Oikos 125: 975–987, 2016
doi: 10.1111/oik.02838

Subject Editor: Werner Ulrich. Editor-in-Chief: Dries Bonte. Accepted 22 September 2015

A new cost-effective approach to survey ecological communities

F. Guillaume Blanchet, Pierre Legendre and Fangliang He

F. Guillaume Blanchet (guillaume.blanchet@math.mcmaster.ca), Dept of Mathematics and Statistics, Hamilton Hall, Room 218, McMaster University, 1280 Main Street West, Hamilton, ON, L8S 4K1, Canada, and Dept of Biosciences, Univ. of Helsinki, PO Box 65 (Viikinkaari 1), FIN-00014 Helsinki, Finland. – FGF and F. He, Dept of Renewable resources, Univ. of Alberta, 751 General Services Building, Edmonton, AB, T6G 2H1, Canada. – P. Legendre, Dépt de sciences biologiques, Univ. de Montréal, C.P. 6128, succursale Centre-ville, Montréal, QC, H3C 3J7, Canada.

Surveying ecological communities often means the tedious work of collecting detailed information on each species within each sampling unit (e.g. trap, transect, quadrat). In this paper, we first argue that presence–absence and abundance data are the two extremes of a spectrum of data formats. By counting individuals of each species within a sampling unit until either a predefined (user-defined) number of individuals is reached or all individuals of the species are counted, all intermediate cases can be generated. By independently correlating each intermediate case with the complete abundance data, we show that it is not necessary to count all individuals to recover the patterns of variation characterizing a community data table. When the same procedure is applied in combination with different distance coefficients such as the Hellinger, chord, χ^2, percentage difference or modified Gower, or the distance between species profiles, an even lower number of individuals per species need to be counted within a sampling unit for the patterns of variation defining a community to be recovered. By applying the same counting procedure to data collected during a pilot study, we show that the maximum number of individuals that need to be counted within a sampling unit for a species can be estimated from a pilot study containing as little as 3% of randomly selected sampling units throughout the complete survey area. An example of how to apply this new counting method is presented, using data from a boreal forest Carabidae community sampled in northwestern Alberta, Canada.

Ecological data collected in the field or obtained from laboratory experiments are the window through which we look to describe the patterns found in nature and understand the processes that generate these patterns. Data collection is undoubtedly the most important step of any ecological study because if data acquisition is incorrectly carried out, data analysis cannot yield correct results. Deciding how ecological data should be obtained is of crucial importance. Therefore, the sampling must be properly designed to address the related ecological questions or hypotheses to be tested.

Community ecologists have proposed many different approaches to sample organisms (Anderson 1965, Martin 1977). The resulting data are usually in the form of either presence–absence or abundance. There are pros and cons for the collection and analysis of either data type. It is usually more time- and cost-effective to obtain presence–absence data than abundance. However, accuracy (the detailed information the data convey) is lost because the information is only about species occurrence. In contrast, abundance data may be tedious to obtain, but the data are more informative, and more knowledge about the ecological processes underlying the community can be gained. As an additional concern, the cost of counting all individuals (abundances) in a community may be overwhelming, e.g. during insect outbreaks. It may also be unethical to count all individuals, for example when species determination requires killing individuals belonging to endangered species.

A number of sampling techniques have been developed in the last century to efficiently collect the most informative data possible. In phytosociology, Braun-Blanquet (1928) proposed a classification system specially designed to study groups of species that are morphologically similar or taxonomically related. It classifies a plant species coverage percentage on a five-level scale (1: $<5\%$ coverage, 2: 5%–25%, 3: 25%–50%; 4: 50%–75%, 5: 75%–100%). The scale of this classification has been refined (McLean and Ivimey-Cook 1951) and modified to adapt it to succession study (Londo 1976). Ideas of new sampling procedures were also proposed for plankton research where samples of hundreds of thousands and even millions of individuals for a single species are common. Frontier and Ibanez (1974) proposed to use a geometric progression to handle these particular data. One issue with these classifications is that a large amount of data needs to be collected in order to make the classification. In this paper, we propose a sampling technique that does not divide abundance into classes; instead, we propose a way to find a threshold beyond which it becomes unnecessary to count individuals. Compared to the classification methods,

the main advantage of our approach is that it is more cost-effective because large abundances of a species in a sampling unit do not need to be fully counted, thus reducing the time devoted to sampling.

Often, individuals of a species are clustered in space or through time. Aggregation of species in space may be the result of animal behaviour, dispersal limitation, or environmental patchiness to which organisms respond (Legendre and Fortin 1989). Through time, species succession and reproductive cycles may also generate clustered patterns (Legendre and Legendre 2012, chapter 12). Clustered patterns of species in space or time usually produce lower α-diversity, compared to species that are randomly or regularly distributed (He and Legendre 2002). If species are aggregated, they are generally found in large abundances in some sampling units (SUs) if the size of a cluster is smaller than or equal to that of a SU. When highly aggregated species are sampled, many individuals are found only in one or a few SUs. In that instance, the information lost by recording only presence–absence data can be very important. The ecological processes that control the abundance of a species may be quite different from the ones controlling its occurrence. As such, if only the presence–absence level is considered for a highly aggregated species, the information lost may be important to understand several ecological aspects of that species. This suggests that both presence–absence and abundance data are important to fully understand the factors influencing the distributions of species. A more complete debate of the importance of accounting for both presence–absence and abundance data in community studies is presented in Blanchet et al. (2014). In this paper, we propose a more cost-effective approach for collecting abundance data.

We first examined whether counting all individuals of a species in each SU is necessary to detect the distribution patterns characterizing a community. To that end, we studied how abundance distributions and aggregation influence the number of individuals of a species found in a SU. We then devised a method to determine a counting threshold, which is the maximum number of individuals per species that needs to be counted within a SU to extract sufficient information to correctly estimate the multivariate variation structure of the community as if all individuals had been counted. Regardless of the species considered, when the counting threshold is reached within a SU, the variation pattern of a community should be similar to that obtained after complete counts.

We constructed an example explaining the counting procedure we are proposing. Table 1 (top) shows the complete abundance of five fictitious species in two SUs. All individuals were counted to obtain these data. Assuming that the patterns defining this fictitious community are apparent if an arbitrary counting threshold of 8 individuals was used, the resulting community data would be the one presented in Table 1 (bottom). Whenever there are 8 or more individuals for a species in a SU, a count of 8 is recorded. For abundances smaller than 8, the total counted abundance is recorded. The counting threshold is thus applied for a species counted in a SU, which means for each cell of a community matrix.

The counting method proposed in this paper aims at finding a balance between presence–absence and abundance data that maximizes cost-efficiency when surveying ecological communities. The goal of this paper is to present a new, efficient way of obtaining data to study species communities. Because abundance and aggregation patterns can vary in many ways, our aim is not to find a universal counting threshold that applies to all communities. Rather, we propose the counting procedure presented in the previous paragraph, which can be used to determine the optimal counting threshold for any particular community of interest. The proposed procedure is validated using simulations. To illustrate how this procedure can be applied to real ecological data, we implemented it for a boreal forest Carabidae assemblage sampled in northwestern Alberta, Canada.

Table 1. Fictitious example illustrating the counting procedure proposed in this paper.

	Complete abundance				
	Species A	Species B	Species C	Species D	Species E
Sampling unit 1	0	2	10	100	900
Sampling unit 2	500	100	9	0	3
	After reaching a counting threshold of 8 individuals				
	Species A	Species B	Species C	Species D	Species E
Sampling unit 1	0	2	8	8	8
Sampling unit 2	8	8	8	0	3

From presence–absence to abundance

Presence–absence and full abundance data are two extremes of a spectrum of data formats characterizing composition and distribution of communities. Intermediate cases can be found by counting individuals of each species within a SU until either a predefined (user-defined) counting threshold is reached or all individuals of a species within the SU are accounted for (see the example in Table 1). By sequentially increasing the counting threshold from one to the largest number of individuals for a species found within a SU, all intermediate cases can be studied from presence–absence to full abundance data. We will refer to the case where all individuals are counted as the 'complete-abundance' count while all cases with counts of fewer individuals will be referred to as 'partial-abundance' counts.

