with replacement from all n sites. Note that the random swapping can be applied to birds' migration since the occupied sites need not preserve their original structures after a long distance flight. It can be used to check the spatial structure of occupied sites differs a lot before and after the migration.

However, in practice, animals are more likely to migrate according to a more structured model. For example, animals with little mobility only move to their adjacent neighborhoods (as in Fewster and Buckland), and animals consuming a lot of water are more likely to move along a river (straight line) or a lake (circle). To demonstrate the influence of such a structured

migration model, we assume that the sites are located in a grid structure, and in particular, we assume that animals move according to an adjacent neighbor (AN) model (Figure 1). Animals move to their neighborhood sites with probability $4(p+q)$ and stay at their original sites with probability $1-4(p+q)$, where $0 < p, q < 1$ and $4(p+q) \leq 1$. Other structured migrations can be defined in the same way. For example, if animals only migrate along a river, say along the horizontal direction, then the probability of staying at the original site becomes $1-2p$.

[Insert Figure 1 here]

Before comparing different migration assumptions, including Fewster and Buckland (2001), we show some properties of the proposed approach for the remainder of this section. To simplify the discussion, we assume that two populations have same numbers of empty and non-empty sites. Let m denote the number of non-empty sites in an n -site map (i.e., with $n-m$ empty sites).

Note that the unrestricted case is related to the classical occupancy problem (see Feller, 1968, for detailed discussion), and the number of non-empty sites after migration satisfies

$$P(x \text{ non-empty sites after migration} \mid m \text{ non-empty sites}) = \binom{n}{x} \sum_{i=0}^x (-1)^i \binom{x}{i} \left(\frac{x-i}{n}\right)^m, 1 \leq x \leq n.$$

From Emigh (1983), we can show that the number of non-empty sites for m out of n non-empty sites after migration satisfies $E(X) = n - n(1 - \frac{1}{n})^m$. As a result, it is immediate that $E(X)/n \rightarrow 1 - e^{-\theta}$ as $m = \theta n \rightarrow \infty$ with $0 < \theta < 1$. In other words, the number of non-empty sites is approximately $(1 - e^{-\theta})n$ after migration. Therefore, the number of matches in the unrestricted case is approximately $(1 - e^{-\theta})^2 n$, which implies that the Jaccard index $J = \frac{(1 - e^{-\theta})^2}{(1 - e^{-\theta}) + (1 - e^{-\theta}) - (1 - e^{-\theta})^2} = \frac{1 - e^{-\theta}}{1 + e^{-\theta}}$. The asymptotic variance can be derived via the delta method.

Example 1. Based on 1,000 simulation runs, sample size $n = 10^6$, and $\theta = 0.1, 0.2, \dots, 0.9$, Table 1 shows means and variances of the Jaccard indices for the unrestricted case. We can see that the means and variances via simulation are very close to those from the delta method, where $\hat{\mu} (\hat{\sigma}^2)$ represents the estimate of mean (variance) from 1,000 simulation runs.

[Insert Table 1 here]

In the AN case, if the animals are at the corners or edges, they have limited (by the boundary) movements and will not be able to move in certain directions. Thus, if the animals are at the corners or edges, we will assume that these animals have a larger probability of staying at their original sites, i.e., transferring the probability of moving outside the boundary

to staying at the original place. The corner and edge effects are expected to be non-ignorable when the number of sites is not large.

Animals in the m non-empty sites can stay at their original spots and move away with probability $1-4p-4q$ and $4p+4q$, respectively. Also, we allow animals to move into the same site (as in the unrestricted case). Therefore, after migration, on average there will be $m-E(X)$ animals staying at their original spots and $E(X)$ animals moving away, where $X \sim B(m, 4p+4q)$ is the number of animals moving away. This means that, among those animals moving away, each would have probability $\frac{m-E(X)}{n}$ of moving to a site already occupied. In other words, after migration there are approximately $m-E(X) \cdot \frac{m-E(X)}{n} = m - \theta m \times c(1-c)$ non-empty sites, given that $m = \theta n$ and $E(X) = (4p+4q)m = cm$.

