Short title: Neutral parameter inference

Title: A novel inference of the fundamental biodiversity number for multiple immigration-limited communities

Champak Beeravolu Reddy.[1*], Pierre Couteron [1] and François Munoz [2]

[1] IRD, UMR AMAP, Bd de la Lironde, TA A-51/PS2, 34398 Montpellier Cedex 05, France

[2] Université Montpellier 2, UMR AMAP, Bd de la Lironde, TA A-51/PS2, 34398 Montpellier Cedex 05, France

*__Corresponding author__:

Champak Beeravolu Reddy

UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro),

Campus International de Baillarguet CS 30 016 F-34988, Montferrier-sur-Lez cedex, France

Tel.: +33 4 99 62 33 31; Fax: +33 4 99 62 33 45

Email: champak_br@yahoo.com

Author contributions: C.B.R., P.C. and F.M. designed the study; C.B.R. performed the research; and C.B.R., P.C. and F.M. wrote the manuscript. The authors declare no conflict of interest.

Keywords: parameter estimation, neutral theory, island biogeography, coalescence theory, community ecology, similarity statistics, tropical trees, spatial scale.

Type of paper: Letter

Word count (abstract): 148

Word count (whole manuscript): 6284

Word count (main text): 4501

No. of references: 50

No. of figures: 4

No. of tables: 0

Abstract

Neutral community theory postulates a fundamental quantity, θ, which reflects the species diversity on a regional scale. While the recent genealogical formulation of community dynamics has considerably enhanced quantitative neutral ecology, its inferential aspects have remained computationally prohibitive. Here, we make use of a generalized version of the original two-level hierarchical framework in order to define a novel estimator for θ, which proves to be computationally efficient and robust when tested on a wide range of simulated neutral communities. Estimating θ from field data is also illustrated using two tropical forest datasets consisting of spatially separated permanent field plots. Preliminary results also reveal that our inferred regional diversity parameter based on community dynamics may be linked to widely used ordination techniques in ecology. This paper essentially paves the way for future work dealing with the parameter inference of neutral communities with respect to their spatial scale and structure.

Introduction

In his now well known contribution, Hubbell (2001) introduced a neutral theory of biodiversity, building upon the mainland-island model borrowed from the theory of island biogeography (MacArthur & Wilson 1967). His theory basically describes a model of interacting communities where the slow and regional scale dynamics such as speciation and extinction occur on a metacommunity level and the relatively faster local demographic events such as the birth and death of individuals are the mainstay of the local community dynamics. Hubbell (2001) further imagined equilibrium situations for these two entities which are dictated by the parameter θ (the fundamental biodiversity number) for the metacommunity dynamics and an immigration parameter I (Etienne & Olff 2004) for the local community dynamics. The interaction between these entities is modelled through the arrival of immigrant individuals from the regional pool of species (i.e. the metacommunity) towards a single or multiple local communities. It can be further interpreted as a measure of the isolation that a local community undergoes from the regional species pool due to some form of limited dispersal or more exactly immigration-limitation (Beeravolu *et al.* 2009). Up till now, much of the interest on neutral models in ecology has been either based on this two-level spatially implicit hierarchical framework (Hubbell 2001; Vallade & Houchmandzadeh 2003; Etienne 2005) or on spatially explicit models which do not make any such hierarchical distinctions (Chave & Leigh 2002; Zillio *et al.* 2005; O'Dwyer & Green 2009). In addition to model structure, details of the metacommunity processes such as speciation have been explored (Haegeman & Etienne 2008; Kopp 2010) though a consensus neutral approach is still lacking in community ecology (Gravel *et al.* 2006) and work is still in progress (Haegeman & Loreau 2011).

One explanation for the discord among ecologists, especially on the practical relevance of the two-level spatially implicit neutral model (hereafter denoted 2L-SINM), has

been the issue of neutral parameter inference. Most of the current methods consist of either fitting a species abundance distribution (Purves & Pacala 2005; Dornelas *et al.* 2006; McGill *et al.* 2006) or finding quantitative point estimates of the neutral parameters (Etienne *et al.* 2006; Munoz *et al.* 2007). In the former, the general binning process of the species information into abundance classes, among other concerns (McGill *et al.* 2007), has raised several issues (Gray *et al.* 2006). At the same time the reliability of point estimates also remains doubtful (Leigh 2007), particularly as I and θ are known to be "hyperbolically correlated" when estimating them simultaneously (Beeravolu *et al. Submitted manuscript*; Etienne *et al.* 2006; Munoz *et al.* 2007; Beeravolu *et al.* 2009; Jabot & Chave 2009). Also, a basic contention underlying all these critiques has been the inherent insufficiency of the abundance information to fully elucidate ecological and evolutionary processes that neutral models aim to combine (Harte 2003).

In this paper, we build upon a recently discussed neutral framework which resembles a decoupling of scales (Levin 1992) and further enhances the spatial structure of the 2L-SINM. This is accomplished by introducing an intermediate level of regional process which generalizes the classical 2L-SINM (Hubbell 2001) into a 3L-SINM (Munoz *et al.* 2008) and enables the relaxation of the speciation-drift (or speciation-extinction) equilibrium assumption on the pool of available immigrant individuals (Beeravolu *et al.* 2009). In a previous paper, Munoz *et al.* (2008) introduced the 3L-SINM and used the steady state results to improve on the independent estimation of the neutral immigration parameter I, whereas we use similar approaches to establish an independent inference of θ for the first time.

Methodological background

The hierarchical SINM

Under the original 2L-SINM model, local communities are defined as panmictic patches of species assemblages sufficiently isolated from each other so as not to receive an immigrant individual directly from another local community (Fig. 1). Besides, from a regional perspective, local communities are presented as immigration-limited samples of the same metacommunity and subject to a top-down arrival of immigrants depending upon the immigration probability m (which is the scaled version of I on the unit scale). One possible generalization of the 2L-SINM can then be defined as a regional scale common to all local communities and whose species composition is allowed to differ considerably from that of the metacommunity, *sensu* Hubbell (2001, pg. 122), which is supposed to be panmictic and at speciation-drift equilibrium (Munoz *et al.* 2008). In more practical terms, immigrant species originating from a common source pool may very well correspond to a particular sub-region of the larger hypothetical metacommunity. For instance, let a mountain range be characterized by a coherent biogeographical history and represent a large-scale metacommunity. If a deep gorge runs through this range at some point, for all practical purposes, we can consider a part of the range to be under some sort of isolation from immigrants coming from the rest of the metacommunity, especially for the case of sessile biota such as plants even though their overall floristic composition remains specific to the mountain biotope.

