Bennett and Gilbert, Beta Diversity

Appendix 2 - Supporting Information

Pairwise Sørensen distance versus Classical β:

Pairwise Sørensen distance ($D_{Sørensen}$) is defined by the number of species jointly present in two samples (J) relative to the total number of species (S) in each sample (Legendre & Legendre, 2012): $D_{Sørensen} = 1 - 2J/(S_1+S_2)$. As shown by Tuomisto (2010), classical β measures can be related to $D_{Sørensen}$; when there are two, equally-weighted samples, $D_{Sørensen} = \frac{\gamma_2 - \bar{\alpha}_2}{\bar{\alpha}_2}$, where γ_2 is the total number of species in the two samples and $\bar{\alpha}_2$ is the mean number of species per sample. Importantly, this metric is related to classical β partitioning for any two sites (β_2) and the scaling exponent (z) of the Arrhenius SAR equation, when area is scaled relative to the size of sample plots so that a single plot is considered an area of one:

$$\beta_2 = \frac{D_{Sørensen}}{1+D_{Sørensen}} \qquad (S1),$$

$$z = Log_2(1 + D_{Sørensen}) \qquad (S2)$$

In Equation S2 above, we see the link between Sørensen distance and the scaling exponent of a species-area relationship, both of which are used in ecological studies to measure species turnover (Anderson et al., 2011). However, we see that when regional diversity is well-modeled by the Arrhenius equation, Sørensen distance allows us to determine the scale-dependence of classic β (Eqn. S2). Indeed, using the same scaling with the Arrhenius equation as was used above, we find that classical β partitioning for a given area is:

$$\beta_{Area} = 1 - Area^{-Log_2(1+\bar{D}_{Sørensen})} \qquad (S3)$$

where $\bar{D}_{Sørensen}$ is the mean Sørensen distance among sample pairs. When the exact shape of the SAR is unknown, Eqns. S2 and S3 provide a qualitative prediction for the scaling of classical β with area, but can only be considered perfectly accurate when two plots are considered (Eqn. S1).

Pairwise Jaccard distance versus classical β

The following equations outline the relationship of pairwise Jaccard distance to β partitioning for any two sites, (β_2) and the scaling exponent (z) of the Arrhenius equation ($S=cA^z$) when area is scaled relative to the size of sample plots.

Jaccard distance for two sites is defined as (Legendre and Legendre, 2012):

$$D_{Jaccard} = 1 - \frac{shared}{\gamma_2} = \frac{2(\gamma_2 - \bar{\alpha}_2)}{\gamma_2} \qquad (S4)$$

We can organize eqn. S4 to define classical β for two sites:

$$\beta_2 = \frac{\gamma_2 - \bar{\alpha}_2}{\gamma_2} = \frac{D_{Jaccard}}{2} \qquad\qquad (S5)$$

Similarly, by solving the Arrhenius equation for two sites, we see that:

$$\gamma_2 = \bar{\alpha}2^z, \; D_{Jaccard} = 2 - \frac{2\bar{\alpha}}{\bar{\alpha}2^z} \qquad\qquad (S6)$$

And,

$$z = log_2\left(\frac{2}{2 - D_{Jaccard}}\right) \qquad\qquad (S7)$$

A similar approach was used to solve for equations S2 and S3 above.

Mathematical predictions for the scale-dependence of multisite distance: the Diserud-Odegaard method

Although a mathematical expectation cannot be derived for Basselga's multisite distance, we can use the complement of the Diserud-Odegaard (2007) similarity metric to examine its scale (in)dependence. As noted by Chao et al. (2012), this metric is standardized by the number of sites sampled. The distance measure, which we term D_{DO}, is (Harrison et al. 1992):

$$D_{DO} = \frac{\gamma/\bar{\alpha} - 1}{N - 1} \qquad\qquad (S8)$$

Using the same approach as above, we set the area equal to the number of plots sampled (N) and assume that γ is well-modeled by the Arrhenius equation. With these assumptions,

$$D_{DO} = \frac{N^z - 1}{N - 1} \qquad\qquad (S9)$$

As N gets large, D_{DO} is approximately $N^{z-1} - 1$. In other words, the non-linear scaling of γ that is common to species-area relationships creates a scale-dependence in D_{DO} that is not corrected by the denominator. We see this scale-dependence in tests of this method (Fig. S15).

Generating a scale-dependent measure of differences in classical β

A scale-dependent measure of differences in classical β should be possible for comparing regions with well-characterized species-area curves. For example, species-area relationships of two regions are well modeled by power functions, the area dependence of region 1 will be:

$$\beta_1 \cong 1 - \frac{\alpha_1}{\alpha_1 A^{z_1}} \qquad\qquad \text{(S10)}$$

Where area is scaled to the size of sample plots so that α_1 is the mean species richness in a single plot and z_1 is the scaling exponent. A scale-dependent measure of classical β could then be defined as:

$$\beta_1 / \beta_2 = \frac{A^{z_1+z_2} - A^{z_2}}{A^{z_1+z_2} - A^{z_1}} \qquad\qquad \text{(S11)}$$

Alternately, for well sampled regions, this scale dependency can be calculated from the data. For example, using mean β of each region for each sample effort, we generate Fig. S19.

Datasets and subsampling procedures

The meadow ecosystem dataset of $605 \times 1\text{m}^2$ plots was collected in 2007 and 2008 from meadow patches in the Garry Oak Ecosystem of southern British Columbia and northern Washington State (see Bennett et al., 2012 for details). To test the effect of area sampled (and γ) on β estimates, the meadow ecosystem dataset was subsampled in increments of 10 plots, from 5 to 605 plots.

The abandoned field dataset of $110 \times 1\text{m}^2$ plots was collected in 2012 at the Koffler Scientific Reserve in southern Ontario, Canada. To test the effect of area sampled (and γ) on β estimates, the abandoned field data were subsampled at increments of 5 plots, ranging from 5 to 110 plots.

The forest dataset was collected in 2002 from Mount St. Hilaire, near Montreal, Canada, and consisted of $85 \times 50 \text{ m}^2$ plots (see Gilbert and Lechowicz, 2005 for details). This dataset was subsampled in increments of 5 plots from 5 to 85 plots.

The diatom dataset of 492 lakes was combined from several datasets of surficial sediments from lakes in eastern North America (see Bennett et al., 2010 for details), and was subsampled at increments of 10 lakes, from 5 to 485 lakes.

The meadow ecosystem and forest plant datasets were initially measured using percent cover estimates and were converted to densities (cf. Gilbert & Lechowicz, 2004), as recommended for randomization of individuals (Crist et al., 2003).

Note that the Chao distance is the Jaccard equivalent described in Chao et al. (2005).

Description of simulated communities

To determine whether the patterns we observed in SD of mean α_{null} in our four datasets were inherent in subsampling from typical species-abundance distributions, we constructed simulated communities as per Kraft et al. (2011), testing for variation in SD α_{null} across a range of sample sizes and species pools (γ_{true} values). We used a lognormal species abundance distribution, with γ_{true} values ranging from 10 to 400, a plot density of 35 individuals per plot (the median of values used by Kraft et al., 2011), and plot numbers of 10 to 100. For each combination of plot number and γ_{true}, we created 100 simulated communities. We then calculated β_{dev} and its components, using 1000 randomizations as per Kraft et al. (2011).