Similarly, there are S nonempty sites in common before and after migration, where $S = m - E(X) + E(X) \cdot \frac{E(X)}{n} = m - cm + c^2\theta m$, $m - E(X)$ is the number of non-migrating sites, and $E(X) \cdot \frac{E(X)}{n}$ is the number of animals moving into originally empty sites. The average Jaccard index would be $E(J) \rightarrow \frac{1-c+c^2\theta}{1+c-c\theta}$ as $m = \theta n \rightarrow \infty$.

Example 2. Similar to Example 1, we can use simulations to check the asymptotic results of AN case. We will simplify the discussion to cases where $\theta = 0.1, 0.2, \dots, 0.5$, and the sample size n varies from 10×10 to 50×50 . Also, since there are several choices of migrating probabilities p and q , our discussion is restricted to $p = q = 0.01$ and 0.05 , or $q = 0$ with $p = 0.01$ and 0.05 . Figures 2 and 3 show the average values and variances of the Jaccard

index obtained via simulation based on 1,000 simulation runs. We can see that both the average and variance appear to gradually converge, as the sample size increases.

[Insert Figure 2 here]

[Insert Figure 3 here]

3. MIGRATION ASSUMPTION AND SIMILARITY VALUES

Since animals can move, we agree with Fewster and Buckland (2001) that it is more realistic to assume that migration occurs when calculating the similarity between two communities. Their approach of swapping mismatches between adjacent sites provides the maximum similarity, while random migration under the UR and AN assumptions provides an alternative approach. The similarity values under the AN assumption is like the average.

Unless we know how the animals move, it is not fair to judge which migration model is more appropriate. The goal of this study is to provide alternative approaches, other than that of Fewster and Buckland. We can compare different migration approaches and provide suggestions for practical uses.

Example 3. Suppose the study regions contain 20 sites (4 by 5, presence/absence data), and there are 4 maps as shown in Figure 4. Our primary interest is to compare Map 1 vs. Map 2

and Map 3 vs. Map 4. Map 1 and Map 2 look very different without considering migration (observed Jaccard's index = 0). If there is a systematic movement from Map 1 to Map 2 (downward), the BAM by Fewster and Buckland (2001) can reflect migration perfectly.

[Insert Figure 4 here]

Table 2 shows the similarity values under random migration. The UR and AN cases are based on 10,000 simulation runs, and the averages and standard errors (s.e.) are shown as average \pm s.e. If the migration probability is very low ($p = .01$ & $q = .01$, total migration probability .08), the similarity value of the AN case for Map 1 vs. Map 2 is no different from 0 (observed Jaccard's index). If there is a much larger migration probability ($p = .1$ & $q = .05$, total migration probability .60), the similarity value of the AN case would be almost identical to that of a UR case. Judging from the observed Jaccard index and similarity values of the AN and UR cases, the BAM is likely to be too optimistic.

Map 3 vs. Map 4 show another possibility. There are 8 nonempty sites each in Map 3 and Map 4, and swapping between mismatches (BAM) produces a similarity value that is identical to the observed Jaccard index. Again, the very low migration probability of the AN case has about the same similarity value as the observed Jaccard index (and BAM). The UR case has a smaller value than the AN case with $p = .1$ & $q = .05$, and the difference between the AN

case and BAM is smaller than that of AN and UR.

[Insert Table 2 here]

Example 3 is designed to show two extreme possibilities. Note that, comparing to completely random permutation of the UR case, the AN case partly preserves the original site structure (depending on the migration probabilities p & q). Therefore, we would expect that the similarity value of the AN case is larger than that of the UR case, unless the two communities are very different. Also, as seen in Example 3, the BAM by Fewster and Buckland (2001) always has larger similarity values than the UR case. This suggests that the similarity values from the BAM and UR cases can be treated as the upper and lower bounds, respectively, of the AN similarity index.

4. EXAMPLES

In this section, two examples will be used to demonstrate the proposed migration index: one example was originally used in Fewster and Buckland (2001), and the other is a woodlark data set from Bowden and Green (1992).

Example 4. The following graphs are from Fewster and Buckland (F&B), which contain maps of 24 sites (6 by 4). We will discuss the AN case first.