Subsequently, the intermediate pool of species under the 3L-SINM can be modelled as a sample drawn from the metacommunity under the influence of non-neutral processes or simply as an immigration-limited sample (see Jabot *et al.* 2008 for a similar approach). Moreover, as for local communities, this intermediate pool can be described analytically using a genealogical approach (for each individual of a local community) similar to the coalescent theory of population genetics (Etienne & Olff 2004). The composition of this intermediate pool is then defined as the set of ancestor species having immigrated at some point of time into the local communities. This 3L-SINM also “collapses” back to the particular case of a

2L-SINM for the case of an immigration-unlimited random sample of the panmictic metacommunity (see Fig. 1). Besides, this collapse and vice versa forms the backbone of Munoz *et al.*'s (2007, Appendix) coalescence simulation strategy of multiple local community samples which is briefly described later (see "Simulating multiple local community samples").

Conditional similarity under the 3L-SINM

Following Munoz *et al.* (2008), we refer to the similarity of individuals belonging to a same species within a local community sample (F_{intra}) using the Simpson concentration (Simpson 1949). Let S be the total number of species among N local community samples. For a given sample k (varying from 1 to N), this index represents the probability of randomly drawing (without replacement) two conspecific individuals (using Munoz *et al.*'s (2008) exact estimator of sample similarity):

$$F_{intra}(k) = \sum_{s=1}^{S} \frac{n_{sk}}{n_k} \frac{n_{sk} - 1}{n_k - 1} \quad (1)$$

where $n_k = \sum_{s=1}^{S} n_{sk}$ and n_{sk} is the number of individuals of the s^{th} species found in the k^{th} sample. We define the time dependant version, $F_{intra}(k,t)$, of the intra sample similarity (or conspecific at time t) and in keeping with the 3L-SINM framework, the theoretical similarity at t between the intermediate (regional) pool of available immigrating species and a given local community k is $F_{inter}(pool,k,t)$. If each local sample is assumed to reasonably approximate the composition of its respective local community, it is possible to derive, as shown in the following, the time independent expectation $F_{intra}(k)$ (Munoz *et al.* 2008, Appendix C).

Let us assume that, between a time step t and t+1, there occurs a random death in the k^{th} local community, such that two randomly chosen individuals may or may not contain the dead individual with probability $2/n_k$ and $1 - 2/n_k$ respectively. Following a coalescent approach (see Etienne & Olff 2004, eqn 1), for every dead individual, the replacing individual is either the offspring of a local individual or an immigrant individual of a lineage currently not present in the community. In the former, the replacement probability is given by m_k and $1 - m_k$ for the latter, where m_k represents the immigration probability into the k^{th} local community. Consequently, the conspecific probability (hereafter denoted CnP) at t+1 of two randomly chosen individuals from a community which include the dead individual is the sum of three different transition probabilities.

When the replacement is an immigrating replacing individual, the CnP of the chosen couple is $F_{inter}(pool,k,t)$ which is the time conditional version of $F_{inter}(pool,k)$. If the replacement is a local event, it could be the descendant of the other individual, with probability $1/n_k$, in which case the CnP would be 1. If the replacing individual belongs to an offspring of an individual from the rest of the community (with probability $1 - 1/n_k$) the CnP is given by $F_{intra}(k,t)$. We could also consider the possibility that the dying individual produces offspring (a modification of Moran's (1958) model), thus adding to the competition for the vacant spot as suggested by Hubbell (1979; 2001), though this wouldn't affect the final result (Munoz *et al.* 2008). In sum, considering all the transition probabilities defined above, we can write the full CnP at t+1 for any two individuals in a local community as (Munoz *et al.* 2008, Appendix C)

$$F_{intra}(k,t+1) = \left(1-\frac{2}{n_k}\right)F_{intra}(k,t) + \frac{2}{n_k}\left\{m_k F_{inter}(pool,k,t) + (1-m_k)\left[\frac{1}{n_k} + \left(1-\frac{1}{n_k}\right)F_{intra}(k,t)\right]\right\}. \quad (2)$$

Here, m_k can be defined either as $I_k/(I_k + n_k)$ or $I_k/(I_k + n_k - 1)$ which corresponds to the unmodified and modified Moran's models respectively and I_k stands for the number of immigrating individuals into k^{th} local community (Etienne & Olff 2004). At steady state, $F_{intra}(k,t+1) = F_{intra}(k,t) = F_{intra}(k)$, which reduces eqn (2) to

$$F_{intra}(k) = 1 - \frac{I_k}{I_k + 1}\left(1 - F_{inter}(pool, k)\right). \quad (3)$$

Let us assume that the k local community samples are far enough from each other in order to represent distinct communities and denote $F_{inter}(k)$ to be the CnP of an individual from the k^{th} community and an individual from one of the $k - 1$ other communities. Consequently, the coalescence approach dictates that $F_{inter}(k)$ is equal to the CnP of their respective ancestors who are distinct immigrating individuals from the regional pool (i.e. $F_{intra}(pool)$) which can be written down as (Munoz *et al.* 2008, eqn 3):

$$F_{inter}(k) = F_{inter}(pool, k) = F_{intra}(pool). \quad (4)$$

Local similarity under the 2L-SINM

In this section, we develop the other key idea which pertains to the model "collapse" from the 3L-SINM to the 2L-SINM (see Fig. 1.) which entails that the results of the previous section also apply to the particular case of the 2L-SINM. For the sake of completeness, we also detail the analytical relationship linking $F_{intra}(k)$ to the immigration parameter I_k and the biodiversity parameter θ for the 2L-SINM (see also Etienne 2005, eqn 8).

Let $F_n(k)$ be the CnP of drawing a sample of n individuals from the k^{th} local community and j the corresponding number of ancestors in the metacommunity for the same sample. In other words, the quantity j corresponds to the number of those individuals who were the first of every lineage to have immigrated into the community sample, which, for a sample of size n, is $j < n$ original community lineages. Accordingly, we write the n-sample CnP as

$$F_n(k) = \sum_{j=1}^{n} p_{meta}(1/j)\, p_{comm}(j/n) \tag{5}$$

where $p_{comm}(j/n)$ is the probability of drawing a sample of n individuals from the k^{th} local community which are the progeny of exactly j different immigrating individuals and $p_{meta}(1/j)$, the probability that all of these j ancestors belong to the same species.