Supplementary Figures
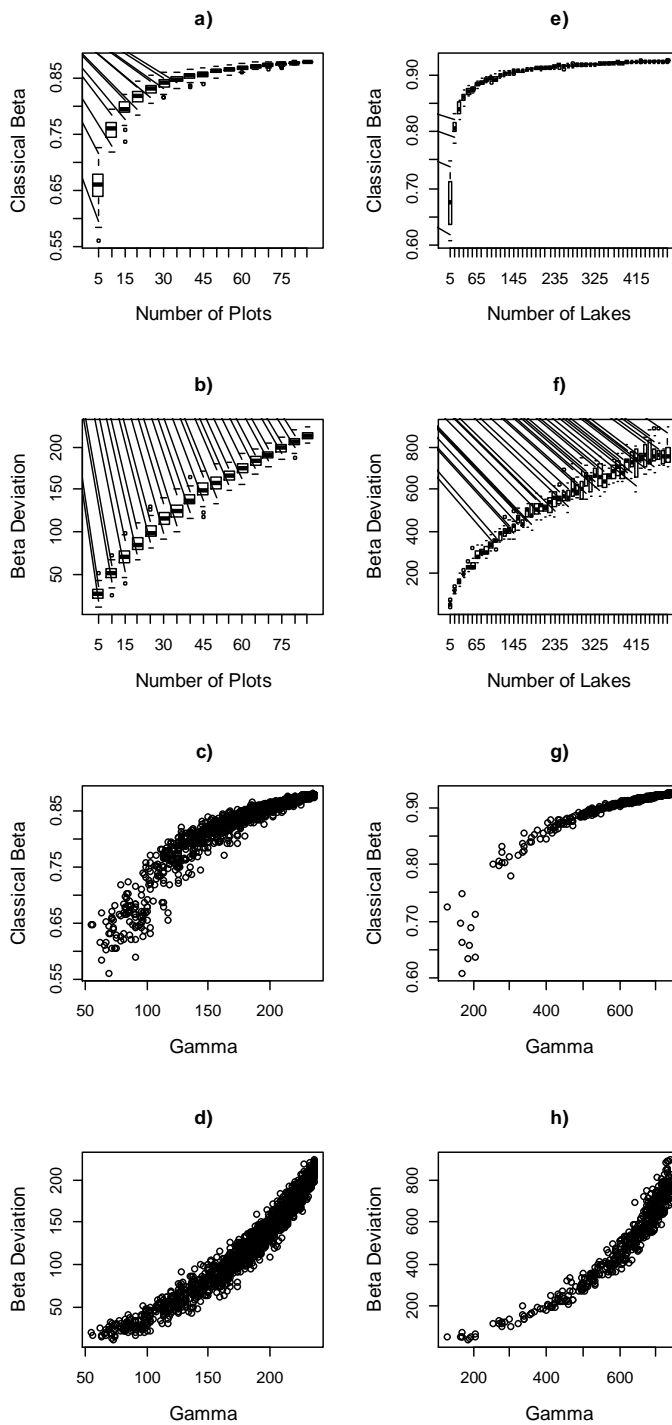
Fig. S1. Classical β versus number of plots (a), β deviation versus number of plots (b), observed β versus γ (c), β deviation versus γ (d) for forest plots; observed β versus number of lakes (e), $β_{dev}$ versus number of lakes (f), observed β versus γ (g), and β deviation versus γ (h) for diatom surveys.
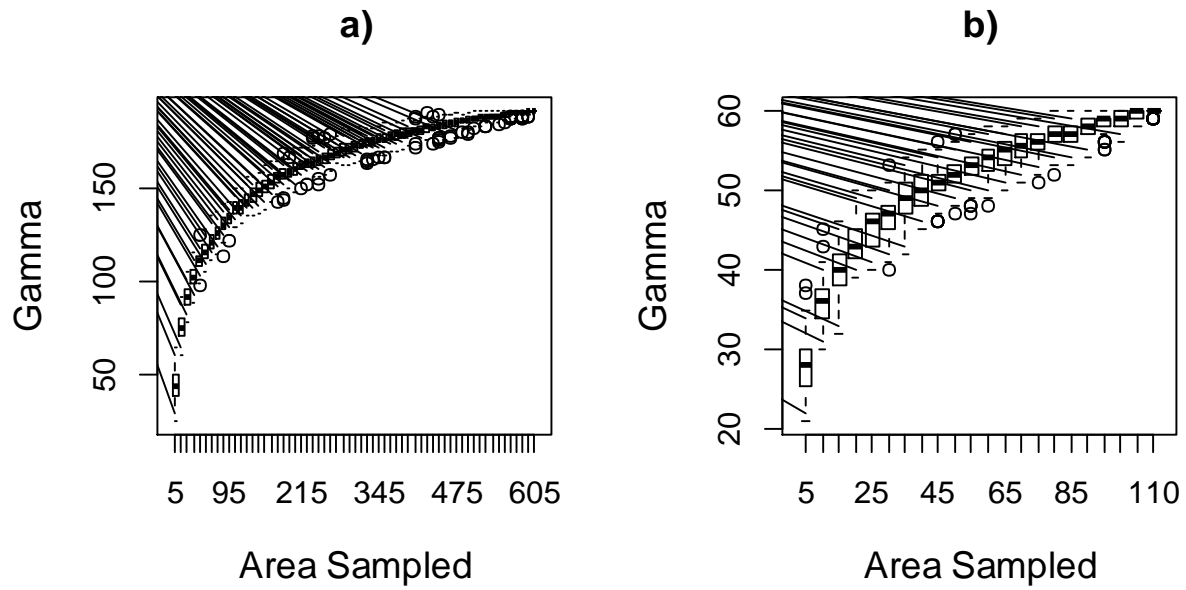
Fig. S2. γ versus area sampled (m^2) for meadow ecosystem plots (a) and early-successional ecosystem plots (b).
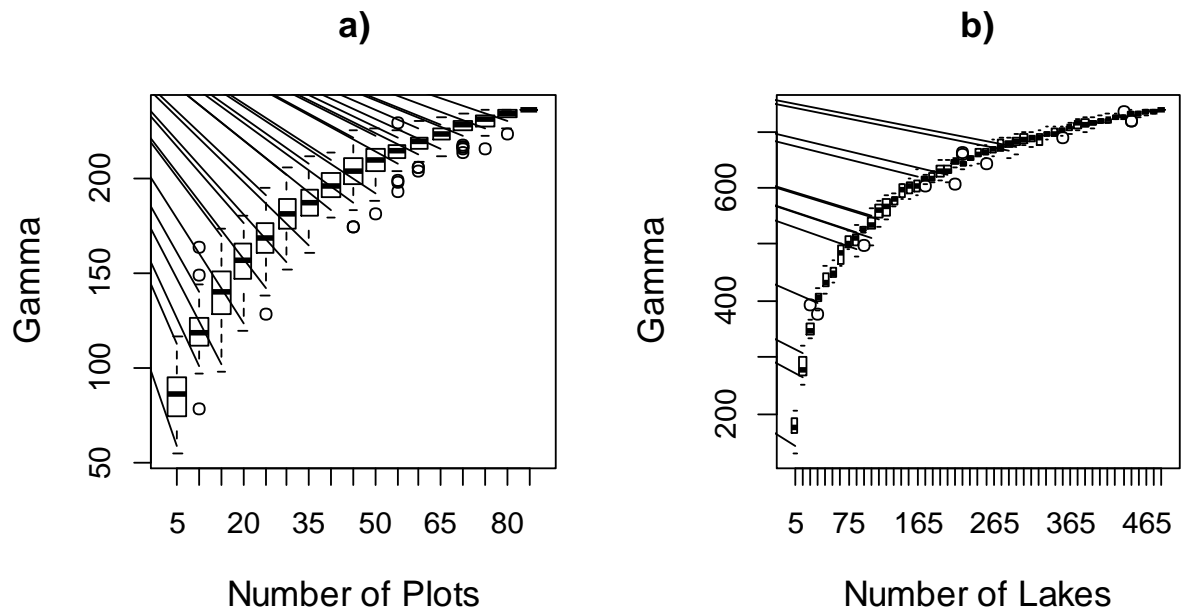
Fig. S3. γ versus number of 50 m² plots for forest plots (a), and γ versus number of lakes for diatom surveys (b).