In this paper we consider that a species is abundant if it is found with high abundance in at least one SU. As a rule of thumb, we consider the abundance of a species to be high if there are more individuals than the number of SUs. Conversely, a scarcely distributed species (or scarce species) can potentially be found in many SUs but its abundance is low in all SUs. Given these definitions, modest variations in abundance do not generally influence the interpretation of the patterns of variation of the abundant species but can importantly impact on the interpretation of the scarce species. Based on this premise, use of partial-abundance instead of complete-abundance counts can effectively produce information about scarce species while the associated loss

of information for common species may be largely inconsequential to understanding community variation patterns. Note that this does not mean that common species are not important in community ecology; it rather reflects the fact that variations in abundance affect common species much less than scarce ones. The challenge, therefore, is to find the lowest counting threshold that efficiently and accurately allows the description of the variation patterns of a community.

Simulating ecological communities

Species-abundance distributions (SAD) and patterns of spatial or temporal aggregation of species vary among communities. To evaluate how these two components influence the efficiency of partial-abundance data for characterizing patterns of variation in a community, we simulated community matrices comprised of 100 SUs and 50 species. Sample size and species richness should not influence the counting threshold required for community patterns to be accurately characterized because these two components do not affect the spatial or temporal aggregation patterns of species or the positively skewed abundance distribution typical of ecological communities.

In our simulations, the species abundances in the community ranged from 1 to 500 individuals (Fig. 1a). A probability was given to each abundance value following a lognormal distribution (Preston 1948) with a standard deviation of 5; this was the smallest value for which a community could be generated where at least one individual of any species was found in all 100 SUs of the community matrix. Because χ^2-based ordination methods commonly used in community ecology (e.g. principal component analysis after χ^2 transformation of the data and correspondence analysis) have trouble handling situations where SUs are found with no individuals for any species, we did not simulate these cases.

For each species of a simulated community, the spatial position of each individual was specified using a Matérn cluster point process (Illian et al. 2008), which depends on 1) a homogeneous Poisson process that characterizes the number and position of cluster centres (black points in Fig. 1d), 2) the average number of individuals within each cluster, and 3) the radii of clusters in space. In our simulation, the intensity of the Poisson process was defined by random selection of an integer value between 1 and the species abundance previously obtained from the lognormal distribution (Fig. 1b–c). This allowed the species to present spatial patterns ranging from aggregated into one patch (when the intensity of the Poisson process was 1) to randomly dispersed where each individual forms a separate spatial cluster (when the intensity of the Poisson process is the number of individuals in a species).

The average number of individuals within each cluster was obtained by dividing the abundance chosen from the lognormal distribution by the intensity of the Poisson process, which on average was the number of clusters generated. Individuals within each spatial cluster were uniformly distributed (Fig. 1b–d).

Cluster radii were generated as described in the caption of Fig. 1. The radii of the spatial clusters were used as surrogates for aggregation levels because each radius define the zone of influence of a cluster. Because clusters can overlap, patches of individuals may have different shapes and sizes (Fig. 1d). A variety of spatial patterns can thus be generated, even if the radii of all clusters were the same, when individuals of a species are grouped into more than one cluster. Unlike the number of clusters and the species abundance, which are related to each other in the Matérn cluster process, the cluster radii are chosen independently.

Because the Matérn cluster process is a random process, the total number of individuals of a simulated species varied around the abundance value defined in the first step of the simulation procedure. We inspected the abundance patterns of all simulated communities to ensure that the random variations resulting from the Matérn cluster process did not make the resulting abundance distributions diverge markedly from the reference lognormal SAD defined in the first step of the simulation. The random variations introduced by the Matérn cluster process only had minor influence on the abundance distributions, hence they did not affect the following steps of the simulations.

We generated a first set of communities where the range of aggregation was broad and another in which individuals were highly aggregated (Table 2). For all species in each set of communities, cluster radii were randomly sampled from a uniform distribution within the range of cluster radii. For each aggregation level, we simulated 1000 communities. The 'spatstat' package for statistical analysis of spatial point patterns (Baddeley and Turner 2005) in the R statistical language (<www.r-project.org>) was used to simulate these communities.

Simulations based on the same parameters were also performed using the broken-stick model (MacArthur 1957) as the reference SAD (Table 2). In the broken-stick model, the probability of finding a species with an abundance ranging from 1 to 500 is defined by randomly cutting a conceptual stick of unit length at 499 random points. The broken stick pieces are then ordered from the longest to the shortest to define the probability of sampling a species with an abundance of 1 through 500. Because the lengths of the pieces in the broken-stick model can vary among iterations, we used the expected stick piece length values to choose the abundance of a species (Barton and Davis 1956). The lognormal distribution and the broken-stick model are commonly used to model SADs, making them relevant choices to define our simulated community abundances.

Finally, we generated a third set of communities where the total abundance of all species was 500, with either a broad range of aggregation of the individuals or highly aggregated individuals, following the procedure described above (Table 2). As previously explained, because the Matérn cluster process is a random process, the exact abundance of each species was not necessarily 500; it often diverged slightly from that value. We also produced communities where the locations of individuals in the unit-size sampling area was defined by drawing values at random from a uniform distribution (minimum = 0, maximum = 1) for the x- and y-coordinates (Table 2). This last set of communities differed from the other two in that the species were not clustered but randomly distributed in the sampling area. These three sets of communities were used to evaluate the importance of the abundance distribution and aggregation patterns in determining a counting threshold.

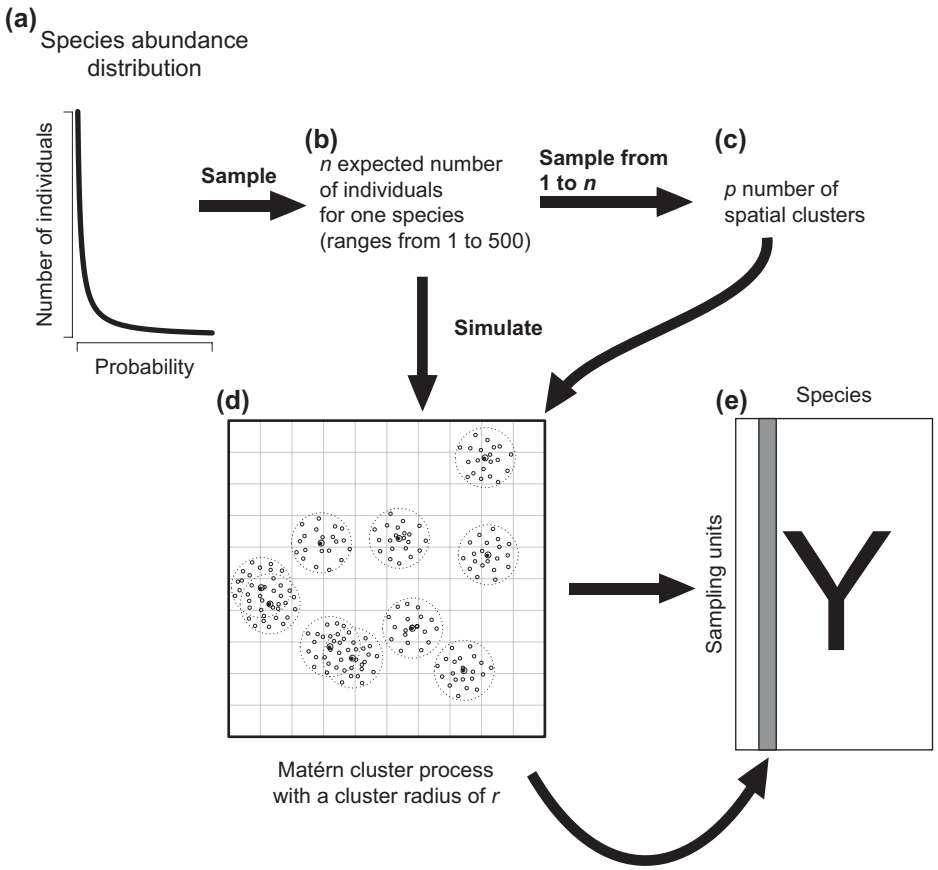

Figure 1. Diagram illustrating the steps followed to generate the abundances for one species of a simulated data table. All simulated species were constructed following this general scheme. (a) First, a species abundance distribution was used as reference, which defines the probabilities of a species to have particular numbers of individuals. Three species abundance distributions were used in all simulations (Table 2). (b) Second, a value was drawn from this species abundance distribution, defining the number of expected individuals n for the simulated species. (c) Third, we defined the number of spatial clusters p in which the species is distributed by randomly sampling a value between 1 and n (uniform distribution). (d) Fourth, in a square map of unit size, we distributed the individuals of the species (open points) using a Mátern cluster point process; the locations of the cluster centres are shown as black points. Cluster radii r were generated to describe species showing a broad range or a high level of aggregation (Table 2). For example, to choose a cluster radius for a highly aggregated species, we sampled a uniform distribution with minimum = 0.01 and maximum = 0.02 (Table 2). A species only had one radius size. (e) Lastly, we divided the sampling area into 100 quadrats (grey lines in d), counted the number of individuals of the species per quadrat and recorded the values in a (sampling units × species) matrix.