The 2L-SINM describes a metacommunity at speciation-drift equilibrium whose sample abundance distribution can be described using well known multinomial formulas from the field of population genetics (see Hubbell 2001, 119). Thus a random sample from the metacommunity containing j individuals belonging to σ different species has the probability

$$p_{meta}(\sigma/j) = \frac{\overline{s}(\sigma, j)\theta^{\sigma}}{\theta^{(j)}} \tag{6}$$

where $\overline{s}(\sigma, j)$ is the unsigned Stirling number of the first kind and $\theta^{(j)}$ is the Pochhammer notation for the rising factorial defined as $\theta^{(j)} := \prod_{i=1}^{j}(\theta + i - 1)$ (Ewens 1972; Tavaré & Ewens 1997, eqn 41.5). This can also be extended to an immigration-limited local community (at

immigration-drift equilibrium) where the immigration process replaces the speciation process and

$$p_{comm}(j/n) = \frac{\overline{s}(j,n)(I_k)^j}{(I_k)^{(n)}}. \quad (7)$$

Thus, eqn (5) can be rewritten as

$$\begin{aligned} F_n(k) &= \sum_{j=1}^{n} p_{meta}(1/j)\, p_{comm}(j/n) \\ &= \sum_{j=1}^{n} \frac{\overline{s}(1,j)\theta}{\theta^{(j)}} \frac{\overline{s}(j,n)(I_k)^j}{(I_k)^{(n)}} \end{aligned} \quad (8)$$

which, for the special case of a conspecific sample consisting of two individuals (i.e. $n = 2$) reduces to (Etienne 2005, eqn 8)

$$F_2(k) = F_{intra}(k) = 1 - \frac{\theta}{\theta+1}\frac{I_k}{I_k+1}. \quad (9)$$

Using eqns (3), (4) and (9), the expression for θ simply reduces to

$$F_{inter}(k) = \frac{1}{\theta+1}. \quad (10)$$

Estimator θ based on the conditional inter-sample similarity

Here, we define a statistic (previously used by Munoz *et al.* 2008) which measures the k^{th} sample's similarity with respect to the rest of the samples thereby providing an estimate of the quantity $F_{inter}(k)$. This can be written in terms of sampling without replacement as

$$\hat{F}_{inter}(k) = \sum_{s=1}^{S} \frac{n_{sk}}{n_k} \frac{n_s - n_{sk}}{N_T - n_k} \tag{11}$$

where $N_T = \sum_{k=1}^{N} \sum_{s=1}^{S} n_{sk}$ and $n_s = \sum_{k=1}^{N} n_{sk}$. Using eqn (10), we can now write down our novel estimator for the biodiversity parameter θ as,

$$\hat{\theta} = E_k\left[\theta_k\right] = E_k\left[\frac{1}{\hat{F}_{inter}(k)} - 1\right] \tag{12}$$

where E_k denotes the expectation over the k local community samples.

While eqn (12) can be applied directly to simulated local community samples with a known theoretical θ, it needs to be adapted for a field dataset which may belong to a single or several metacommunities. In the following we attempt to identify a subset of the field dataset which is most likely to correspond to a 2L-SINM framework (i.e. a speciation-drift equilibrium at the metacommunity level). One possible way to go about this task is to measure the spread of the $\hat{\theta}_k$ distribution and attempt to reduce it by using a sequential elimination scheme which opts out field samples one at a time.

Let us assume that our field dataset consists of Y community samples of variable size. Our method consists of calculating the $\hat{\theta}_Y$ value of a given dataset along with a measure of statistical deviation (denoted COD_Y, see below) and repeating the same by randomly pulling

out one sample at a time and computing $\hat{\theta}_{Y-1}$ and COD_{Y-1}. Among the Y values of $\hat{\theta}_{Y-1}$ thus obtained, we eliminate the sample whose absence produced the smallest COD_{Y-1} value and then proceed with the sequential elimination scheme with the remaining Y–1 samples. We define our coefficient of deviation or COD (a robust analogue of the coefficient of variation) as the ratio of the mean absolute deviation (MAD) over the average where

$$COD_k\left[\hat{\theta}_k\right] = \frac{MAD_k\left[\hat{\theta}_k\right]}{E_k\left[\hat{\theta}_k\right]} = \frac{E_k\left[\left|E_k\left[\hat{\theta}_k\right] - \hat{\theta}_k\right|\right]}{E_k\left[\hat{\theta}_k\right]}. \quad (13)$$

The MAD is a well known robust statistic of the sample variation when the population distribution is unknown or for highly skewed curves commonly known as heavy-tailed distributions. For a normal distribution, the standard error can be roughly calculated as 1.253 times its MAD value (MathWorks 2008). This descriptor should not be confused with the closely related median absolute deviation (also known as MAD) which uses a median instead of a mean. The elimination scheme was finally stopped with the appearance of a consistent asymptotic pattern of the sequentially obtained COD values which indicated a stable estimate for a network of samples.

Applications

Simulating multiple local community samples

We simulated steady state local community samples following the modification of the sequential construction scheme of Etienne (2005, Appendix S2). The simulation algorithm basically follows the 3L-SINM structure with m for the intermediate pool set to 1 (refer to Fig. 1.) thereby producing samples strictly congruent with the 2L-SINM. Accordingly, we

create an explicit link between the ancestry information of the community samples and a large predefined metacommunity (see Munoz *et al.* 2007, Appendix).

Every simulation consists of a set of N local community samples, each having a randomly chosen immigration parameter I_k (varying from 3 to 300) and sample size n_k (varying from 200 to 600). Our criterion for the sample size corresponds approximately to the size of a hectare of tropical forest (i.e. n_k = 400 trees above 10 cm of diameter) typically found in field studies involving several permanent sampling plots (Pyke *et al.* 2001; Ramesh *et al.* 2010b). We also varied the number of samples (i.e. N) by simulating 5, 10, 20, 30 and 50 samples. These simulations also need a biodiversity parameter to be defined for which we simulated sets of scenarios where θ = (10, 50, 100, 200, 300). In the following, we shall to a simulated sampling protocol (or SSP hereafter) as a simulation generated using the information provided by the couplet (N, θ) as the other parameters are chosen to vary at random (i.e. I_k and n_k). Thus, we have considered a total of 5 (values for N) x 5 (values for θ) = 25 SSPs, each of which was in turn replicated 200 times whereby we obtained a grand total of 5000 estimates (denoted $\hat{\theta}$) of the theoretical biodiversity parameter θ. We assessed the performance of the estimation of θ by studying the histograms of the Relative Bias (RB) given by $(\hat{\theta}-\theta)/\theta$ and the COD (cf. eqn (13)).