[Insert Figure 5 here]

The results of the AN case are shown in Table 3 and the observed Jaccard indices are in the second row. The similarity values obtained from comparing the reference map vs. map A, reference map vs. map B, and map A vs. map B are $5/8 \approx 0.6250$, $4/7 \approx 0.5714$, and $5/17 \approx 0.2941$, respectively. These numbers suggests that maps A and B have the smallest similarity, without considering migration.

[Insert Table 3 here]

The entries in each cell of Table 3, in the form $a \pm b$, represents that the average (a) and standard error of simulation (b), respectively. The third, fourth, and fifth rows of Table 3 are the expected Jaccard values if low, medium, and high probabilities of migration are considered, respectively. In particular, the second to fourth columns can be treated as the self-similarity for the reference map, map A, and map B, respectively. To simplify the discussion, let $J(X,Y)$ denote the observed Jaccard index of map X vs. map Y. Since $J(\text{Ref}, A)$ and $J(\text{Ref}, B)$ are closest to and within the range of two times the standard errors of that of $J(\text{Ref}, \text{Ref})$ with $p = .04$ and $q = .02$, we do not reject the hypothesis that maps A and B have

the identical presence/absence structure as the reference map before the migration. Similar interpretations can be applied to the last two columns.

The last three columns of Table 4 are the average similarity values under different migration assumptions and can be treated as the similarity values for the next season. The observed Jaccard indices of the last three columns are also closest to and within the range of the two times standard errors of values in the case of low migration probability. For average values in Table 4, the similarity values after migration are all smaller than those before migration. This indicates that the reference map, map A, and map B are similar when considered in pairs, and any migration reduces their similarity values.

The fourth row of Table 4 shows the Jaccard indices corresponding to BAM. The computation follows Fewster and Buckland, except that the Jaccard index is used. At first glance, it looks like that the reference map and map A are almost identical (judging from BAM) and map A and map B are very different (based on Jaccard's index). However, the similarity values for the three types of migration assumption together lead to somewhat different conclusions. Judging from the results of our three approaches, the reference map and map A are the closest, but other pairs of maps also have close similarity values. There is no difference in the similarity values between map A vs. map B and the reference map vs. map B. Also, although the reference map and map A are somewhat similar and map A and map B are not very similar, this does not imply that the similarity index of reference map vs. map B is

small as well.

[Insert Table 4 here]

Example 5. Because the AN case seems more realistic, we will apply the AN approach to a real data set of larger size, from Bowden and Green (1992). The data considered are the woodlark data and can be found in Fewster and Buckland (2001). Originally, there are 497 records of the woodlark (presence/absence) annually for year 1986~1990. Since the study region is more dense in the area with coordinates $75 < X < 85$ and $80 < Y < 90$, we focus the discussion on this area. The coordinates of all woodlark records are shown in Figure 6. Also, to simplify the calculation, the study area will be divided into a site of 10 km by 10 km, i.e., 100 square cells. Across the years, the number of non-empty cells is between 20 and 30.

The woodlark study area is much larger than previous two examples. Since the focus is on the AN case, we shall duplicate the simulation in Table 3 and see if there are any differences compared to the previous results. In particular, we consider the self-similarity and similarity between any two years for the AN approach, similar to that in Table 3. Table 5 lists the simulation results for 10,000 runs.

[Insert Figure 6 here]

The observed Jaccard index between 1986 and 1987 lies in the confidence interval of the AN similarity values between 1986 and 1986 with a high migration assumption (i.e., $p = .1$ and $q = .05$). This suggests that the map of 1987 can be treated as the result of AN-type migration from 1986. Likewise, 1988 can be treated as the result of migration from 1987 with $p = .1$ and $q = .05$. This might suggest that, if the AN case is the true migration model, the woodlark has a large probability of migration.

[Insert Table 5 here]

We can also compare the similarity values under different migration assumptions for the woodlark data as well (Table 6). Similar to the previous two examples, the results of the AN case always lie between those of the UR case and BAM. Also, the average and observed Jaccard indices of 1987 vs. 1988 (1986 vs. 1988) are the largest (smallest) from the AN case. It seems that the AN case can preserve the same order as the observed Jaccard index, and this is also the case in Table 4. Unlike the AN case, the BAM can produce similarity values which have different rankings than the observed Jaccard index.