The sampling area was divided into 100 non-overlapping SUs of equal size using a regular grid (Fig. 1d) and the number of individuals of each species in each SU was counted for all simulated communities (Fig. 1e). This count provided the complete-abundance community matrix. Although in these simulations the SUs completely covered the study area, this condition is not necessary for the counting approach we are proposing. In the 'Ecological illustration' section, we applied our procedure to an experimental research area where the SUs only covered a small fraction of the study area.

Table 2. Components used in the simulation of ecological communities.

Species abundance distribution	Range of cluster radii for the Matérn cluster process
Log-normal distribution (sd = 5) (Preston 1948)	0.01–0.5 (broad range of aggregation) 0.01–0.02 (highly aggregated)
Broken stick model (Barton and Davis 1956, MacArthur 1957)	0.01–0.5 (broad range of aggregation) 0.01–0.02 (highly aggregated)
500 individuals for all species	0.01–0.5 (broad range of aggregation) 0.01–0.02 (highly aggregated) Individuals were uniformly distributed across the sampling area

Correlation of all partial-abundance with the complete-abundance data

To evaluate how much information is included in the increasingly precise partial-abundance data, we used the RV coefficient (Escoufier 1973, Robert and Escoufier 1976) to correlate the partial-abundance community matrices with the complete-abundance community matrix. The RV coefficient is defined as:

$$RV(\mathbf{X},\mathbf{Y}) = \frac{tr\left(\mathbf{X}\mathbf{X}^t\,\mathbf{Y}\mathbf{Y}^t\right)}{\sqrt{tr\left(\left(\mathbf{X}\mathbf{X}^t\right)^2\right)\,tr\left(\left(\mathbf{Y}\mathbf{Y}^t\right)^2\right)}} \quad (1)$$

where $\mathbf{X}$ and $\mathbf{Y}$ are two column-centred matrices with the same number of rows, t is the transpose of a matrix, and tr() the trace of a matrix. RV measures the co-inertia between

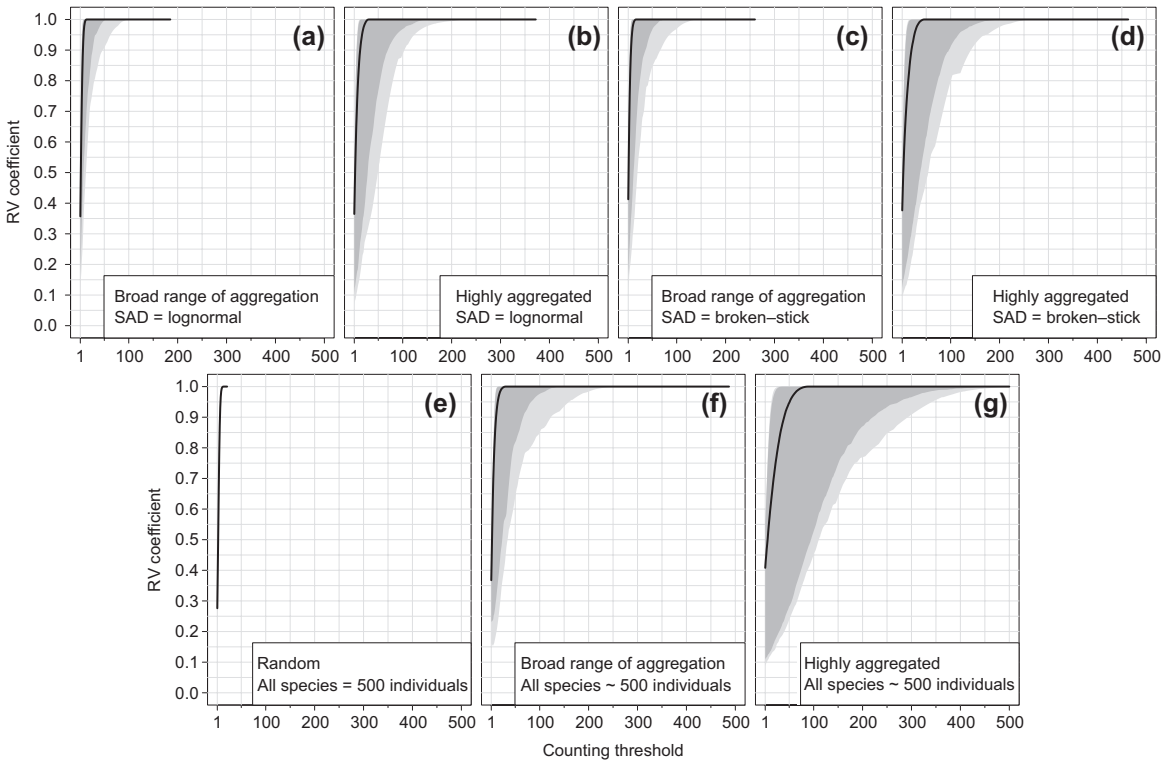

Figure 2. Simulation results of the multivariate correlation (using RV coefficient) between increasingly precise partial-abundance community data and the complete-abundance community data using raw data. The counting threshold (abscissa) is the maximum number of individuals counted for a species within a sampling unit. The ordinate represents the RV coefficients. Each panel presents the results of a set of simulated communities. In each panel, the leftmost result presents the RV coefficient associated to presence–absence data whereas the rightmost result shows the RV coefficient associated to the complete-abundance data. Light grey areas represent the 99% empirical confidence intervals of the simulation results (constructed using the 5th and 995th largest RV coefficients associated to each increasingly precise partial-abundance data, over 1000 simulations), the dark grey areas the 95% empirical confidence intervals (constructed using the 25th and 975th largest RV coefficients associated to each increasingly precise partial-abundance data), and the black lines are the medians per count threshold value.

two data matrices (i.e. the sum of the squared covariances between the two sets of variables) normalized by the total inertia in each matrix, which produces a coefficient ranging from 0 (no correlation) to 1 (perfect correlation). RV is the multivariate extension of the squared Pearson correlation coefficient. Note that although the relationships of a species with other species and the environment may be highly non-linear, it does not prevent the RV coefficient, a statistics derived from linear algebra, to efficiently capture the relationships among partial and complete-abundances because this relationship is linear.

In this paper, we exclusively use the lower bound of the 99% confidence interval to elaborate our conclusions. By using only these extreme scenarios of our simulation results, we minimize the impact of losing important information by counting too few individuals.

Figure 2 presents the correlation results between the increasingly accurate partial-abundance data (abscissa) and the complete-abundance data for the seven different sets of simulated communities. In this figure, the first four panels (a-d) highlight the results obtained from the two SADs (lognormal distribution and broken-stick model) while the remaining three panels (e-g) present communities where species are all abundant (~500 individuals) and the aggregation patterns range from random to highly aggregated. The most striking result was that the stronger the aggregation of species, the more individuals had to be counted to reach the same RV coefficient, compared to communities where species had a wider range of aggregation or were randomly distributed; compare panels (a) and (b), (c) and (d), (f) and (g) of Fig. 2. Another noteworthy observation was that in all simulated communities, including those in which individuals were randomly distributed in the study area (Fig. 2e), the number of individuals that needed to be counted to reach a high RV coefficient (e.g. > 0.9) was much smaller than the maximum number of individuals per species per SU found in the complete-abundance data. In Fig. 2, this is illustrated by the long horizontal line showing an RV coefficient $= 1$ found for all sets of communities. The length of the line depends on the maximum number of individuals found for a single species at one SU in the complete-abundance data.

These results confirm our hypothesis that the effort placed on counting all individuals is not necessary to reconstruct and analyse the variance of communities. Regardless of the SAD used, one does not have to count all individuals to identify the variation patterns defining a community. We want to stress that this result has major implications because it means that all studies where the interest is to understand how species in a community vary across the sampling units (generally either in space or through time) can be carried

out just as efficiently with a fraction of the data. This is an important observation because it implies that the counting approach we are proposing is not limited to community data, it can be applied to virtually any type of multivariate count data. In more statistical terms, all analyses of multivariate data currently carried out using simple or canonical ordinations can be performed using a reduced amount of information, with as much resolution as if all the data were used. For the remainder of this paper we will continue to present our results and interpretation in terms of ecological communities, species, and sampling units.

