Palaeontologia Electronica
http://palaeo-electronica.org

Correlations and co-occurrences of taxa: the role of temporal, geographic, and taxonomic restrictions

Aleksi Kallio*, Kai Puolamäki*, Mikael Fortelius, and Heikki Mannila

ABSTRACT

Correlation between occurrences of taxa is a fundamental concept in the analysis of presence-absence data. Such correlations can result from ecologically relevant processes, such as existence and evolution of species communities. Correlations are typically quantified by some sort of similarity index based on co-occurrence counts. We argue that the individual values of a similarity index are not useful as such: rather, we have to be able to estimate the statistical significance of the index value. Secondly, we argue that before computing the correlations one has to carefully select what is the underlying base set of locations for which the co-occurrence counts, similarity indices, and their significance is computed. We demonstrate base set selection with synthetic examples and conclude with an analysis of real data from a large database of fossil land mammals.

Aleksi Kallio. CSC - IT Center for Science Ltd. P.O. Box 405, FI-02101 Espoo, Finland. aleksi.kallio@csc.fi
Kai Puolamäki. Aalto University, Department of Information and Computer Science and HIIT Helsinki Institute for Information Technology. P.O. Box 15400, FI-00076 Aalto, Finland. kai.puolamaki@tkk.fi
Mikael Fortelius. University of Helsinki, Department of Geosciences and Geography, P.O. Box 64, 00014 University of Helsinki, Finland. mikael.fortelius@helsinki.fi
Heikki Mannila. Aalto University, Department of Information and Computer Science and HIIT Helsinki Institute for Information Technology. P.O. Box 15400, FI-00076 Aalto, Finland. heikki.mannila@aalto.fi
* These authors contributed equally to the work.

KEYWORDS: correlation; co-occurrence; base set; similarity index; statistical significance

INTRODUCTION

Presence-absence data indicate for a collection of locations and taxa which taxa are present in given locations. The locations can be, for example, fossil sites, with a known age, or map grid cells associated with observations of present-day mammals. Fundamental concepts in the analysis of presence-absence data are co-occurrence and correlation of pairs of species. Are two taxa occurring together more or less often than they should on the basis of pure chance?

A correlation between two species does not, by itself, necessarily carry any ecological meaning (for discussion, see Schluter (1984)). Correlation

PE Article Number: 14.1.4A

Submission: 15 October 2009. Acceptance: 28 November 2010

Kallio, Aleksi, Puolamäki, Kai, Fortelius, Mikael, and Mannila, Heikki, 2011. Correlations and co-occurrences of taxa: the role of temporal, geographic, and taxonomic restrictions. *Palaeontologia Electronica* Vol. 14, Issue 1; 4A:14p;
http://palaeo-electronica.org/2011_1/222/index.html

can be explained by trivial reasons, such as prevalence of species (Manel et al. 2001). However, a statistically significant correlation can be due to some ecologically relevant process such as the existence and evolution of species communities. Correlations can be used as an input or a starting point for more complicated analysis, such as cluster analysis ("find clusters of highly correlated species") or multidimensional scaling ("find a projection of species to a plane such that the correlated species are near to each other and uncorrelated species are far away"). In some studies the complete presence-absence matrix is analysed, but here we focus on co-occurrence and correlation of pairs of species. There exists a good body of work on the statistical questions related to the analysis of binary presence-absence matrices, and we refer the interested reader to Gilpin and Diamond (1984), Connor and Simberloff (1984) and Zaman and Simberloff (2002).

The first step in analysis is often to find out whether there exists any correlation between two species. The second step is to find a sound explanation for the correlations, or to apply a more advanced method. This paper focuses on this first step in a general setting. The ecological interpretation or detailed analysis of the causes of the correlation always depends on the context of the data set.

Co-occurrences are traditionally quantified by various similarity indices. There are many such indices (Shi 1993; Hubálek 1982; Archer and Maples 1987; Maples and Archer 1988). Most of the similarity indices can be computed using a contingency table (Table 1).

Typical similarity indices are those suggested by Jaccard (Jaccard 1912), Dice (Dice 1945; Sørensen 1948), Kulczynski (Kulczynski 1927) and Ochiai (Ochiai 1957); these four indices were recommended by Hubálek (1982) who evaluated 43 similarity indices for presence-absence data.

TABLE 1. Contingency table for two taxa A and B. Here a is the count of locations where both species A and B occur, b is the count of locations where A occurs but B doesn't, c is the count of locations where B occurs but A doesn't, and d is the count of locations where neither A nor B occur. We denote by n the total number of locations, i.e., n=a+b+c+d. The occurrence of A and B is denoted by A=1 and B=1, and non-occurrence by A=0 and B=0, respectively.

	B=1	B=0
A=1	*a*	*b*
A=0	*c*	*d*

Notice that the similarity indices are often used to measure taxon similarity between locations (samples), while we use the indices here to measure similarity between species based on their presence or absence at various locations. These indices can be computed using the counts in the contingency table:

- Jaccard: *a/(a+b+c)*
- Dice: *2a/(2a+b+c)*
- Kulczynski: *(a/(a+b)+a/(a+c))/2*
- Ochiai: *a/sqrt((a+b)(a+c))*

It is worth noting that none of these indices depend on the number of locations with no occurrences *d*, or on the total count of locations *n*. All indices range from 0 to 1, with the value 1 taking place when the taxa always co-occur (*b=c=0*).

As we are analysing real data, the data will always contain noise, i.e., the counts will have errors. Species may sometimes be incorrectly labeled present, but more often present species are incorrectly labeled absent, due to them not being detected or neglected for some other reason (Rosen and Smith 1988, Upchurch and Hunn 2002). This pseudo-absence is one of the error sources that have to be accepted in data analysis, emphasising the need for rigorous statistical framework capable of dealing with uncertainty.