a)

c)

b)

d)
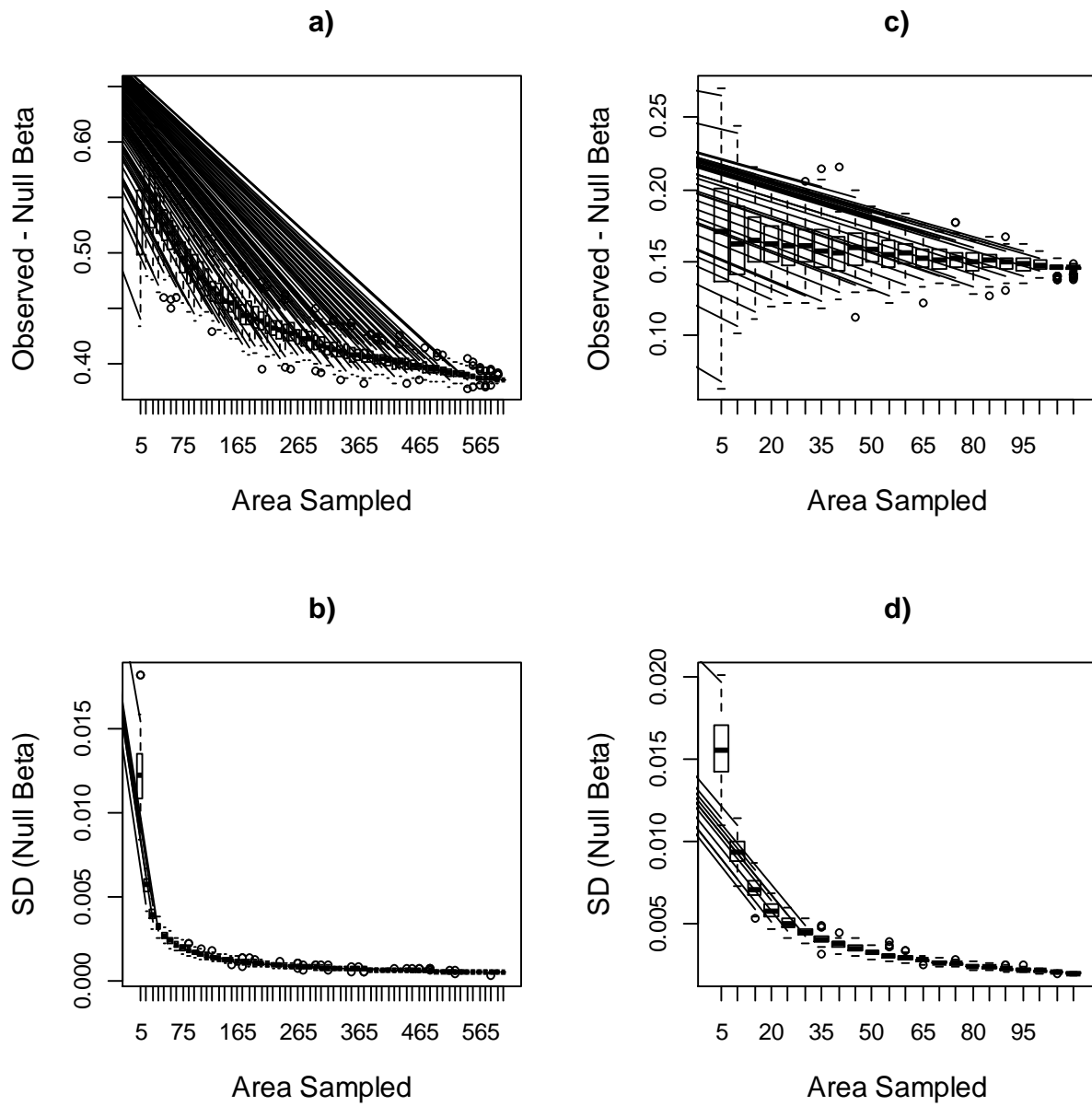
Fig. S4. Observed β minus mean β_{null} (a), and standard deviation of mean β_{null} (b) versus area sampled for meadow ecosystem; observed β minus mean β_{null} (c), and standard deviation of mean β_{null} (d) versus area sampled for early successional ecosystem.

a)

c)

b)