If we assume that the minimum RV coefficient required between partial and complete-abundance count should be at least 0.9 to give an acceptable representation of the community patterns, we can evaluate the cost of counting partial-abundance data with a good level of accuracy. Note that an RV coefficient of 0.9 is way above the 0.05 threshold commonly used to evaluate the level of significance of a model. Although the 0.9 RV coefficient threshold is arbitrary, it describes a very strong correlation between partial and complete-abundance data. In the most optimistic case, where all species are composed of 500 individuals uniformly distributed over the sampling area (Fig. 2e), an RV coefficient of 0.9 is reached with a counting threshold of 7 individuals, an RV coefficient of 0.95 is attained with a counting threshold of 8 individuals, and a 0.99 RV coefficient requires a counting threshold of 11 individuals. These results are interesting because they show that with randomly distributed individuals in space or time, it is possible to be very cost-effective when counting individuals.

At the other end of our simulation spectrum, when species are all abundant (i.e. composed of ~500 individuals) but highly aggregated (Fig. 2g), to reach a 0.9, 0.95 and 0.99 RV coefficient between partial and complete-abundance data, counting thresholds of 293, 343 and 430 individuals are needed, respectively (lower bounds of the 99% confidence intervals).

The results presented in Fig. 2 show that for the ecological situation we simulated, aggregation is the dominant factor increasing the number of individuals that need to be counted to reach a predefined RV coefficient. Thus, because species are known to aggregate in space and time, we must ask: Is this procedure applicable to natural community data?

Correlating partial to complete-abundance data using ecologically meaningful distances

Multivariate analyses of communities are rarely carried out on raw count data because using raw count data is equivalent to performing an ordination (or a clustering) analysis based on Euclidean distances among SUs. Euclidean distance is appropriate to answer ecological questions focusing on phenomena that cause changes in total abundances, such as disturbances or predation, but it is ill-adapted for other types of ecological questions such as β-diversity assessment (Anderson et al. 2011, Legendre and De Cáceres 2013). Numerous other distances have been proposed to study patterns in community data resulting from habitat variation among SUs. Legendre and Legendre (2012, chapter 7) described many distances specifically designed for modelling a variety of ecological data.

Currently, community data are almost always analysed with tools that use distances other than the Euclidean to extract ecological patterns. A typical example is the widespread use of correspondence analysis (CA, Greenacre 2007) and its canonical counterpart CCA (ter Braak 1986), which involves the χ^2-distance. It is important to note that the χ^2-distance artificially gives high weights to rare species occurring only at SUs where few or no other species occur (Greenacre 2013). For this reason, prior to choosing an analysis that relies on the χ^2-distance (e.g. CA and CCA), it is important to check the data to make sure this situation does not occur.

Aside from the χ^2-distance, other distances are commonly used to analyse the variation of community composition data. In these instances, it becomes relevant to study how partial-abundance counts can accurately characterize the community patterns defined by complete-abundance data using distances other than the Euclidean. We can then evaluate if by using different distances the information in the complete-abundance data can be recovered by counting a smaller number of individuals of each species.

Six distances commonly used with community composition data were considered: the Hellinger (Rao 1995), chord (Orlóci 1967), species profiles (Legendre and Gallagher 2001), χ^2 (Lebart and Fénelon 1971), percentage difference (Odum 1950), and modified Gower using base 2 logarithms (Anderson et al. 2006). All these distances are well adapted to the analysis of community composition data (Legendre and Legendre 2012, chapter 7). The χ^2-distance is widely used in ecology because it is the basis for CA and CCA. The percentage difference (also known as the Odum or, incorrectly, the Bray–Curtis distance) has been shown by Faith et al. (1987) to be well adapted to extract ecological patterns. Anderson et al. (2006) applied their transformation to the asymmetrical form of the Gower distance coefficient and called the combination the modified Gower dissimilarity (or distance). Anderson et al. (2006) transformed all abundances in a community matrix using 1 + (logarithm of the abundance), with the exception of 0s, which remain unchanged. When calculating the modified Gower distance, an increase in the base of the logarithm decreases the emphasis on abundances. For this reason, we chose the modified Gower using base 2 logarithms because any larger base of logarithm would give less importance to abundant species and thus make it easier to find a higher correlation between partial and complete-abundance data.

It is also possible to transform a community matrix in such a way that a distance other than the Euclidean is preserved in principal component ordinations. Two different approaches were followed to transform the community data (partial and complete), depending on the distance used. 1) Legendre and Gallagher (2001) have shown that the Hellinger, chord, species profiles, and χ^2-transformations can be applied directly to a community matrix using pre-transformations, without calculating a distance. A pre-transformation is a transformation applied to a community matrix before any analyses are carried out; the transformation changes the distance preserved between SUs in analyses involving linear models, such as principal component analysis (PCA), redundancy analysis (RDA), or *K*-means partitioning. Calculating the Euclidean distance

of a pre-transformed community matrix yields a symmetric distance matrix where the distance between each pair of SUs is the distance corresponding to the pre-transformation. For example, the Euclidean distance of chord-transformed data yields the chord distance matrix. We thus pre-transformed all partial-abundance community data and correlated them to the pre-transformed complete-abundance community matrix using the RV coefficient. 2) The percentage difference and modified Gower distance using base 2 logarithms cannot be obtained by pre-transforming a community matrix. To compare partial and complete-abundance for these two distances, we first calculated the distance matrices for all partial-abundance community matrices and the complete-abundance community matrix. We then performed a principal coordinate analysis (PCoA, Gower 1966) independently on each distance matrix (partial and complete). We used all the eigenvectors of each partial-abundance community data and correlated them with all the eigenvectors from the complete-abundance data using RV coefficients. PCoA is not used here as a dimension reduction tool, it is used to transform a distance matrix into a matrix with the same format as the original community matrix but where the species (columns of the community matrix) are replaced by eigenvectors. Performing a PCoA on percentage difference matrices, which are non-Euclidean, may generate complex eigenvectors (Legendre and Legendre 2012, section 9.3.4), which are difficult to handle. To ensure that no complex eigenvectors would be generated, we square-rooted all percentage difference distance matrices; after this transformation, the matrices are metric and Euclidean (Legendre and Legendre 2012, section 7.4.2), which ensures that no complex eigenvectors are generated in PCoA. Applying a square-root transformation to a modified Gower distance matrix, however, may not make it Euclidean. For this reason, we added a constant equal to the largest positive eigenvalue to all values of each modified Gower distance matrix to ensure that it becomes Euclidean and that no complex eigenvectors were generated in PCoA (Gower and Legendre 1986, Legendre and Legendre 2012, section 9.3.4). This procedure is known as the Cailliez correction (Cailliez 1983). All of the calculations presented in this paragraph were performed with the vegan package (Oksanen et al. 2015), with the exception of the PCoA, which was carried out with the stats package (<www.r-project.org>). All calculations were performed within the R statistical language.

To compare the different distances, we focused on simulated communities where the abundances of all species were large (~500 individuals) and species were highly aggregated (Fig. 2g). We focused on this set of communities instead of any other because it required the most individuals to reach high RV coefficients between partial and complete-abundance data when raw count data were used.

The results in Fig. 3 show that regardless of the transformation used, the 0.9, 0.95 and 0.99 RV coefficients between partial and complete-abundance were reached with much fewer individuals than when raw community data were used. Of the distances compared, the distance between species profiles (Fig. 3, green) required the largest number of individuals to reach the same RV coefficients between partial and complete-abundance, compared to the other distances. To reach a 0.99 RV coefficient, a counting threshold of at least 350 individuals was needed. Although it is the worst result among the compared distances, it is still more efficient than using raw count data (compare to Fig. 2g) where counting 430 individuals was required.

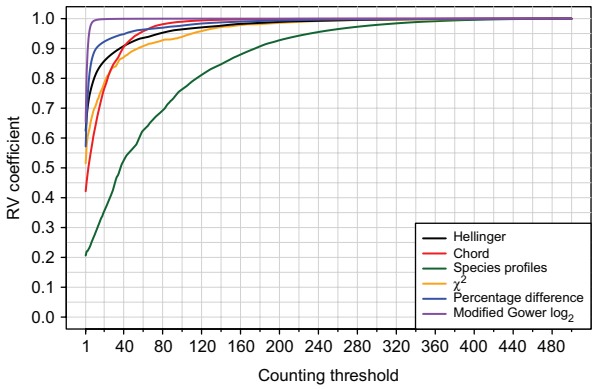

Figure 3. Simulation results of the multivariate correlation (using RV coefficient) between increasingly precise partial-abundance community data and the complete-abundance community data. All species in the simulated communities were composed of ~500 individuals and species were highly aggregated in the sampling area. The leftmost result presents the RV coefficient associated to presence–absence data for each distance whereas the rightmost result shows the RV coefficient associated to the complete-abundance data. The counting threshold (abscissa) is the maximum number of individuals counted for a species within a sampling unit. The ordinate represents the RV coefficients. The curves represent the lower bound of the 99% empirical confidence intervals of the simulation results, constructed using the 5th largest RV coefficients associated to each increasingly precise partial-abundance data.