[Insert Table 6 here]

From the previous examples, it seems that the UR similarity values are always not larger than the AN similarity values. Heuristically, this seems reasonable: the UR migration model is the least restrictive, and thus might lead to a minimal similarity value. These heuristics are incorrect however, and in fact there is no dominant relationship between the UR and AN values, as we can see from the following example.

Example 6. Suppose we have a site of 5 by 5 cells and only the middle 4 cells are occupied (Figure 7). Also, we assume that only movements parallel to the axes (i.e. not diagonally) are allowed, and let $p = 0.25$, i.e. animals in the occupied cells must move. Then, the occupied cells after migration have no overlap with those before migration (i.e., the Jaccard index is 0). However, the Jaccard index between before and after migration under the UR assumption is 0.0977 ± 0.1147 (average \pm s.e.), under 1,000 simulation runs. This means that the UR Jaccard index can be larger than the AN Jaccard index.

[Insert Figure 7 here]

5. DISCUSSION

In this paper, we study the influence of migration on a similarity index for binary (presence/absence) data. Based on the Jaccard index, we proposed an approach by taking all

possible migrations into account. Our approach is like computing the expected Jaccard index, compared to the approach of Fewster and Buckland (2001) which is based on finding the maximum index.

In particular, we considered two cases of stochastic migration: unrestricted (UR) and adjacent neighbor (AN) cases. In the UR case, animals are allowed to move to any place, while the animals' movements are limited to some predefined patterns in the AN case (Figure 1). Therefore, in general the similarity value of UR case is usually smaller than that of AN case, and the UR case can provide a lower bound of the similarity value. However, the relationship between the UR case and the AN case depends on the values of p and q .

On the other hand, the similarity of BAM is like taking the maximum and it shall generally be larger than that of the AN case. However, since our approach and the approach of Fewster and Buckland have different interpretations, the AN Jaccard indices and the BAM Jaccard indices do not necessarily give the same conclusions.

As mentioned in section 2 when we introduced the AN movement model, we think that structured migration models may be more reasonable in practice. The use of the AN movement is intended to show how we compute the Jaccard index under a structured model assumption. We do not wish to imply that the AN movement model is the most likely or the only structured model. Depending on the specific system under study, it would be appropriate to apply various structured models and adapt our approach to determine the average Jaccard

index for judging the similarity of two populations in that system. In addition, we should note that our efforts have focused on Jaccard's index. Depending on the needs, it could be appropriate to employ other indices.

ACKNOWLEDGEMENT

The authors are grateful for the funding from National Science Council in Taiwan, NSC Project 92-2118-M-004-004, and insightful comments from the anonymous reviewer, which helped us to clarify the context of our work.

REFERENCES

- Bowden, C.G.R. and Green, R.E. (1992). The Ecology and Management of Woodlarks on Pine Plantations in the Thetford and Sandlings Forests. The Lodge, Sandy, Bedfordshire: RSPB Research Department. Research report.
- Feller, W. (1968). An Introduction to Probability Theory and its Applications, vol. 1 (3rd ed.) New York, Wiley.
- Emigh, T.H. (1983). On the Number of Observed Classes from a Multinomial Distribution. *Biometrics* 39: 485-491.
- Perry, J.N. and Hewitt, M. (1991). New Index of Aggregation for Animal Counts. *Biometrics* 47: 1505-1518.
- Fewster, R.M. and Buckland, S.T. (2001). Similarity indices for spatial ecological data.

Biometrics 57: 495-501.

Yue, J.C., Clayton, M.K., and Lin, F. (2001). A Nonparametric Estimator of Species Overlap.

Biometrics 57: 743-749.

Yue, J.C. and Clayton, M.K. (2005). An Overlap Measure based on Species Proportions.

Communications in Statistics: Theory Methods 34: 2123-2131.