d)
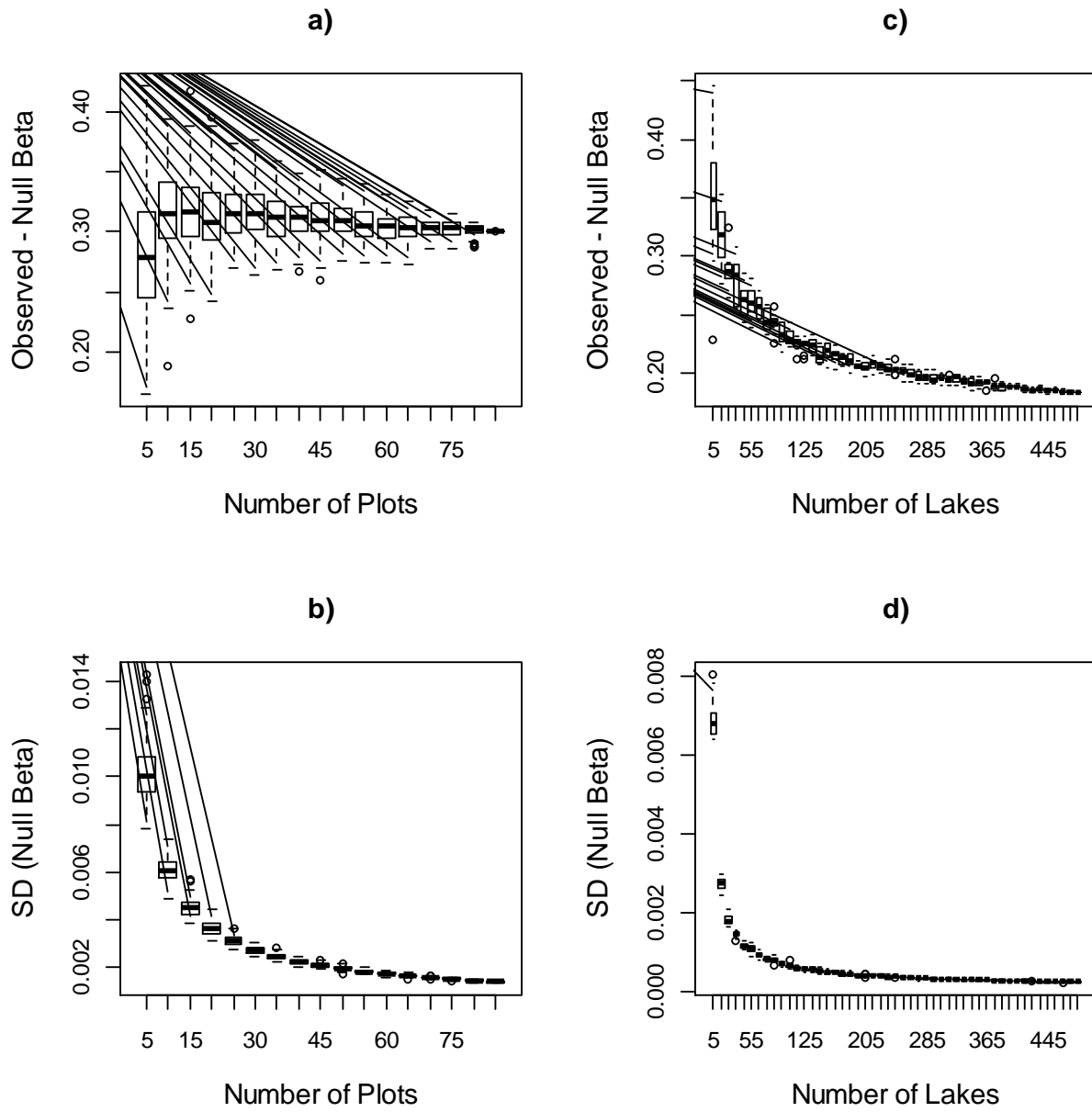
Fig. S5. Observed minus null β (a), and standard deviation of null β (b) versus number of plots for forest dataset; observed minus null β (c), and standard deviation of null β (d) versus number of lakes for diatom dataset.