The χ^2 (Fig. 3, orange) and Hellinger (Fig. 3, black) distances also required many individuals to reach a predefined RV coefficient. To attain the 0.99 RV coefficient, a counting threshold of 224 individuals was needed for the χ^2-distance and of 207 for the Hellinger distance. The percentage difference (Fig. 3, blue) and chord (Fig. 3, red) distances were more efficient, requiring counting thresholds of 160 and 102 individuals, respectively, to reach a 0.99 RV coefficient between partial and complete-abundance data. For these simulated data, the best result was obtained from the modified Gower distance (Fig. 3, purple): a counting threshold of merely 9 individuals was needed to reach a 0.99 RV coefficient. The counting thresholds were calculated by referring to the lower bounds of the 99% confidence interval.

The results found with the other sets of simulated communities are presented in the Supplementary material Appendix 1. They yield the same conclusion as discussed in the two previous paragraphs, although a lower counting threshold was needed when any of the other six sets of simulated communities was used.

The results that stem from Fig. 2 and 3 clearly show that few individuals need to be counted in a community for the variation patterns to be identified. However, these results are not helpful when planning a survey because they do not suggest a counting threshold before all sites have been sampled.

Pilot study: the basis for a new sampling procedure

The generality of the results presented in Fig. 2 and 3 makes it possible to apply the same procedure to a reduced

number of randomly selected sampling units, in a pilot study, to estimate the counting threshold required for a sample that provides a good representation of the actual community. Pilot studies have been used to evaluate the cost in time and money to perform a survey or an experiment. In this paper, a pilot study is a study used as a reference to estimate a counting threshold. It can result from data collected during a previous sampling year, or it can include SUs selected at random among the possible sampling sites (e.g. using a random number generator to identify geographical coordinates to locate the pilot SUs within the study area), or in a systematic design in the sampling area of an on going study.

The size of the pilot study is important to infer a meaningful counting threshold. A pilot study that includes two SUs is unlikely to yield the same counting threshold as one that comprises ten SUs. From the results in Fig. 2 and 3, we know that it is possible to estimate community patterns by counting a fraction of all individuals. In this section, our goal is to evaluate the minimum number of SUs that need to be randomly sampled in a pilot study to ensure that the counting threshold associated to a particular RV coefficient can be reached.

To ensure that our simulation results can be used as a reference for studies on real communities, we estimated the counting threshold of the pilot study data by constructing a 99% empirical confidence interval from the RV coefficient correlating partial and complete-abundance of the full survey data. We consistently used the lower bound of the confidence interval. In other words, we referred to extreme cases where the number of individuals to count is large. Also, the choice of the SUs in the pilot study may considerably influence the estimation of the counting threshold, especially when the number of SUs included in a pilot study is small.

For our simulation results to be applicable to a wide range of studies, we randomly sampled SUs to be included in the pilot study 100 times. Using the communities previously simulated, we randomly generated pilot studies that included 3% of the SUs. Referring to the lower bound of the 99% confidence interval of the randomly sampled pilot studies, we correlated the increasingly accurate partial-abundance data with the complete-abundance data using RV coefficients. This is the same procedure as in the previous sections but using only information from randomly sampled pilot data. Note that the number of species in a pilot study will not affect the estimation of the counting threshold because the procedure we propose focuses on variation at the individual level. We then compared the 0.9, 0.95 0.99, 0.999 and 0.9999 RV coefficients calculated from the pilot data with the 0.9, 0.95, 0.99, 0.999 and 0.9999 RV coefficients computed from the full-survey data. We repeated the procedure using pilot studies that included 5%, 10%, 15%, …, up to 95% of the SUs. This procedure was carried out for each set of simulated communities and using all distances considered in the previous section.

In our simulations, we know the abundance and aggregation patterns of the sampled species because these parameters formed the basis of our artificial communities. However, such patterns are difficult to evaluate using only data obtained from a pilot study. For this reason, in our interpretation of Fig. 4 and Supplementary material Appendix 2 Fig. A2.1–A2.7, we consistently selected the number of SUs where the survey-wide RV coefficient calculated between partial and complete-abundance data was exceeded. This ensured that the survey-wide RV coefficients between partial and complete-abundance were reached even in the most difficult scenarios.

Figure 4 presents the worst scenarios we simulated for the different distances compared. For the Hellinger, chord, χ^2 and percentage difference distances, these results were obtained from the set of communities where the abundance of each species was large (~500 individuals for each species) and species had a broad range of aggregation levels. For the distances between species profiles and modified Gower distance calculated with base 2 logarithms, the worst scenario was obtained with the set of communities where the abundance of each species was large (~500 individuals for each species) and species were highly aggregated. Focusing on the worst cases makes our interpretation of these results more conservative and makes these simulation results applicable to a broader range of studies. From these results, the modified Gower distance shows a clear advantage over the other dissimilarities. It is the only distance where by using a pilot study that includes only 3% of the SUs, the survey-wide 0.95 RV coefficient calculated between partial and complete-abundances can be attained. The RV coefficient calculated between partial and complete abundances within the pilot study data needs to be at least 0.9999 for the survey-wide 0.95 RV coefficient to be reached when using the modified Gower distance. To reach a survey-wide 0.99 RV coefficient, a pilot study covering at least 15% of the study area needed to be surveyed with a pilot study where a 0.999 RV coefficient was used as a reference. If a survey-wide RV coefficient above 0.999 is required, the modified Gower is the only distance that is worth using because it is the only dissimilarity that makes it possible to reach this very high level of accuracy with a pilot study.

The percentage difference is the next best choice of distance after the modified Gower distance. We can expect that by using 35% of the study area it is possible to reach a survey-wide 0.95 RV coefficient if we refer to the pilot study 0.9999 RV coefficient.

The chord distance is also interesting because with a pilot study that includes 75% of the survey area, a survey-wide 0.99 RV coefficient can be obtained when referring to the 0.9999 RV coefficient calculated from the pilot study. The Hellinger- and χ^2-distances and the distance between species profiles required that a pilot study included, respectively, 65%, 75% and 85% of the SUs to reach a survey-wide 0.95 RV coefficient using the 0.9999 RV coefficient obtained from the pilot study data.

As for the Euclidean distances, if one needs to use it to extract community patterns, it is preferable to use a pilot study that is at least as large as the surveyed area. For example, using data collected in the same study area during a previous year on the same group of organisms could serve as a reference pilot study. In the 'Ecological illustration' section, we show how data from a previous year can be used as a pilot study to estimate a counting threshold.

The results discussed above represent the worst-case scenarios of our simulations. Because species abundance distributions are always positively skewed for ecological communities, one can refer to the results obtained from