We define two species to be uncorrelated if they occur independently of each other. More formally, the two species are uncorrelated if the contingency table approximately obeys the product of the marginal distributions of species, that is, *a is approximately equal to (a+b)(a+c)/n*, *b is approximately equal to (a+b)(b+d)/n*, *c is approximately equal to (c+d)(a+c)/n,* and *d is approximately equal to (c+d)(b+d)/n*. A similarity index that directly measures this correlation can be formally defined using the hypergeometric distribution function *phyper(a,a+b,c+d,a+c)*.[1] The application of hypergeometric distribution in constructing a similarity index that measures the faunal similarity between locations has been discussed in Raup and Crick (1979). The hypergeometric distribution function gives directly p-value of the one-tailed Fisher's Exact test. A p-value close to zero corresponds to a strong negative correlation, while a value close to unity corresponds to a strong positive correlation (i.e., the species tend to co-occur), and a value

1. The hypergeometric distribution corresponds to a process where k balls are drawn in random from an urn with m white and n black balls. *phyper(q,m,n,k)* denotes the probability of drawing at most *q* white balls.

near ½ corresponds to lack of correlation (i.e., the species occur independently of each other). An essential property of Fisher's Exact test, or the definition of correlation in general, is that if we add locations where neither of the species occur (i.e., *d* grows large), we will obtain a strong positive correlation (the p-value tends to 1 as *d* grows large).

The four similarity indices described earlier are fundamentally different from the p-value of Fisher's Exact test. The value of Jaccard, Dice, Kulczynski, or Ochiai index carries little information about correlation (except when the indices are exactly one), that is, whether the species occur independently or not; see Table 2 for an example. In the table high values of Jaccard and like indices are simply due to the fact that both of the species A and B are quite common (they both occur on 90% of the find sites), and, therefore, the co-occurrence count is high even though the species occur independently of each other. These indices are useful, for example, in comparing the relative co-occurrences of several pairs of species, but they do not tell about correlation of two species.

To study the existence of correlation, as discussed above, it is essential to take into account the number of locations where neither of the species occur. This number is related to the choice of the base set: if we study the correlation of, say, two African species we will typically obtain very different results if we take into account only the African locations, or if we take into account all locations across the globe. By base set we mean the set of locations, which we include in our study, that is, the locations which we use to compute the contingency matrix of Table 1. As discussed later in more detail, if we take into our study all locations across the globe we will usually obtain a strong positive correlation (p-value of Fisher's Exact Test that is close to one), due to the fact that there are many locations where neither of the species occurs because both species exist in Africa only.

TABLE 2. Contingency table which obeys the marginal distribution of species, i.e., the species A and B are uncorrelated. As expected, Fisher's Exact test and its mid-P variant, described later, imply no or weak correlation (the values of indices being 0.74 and 0.53, respectively). Jaccard, Dice, Kulczynski, and Ochiai, on the other hand, all output a high value of similarity (0.82, 0.90, 0.90, and 0.90, respectively).

	B=1	B=0
A=1	81	9
A=0	9	1

The first main argument in this paper is that as an initial step in any analysis involving co-occurrences of species it is often necessary to ascertain whether there is a statistically significant negative or positive correlation between two given species. For this purpose, as discussed above, most of the traditional association similarity indices are useless as such. A certain value of an association similarity index such as Jaccard does not imply existence of negative or positive correlation. Our second main argument is that before computing the correlations it is essential to select the base set properly, depending on the effect we want to study. As discussed above we can, for example, almost always obtain a positive correlation by adding to our study locations in which neither of the species occur. In this work we give principled guidelines on how to select the base set based on the effects we want to study.

STATISTICAL SIGNIFICANCE

An observed positive or negative correlation may arise from purely random effects. Statistical significance testing methodology gives a way of determining whether an observed correlation is just because of random occurrences, or whether it is a real phenomenon, i.e., statistically significant.

The ingredients of statistical significance testing are given by the null hypothesis and the test statistic. The null hypothesis describes the case when there is no correlation. In our case an obvious choice for a statistical significance test is Fisher's Exact Test, mid-P variant (Berry and Armitage 1995).

In Fisher's Exact Test the null hypothesis is that the contingency table has been sampled uniformly and randomly from a set of contingency tables having fixed marginal species counts (*a+b*, *a+c*, *b+d,* and *c+d*, respectively). The count *a* is used as a test statistic. The p-value of the one-tailed Fisher's Exact Test is defined to be the probability that the value of the test statistic is at least as extreme in the given direction under the null hypothesis; the lower-tail p-value can be expressed using the hypergeometric distribution function as *phyper(a,a+b,c+d,a+c)*, as discussed earlier. The mid-P variant addresses the fact that the one-tailed Fisher's Exact Test is slightly too conservative; the standard test is modified slightly such that the sum of lower-tail and upper-tail p-values are guaranteed to add up to unity (see Berry and Armitage (1995) for details and discussion). As

usual, we define a correlation to be significant if the p-value is at most some predefined value, such as *p≤0.05*. Otherwise we declare the correlation not statistically significant. In the remainder of this paper we use the significance limit of 0.05 and therefore declare a correlation significant if and only if the respective p-values satisfy *p≤0.05*, and by Fisher's Exact Test, mid-P variant, we mean the lower-tail variant unless otherwise noted. In a proper statistical significance testing method, such as Fisher's Exact Test, mid-P variant, the probability of a false positive (an event where the null hypothesis is rejected even if it holds) is at most 0.05.

Because of using the count *a* as a test statistic, i.e., considering lower counts more significant, the lower-tail p-value measures the significance of negative correlation. The p-value is small when the count *a* is exceptionally small compared to the marginal sums, i.e., the null hypothesis. Positive correlation can be measured with the upper-tail p-value, which can be computed by *1-p*, where *p* is the p-value for negative correlation.