9

a) **b)**
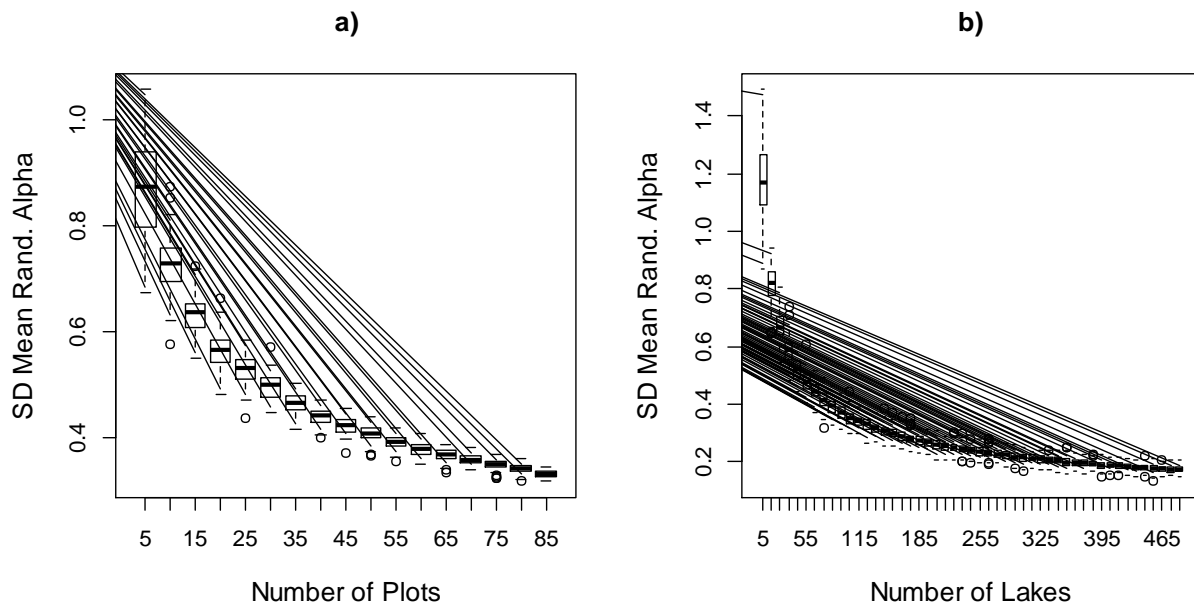
Fig. S6. Standard deviation (SD) of mean randomized α versus sampling effort, for forest plots (a), and diatom surveys (b).
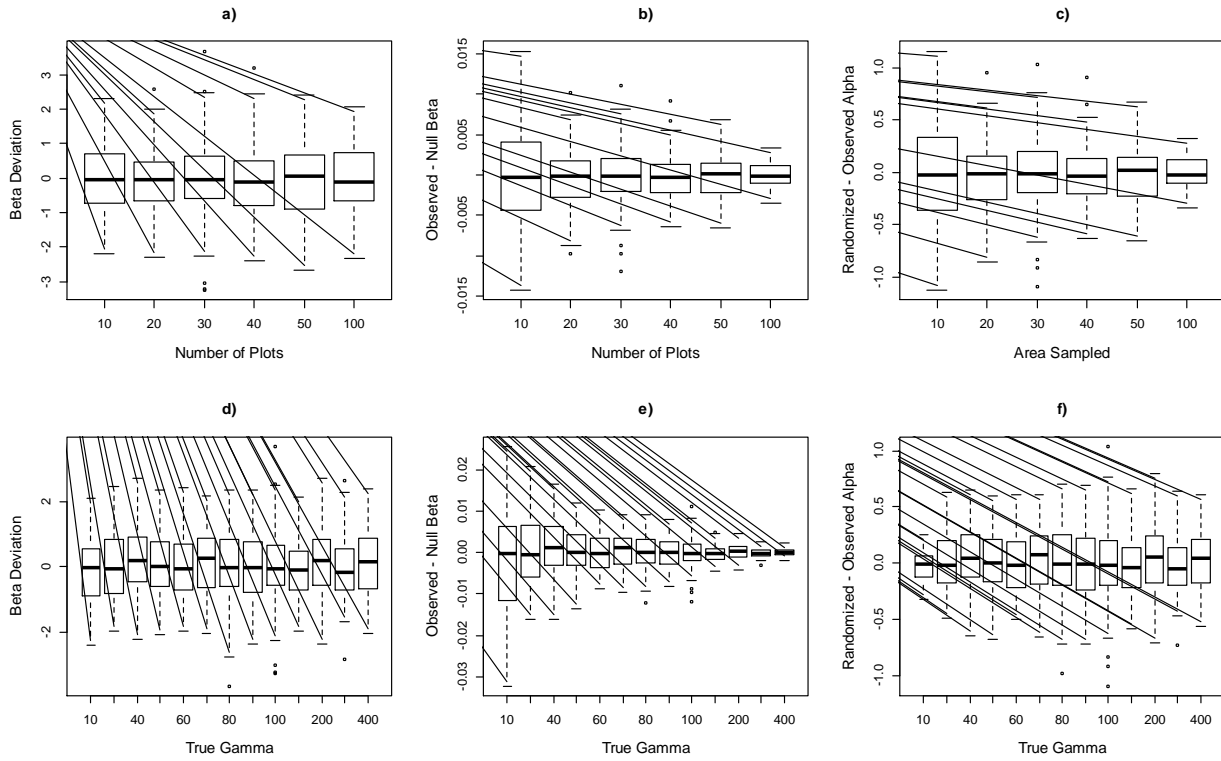
Fig. S7. β_{dev} and numerators of Eqns. 2 and 2a,b for simulated data as described above. β_{dev} (a), observed minus mean null β (b), and mean randomized minus observed mean α (c) versus number of plots for datasets with γ_{true} of 100; and β_{dev} (d), observed minus mean null β (e), and randomized minus observed mean α (f) versus γ_{true} for datasets with 30 simulated plots. In all cases the numerators of Eqns. 2 and 2a,b centred on zero because the randomization process in Kraft et al. (2011) shuffled data (while preserving the abundance distribution) for data that were already randomly-generated within the constraints of the abundance distribution.
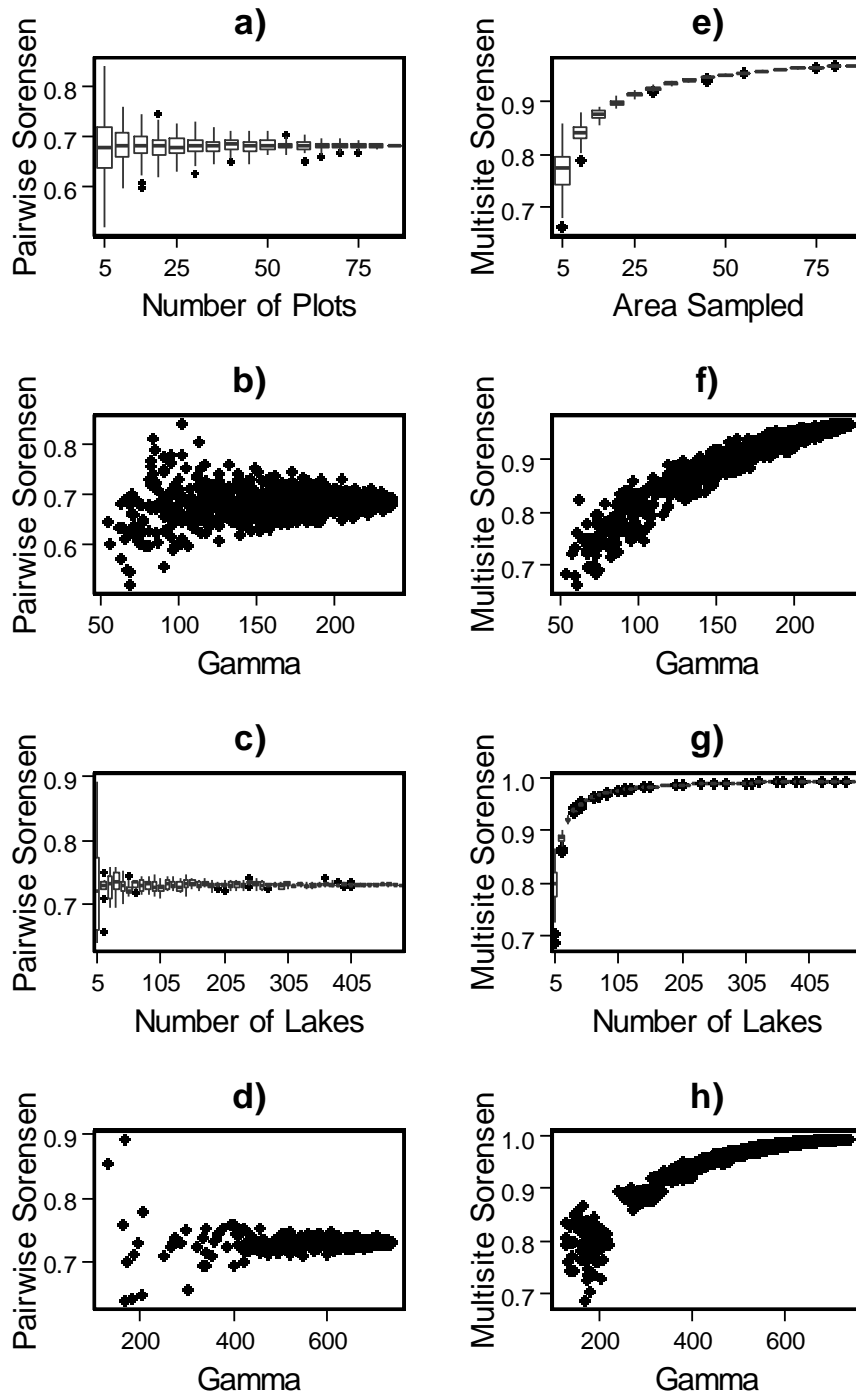
Fig. S8: Mean pairwise Sørensen distance versus number of 50 m^2 plots sampled (m^2) (a) and γ (b), for forest plots; mean pairwise Sørensen distance versus number of lakes sampled (m^2) (c) and γ (d) for diatom dataset; multiple-site Sørensen versus number of 50 m^2 plots sampled (e) and γ (f), for forest plots (a); and multiple-site Sørensen versus number of lakes sampled (m^2) (g) and γ (h) for diatom dataset.
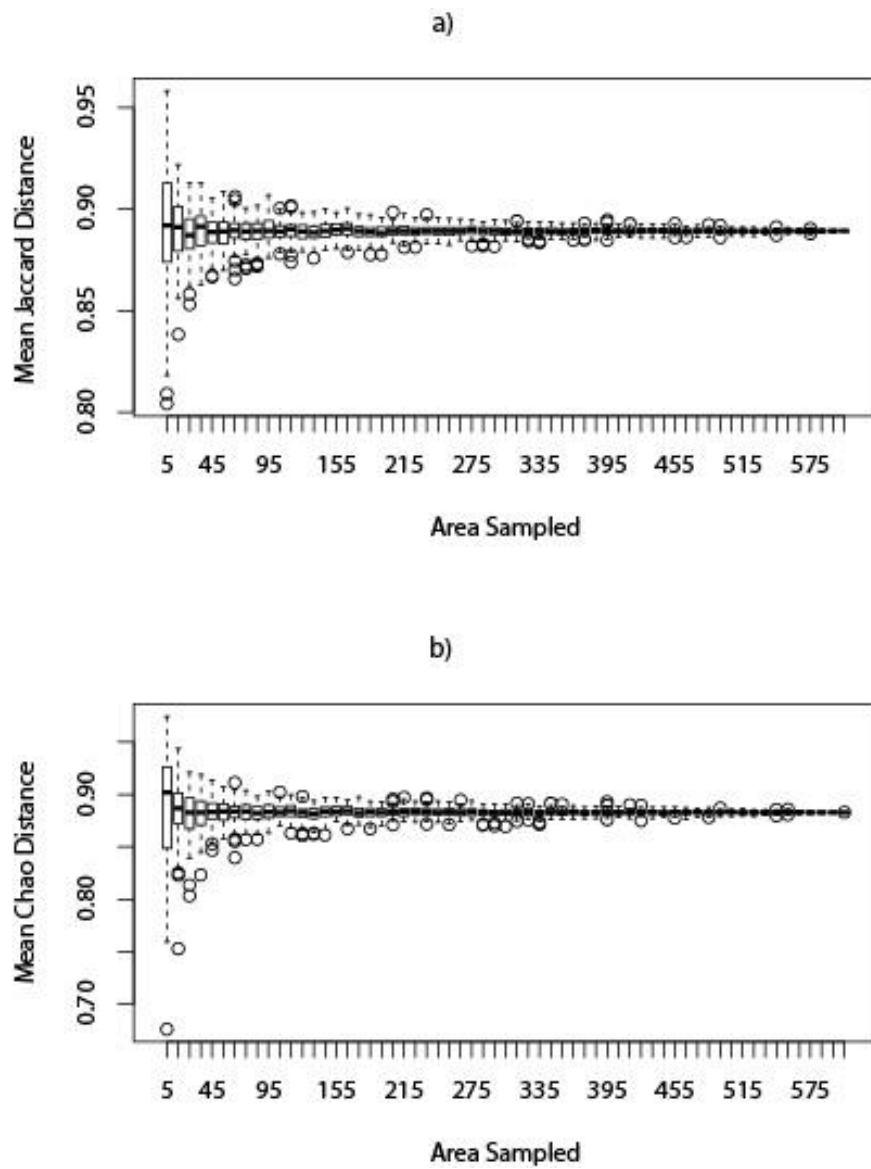
a)