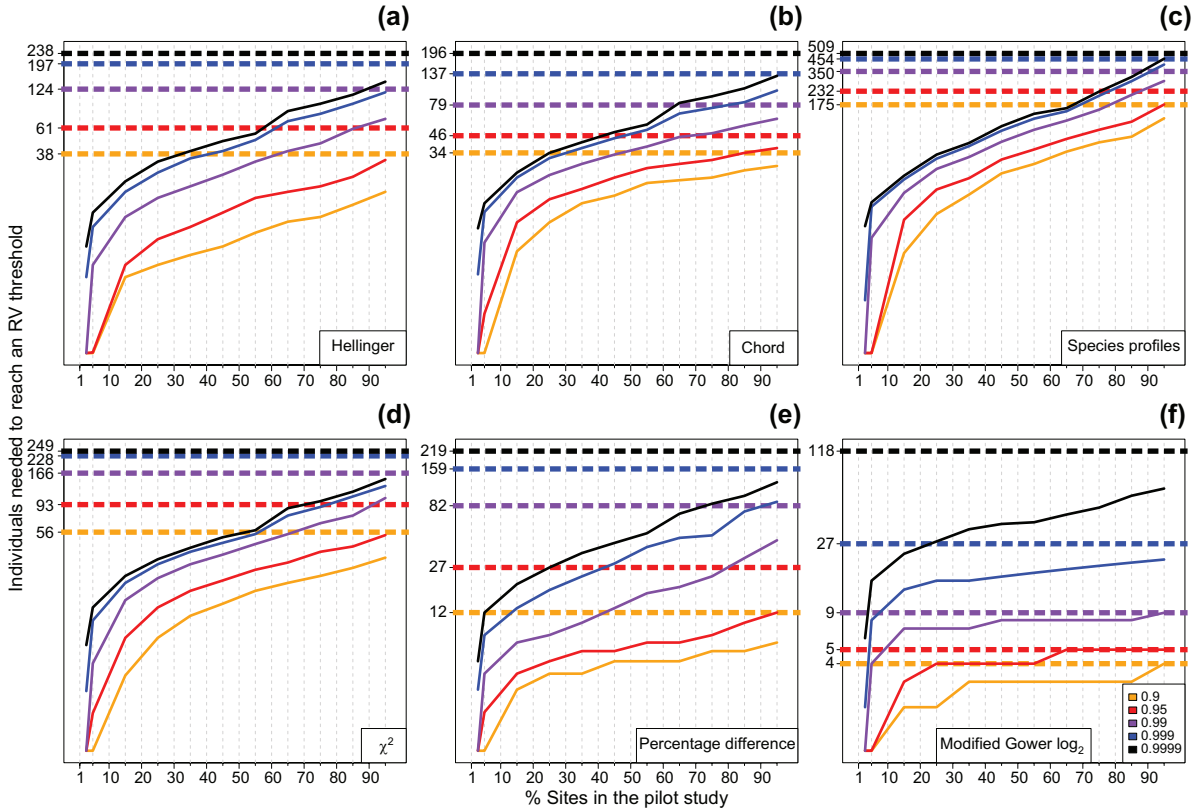

Figure 4. Percentage of the sites required in a pilot study (abscissa) to accurately estimate the number of individuals that need to be considered when sampling partial abundances. In this figure, we focus on six distances and show the number of individuals (ordinate) needed to meet RV coefficients of 0.9, 0.95, 0.99, 0.999 and 0.9999 calculated between partial and complete-abundance data. The ordinates were log-transformed for visual clarity. The species in the simulated communities were all composed of ~500 individuals and the range of their spatial aggregation level was broad for the Hellinger (a), chord (b), χ^2 (d) and percentage difference distances (e) while the spatial aggregation level was high for the distance between species profile (c) and modified Gower distance with base 2 logarithms (f). The survey-wide RV coefficients are represented by dotted lines; they are the lower bounds of the 99% confidence intervals of the simulations results presented in Fig. 3 and Supplementary material Appendix 1. The full lines represent the RV coefficients between partial and complete-abundance calculated using the pilot study data. TTo obtain the pilot study RV coefficient, the number of sampling units associated to the percentage (%) of sites to be included in the pilot study was randomly sampled 100 times. From this sample, the lower bound of the 99% confidence interval was used to obtain the pilot study RV coefficient. This procedure was carried out for pilot studies that included 3%, 5%, 10%, ..., to 95% of the study area."

simulated communities where the species abundance distributions follow a lognormal distribution or a broken-stick model (Supplementary material Appendix 2 Fig. A2.1–A2.7). However, these simulations should only be used as references if all species sampled in a community are used. In any case, it is preferable to refer to the scenario where the number of SUs to sample is the largest, as we did in this section.

Ecological illustrations – boreal forest ground beetles (Carabidae)

To illustrate how this new method can be applied to real ecological data, we used data about boreal Carabidae. In that study, 196 sites were sampled using pitfall traps (Spence and Niemelä 1994) in a near-regular grid of 70 km^2 of mature boreal forest at the Ecosystem Management Emulating Natural Disturbances (EMEND) research site in northwestern Alberta, Canada (Bergeron et al. 2011, 2012, Blanchet et al. 2013). The data include 9869 individuals pertaining to 45 carabid species. *Calathus advena* was the most abundant at any single site with 128 individuals.

By using data from all sites, we first estimated the counting threshold for future studies assuming that the only information available is the species abundance gathered from the 196 sites presented above. We then estimated the counting threshold that would be needed to extract the patterns found in this carabid assemblage. We made both estimations using all distances considered previously.

When sampling is carried out using pitfall traps, it is common to correct for disturbances (e.g. flooding of traps) by dividing the abundance of each species by the number of days a trap was active. The procedure proposed in this paper is unaffected by such normalization because the time for which a trap was active remains constant regardless of the number of individuals of a species counted in a trap. In other words, the normalization does not affect the calculation of the counting threshold. For this reason, we can omit any normalizing procedure applied on the SUs when estimating the counting threshold.

Using all beetles sampled at the 196 sites as a pilot study, we estimated that with a counting threshold of 10 individuals, a 0.9 RV coefficient can be reached with all distances compared in this study except for the Euclidean distance

(Table 3). This counting threshold of 10 individuals can be used for any future carabid study carried out on the EMEND landscape that includes up to 196 sites. However, when for example the Hellinger and modified Gower (log base 2) distances are used, an RV coefficient higher than 0.95 can be reached by counting up to 10 individuals per species. If the χ^2-distance is used, an RV coefficient of at least 0.99 can be attained with a counting threshold of 10 individuals. Note that this result is an artefact of the χ^2-distance, which gives high weights to rare species occurring only at SUs where few or no other species occur, as explained in the 'Correlating partial to complete-abundance data using ecologically meaningful distances' section.

To illustrate the implication of the previous result for ecologists, we compared two PCAs carried out with the Hellinger pre-transformation using Procrustes analysis (Gower 1971). The first PCA was carried out on the complete-abundance data while the other was performed on partial-abundance data with a counting threshold of 10 individuals. The comparison of the first two PCA axes is presented in Fig. 5. In a nutshell, although there are some differences between the partial and complete-abundance data, the differences are minimal and would not influence in any way the interpretation of the results. This means that any ecological question where the variation in the data needs to be interpreted (e.g. using ordinations) can be answered with partial-abundance data, as it can with complete-abundance data, as long as the procedure described in this paper to choose a good RV threshold is followed. For this example, the RV coefficient was 0.967. With a counting threshold of 10 individuals, 58.76% of the individuals were counted.

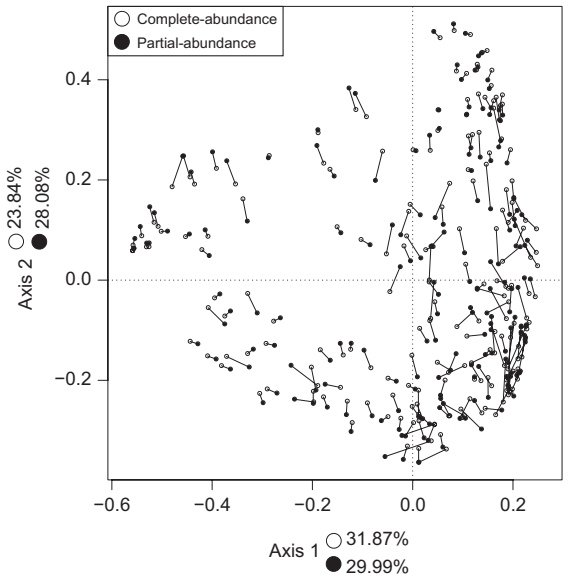

Figure 5. Procrustes plot comparing the first two axes of a principal component analysis (PCA) carried out on the Hellinger transformed data using complete-abundances as the target dataset (open circles) and the Hellinger transformed data using partial-abundance with a counting threshold of 10 individuals as the rotated dataset (black circles). Line segments show the differences in positions of SUs between the partial and complete-abundance data. The RV coefficient between the partial and complete abundance was 0.967.

As can be logically expected, when more individuals are counted, a higher RV coefficient can be reached. For example, with a counting threshold of 21 individuals, a 0.99 RV coefficient can be attained if the Hellinger and modified Gower (log base 2) distances are used.

These counting thresholds can be translated into cost-effectiveness by evaluating the number of individuals that would be counted in total to reach a predefined RV coefficient if the same 196 sites were sampled. For example, to reach a 0.9 RV coefficient with the modified Gower distance (log base 2), a total of 3513 individuals would have to be counted. This is 35.6% of all the individuals counted for the full survey. This evaluation of cost-effectiveness shows it is possible with real ecological data to be more efficient, by counting only a subset of all the individuals.