Inferring θ using field data

Apart from simulations, we estimated $\hat{\theta}$ using two tropical forest datasets each consisting of multiple small permanent field plots. Both these datasets consist of the abundance information of individuals above 10 cm of diameter at breast height which have been identified to the species level. The first dataset consists of 50 plots (mainly 1ha) along the Panama Canal Watershed (PCW) area set up by the CTFS (Pyke *et al.* 2001), a subset of which was previously used by Jabot *et al.* (2008) for inferring immigration parameters. This

freely available dataset is part of a larger study (the Marena dataset) and the field plots used in this paper are originally referred to using the following numbering 1-41, C1-C4 and S0-S4 (see Condit *et al.* 2002, Appendix; Chave *et al.* 2004, Appendix B). Our second dataset also consists of 50 field plots (each of 1ha in size) of the wet evergreen forest type from the Western Ghats (WG) region of South India (extracted from Ramesh *et al.* 2010a). This dataset has been discussed previously by Munoz *et al.* (2007, Appendix) for the purpose of inferring neutral parameters.

Results

We studied the RB (relative bias) and the COD (coefficient of deviation, see eqn (13)) histograms for the various SSPs (simulated sampling protocols). The 5000 SSP estimates of $\hat{\theta}$ which make up the histograms were obtained in the matter of a few minutes using the MATLAB® software (MathWorks 2008). In both the five and fifty SSPs (Figs. 2 and 3 respectively), the best fit normal curve clearly emphasized the increasing symmetry of the distribution of the RB of $\hat{\theta}$ with increasing theoretical θ and suggested that the estimator is unbiased. In general, the RB distribution of $\hat{\theta}$ showed a tendency to be skewed for low values of θ while it became more symmetric and always remained centred around zero as θ increased (see also Figure S1 in the Supporting Information). At the same time, the COD distribution was skewed for a low number of samples (e.g. N = 5, 10) and a high theoretical θ and vice versa for N > 10, while the COD skewness varied little (for low θ) compared to the RB skewness (Figure S1). However, note that the COD histograms (Figs. 2 and 3) rarely exceed a maximum value of 0.2, which was used as a benchmark when estimating θ on field data for large N.

We also applied our estimator to the two tropical forest datasets presented above (Fig. 4). When using all the 50 field samples of the PCW data, we obtained a high COD_{50} value (≈

0.8) in comparison to the COD histogram for the N = 50 strictly neutral SSPs (Fig. 3). In contrast, the COD_{50} for the WG data (≈ 0.2) was well within the range (not shown). Subsequently, we applied the sequential elimination scheme on both these datasets (cf. previous section) in order to identify the network of plots composing the ideal (i.e. panmictic) 2L-SINM metacommunity. The estimation of the neutral biodiversity parameter for the WG dataset proved to be comparatively stable while its respective COD values fluctuated slightly (between 0.1 – 0.2) below the maximum value observed for strictly neutral simulations. Furthermore, our estimation of $\hat{\theta}$ for the WG data was well bounded by the values 62.33 and 50.99 which are those found by Munoz *et al.* (2007) for the very same plots. For the PCW dataset, $\hat{\theta}$ estimates reached a qualitatively stable estimate which corresponded to a COD < 0.2. This was obtained after having sequentially eliminated 15 plots (COD < 0.2), though we continued to eliminate samples in order to check for its stability (Fig. 4). Besides, a closer look at the remaining PCW plots revealed that the first eight eliminated plots were part of the Outer PCW region (numbered 31-39, see Pyke *et al.* 2001, Fig. 1).

Discussion

In this paper, we have basically introduced a new estimator for multiple field samples of the neutral biodiversity parameter θ, first formulated by Hubbell. Subsequently, this estimator has been tested on wide-ranging simulations of multiple neutral local community samples at migration-drift equilibrium. A general conclusion from our simulation study is that the relative bias of $\hat{\theta}$ seems to be relatively well distributed around zero with a progressive tightening of the same as θ increased. This property is highly desired given that currently available estimators of θ for a single local community sample present an increasing bias with increasing θ (Munoz *et al.* 2007). As for the estimation variance, measured using the COD values, we note that the number of community samples used (i.e. N) is an important factor as

a fewer number (e.g. 5 and 10) leads to an increase in the spread of the COD distribution (Fig. 2). Our results can also be seen as a significant improvement in contrast to likelihood approaches (Etienne 2009b) which become computationally intractable for more than five samples of small size like the ones used in this paper (Beeravolu *et al. Submitted manuscript*).

However, the technique presented here needs to be further compared to existing methods (Munoz *et al.* 2007; Etienne 2009a) that estimate θ from a single panmictic sample and also make use of the analytical expectations for the Ewens Multivariate Distribution (Ewens 1972) of population genetics literature. In particular, by randomly sampling an individual from several spatially separated samples these authors (Munoz *et al.* 2007; Etienne 2009a) constitute a metacommunity sample which is repeated a number of times in order to hone their estimates of $\hat{\theta}$. But this approach can be unreliable when the number of local community samples is small as it would provide a (relatively small) metacommunity sample with insufficient information for inferential purposes. Though, an added advantage of using the Ewens estimation for θ is the availability of an analytical expression for the induced bias which decreases with increasing sample size and increases with an increase in θ (Donnelly & Tavaré 1995, 414; Tavaré & Ewens 1997, 236).

Moreover, as our estimates using the WG data seem to be relatively robust to the sequential elimination scheme and coherent with Munoz *et al.*'s (2007) estimates, the 50-sample evergreen forest dataset from the WG seems to strongly corroborate a 2L-SINM at this particular sampling scale. At the same time, Ramesh *et al.* (2010b) have found that some environmental variables had a strong predictive power on the plots' floristic composition. This implies that while the data seem to agree with the neutral model, violations from neutral assumptions might not hinder a sound estimation of $\hat{\theta}$ as a phenomenological descriptor of the overall diversity of the region. Instead, estimating $\hat{\theta}$ may be a good basis to compare the overall diversity between biogeographic regions as Fisher's α is known to be asymptotically

identical to θ (Hubbell 2001, 165). This approach could also be extended to other forest types and biogeographic regions in order to identify the extents of the different metacommunities as discussed above and measure their relative diversity.