b)

Fig. S9: Mean Jaccard (a) and Chao (b) distances with area sampled for meadow ecosystem (area increases with the number of plots, all plots $1m^2$). There is no significant relationship for either distance measure (both $p>0.4$).
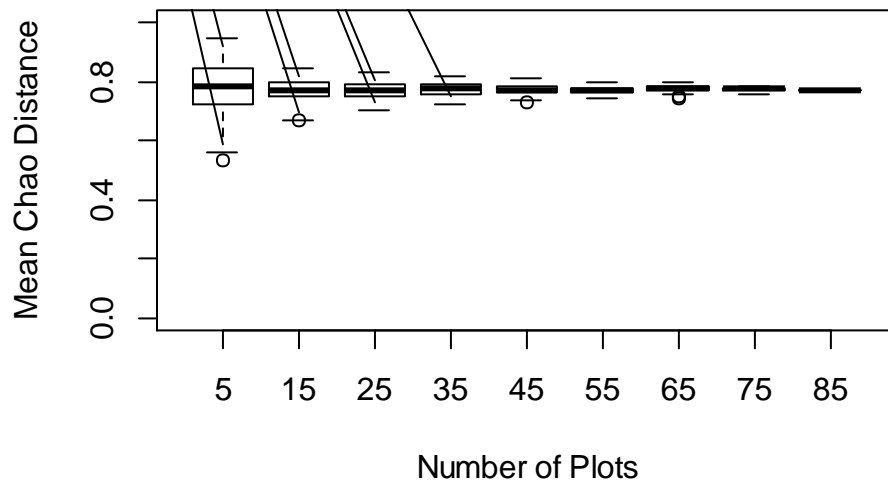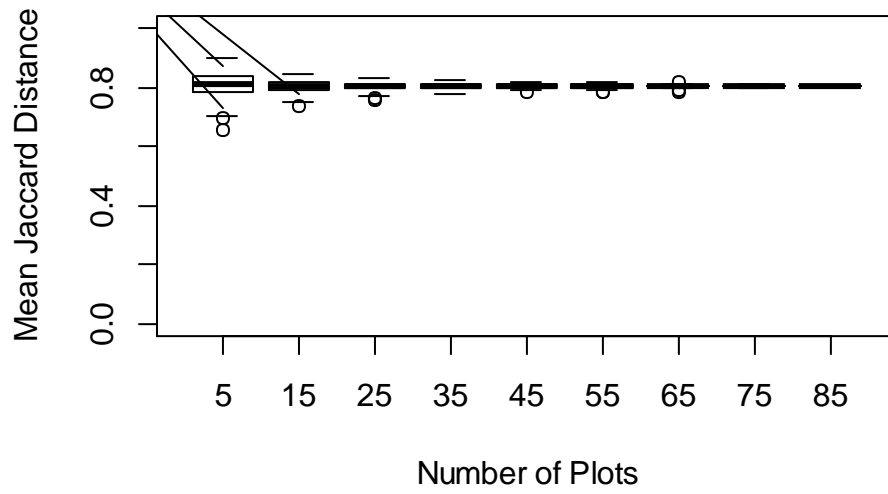
13

a)

b)
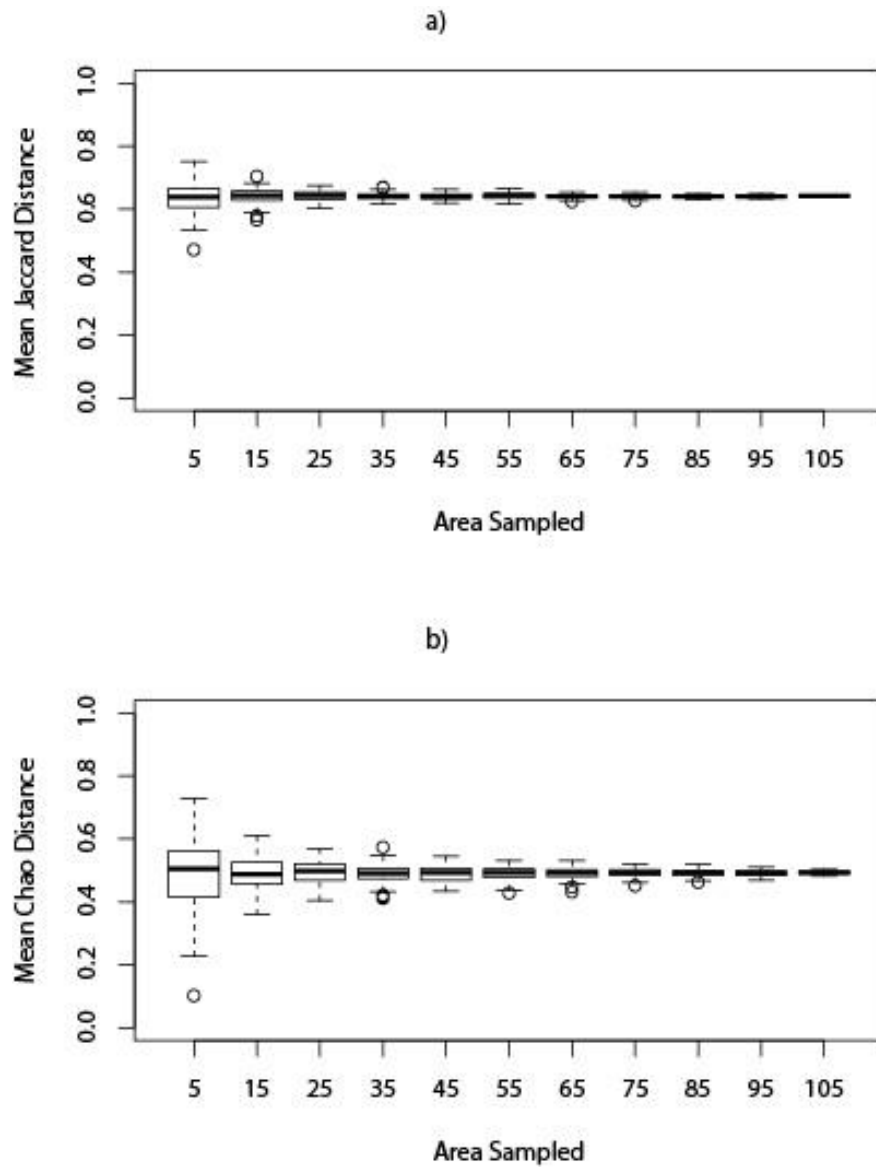
Fig. S10: Mean Jaccard (a) and Chao (b) distances with area sampled for the early successional dataset (area increases with the number of plots, all plots 1m^2). There is no significant relationship for either distance measure (both p>0.6).