Table 1. Estimated \hat{J} and its variance for varying values of θ , given $n = 10^6$

θ value	Unrestricted	
	$\mu(\sigma^2)$	$\hat{\mu}(\hat{\sigma}^2)$
0.1	.05 (.25)	.05 (.25)
0.2	.10 (.25)	.10 (.24)
0.3	.15 (.25)	.15 (.24)
0.4	.20 (.24)	.20 (.25)
0.5	.25 (.24)	.25 (.24)
0.6	.29 (.23)	.29 (.22)
0.7	.34 (.23)	.34 (.23)
0.8	.38 (.22)	.38 (.23)
0.9	.42 (.22)	.42 (.22)

Note: The results are from 1,000 simulation runs

Table 2. Averages and standard errors of the data in Example 3

	Site 1 vs. Site 2	Site 3 vs. Site 4
AN ($p = .01, q = .01$)	.040±.045	.324±.059
AN ($p = .1, q = .05$)	.206±.096	.278±.113
Unrestricted	.199±.090	.172±.097
BAM (Jaccard)	1.0	1/3 \approx .333
Observed Jaccard	0	1/3 \approx .333

Note: The AN and unrestricted cases are from 10,000 simulation runs.

Table 3. Averages and standard errors of the F&B and the AN case

	Ref. vs. Ref.	A vs. A	B vs. B	Ref. vs A	Ref. vs. B	A vs. B
Jaccard's Index	1	1	1	.63	.57	.29
p =.01, q =.005 (P(move)=.06)	.87±.11	.88±.10	.81±.16	.57±.07	.51±.08	.29±.05
p =.04, q =.02 (P(move)=.24)	.61±.13	.64±.13	.50±.17	.47±.10	.39±.11	.28±.08
p =.1, q =.05 (P(move)=.60)	.43±.12	.46±.13	.30±.13	.38±.11	.29±.11	.25±.10

Note: 10,000 simulation runs

Table 4. Averages and standard errors of the F&B data

	Ref vs. A	Ref vs. B	A vs. B
AN ($p = .1, q = .05$)	.355±.103	.305±.106	.219±.090
Unrestricted	.204±.080	.179±.082	.179±.082
BAM (Jaccard)	1.0	4/7 \approx .571	4/7 \approx .571
Observed Jaccard	5/8 = .625	4/7 \approx .571	5/17 \approx .294

Note: The AN and unrestricted cases are from 10,000 simulation runs

Table 5. Averages and standard errors of the woodlark data and AN case

	'86 vs. '86	'87 vs. '87	'88 vs. '88	'86 vs. '87	'86 vs. '88	'87 vs. '88
Jaccard's Index	1	1	1	.45	.40	.52
p =.01, q =.005 (P(move)=.06)	.89±.08	.90±.08	.91±.07	.43±.04	.38±.03	.49±.04
p =.04, q =.02 (P(move)=.24)	.65±.10	.68±.10	.69±.10	.35±.06	.33±.06	.42±.06
p =.1, q =.05 (P(move)=.60)	.37±.08	.40±.08	.43±.08	.28±.06.	.26±.06	.31±.07

Note: 10,000 simulation runs

Table 6. Averages and standard errors of the woodlark data

	'86 vs. '87	'86 vs. '88	'87 vs. '88
AN ($p = .1, q = .05$)	.275±.065	.259±.058	.309±.065
Unrestricted	.138±.050	.141±.050	.136±.050
BAM (Jaccard)	5/7 \approx .714	18/31 \approx .581	18/29 \approx .621
Observed Jaccard	.455	.400	.517