To pursue this idea further, we can use the distribution of the $\hat{F}_{inter}(k)$ values in order to identify the local communities whose top-down linkage to a common metacommunity may be unlikely. Conversely, it is also a simple approach which delineates a subset of samples that look "floristically homogeneous" with respect to the 2L-SINM. Such a group of field plots whose taxonomic composition complies with the 2L-SINM of a metacommunity could be used to spatially delimit or map forest types over a regional sampling design. While the elimination technique presented in this paper can be seen as a simple top-down clustering scheme, it produces results which bear a close resemblance to well known bottom-up ordination schemes such as hierarchical agglomerative clustering. For example, Chust *et al.* (2006) used spatially explicit information in relation to a field plot's ecology such as its elevation, remotely sensed data and the geographical distance to predict the forest types for almost all of the PCW data used in this paper. They perform a hierarchical agglomerative clustering of the sample abundances (or occurrences) using a proportional-link linkage algorithm and further extrapolate each cluster's spatially explicit characteristics using a multiple regression model thereby mapping a forest type. The most "distant" clusters found in their study (Chust *et al.* 2006, Appendix 2) correspond perfectly to some of the first few plots eliminated using our sequential elimination technique. Although, note that Chust *et al.* (2006, Figs. 1 and 4) exclude some of the plots used in this paper (numbered 38 and 39) and instead use other plots (numbered P1, P2, G1, G2 and SH) which were absent from the present (see Chave *et al.* 2004, Appendix B for a corrected account of the PCW field plot numbering).

For the case of a single large local community sample, Jabot & Chave (2009), echoing Harte (2003), contend that species abundances contain a limited amount of useful information

and supplement it by the use of species phylogenetic information in order to resolve inconsistencies in the estimation of neutral parameters. Our results set in a more complex context raise the question whether species phylogenies are truly needed when multiple spatially separated field samples are available. Nevertheless, the $F_{inter}(k)$ statistic (eqn (11)) presented here can easily be developed into an efficient Bayesian framework (*sensu* Jabot & Chave 2009), which remains a very powerful method, for injecting additional information such as species phylogenies or even the demographic history of the communities (Beaumont & Rannala 2004). Moreover, recent developments in the field of theoretical population genetics make use of a comparable inter-sample similarity metric (Gaggiotti & Foll 2010) under the island model of Wright (1931), which is a classic model of population subdivision. Interestingly, note that the island model comes conceptually close to the mainland island model of MacArthur & Wilson (1967) for the case of an infinite number of islands, in which case it is known as the continent-island model (Wilkinson-Herbots 1998, 574) or an island-mainland metapopulation (Rannala & Hartigan 1995) although there are subtle differences to be taken into account (in terms of demographic assumptions).

Finally a major weakness of almost all neutral approaches is that it is an equilibrium theory which nevertheless has greatly facilitated its mathematical development. Though, truly dynamic neutral modes are desperately lacking in community ecology (but see Leigh *et al.* 1993; Gilbert *et al.* 2006) and some initial steps have been taken in this direction (Vanpeteghem & Haegeman 2010), much needs to be done before we are able to infer the parameters of a dynamic model from field data. However, the main improvement presented in our paper is a simple and computationally efficient approach for estimating the biodiversity parameter θ (in the case of a multiple sample 2L-SINM framework) which is in many ways complementary to the estimation of the multiple sample I_k parameter (Munoz *et al.* 2008).

Acknowledgements

We thank Raphaël Pélissier, Olivier Hardy and Tim Keitt for helpful suggestions on a previous draft. C.B.R. is extremely grateful for the unwavering support and funding from Aurélie Loiseau towards the end of his Ph.D.

References

Beaumont M.A. & Rannala B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5, 251-261.

Beeravolu C.R., Couteron P., Pélissier R. & Munoz F. (2009). Studying ecological communities from a neutral standpoint: A review of models' structure and parameter estimation. *Ecol. Model.*, 220, 2603-2610.

Beeravolu C.R., Munoz F. & Couteron P. On the simultaneous maximum likelihood inference of neutral parameters from several local community samples. *Submitted manuscript*.

Chave J., Condit R., Aguilar S., Hernandez A., Lao S. & Perez R. (2004). Error propagation and scaling for tropical forest biomass estimates. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359, 409.

Chave J. & Leigh E.G. (2002). A spatially explicit neutral model of beta-diversity in tropical forests. *Theor. Popul. Biol.*, 62, 153-168.

Chust G., Chave J., Condit R., Aguilar S., Lao S. & Perez R. (2006). Determinants and spatial modeling of tree beta-diversity in a tropical forest landscape in Panama. *Journal of Vegetation Science*, 17, 83-92.

1 Condit R., Pitman N., Leigh E.G., Chave J., Terborgh J., Foster R.B., Nunez P., Aguilar S.,
2 Valencia R., Villa G., Muller-Landau H.C., Losos E. & Hubbell S.P. (2002). Beta-
3 diversity in tropical forest trees. *Science*, 295, 666-669.

4 Donnelly P. & Tavaré S. (1995). Coalescents and genealogical structure under neutrality.
5 *Annu. Rev. Genet.*, 29, 401-421.

6 Dornelas M., Connolly S.R. & Hughes T.P. (2006). Coral reef diversity refutes the neutral
7 theory of biodiversity. *Nature*, 440, 80-82.

8 Etienne R.S. (2005). A new sampling formula for neutral biodiversity. *Ecol. Lett.*, 8, 253-260.

9 Etienne R.S. (2009a). Improved estimation of neutral model parameters for multiple samples
10 with different degrees of dispersal limitation. *Ecology*, 90, 847-852.

11 Etienne R.S. (2009b). Maximum likelihood estimation of neutral model parameters for
12 multiple samples with different degrees of dispersal limitation. *J. Theor. Biol.*, 257,
13 510-514.

14 Etienne R.S., Latimer A.M., Silander J.A. & Cowling R.M. (2006). Comment on "Neutral
15 ecological theory reveals isolation and rapid speciation in a biodiversity hot spot".
16 *Science*, 311.

17 Etienne R.S. & Olff H. (2004). A novel genealogical approach to neutral biodiversity theory.
18 *Ecol. Lett.*, 7, 170-175.

19 Ewens W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3,
20 87-112.

1 Gaggiotti O.E. & Foll M. (2010). Quantifying population structure using the F-model.
2 *Molecular Ecology Resources*, 10, 821-830.

3 Gilbert B., Laurance W.F., Leigh E.G. & Nascimento H.E.M. (2006). Can neutral theory
4 predict the responses of amazonian tree communities to forest fragmentation? *Am.*
5 *Nat.*, 168, 304-317.