Assume now that we want to plan a survey where 200 sites are sampled to study the carabid assemblage on the EMEND landscape, as was the original plan for the carabid study (Bergeron et al. 2011), but that no previous data is available to evaluate a counting threshold. To estimate the counting threshold, we first have to decide the particular distance measure with which all analyses will be performed. We chose the percentage difference distance for this example because Bergeron et al. (2011) used it to analyse the same data. Referring to the simulation results in Fig. 4e, we know that 20 randomly selected sites (10% of the sampling area) are required to estimate the counting threshold for a survey-wide 0.9 RV coefficient between partial and complete-abundance data by referring to the pilot study 0.9999 RV coefficient calculated between partial and complete-abundance data. Because the results in Fig. 4 present extreme cases, it is highly unlikely that the 20 sites chosen in the pilot study will not reach the minimum number of individuals required to reach a survey-wide 0.9 RV coefficient. In fact, it is likely that the pilot study will show numbers of individuals larger than the minimum required. As an example, if we randomly choose 20 sites in the study area 1000 times and evaluate the counting threshold by referring to a 0.9999 RV coefficient for all iterations, we can estimate that 70 would be the average number of individuals necessary as our counting threshold. This counting threshold is much larger than the 9 individuals required when we have information from the whole data set (Table 3). It is comforting that the pilot study is proposing a number of individuals much larger than the minimum number required if the survey-wide data were considered.

Discussion

Counting partial-abundance data is a cost-effective approach for sampling ecological communities. Because of its flexibility, the approach proposed in this paper makes it possible for researchers to decide the accuracy they want to have in the data they collect and then to reduce (or increase) the sampling effort to achieve this accuracy. We would like to reemphasize here that an RV coefficient of at least 0.9 should be used to have a high enough level of accuracy and that although the 0.9 RV coefficient is an arbitrary threshold it describes very strong correlation.

Table 3. Counting threshold (top values of each table cell) and percentage of the whole data needed to reach a predefined RV coefficient (lower values of each table cell) for the ecological illustration on Carabidae.

	RV threshold				
	0.9	0.95	0.99	0.999	0.9999
Euclidean	49 95.26%	66 97.68%	96 99.30%	116 99.85%	124 99.96%
Hellinger	4 35.60%	8 52.80%	21 78.83%	48 95.02%	85 98.95%
Chord	9 55.91%	15 69.90%	35 90.14%	74 98.28%	102 99.48%
Distance between species profiles	10 58.76%	16 71.59%	37 91.13%	78 98.56%	105 99.57%
χ^2	3 29.49%	5 40.79%	8 52.80%	16 71.59%	34 89.61%
Percentage difference	9 55.91%	16 71.59%	34 89.61%	68 97.84%	101 99.45%
Modified Gower $\log_2$	4 35.60%	5 40.79%	13 65.97%	28 85.63%	46 94.51%

In our simulations, we showed that it was possible to estimate a counting threshold by using as few as 3% of the SUs. Although our results lead us to believe that, up to a certain extent, a larger pilot study points to a smaller counting threshold, it is left at the discretion of the researcher to consider a larger number of SUs in the pilot study. However, for surveys where the number of SUs to be sampled is small, a pilot study should include a minimum of 5 SUs to ensure that the chance of sampling too few individuals is low.

Although we tried to make our simulations as general as possible, it was not possible to simulate all possible cases found in nature (Milligan 1996). Of the choices we made, we decided to simulate species communities so that it would never happen for an SU to be empty. Although it can be common for such SUs to occur, it is technically challenging to deal with these data because commonly used approaches in community ecology such as CA or CCA do not handle such data well. This problem is beyond the scope of the paper.

Pilot studies are at the core of the procedure we are proposing in this paper. If in the pilot study the counting threshold calculated seems too low, considering more SUs in the pilot study should improve the estimation of the counting threshold. Because the information gained in the pilot study may be included in the full study, the cost of considering additional SUs in the pilot study is not as important as it would be if it were carried out independently from the survey-wide study. Moreover, if the SUs considered in the pilot study present a surprisingly low number of individuals, these SUs should be quick to count compared to SUs with larger abundances, making the effort to include new SUs in the pilot study less of a constraint.

Counting partial-abundance data in a pilot study may be easy or difficult depending on the survey design and the organisms sampled. For example, if individuals are collected in the field and brought to a laboratory for sorting, identification, and counting, it is easy to re-evaluate the counting threshold with a minimum of effort because all pilot study SUs are readily available. Insects, mites, and spiders are typical examples of organisms that allow such flexibility because they are usually sampled in traps, and when sorting is carried out all trap contents are easily accessible. Conversely, if organisms need to be recorded in the field (e.g. trees or birds), a pilot study would need to be carried out before the full-scale survey begins. However, as explained in the previous paragraph, the time spent on the pilot study is not usually lost because the data collected while carrying out the pilot study can often be included in the final data set. Moreover, as we have shown, the pilot study will make it possible to be much more cost-effective when surveying.

We have also shown that the distance function used to analyse the data can have tremendous impact on the number of sites to consider in a pilot study. In that instance, it becomes important to choose the distance with which all analyses will be carried out before sampling the community. For simple and canonical ordinations, it is common for researcher to be ambivalent about the choice of dissimilarity to use. This confusion is justified at least for canonical ordinations where the differences obtained by using one distance or another are usually minor (Blanchet et al. 2014).

Our recommendation to users is to first determine which distances one wants to consider. Then, using survey results from a pilot study, compute the counting thresholds corresponding to the selected value of RV corresponding to the different distances in the selected set, and choose for the full-scale study the distance that provides the lowest counting threshold. The detailed properties of the different distances, in terms of their ability to reproduce the variation of the complete data in statistical analyses of community variation, remain to be investigated in further simulation studies involving different types of communities. Statistical and other properties of dissimilarity coefficients that are of importance for the study of β-diversity have been described by Legendre and De Cáceres (2013).

It is likely that in the pilot study as well as for the entire survey, some species have been missed. The distances we compared in this paper are known not to be sensitive to rare species, so that if a species is absent from the sampling or only a few individuals were found, the counting threshold will not be affected. The χ^2-distance is the only exception because it is known to give high weight to rare species (Legendre and Legendre 2012, p. 308), making the use of this distance conceptually problematic when estimating a counting threshold. This distance also lacks an important property for β-diversity assessment: with that distance, two sites without species in common do not necessarily have the largest dissimilarity.

In addition, because the counting procedure we proposed is carried out on individuals, species missed in the pilot study can be accounted for in the full survey without any associated counting problems. This also means that if the interest of a study is to estimate species richness, the counting approach proposed here does not prevent such undertaking.

Nowadays, studies in community ecology are focussing not only on species richness, but also on the processes of the variation from site to site. Ordinations allow ecologists to study the main axes of variation in community data. Canonical ordinations and multivariate analysis of variance allow

them to test specific hypotheses about the processes involved in community variation. That variation can be decomposed and studied at different spatial/temporal scales by canonical analysis on spatial/temporal eigenvectors (Dray et al. 2006, Legendre and Gauthier 2014). New indices decompose the dissimilarity among sampling units into replacement and richness difference components (Legendre 2014). All these new methods of analysis can be applied to complete-abundance and partial-abundance data.

The counting procedure we propose in this paper applies to a broad range of ecological surveys, but not to all of them. If the group of studied species are challenging to identify because little is known about this group or because extensive manipulations of each individuals is required to make sure the species was well identified, it may become irrelevant to use the proposed counting approach. In such circumstances, all individuals would need to be considered anyway. For example, the counting procedure cannot be applied to insect groups that can only be accurately identified through a dissection of their genitals because each animal needs to be handled individually for identification. If they are not dissected, they cannot be identified properly. However, in many cases trained researchers can readily identify organisms to the species level without having to extensively manipulate each individual. In a nutshell, our counting approach can be used for any group of organisms where individuals can easily be identified and counted.

A situation where our counting approach should not be used is when the goal of the study is to assess species abundance patterns for a group of species, e.g. through species-abundance distributions. Because our counting procedure limits the number of individuals to account for, if used in such a context, it would invariably lead to incorrect results for abundant species which have not been sampled to their full extent.

The counting procedure we proposed in this paper does not impose any constrain on the analyses carried out on the data aside from the distance used to define the counting threshold. The reason why this statement can be made with such certainty is because we used the RV coefficient to make the comparisons. As explained in the section 'Correlation of all partial-abundance to the complete-abundance data', the numerator of the RV coefficient is the sum of the squared covariances between two data matrices. So, if for example the RV coefficient between the partial and complete-abundance community data is 0.9, any data analyses focussing on variation in species abundance across sites using the partial-abundance data will be within a 90% confidence interval of the result that would be obtained with the complete-abundance data. This means that as long as the distance used to define the counting threshold is preserved, any ordination (simple or canonical), or variations of these methods, can be applied without losing interpretability of the results. It is however important to stress that since partial abundances are used, variation due to the counting threshold should be taken into account, as explained above.