Hence we can test two hypotheses, one for both positive and negative correlation. However having two hypotheses per species pair must be taken into account in a multiple hypothesis correction step, which is described later. If we are not interested in the direction of the correlation, i.e., we are only looking for extreme correlations, we should use a two-tailed p-value. It can be easily computed by *2 min(p, 1-p)*, where *p* is the one-tailed p-value (see Dudoit et al. 2003). A less conservative two-tailed p-value can also by computed by taking all contingency tables with given marginals sums, selecting those with probability equal to or less than the observed table, and summing up their probabilities; this approach does not, however, generalize in a straightforward way to a situation where we want to obtain one-tailed p-values.

This statistical test can be used with all association similarity indices if we want to test whether there is a deviation from the independence of the species assumption. Significance testing can therefore be done independently of the selected association similarity index.

Besides simplicity, the strength of Fisher's Exact Test is that it produces valid results regardless of the sample size. Pearson's Chi-square Test might be used instead of Fisher's Exact Test, but it assumes a sufficiently large sample size. In many presence-absence data sets sample sizes are not large enough to claim them sufficiently large with confidence. Besides analytical tests, significance of similarity indices can also be tested using the Monte Carlo methods. Monte Carlo methods rely on computational power to generate random samples and to calculate empirical p-values by comparing statistics in real data and in random samples drawn from a null distribution. Monte Carlo methods do not require cumbersome analytical treatment and they make it possible to derive significance estimates when no analytical solution is known. For the independence of the species null hypothesis we know the analytical solution, so Monte Carlo methods are not needed.

Statistical significance testing is further complicated by the fact that typically there is not only one pair of species, but several pairs of species of whose correlations we want to study. Multiple tests may result in false negatives. For example, assume that there are seven species. Then there are 21 pairs of species and an equal number of positive correlations to test for significance. We can test each of the 21 individual correlations for significance using Fisher's Exact Test, as described above. We call the p-values produced by these tests unadjusted p-values. It follows that even if the null hypothesis is true (i.e., there are really no correlations for any of the pair of species) we would declare on average about one of the 21 correlations significant by random chance alone. This effect is because we are rejecting null hypotheses at level *0.05 = 1/20*. Statistical test controls the probability of falsely rejecting a single null hypothesis, but when the test is repeated, the probability of a mistake increases unless further control procedures are used. Also if we are testing both for positive and negative correlation, there are two hypotheses for each species pair, effectively doubling the number of simultaneous hypotheses.

There are several ways to construct a multiple hypothesis testing method (see Dudoit et al. 2003 for a review and references) that corrects for this effect. Multiple hypothesis testing methods take as an input the unadjusted p-values, in our case those produced by the mid-P variants of the one-tailed Fisher's Exact Tests. The multiple hypothesis testing methods output adjusted p-values, one for each of the correlations. A null hypothesis is then rejected if the respective adjusted p-value is at most the chosen level, in our case 0.05. The simplest and the most well known of the methods is the Bonferroni correction, where the adjusted p-values are obtained by multiplying the respective unadjusted p-values by the total number of hypothesis (in this case, correlations) to be tested. The Bonferroni correction, while proper, is however

excessively conservative and therefore lacks power.

We use false discovery rate (FDR) adjustment and the Benjamini-Hochberg method (Benjamini and Hochberg 1995) to obtain the adjusted p-values and use the adjusted p-values to find out significant correlations. We declare a correlation significant if the respective adjusted p-value is at most 0.05. The derivation of the Benjamini-Hochberg method is somewhat involved, but it can be implemented only by few lines of program code to compute the adjusted p-values out of the unadjusted p-values. The Benjamini-Hochberg method guarantees that the expected fraction of false positives among all correlations declared significant is at most 0.05 (Benjamini and Hochberg 1995).

Summarizing, we compute an unadjusted p-value for each of the correlations using the one-tailed Fisher's Exact Test, mid-P variant. We then apply the Benjamini-Hochberg method to these unadjusted p-values to obtain the adjusted p-values. We declare a correlation significant if the respective adjusted p-value is at most 0.05. A model software implementation of the method is presented in the Appendix.

CORRELATIONS AND BASE SET

As explained in the introduction, the selection of a proper base set is crucial in the analysis of correlations. In selecting the base set it is important to understand for what purpose the correlations are used: we use the correlations as indicators of some effects between pairs of species. The selection of the base set can be used to choose the effects for which we want to test.

The situation is analogous to the design of controlled experiments. We would like to design an experiment such that only the variables that we are interested in affect the results. For example, if we would like to rule out the effects of large scale geography we would choose the locations so that we can compare only nearby sites. In analyzing ecological data we usually have the data given and cannot choose how the experiment is designed (for example, where and when the find sites are located). Lacking the ability to design the experiment we use base set selection to control the variables of whose effects we want to study. The price we pay is the reduction of power: with a proper selection of the base set we can (as explained below) control to some degree the variables that we want to study, but the more we restrict the base set the less locations it will include and the statistical test will be correspondingly less powerful.

For example, consider global presence-absence data where we have a set of present-day locations. Assume that we study the correlation between two species that occur only within Africa.

1. We can take all locations in the base set. We are likely to obtain a significant positive correlation, because in most of the locations (all locations outside Africa) neither of the species occur. The correlation is real, but the reason for it is trivial: the species are clearly not independently distributed, because both of them occur only in Africa, and therefore they are correlated.
2. We can use all African locations as a base set. If we choose some pre-defined set of locations (such as Africa) as a base set we can exclude the effect of this set of locations to the correlation. In this case, if we use Africa as a base set, and if we still observe a statistically significant correlation, we know that the correlation must be due to some other reason than both of the species occurring only in Africa. In other words, we have controlled the experiment so that the effect of Africa does not affect the results. In many applications, a more reasonable choice than a continent could be, e.g., the area covered by a given biome. As a result, if we would observe a correlation, it would be due to some other reason than the biome only.
3. We can use the union of the areas of occurrence of the species as a base set. Often, there may be no straightforward pre-defined area (such as Africa or a given biome) that we could use as a base set. If this is the case, one can use the locations within the union of the areas of occurrence of the two species as a base set. This choice guarantees that any observed correlation is due to some effect that takes place within the area of occurrence of the two species. Notice that this choice is closest in spirit to the Jaccard and like indices, which ignore the locations in which neither of the species occur. As discussed before, Jaccard and like indices are however not proper indicators for correlation: they can give high co-occurrence counts even when there is no correlation, like in the example of Table 2.
4. We can use the intersection of the areas of occurrence as a base set. If we want to rule out the effect of the large-scale areas of occurrences altogether, then we can use as a base set the locations within the intersection