Note: AN and unrestricted cases are from 10,000 simulation runs

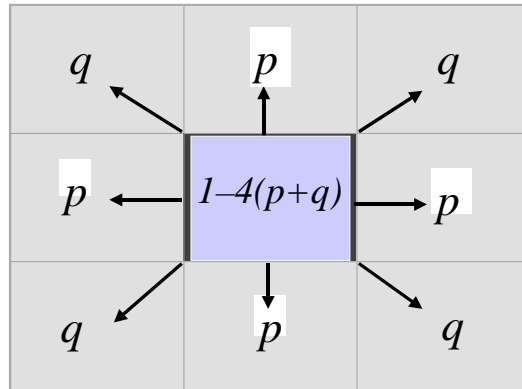


Figure 1. Adjacent neighbor migration

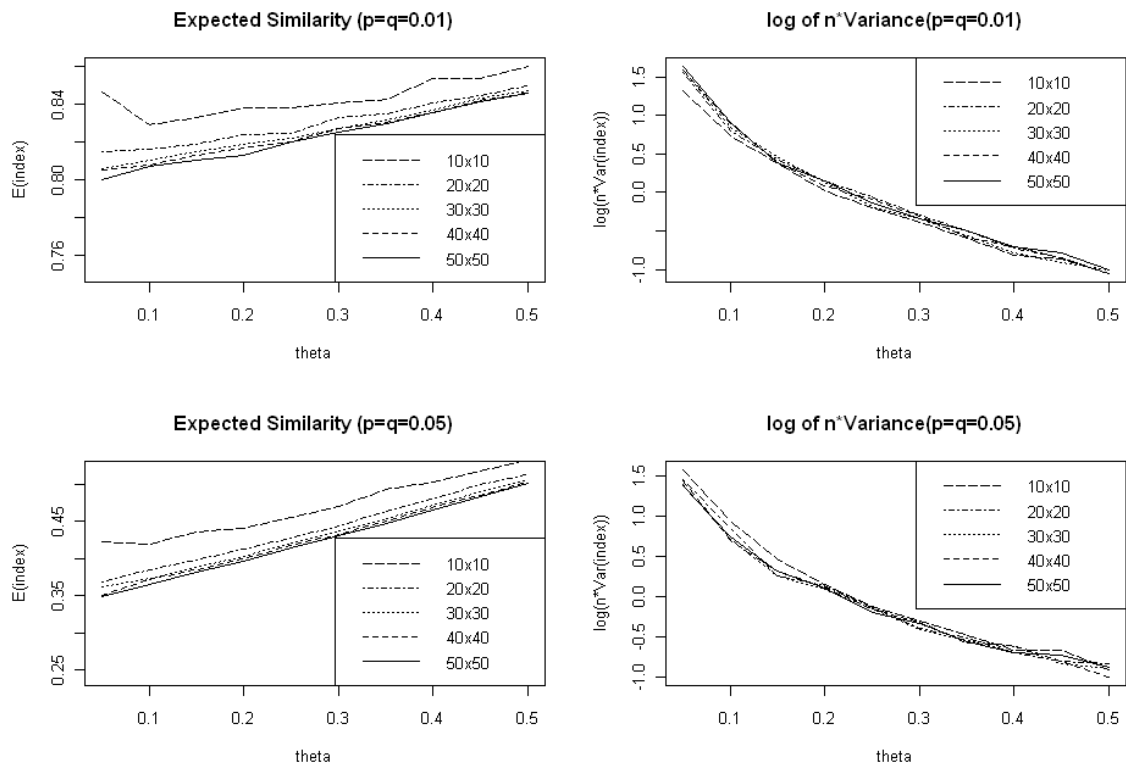


Figure 2. Estimated \hat{J} and its variance for varying sample sizes, varying values of θ , and migrating probabilities $p = q$ (10,000 simulation runs).