Fig. S11: Mean Jaccard (a) and Chao (b) distances number of plots for forest plots. There is no significant relationship for either distance measure (both p>0.28).
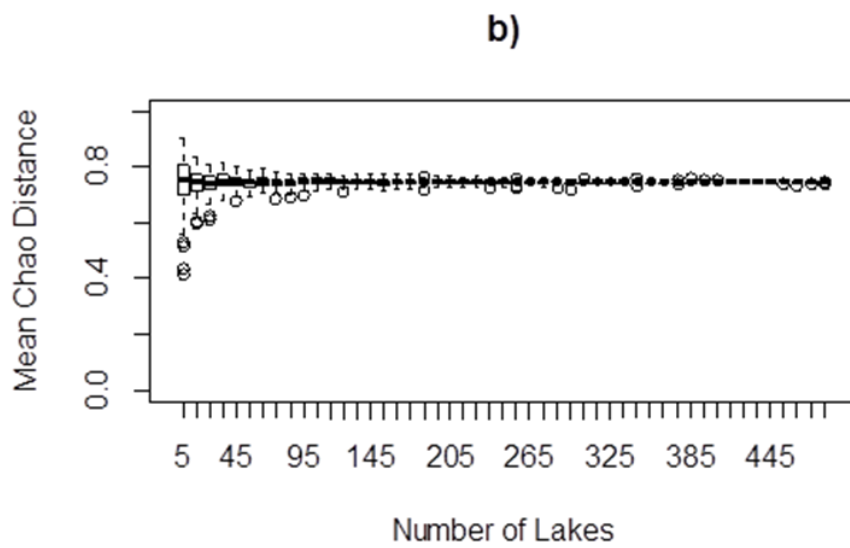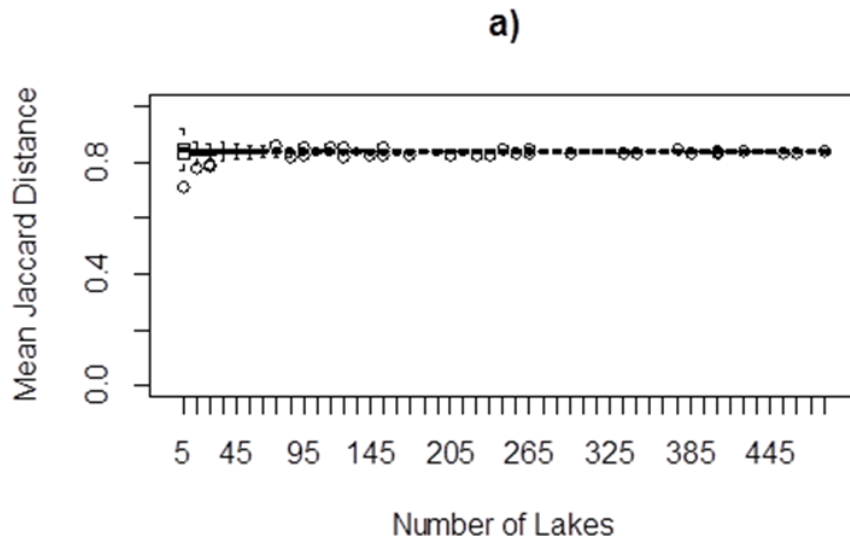
Fig. S12: Mean Jaccard (a) and Chao (b) distances with number of lakes sampled for diatom surveys. There is no significant relationship for either distance measure (both p>0.4).

a - Mean Sorensen

b - Sorensen mean distance to centroid

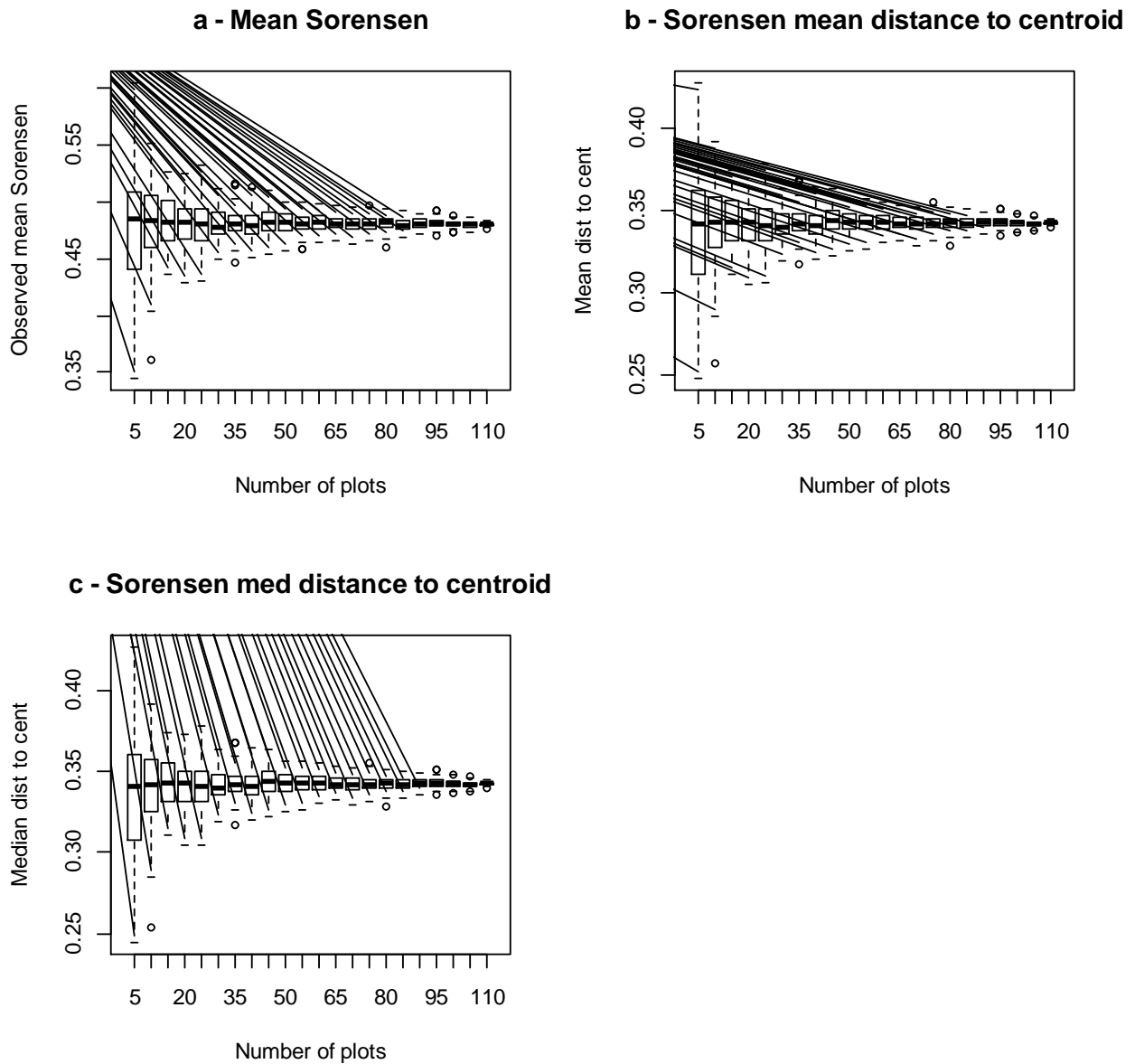**c - Sorensen med distance to centroid**