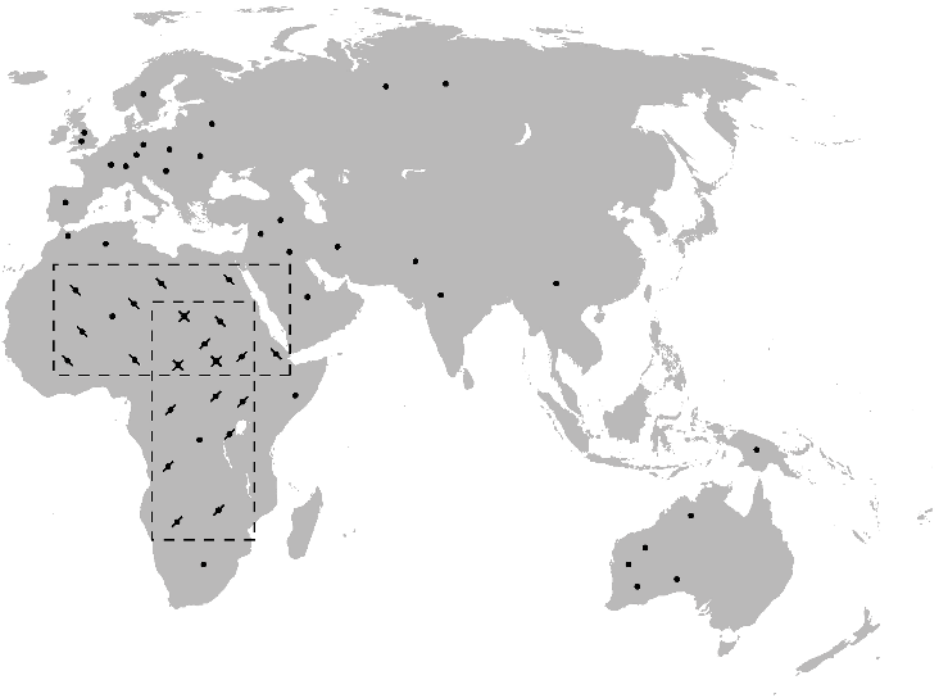

FIGURE 1. Part of a world map with locations marked as dots, occurrences of species A marked as backward slashes and occurrences of species B marked as forward slashes. The dashed rectangles show areas of occurrence for both species. We have enlarged the area covered by the rectangles slightly for visual clarity. The union or the intersection of the areas can be used for selecting the base set of locations.

of the areas of the occurrences of the two species. If we observe a correlation it must be due to a reason not related to the areas of occurrences.

In this paper, we use the smallest rectangle that can be used to enclose all occurrences of a species as an area of occurrence. The union or intersection of these areas is then used to define the base set of locations. Rectangles are not rotationally invariant, which means that their definition is dependent on the direction of the coordinate axes. For improved precision, smallest rectangles can be replaced with more advanced structures, such as convex hulls, i.e., minimal polygons containing all the locations where the species occurs.

The fossil data consist of locations that have a specific age in addition to the spatial location. We can define the lifetime of the species as the time interval between and including the earliest and the latest occurrence of the species. If we take all locations from all times into account the results are easily dominated by the trivial fact that the lifetimes differ for most pairs of species. Typically, we want to exclude the effect of time from the analysis. Therefore, we can use as a base set the locations within the intersection of the lifetimes of the respective species. It follows that if we observe any correlation, it must be due to some reason that is not related to the lifetimes of the species.

To demonstrate the effect of different base set selection criteria, we present synthetic data for two African species A and B. Using different geographic selection criterias, we calculate Fisher's Exact Test, mid-P, and as an example of a typical association similarity index, the Jaccard index. For statistical significance we use the level 0.05, without multiple hypothesis testing correction.

In Figure 1 occurrences of both species are presented in part of a world map. We first concentrate on the whole map and calculate occurrences in Table 3. As can be seen from Table 3, most of the locations do not contain either of the species, and there is no statistically significant correlation. In Table 4 we have restricted our base set to Africa. The two tables are identical, except for the case of *A=0/B=0*, i.e., neither of species are present. A typical similarity index such as the Jaccard index gives identical values on both cases. However, the

TABLE 3. Contingency table for all locations. Fisher's Exact Test, mid-P, gives the p-value of 0.622, implying positive correlation, which is however not statistically significant. Jaccard index is 0.143.

	B=1	B=0
A=1	*3*	*9*
A=0	*9*	*34*

TABLE 4. Contingency table for African locations. Fisher's Exact Test, mid-P, gives the correlation of 0.044, implying statistically significant negative correlation. Jaccard index is 0.143, the same as in Table 3.

	B=1	B=0
A=1	*3*	*9*
A=0	*9*	*6*

TABLE 5. Contingency table for union of locations. Fisher's Exact Test, mid-P, gives the correlation of 0.005, implying statistically significant negative correlation. Jaccard index is 0.143, still same as in Tables 3-4.

	B=1	B=0
A=1	*3*	9
A=0	9	2

TABLE 6. Contingency table for intersection of locations. Fisher's Exact Test, mid-P, gives statistically non-significant negative correlation of 0.333. Jaccard index is 0.500.