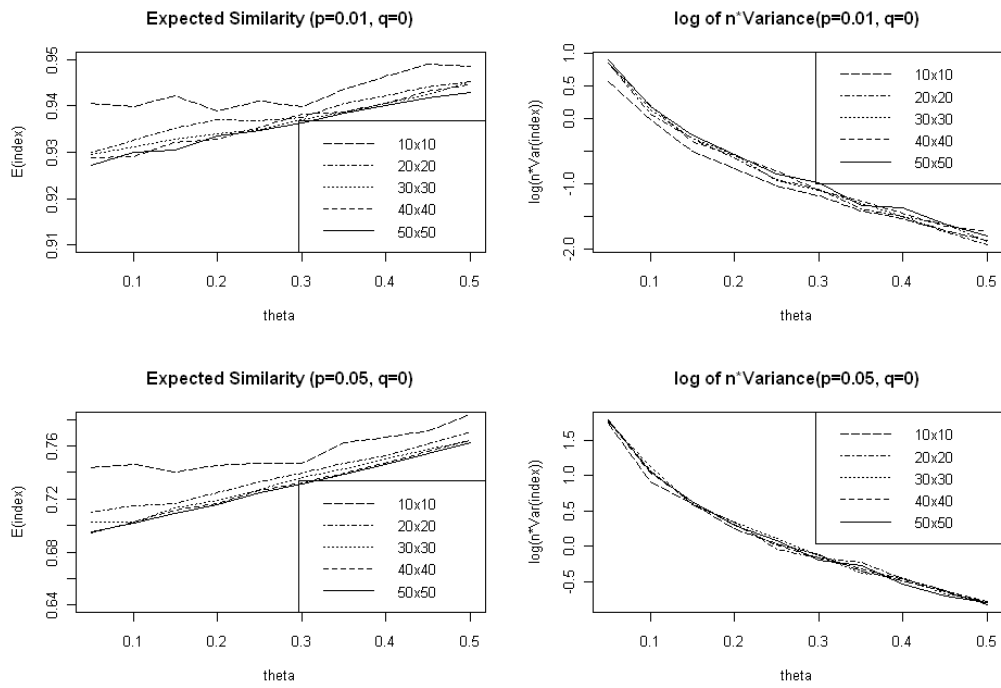


Figure 3. Estimated \hat{J} and its variance for varying sample sizes, varying values of θ , and migrating probabilities $q = 0$ (10,000 simulation runs).

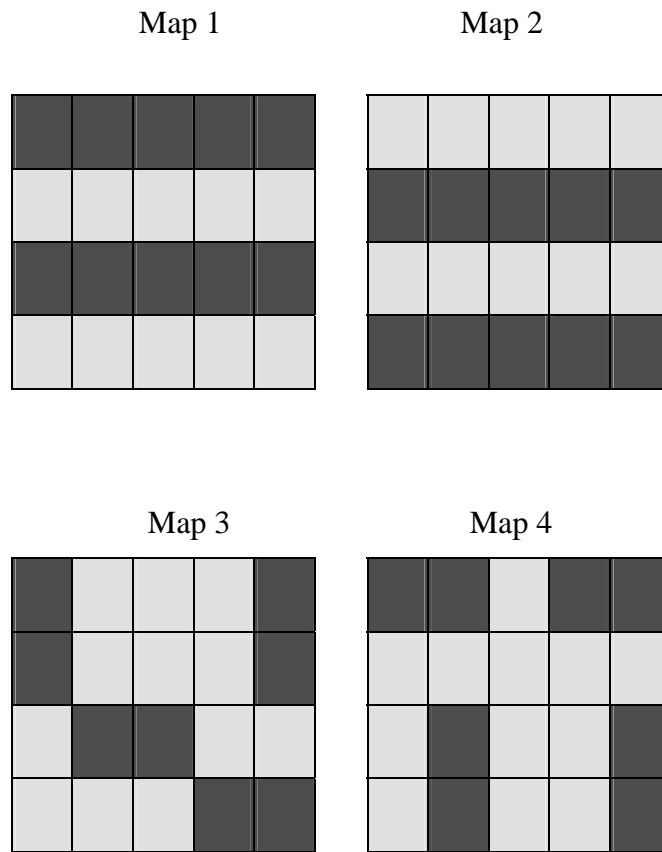




Figure 4. Data locations of Example 3

Note: Areas with darker color  indicates presence and  indicates absence.