6 Gravel D., Canham C.D., Beaudet M. & Messier C. (2006). Reconciling niche and neutrality:
7 the continuum hypothesis. *Ecol. Lett.*, 9, 399-409.

8 Gray J.S., Bjorgesaeter A. & Ugland K.I. (2006). On plotting species abundance distributions.
9 *J. Anim. Ecol.*, 75, 752-756.

10 Haegeman B. & Etienne R.S. (2008). Relaxing the zero-sum assumption in neutral
11 biodiversity theory. *J. Theor. Biol.*, 252, 288-294.

12 Haegeman B. & Loreau M. (2011). A mathematical synthesis of niche and neutral theories in
13 community ecology. *J. Theor. Biol.*, 269, 150-165.

14 Harte J. (2003). Tail of death and resurrection. *Nature*, 424, 1006-1007.

15 Hubbell S.P. (1979). Tree Dispersion, Abundance, and Diversity in a Tropical Dry Forest.
16 *Science*, 203, 1299-1309.

17 Hubbell S.P. (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton
18 University Press, Princeton, N.J.

19 Jabot F. & Chave J. (2009). Inferring the parameters of the neutral theory of biodiversity
20 using phylogenetic information and implications for tropical forests. *Ecol. Lett.*, 12,
21 239-248.

Jabot F., Etienne R.S. & Chave J. (2008). Reconciling neutral community models and environmental filtering: theory and an empirical test. *Oikos*, 117, 1308-1320.

Kopp M. (2010). Speciation and the neutral theory of biodiversity. *Bioessays*, 32, 564-570.

Leigh E.G. (2007). Neutral theory: a historical perspective. *J. Evol. Biol.*, 20, 2075-2091.

Leigh E.G., Wright S.J., Herre E.A. & Putz F.E. (1993). The decline of tree diversity on newly isolated tropical islands: a test of a null hypothesis and some implications. *Evol. Ecol.*, 7, 76-102.

Levin S.A. (1992). The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology*, 73, 1943-1967.

MacArthur R.H. & Wilson E.O. (1967). *The theory of island biogeography*. Princeton University Press, Princeton, N.J.

MathWorks (2008). Matlab R2008a, version 7.6.0.324. In. The MathWorks, Inc. Natick, MA.

McGill B.J., Etienne R.S., Gray J.S., Alonso D., Anderson M.J., Benecha H.K., Dornelas M., Enquist B.J., Green J.L., He F.L., Hurlbert A.H., Magurran A.E., Marquet P.A., Maurer B.A., Ostling A., Soykan C.U., Ugland K.I. & White E.P. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.*, 10, 995-1015.

McGill B.J., Maurer B.A. & Weiser M.D. (2006). Empirical evaluation of neutral theory. *Ecology*, 87, 1411-1423.

Moran P.A.P. (1958). Random processes in genetics. *Proc. Camb. Phil. Soc.*, 54, 60-71.

Munoz F., Couteron P. & Ramesh B.R. (2008). Beta Diversity in Spatially Implicit Neutral Models: A New Way to Assess Species Migration. *Am. Nat.*, 172, 116-127.

Munoz F., Couteron P., Ramesh B.R. & Etienne R.S. (2007). Estimating parameters of neutral communities: From one single large to several small samples. *Ecology*, 88, 2482-2488.

O'Dwyer J.P. & Green J.L. (2009). Field theory for biogeography: a spatially explicit model for predicting patterns of biodiversity. *Ecol. Lett.*, 13, 87-95.

Purves D.W. & Pacala S.W. (2005). Ecological drift in niche structured communities: neutral pattern does not imply neutral process. In: *Biotic Interactions in the Tropics* (eds. Burslem D, Pinard M & Hartley S). Cambridge University Press Cambridge, pp. 107–138.

Pyke C.R., Condit R., Aguilar S. & Lao S. (2001). Floristic composition across a climatic gradient in a neotropical lowland forest. *Journal of Vegetation Science*, 12, 553-566.

Ramesh B.R., Swaminath M.H., Patil S.V., Pélissier R., Venugopal P.D., Aravajy S., Elouard C. & Ramalingam S. (2010a). Forest stand structure and composition in 96 sites along environmental gradients in the central Western Ghats of India (Ecological Archives E091-216). *Ecology*, 91, 3118-3118.

Ramesh B.R., Venugopal P.D., Pélissier R., Patil S.V., Swaminath M.H. & Couteron P. (2010b). Mesoscale Patterns in the Floristic Composition of Forests in the Central Western Ghats of Karnataka, India. *Biotropica*, 42, 435-443.

Rannala B. & Hartigan J.A. (1995). Identity by Descent in Island-mainland Populations. *Genetics*, 139, 429-437.

1 Simpson E.H. (1949). Measurement of Diversity. *Nature*, 163, 688-688.

2 Tavaré S. & Ewens W.J. (1997). Multivariate Ewens distribution. In: *Discrete Multivariate*
3 *Distributions* (eds. Johnson NL, Kotz S & Balakrishnan N). Wiley New York, pp.
4 232-246.

5 Vallade M. & Houchmandzadeh B. (2003). Analytical solution of a neutral model of
6 biodiversity. *Phys. Rev. E*, 68.

7 Vanpeteghem D. & Haegeman B. (2010). An analytical approach to spatio-temporal
8 dynamics of neutral community models. *J. Math. Biol.*, 1-35.

9 Wilkinson-Herbots H.M. (1998). Genealogy and subpopulation differentiation under various
10 models of population structure. *J. Math. Biol.*, 37, 535-585.

11 Wright S. (1931). Evolution in Mendelian Populations. *Genetics*, 16, 97-159.

12 Zillio T., Volkov I., Banavar J.R., Hubbell S.P. & Maritan A. (2005). Spatial scaling in model
13 plant communities. *Physical review letters*, 95, 98101-98101.