	B=1	B=0
A=1	*3*	*1*
A=0	*2*	0

statistical significance as calculated with Fisher's Exact Test, mid-P variant, is different: in Table 3 correlation was non-significant positive correlation, but in Table 4 it is significant negative correlation.

We look into this negative correlation more closely. Figure 1 shows areas of both species as dashed rectangles. The base set can be selected from these areas by either looking at them both, i.e., the union of areas, or by looking only into the intersection of the two areas, i.e., where we have evidence for both of species occurring. Table 5 shows occurrence counts in the case of union of areas.

Looking at Table 5, we see significant negative correlation as can be read from the very low p-value. It could be argued that the two species are dissimilar, maybe suggesting an interaction that would not allow the two species to co-exist. However, when looking at the base set of intersecting areas, as reported in Table 6, we cannot see significant correlation.

Table 6 is most suited for analysing interactions between species, and it does not support hypothesis of dissimilarity, possibly due to lack of data. Dissimilarity in Table 5 could be explained by geography, as the two species have different areas of occurrence. In the area where both species occur there is no strong evidence of interaction as seen from Table 6.

What we see in the example above is fundamentally related to spatial autocorrelation: the probability of occurrence and co-occurrence of the species depends on the geographic locations of the occurrences. Fisher's Exact Test does not account for spatial autocorrelation directly. Instead, our approach is to use base set selection to control the spatial effects. By making this choice explicit we do not hide uncertainties related to spatial autocorrelation and make the analysis process easier to understand and evaluate. In the example case, we examined how accounting for spatial effects changed the results significantly.

Summarizing the above discussion, before we can choose a base set, we have to decide which effects we want to study. The answer (whether there is a correlation or not) depends on which types of effects we want to study. Naïve selection of a base set leads to trivial results. For example, if we select all locations as a base set then any observed correlation may be simply due to different areas of occurrences and differences in the lifetimes of the species. Typically we are not interested in these variables because they are trivial to notice and understand even without any correlation analysis. Therefore, we need to bound the base set such that the effect of known or uninteresting variables is eliminated.

CORRELATIONS AND TAXONOMIC LEVEL

We can also constrain the base set by using taxonomic information instead, or in addition to, geographic and temporal restrictions. For example, if we are studying the correlations between species we can take into base set only the locations in which there is at least one representative of the order (or family) of each of the two species. This way, if we observe a correlation, it is not because the respective species exist in a given order (or family). A purely practical reason for applying such a constraint is that many data sets are compiled from literature that is typically organised taxonomically, potentially creating a pseudo-absence effect. Therefore, the dogs and horses might be known from a site, but not the deer, even though they were in fact present, but not relevant for the study

that produced the data. By using taxonomic criteria, we have evidence for each included location that some representatives of the order or family of each of the two species is present. Therefore, it is more unlikely that there is pseudo-absence or irrelevant absence effects in that location.

In the example analysis below, where we analyze fossil find sites of large land mammals, we have in fact already implicitly used a kind of taxonomic restriction. Imagine that in our full data set we had find sites having large and small land mammals, respectively. Because we only want to study the effects within large land mammals, we select into the base set only find sites in which large land mammals occur. This way we can rule out the possibility that any correlation we might observe would be due to the differences of large and small land mammals.

EXAMPLE ANALYSIS

This example analysis of correlations is based on the 2007 version of Neogene of the Old World Database of Fossil Mammals NOW (Fortelius 2007). The database contains information about Eurasian Miocene to Pleistocene land mammal taxa and localities. An extensive database collected in international collaboration is a good example of the importance of base set selection, as it has been collected from various data sources and from studies looking at completely different research questions. In other words, the data have not been collected to answer the questions we are about to analyze, and therefore, it is essential to carefully select the subset of the data that is relevant and as unbiased as possible for answering our question.

NOW data were preprocessed by including only large mammals that were present at 10 or more sites. Sites were filtered by including only sites with 10 or more genera. The preprocessing resulted in a base set of 217 sites and 169 genera. Without filtering the data set would have contained more marginal species and locations, lowering the number of statistically significant results. The justification for the filtering is that we select into our base set only those species and locations that have been studied more widely. This will prevent biases such as having a set of findings from a very exotic and tightly focused research programme distort the results of our general correlation studies.

We use the data to show the effect of the previously described filtering criteria. For presenting the results, we use shorthand notation for the restrictions. Keyword **GeoUnion** is used for data that have been filtered using the geographic criteria where all locations within the combined area (union of areas) of the two species are included as a base set. Keyword **GeoIntersection** is used when only locations on the shared area (intersection of areas) of the two species are included. When no georaphic restriction is used we use the keyword **NoGeo**.

For temporal restrictions keyword **Time** is used when we apply restriction that selects only sites with MN units in which both of the genera existed and **NoTime** for data that have not been filtered with this restriction.

For taxonomic criteria, keyword **Family** is used for data that have been filtered by similarity at the family level, that is, for each pair of species, we take into base set only the find sites in which there is at least one representative from the families of both of the two species. This way we can rule out the distribution patterns of the families as explanatory factors. Keyword **Order** has been used for data that have been filtered by similarity at the order level, respectively, and keyword **NoTaxonomic** for data that have not been filtered with this restriction.

For all pairs of taxa we calculated correlation and p-value for the correlation using Fisher's Exact Test, mid-P variant. Both positive and negative correlations were tested. Multiple testing correction was conducted using the Benjamini-Hochberg method, with false discovery rate controlled to 0.05. Total numbers of significant correlations for the NOW data set are given in Table 7.

In Table 7 the most obvious difference between counts of significant correlations is observed when the temporal restriction is applied. Without the restriction a vast amount of correlations are seen as the database contains large numbers of species that have lived at different times. As the counts for **NoTime** are about two orders of magnitude larger than for **Time**, it is obvious that the trivial temporal effect dwarfs other ecologically more interesting effects and therefore should be taken care of by using the temporal restriction.