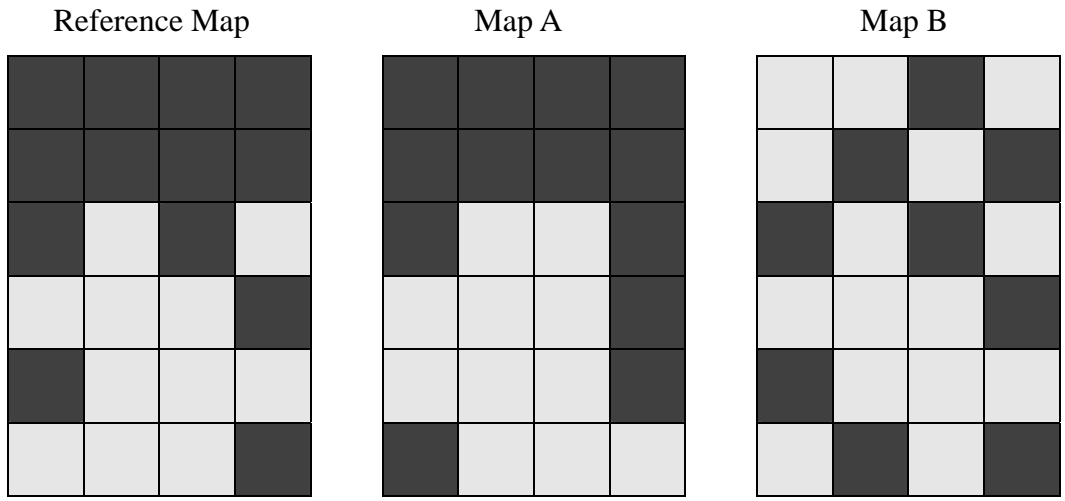


Figure 5. Maps of 24 sites in Fewster and Buckland

Note: indicates presence and indicate absence.

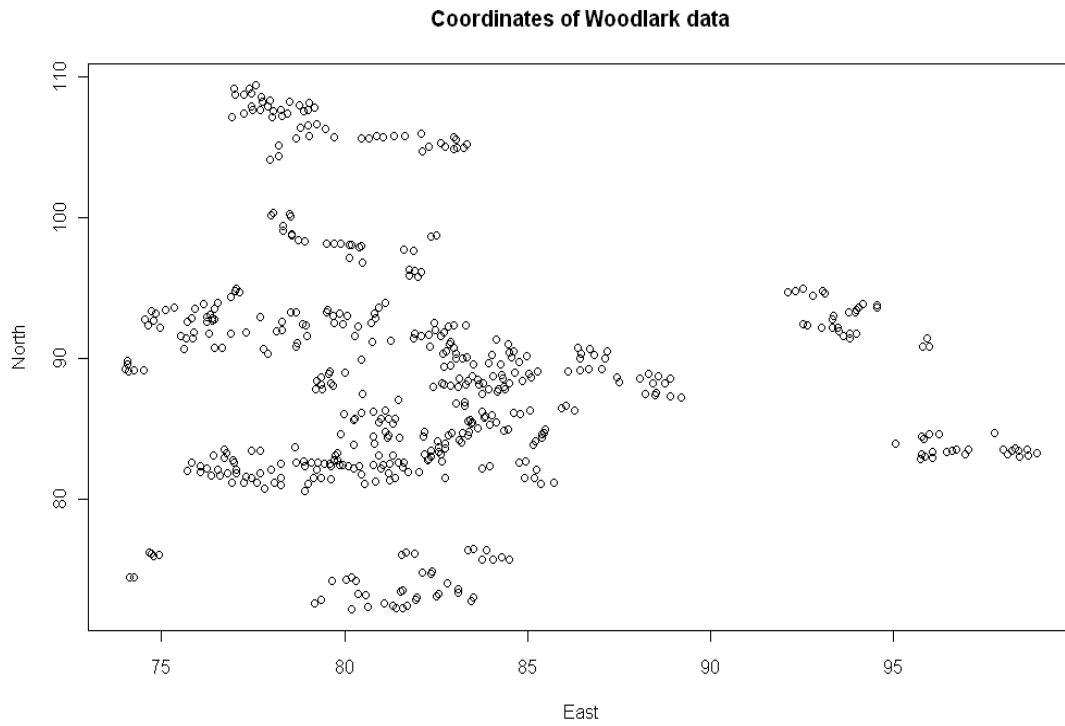


Figure 6. The locations of woodlark data (Bowden and Green, 1992)

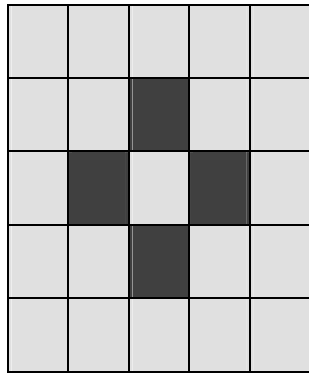


Figure 7. Data locations of Example 6

Note: Areas with darker color  indicates presence and  indicates absence.