The method proposed in this paper builds on the idea that for a sampling unit where there are a lot of individuals, it is not the absolute 'true' number of individuals found at the SU that is important, but the fact that there are many. The counting approach we are proposing focuses on finding a way to define how many individuals are enough to still gain unbiased knowledge about the variation among sampling units. Our counting procedure also accounts for the species that occur scarcely throughout the sampling area. Just knowing where these scarce species occur would likely lead to the same conclusions as knowing their abundance, simply because of their scarce distribution. Blanchet et al. (2014) discussed the importance of analysing presence–absence and abundance community data independently because it often produces a better understanding of the ecology of the species.

In this paper, we showed that using prior information from a pilot study to evaluate community patterns can be useful to increase cost-effectiveness while minimizing the loss of information. The proposed counting procedure has the potential to be applied to numerous types of studies within and outside the scope of ecology. In ecology, it can be valuable for large-scale monitoring studies such as the Alberta Biodiversity Monitoring Institute project (Boutin et al. 2009). Another interest of the method proposed in our paper is that governmental surveys could be carried out at lower costs, and thus could be made more extensive through space and time with the same budget constraints. In some studies involving organisms whose sizes differ greatly, biomass data can be used instead of abundances, and it is possible to evaluate a counting threshold on biomass data using the approach described in this paper. Our approach is also applicable to landscape genetics where gene (or marker, etc.) frequencies in local populations are used instead of species frequencies. Since the lab work is costly in genetic studies, our approach could lead to important savings of technician time and materials. Although our counting approach has been presented mainly in the context of terrestrial surveys, it can be applied as well to aquatic communities (e.g. counts of landings in fisheries studies). Outside the scope of ecology, it can be applied to any situation where many count variables are describing a number of SUs.

Acknowledgements – We are grateful to John R. Spence, Dave W. Roberts, Mark A. Lewis and Ilkka Hanski for insightful comments on early drafts of the manuscript. This research was supported by NSERC grants to F. He and P. Legendre. The procedures presented in this paper have been implemented in the countComm R package which can be found at the following webpage: <www.numericalecology.com>.

References

Anderson, M. J. et al. 2006. Multivariate dispersion as a measure of beta diversity. – Ecol. Lett. 9: 683–693.

Anderson, M. J. et al. 2011. Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist. – Ecol. Lett. 14: 19–28.

Anderson, R. M. 1965. Methods of collecting and preserving vertebrate animals. – Dept of the Secretary of State.

Baddeley, A. and Turner, R. 2005 spatstat: an R package for analyzing spatial point patterns. – J. Stat. Soft. 12:1–42. R package ver. 1.41-1.

Barton, D. E. and Davis, F. N. 1956. Some notes on ordered random intervals. – J. R. Stat. Soc. B Met. 18: 79–94.

Bergeron, J. A. C. et al. 2011. Landscape patterns of species-level association between ground-beetles and overstory trees in boreal forests of western Canada (Coleoptera, Carabidae). Proc. Symp. honoring the careers of Ross and Joyce Bell and

their contributions to scientific work, Burlington, VT, June 2010 (ed. T. L. Erwin). – ZooKeys 147: 577–600.
Bergeron, J. A. C. et al. 2012 Ecosystem classification and inventory maps as surrogates for ground beetle assemblages in boreal forest. – J. Plant Ecol. 5: 97–108.
Blanchet, F. G. et al. 2013. Landscape effects of disturbance, habitat heterogeneity and spatial autocorrelation for a ground beetle (Carabidae) assemblage in mature boreal forest. – Ecography 36: 636–647.
Blanchet, F. G. et al. 2014. Consensus RDA across dissimilarity coefficients for canonical ordination of community composition data. – Ecol. Monogr. 84: 491–511.
Boutin, S. et al. 2009. A new approach to forest biodiversity monitoring in Canada. – Forest. Ecol. Manage. 258: S168–S175.
Braun-Blanquet, J. 1928. Pflanzensoziologie. Grundzüge der Vegetationskunde. – Biologische Studienbücher 7.
Cailliez, F. 1983. The analytical solution of the additive constant problem. – Psychometrika 48: 305–308.
Dray, S. et al. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). – Ecol. Modell. 196: 483–493.
Escoufier, Y. 1973. Le traitement des variables vectorielles. – Biometrics 29: 751–760.
Faith, D. et al. 1987. Compositional dissimilarity as a robust measure of ecological distance. – Vegetatio 69: 57–68.
Frontier, S. and Ibanez, F. 1974. Utilisation d'une cotation d'abondance fondée sur une progression géometrique, pour l'analyse des composantes principales en écologie planctonique. – J. Exp. Mar. Biol. Ecol. 14: 217–224.
Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. – Biometrika 53: 325–338.
Gower, J. C. 1971. Statistical methods of comparing different multivariate analyses of the same data. – In: Hodson, F. R. et al. (eds), Mathematics in the archaeological and historical sciences. Edinburgh Univ. Press, pp. 138–149.
Gower, J. C. and Legendre, P. 1986. Metric and Euclidean properties of dissimilarity coefficients. – J. Classification 3: 5–48.
Greenacre, M. 2007. Correspondence analysis in practice, 2nd edn. – Chapman and Hall.
Greenacre, M. 2013. The contributions of rare objects in correspondence analysis. – Ecology 94: 241–249.
He, F. and Legendre, P. 2002. Species diversity patterns derived from species-area models. – Ecology 83: 1185–1198.
Illian, J. et al. 2008. Statistical analysis and modelling of spatial point patterns. – Wiley.
Lebart, L. and Fénelon, J.-P. 1971. Statistique et Informatique Appliquées. – Dunod.

Supplementary material (available online as Appendix oik-02838 at <www.oikosjournal.org/appendix/oik-02838>). Appendix 1–2.

Legendre, P. 2014. Interpreting the replacement and richness difference components of beta diversity: replacement and richness difference components. – Global Ecol. Biogeogr. 23: 1324–1334.
Legendre, P. and Fortin, M.-J. 1989. Spatial pattern and ecological analysis. – Vegetatio 80: 107–138.
Legendre, P. and Gallagher, E. 2001. Ecologically meaningful transformations for ordination of species data. – Oecologia 129: 271–280.
Legendre, P. and Legendre, L. 2012. Numerical ecology, 3rd English edn. – Elsevier.
Legendre, P. and De Cáceres, M. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. – Ecol. Lett. 16: 951–963.
Legendre, P. and Gauthier, O. 2014. Statistical methods for temporal and space–time analysis of community composition data. – Proc. R. Soc. B 281: 20132728.
Londo, G. 1976. The decimal scale for relevés of permanent quadrats. – Vegetatio 33: 61–64.
MacArthur, R. H. 1957. On the relative abundance of bird species. – Proc. Natl Acad. Sci. USA 43: 293–295.
Martin, J. E. H. 1977. The insects and arachnids of Canada. Part 1. – Agriculture Canada.
McLean, R. C. and Ivimey-Cook, W. R. 1951. Textbook of theoretical botany. – Longmans, Green.
Milligan, G. W. 1996. Clustering validation: results and implications for applied analyses. – In: Arabie, P. et al. (eds), Clustering and classification. World Scientific, pp. 341–375.
Odum, E. P. 1950. Bird populations of the Highlands (North Carolina) plateau in relation to plant succession and avian invasion. – Ecology 31: 587–605.
Oksanen, J. et al. 2015. vegan: community ecology package. R package ver. 2.3-0. – <http://CRAN.R-project.org/package=vegan>.
Orlóci, L. 1967. An agglomerative method for classification of plant communities. – J. Ecol. 55: 193–206.
Preston, F. W. 1948. The commonness, and rarity, of species. – Ecology 29: 254–283.
Rao, C. R. 1995. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. – Qüestiió 19: 23–63.
Robert, P. and Escoufier, Y. 1976. A unifying tool for linear multivariate statistical methods: the RV-coefficient. – J. R. Stat. Soc. C 25: 257–265.
Spence, J. R. and Niemelä, J. K. 1994. Sampling carabid assemblages with pitfall traps – the madness and the method. – Can. Entomol. 126: 881–894.
ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. – Ecology 67: 1167–1179.