When considering geographic restrictions, we see that the three cases **GeoIntersection**, **GeoUnion**, and **NoGeo** all have a different level of strictness. Typically there are no grounds for considering geographic locations that for some reason have not been reachable by the other species, suggesting that at least the **GeoUnion** restriction should be applied and that the correlations from **NoGeo** might stem from some obvious or uninteresting geographic effects. If we are interested in

TABLE 7. Number of significant correlations (positive correlations + negative correlations) in the NOW data set with 18 different restriction parameter combinations. False discovery rate is controlled to 0.05 using the Benjamini-Hochberg method.

Number of correlations	Time			NoTime		
	GeoIntersection	GeoUnion	NoGeo	GeoIntersection	GeoUnion	NoGeo
Family	7	15	19	481	690	797
Order	5	14	33	958	1268	1629
NoTaxonomic	9	14	49	986	1253	1645

intraspecies dynamics, it is advisable to use the stricter **GeoIntersection** restriction to include only areas where we have evidence of both species existing.

Finally, the selection of taxonomic level yields different counts of significant correlations. When interested in interactions between orders, criteria **Order** might offer the right level of inspection, as it filters out effects stemming from possibly different orders of the two species. Similarly, the criteria **Family** can be used for studying interactions between families. It can also be used for studying interactions between orders, but is not optimal for that, as it filters out more locations. **NoTaxonomic** is a good choice if we know that data do not contain taxonomic biases, because no filtering yields in the largest number of locations and hence the best statistical power. The right level of filtering should be decided based on the question that is being studied.

It is important to bear in mind that we only see correlations between species and not their interpretation. The existence of a correlation often implies an ecologically relevant process, assuming that the base set is selected appropriately. However, a correlation does not directly carry any ecological meaning, but instead only states that the occurrence of the two species in the base set cannot be simply explained by random effects only. Correlations are an invaluable way for generating hypothesis and finding interesting aspects of the data set for further examination. It is the task of the next analysis step to validate correlations and to find sound ecological explanations for them.

DISCUSSION

To complete our analysis we briefly discuss some of the correlations produced with the methods described in this paper and their ecological interpretation. We look at two sets of correlations that both are produced using criteria **Time** and **NoTaxonomic**, but with different geographic base sets. In Figure 2 we show all significant positive and negative correlations when looking only at the union of the areas covered by the two species (**GeoUnion**), and in Figure 3 all significant positive and negative correlations between the species without using a geographic filtering criteria (**NoGeo**). In the figures, species are presented as ellipses, positive correlations as bold lines, and negative correlations as dashed lines.

The correlation patterns observed are clearly very sensitive to geographic restrictions. The geographically restricted set in Figure 2 shows mostly correlations that can be plausibly explained by ecology. Thus, *Stephanorhinus* and *Cervus* are two of the main genera of interglacial (warm) assemblages of the Pleistocene Ice Age, while *Mammuthus* and *Equus* characterise the glacial (cold) assemblages that alternate with the interglacial ones. The biogeography of these assemblages was dynamic (Koenigswald 2007), with alternating expansion from refugia, so the correlations are not

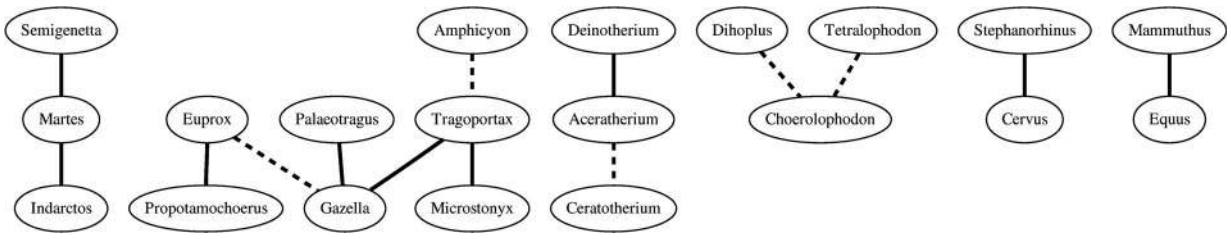

FIGURE 2. Correlations between genera with geographic filtering (**GeoUnion**). Genera are presented as ellipses, statistically significant positive correlations as bold lines, and negative correlations as dashed lines.