14

15

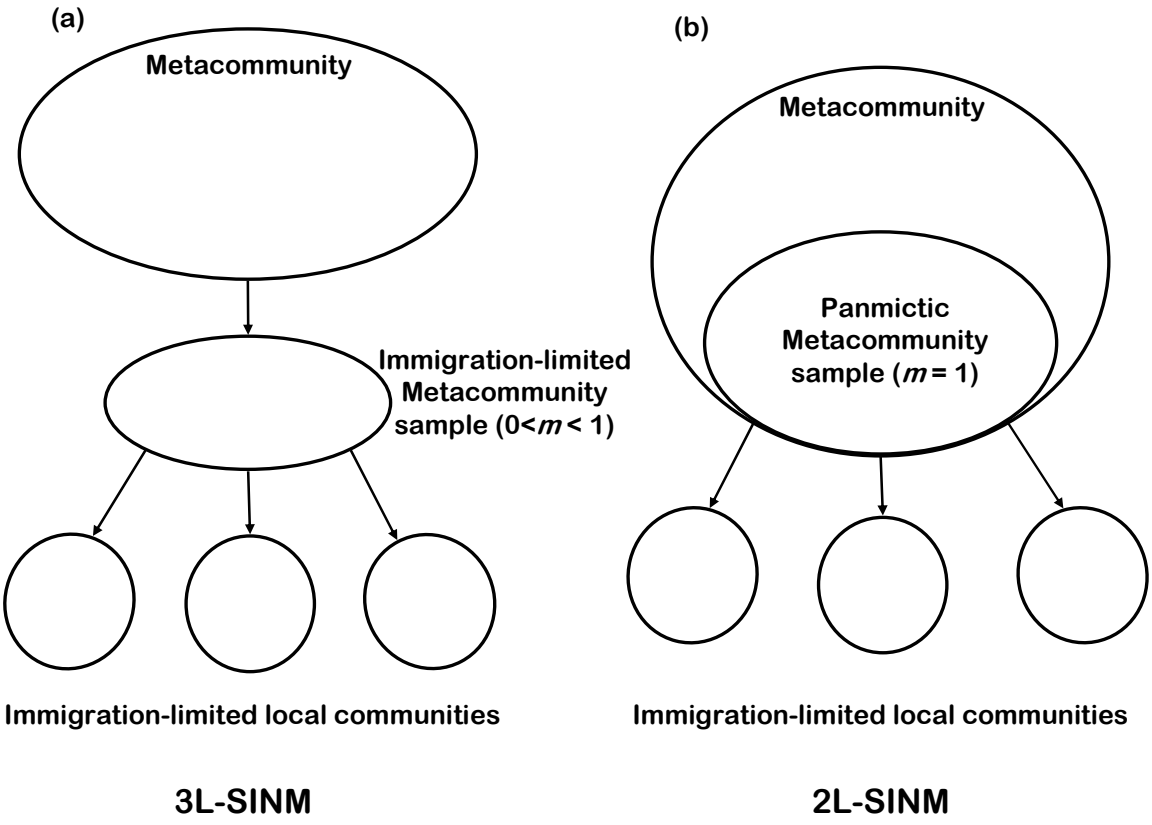
16

1 **Figures**

2

3 Figure 1

4



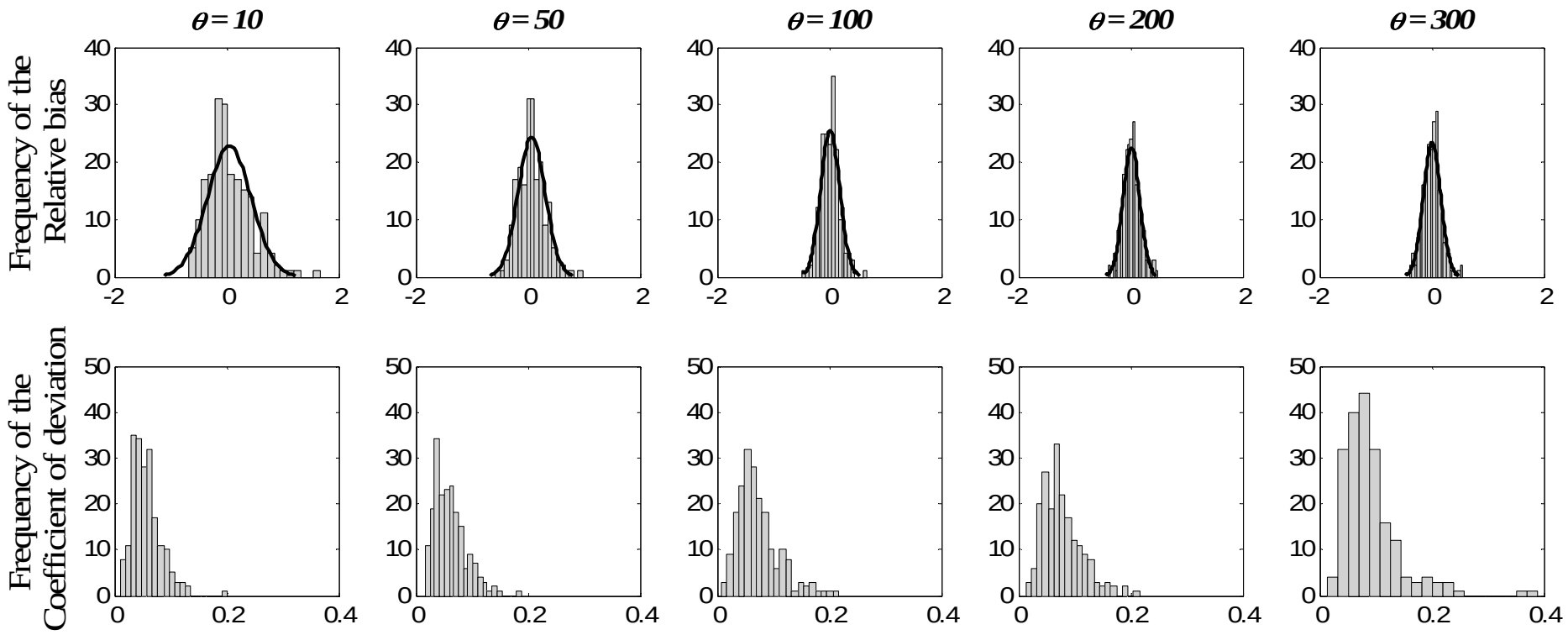
5

6

7

1 Figure 2

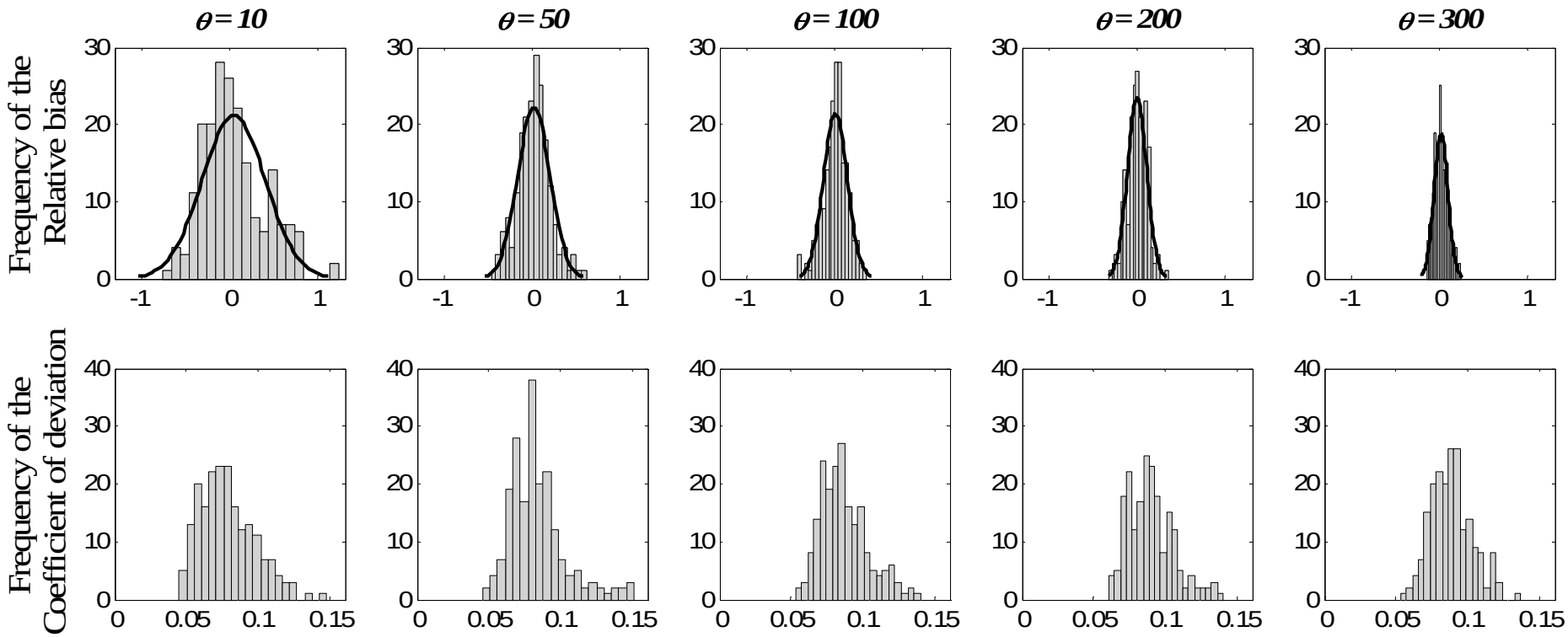
2



3

1 Figure 3

2

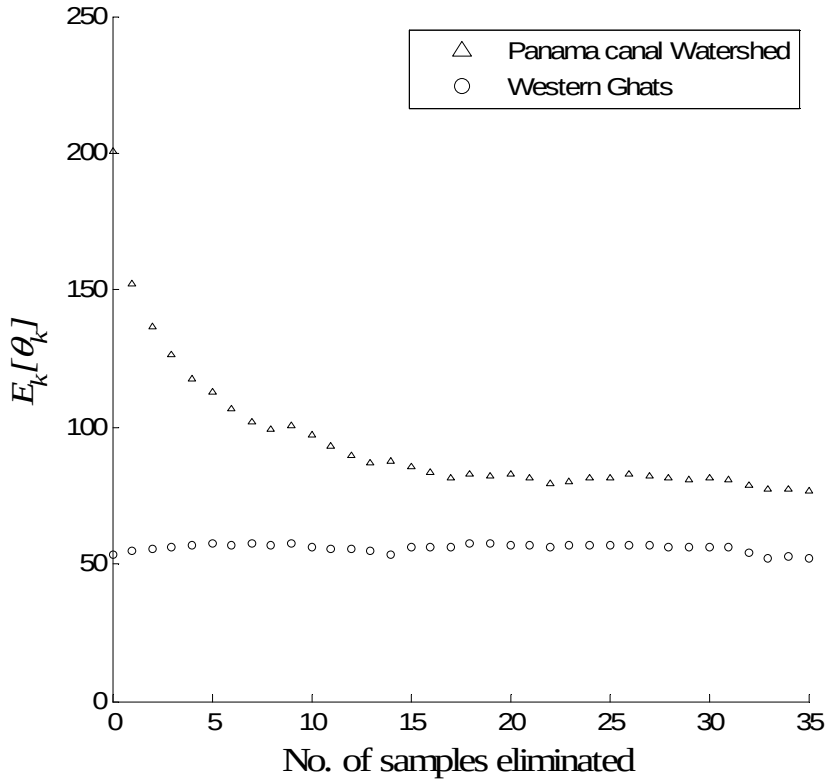


3

4

1 Figure 4

2



3

4

Figure legends

Figure 1 The hierarchical coupling of scale oriented processes in spatially implicit neutral models (SINM). Processes such as speciation and extinction of species exist in the metacommunity and the relatively faster demographic processes take place at the local community level. For Hubbell's (2001) two level model (2L-SINM) an ideal panmictic metacommunity sample can be considered to be the source pool of multiple local communities. Without loss of generality, we can assume that the pooled species composition of the dispersal-limited local community samples is itself a immigration-limited metacommunity sample (the 3L-SINM) corresponding to some sub-region of the same which gives rise to a three level hierarchical model (Beeravolu *et al.* 2009). The value of the parameter m then acts as the degree of isolation from the metacommunity.

Figure 2 The distribution of the relative bias and the coefficient of deviation values of $\hat{\theta}$ (the estimated biodiversity parameter) for a sampling protocol containing 5 simulated plots each and for five different theoretical values of θ found in the literature. The solid black line is the best fit normal curve and emphasizes the increasing symmetry of the distribution of the relative bias with increasing θ.

Figure 3 The distribution of the relative bias and the coefficient of deviation values of $\hat{\theta}$ (the estimated biodiversity parameter) for a sampling protocol containing 50 simulated plots each and for five different theoretical values of θ found in the literature. The solid black line is the best fit normal curve and emphasizes the increasing symmetry of the distribution of the relative bias with increasing θ.

Figure 4 Estimating $\hat{\theta}$ from field data using eqn (13) and following a sequential elimination scheme (see main text) in order to identify the network of plots composing the ideal 2L-SINM metacommunity. The sample elimination criterion is determined by the maximum decrease in the coefficient of deviation due to the absence of a particular plot. Here we sequentially eliminated 35 samples which appeared sufficient to highlight the asymptotic pattern which indicated the stable $\hat{\theta}$ estimate for a network of plots.