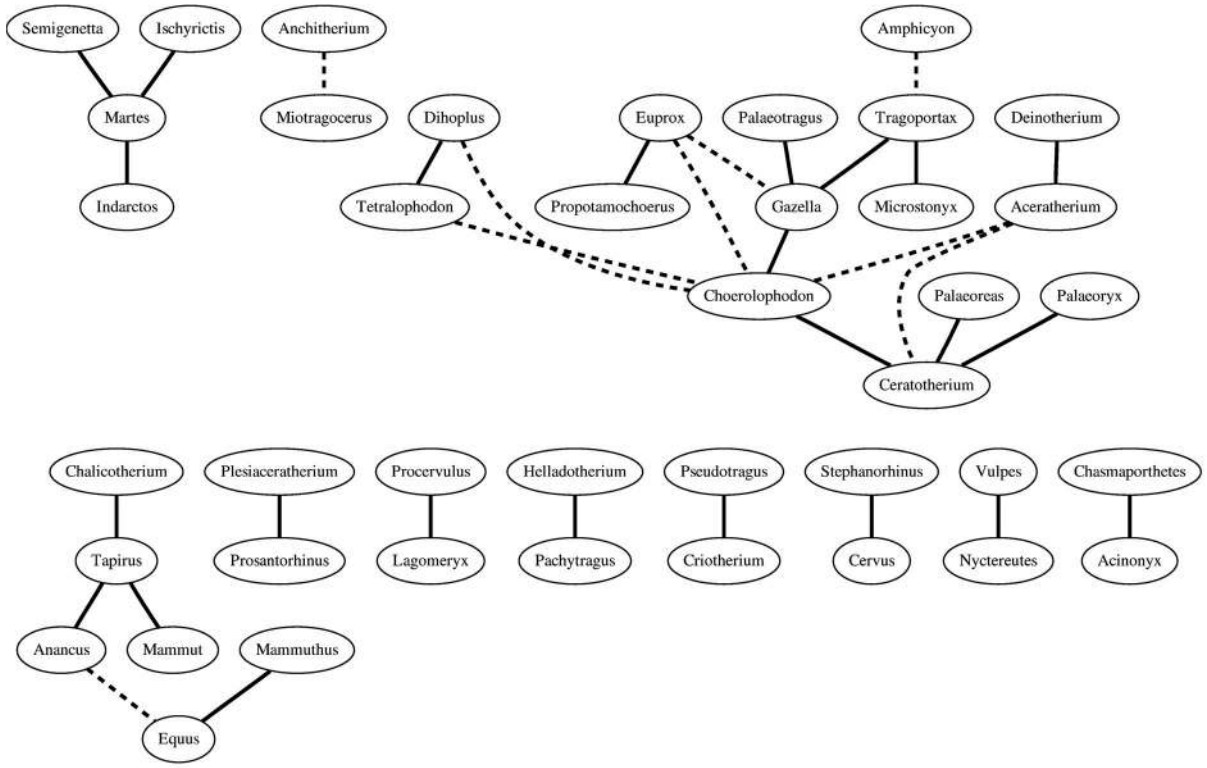

FIGURE 3. Correlations between genera without geographic filtering (**NoGeo**). Genera are presented as ellipses, statistically significant positive correlations as bold lines and negative correlations as dashed lines.

likely to be driven by geography. Among the geologically older genera the cluster around *Gazella* is made up of other open-adapted taxa (*Palaeotragus*, *Tragoportax*, *Microstonyx*), with negative correlations to forest taxa (*Euprox*, and indirectly *Propotamochoerus* and *Amphicyon*). The genera negatively correlated to the open-adapted *Ceratotherium* and *Choerolophodon* are also primarily forest taxa (*Dihoplus*, *Aceratherium*, *Tetralophodon* and, with some reservations, *Deinotherium*). The carnivore cluster around *Martes* also represents a forest setting.

As expected, the correlations in the geographically unrestricted case in Figure 3 appear in many cases to be due primarily to distribution patterns. The cluster around *Tapirus* as well as several pairs of genera belong to this group, for example *Pleasiaceratherium-Prosanthorhinus*, *Procervulus-Lagomeryx*, *Helladotherium-Pachytragus*, and *Pseudotragus-Criotherium*. Among negatively correlated pairs, *Euprox-Gazella* and *Euprox-Choerolophodon* belong to this category. The main reason for these geographic associations is, however, not so much random spatial patterns as a strong underlying and ultimately climatological forcing: the distribution of the genera reflects the distribution of the ecological associations ("chronofaunas") to which they belong (Eronen et al. 2009), essentially the forested western and central European faunas versus the open woodlands of eastern Europe and western Asia (Fortelius et al. 1996).

Some pairs in the geographically unrestricted set appear to be related to other factors than geography. The negative correlation of *Anchitherium-Miotragocerus* is likely caused by nearly non-overlapping temporal distributions. Two pairs of widely distributed carnivores are unlikely to be explained by geographic distribution, unless by chance, but appear instead to be related through foraging behaviour to habitat: *Vulpes* (red fox) and *Nyctereutes* (raccoon dog) are both short-legged generalists strongly associated with vegetation cover, while *Acinonyx* (cheetah) and *Chasmaporthetes* (cheetah-like hyaena) represent extremely long-legged pursuit predators in open habitats.

When looking at the data in Figure 2 and Figure 3, it is important to keep in mind that it reports results that are statistically significant, but the

absence of a correlation is not significant. Our statistical methodology allows us to report results that are backed up by data, but in interpretation the absence of a correlation should not be considered as a significant result. Absence of a correlation means that there is not enough evidence for the correlation in the data, but it does not mean that the two species must be non-related. Correlations are also not transitive: significant correlations of A and B, and of B and C, do not necessarily imply a significant correlation of A and C.

CONCLUSIONS

Co-occurrence and correlation of pairs of species is an important element in ecological data analysis. When using the co-occurrence indices, it is, however, important to understand that it matters both how the correlation is computed and how the base set is selected. If the base set is selected improperly the observed correlation can be due to some relatively trivial reason, such as both species existing on the same continent. We show how to apply spatial, temporal, and taxonomic criteria to select a proper base set. Similarity indices such as the Jaccard index sidestep this problem by ignoring locations in which neither of the species exist, but as a result, these indices are not suitable indicators of existence or non-existence of correlations.

REFERENCES

Archer, A.W. and Maples, C.G. 1987. Monte Carlo simulation of selected binomial similarity coefficients (I): effect of number of variables. *PALAIOS*, 2:609-617.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289-300.

Berry, G. and Armitage, P. 1995. Mid-P confidence intervals: a brief review. *The Statistician*, 44(4):417-423.

Connor, E.F. and Simberloff, D. 1984. Neutral models of species' co-occurrence patterns, p. 316–331. In Strong, D.R. Jr., Simberloff, D., Abele, L.G., and Thistle, A.B. (eds.), *Ecological Communities: Conceptual Issues and the Evidence*. Princeton University Press.

Dice, L.R. 1945. Measurement of the amount of ecological association between species. *Ecology,* 26:297-302.

Dudoit, S., Shaffer, J.P., and Boldrick, J.P. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71-103.

Eronen, J.T., Mirzaie Ataabadi, M., Micheels, A., Karme, A., Bernor, R.L., and Fortelius, M. 2009. Distribution history and climatic controls of the Late Miocene Pikermian chronofauna. *Proceedings of the National Academy of Sciences (USA)*, 106(29):11867-11871.

Fortelius, M. (coordinator) 2007. Neogene of the Old World database of fossil mammals (NOW). University of Helsinki.
www.helsinki.fi/science/now/.

Fortelius, M., Werdelin, L., Andrews, P., Bernor, R.L., Gentry, A., Humphrey, L., Mittmann, L., and Viranta, S. 1996. Provinciality, diversity, turnover and paleoecology in land mammal faunas of the later Miocene of western Eurasia, p. 414-448. In Bernor, R., Fahlbusch, V., and Mittmann, W. (eds.), *The Evolution of Western Eurasian Neogene Mammal Faunas*. Columbia University Press.

Gilpin, M.E. and Diamond, J.M. 1984. Are species co-occurrences on islands non-random, and are null hypotheses useful in community ecology?, p. 297-315. In Strong, D.R. Jr., Simberloff, D., Abele, L.G., and Thistle, A.B. (eds.), *Ecological Communities: Conceptual Issues and the Evidence*. Princeton University Press.

Hubálek, Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Reviews of the Cambridge Philosophical Society,* 57(4):669-689.

Jaccard, P. 1912. The distribution of flora in the alpine zone. *The New Phytologist*, 11(2):37-50.

Koenigswald, W.v. 2007. Mammalian faunas from the interglacial periods in Central Europe and their stratigraphic correlation, p. 445-554. In Sirocko, F., Claussen, M., Litt, T., and Sanchez-Goni, M.F. (eds.), *Developments in Quaternary Science 7, The Climate of Past Interglacials*, Elsevier, Amsterdam.

Kulczynski, S. 1927. Die Pflanzenassoziationen der Pieninen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles B*, 57-203.

Manel, S., Williams, H.C., and Ormerod, S.J. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38(5):921-931.

Maples, C.G. and Archer, A.W. 1988. Monte Carlo simulation of selected binomial similarity coefficients (II): Effect of Sparse Data. *PALAIOS*, 3:95-103.

Ochiai, A. 1957. Zoographic studies on the soleoid fishes found in Japan and its neighboring regions. *Bulletin of the Japanese Society of Scientific Fisheries,* 22:526-530

R Development Core Team. 2008. R: a language and environment for statistical computing.
www.r-project.org/

Raup, D.M. and Crick, R.E. 1979. Measurement of faunal similarity in paleontology. *Journal of Paleontology*, 53(5):1213-1227.

Rosen, B.R, and Smith, A.B. 1988. Tectonics from fossils? Analysis of reef-coral and sea-urchin distributions from late Cretaceous to Recent, using a new method. *Geological Society London Special Publications* 37:275.

Schluter, D. 1984. Variance test for detecting species associations, with some example applications. *Ecology*, 65(3):998-1005.

Shi, G.R. 1993. Multivariate data analysis in palaeoecology and palaeobiogeography: a review. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 105(3-4):199-234.

Sørensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 5:1-34.

Upchurch, P. and Hunn, C.A. 2002. "Time": the neglected dimension in cladistic biogeography? *Geobios,* 35:277-286.

Zaman, A. and Simberloff, D. 2002. Random binary matrices in biogeographical ecology—instituting a good neighbor policy. *Environmental and Ecological Statistics* 9(4):405-421.

APPENDIX

We show below a short program code, written in R (R Development Core Team 2008), that takes the presence-absence data as an input and outputs the statistically significant positive, negative, and extreme correlations. In this example, for simplicity, the base set is the same for all pairs of species.

```
# Input:
# n     Number of locations.
# m     Number of species.
# X     nXm matrix such that X[i,j]=1 if species j occurs at locality i,
#       X[i,j]=0 otherwise.
# alpha Significance threshold; we use alpha=0.05.
#
# Output:
# Cneg  mXm Boolean matrix such that Cneg[i,j]=TRUE if there is a
#       significant negative correlation between species i and j,
#       Cneg[i,j]=FALSE otherwise.
# Cpos  mXm Boolean matrix such that Cpos[i,j]=TRUE if there is a
#       significant positive correlation between species i and j,
#       Cpos[i,j]=FALSE otherwise.

# We use the R function for hypergeometric distribution to obtain
# the p-values of the Fisher's Exact Test, mid-P-variant.
#
# Input:
# x Two-dimensional contingency matrix.
#
# Output:
# One-tailed p-value of the Fisher's Exact Test mid-P.

fisher.test.midp <- function(x) {
  ( dhyper(x[1,1],x[1,1]+x[1,2],x[2,1]+x[2,2],x[1,1]+x[2,1])/2
  +(if(x[1,1]>0)
    phyper(x[1,1]-1,x[1,1]+x[1,2],x[2,1]+x[2,2],x[1,1]+x[2,1])
    else 0) )
}

# The p-values related to the correlations using the one-tailed Fisher's
# Exact Test mid-P

P <- matrix(0.5,nrow=m,ncol=m)
for(i in 1:(m-1)) {
  for(j in (i+1):m) {
    baseset <- 1:n # Here you can choose an appropriate base set for each
                   # pair of species i.e. only localities whose index is included
                   # in vector baseset are used to compute correlations
    P[i,j] <- P[j,i] <- fisher.test.midp(table(factor(X[baseset,i], levels=c(0,1)),
      factor(X[baseset,j], levels=c(0,1))))
  }
}

# One-tailed p-values for negative, positive and extreme correlations, respectively.

Pneg <- P
Ppos <- 1-P
Pext <- 2*pmin(P, 1-P)

# Adjust the p-values using the Benjamini-Hochberg method and pick the p-values
# that are at most alpha.
```

```
Cneg <- as.matrix(p.adjust(as.dist(Pneg),method="BH")) <= alpha
Cpos <- as.matrix(p.adjust(as.dist(Ppos),method="BH")) <= alpha
Cext <- as.matrix(p.adjust(as.dist(Pext),method="BH")) <= alpha
```