Fig. S13: Example of near-identical patterns for different measures of variation in Sørensen distances, early successional ecosystem plots. Mean Sørensen distance (a), mean distance to centroid (b) and median distance to centroid (c) all show no trend with number of plots sampled. Mean and median distances to centroid are corrected as per Stier et al. (2013). These patterns were repeated across all datasets.
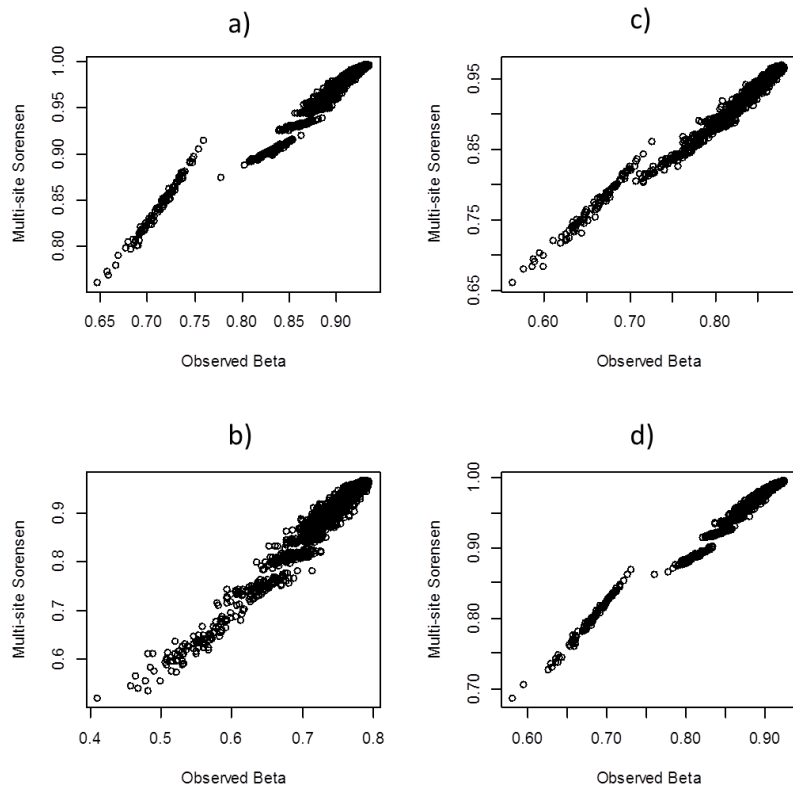
Fig. S14: Scatterplots of multiple-site Sorensen versus observed beta for meadow ecosystem (a; $r^2 = 0.97$, P<0.0001), early-successional ecosystem (b; $r^2 = 0.97$, P<0.0001), forest plots (c; $r^2 = 0.98$, P<0.0001) and diatom dataset (d; $r^2 = 0.99$, P<0.0001).
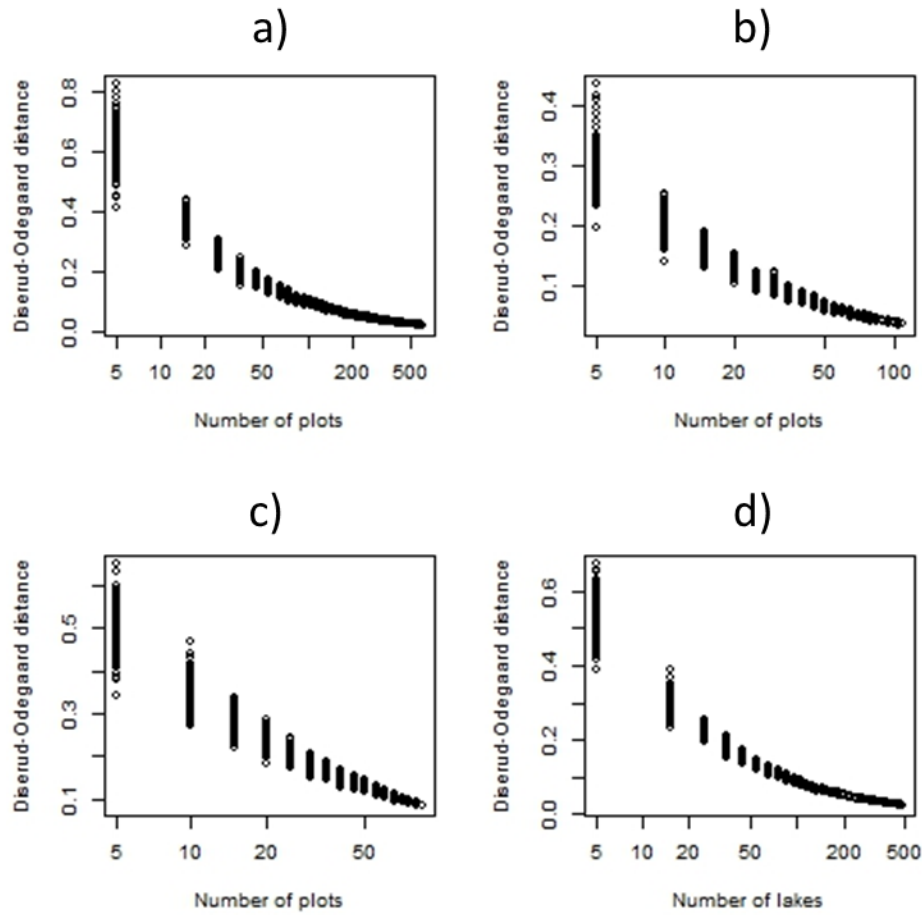
Fig. S15: Multiple-site Diserud & Ødegaard (2007) (DO) index versus number of plots for meadow ecosystem (a), early successional ecosystem (b), forest plots (c); and multiple-site DO index versus number of lakes for diatom dataset (d).
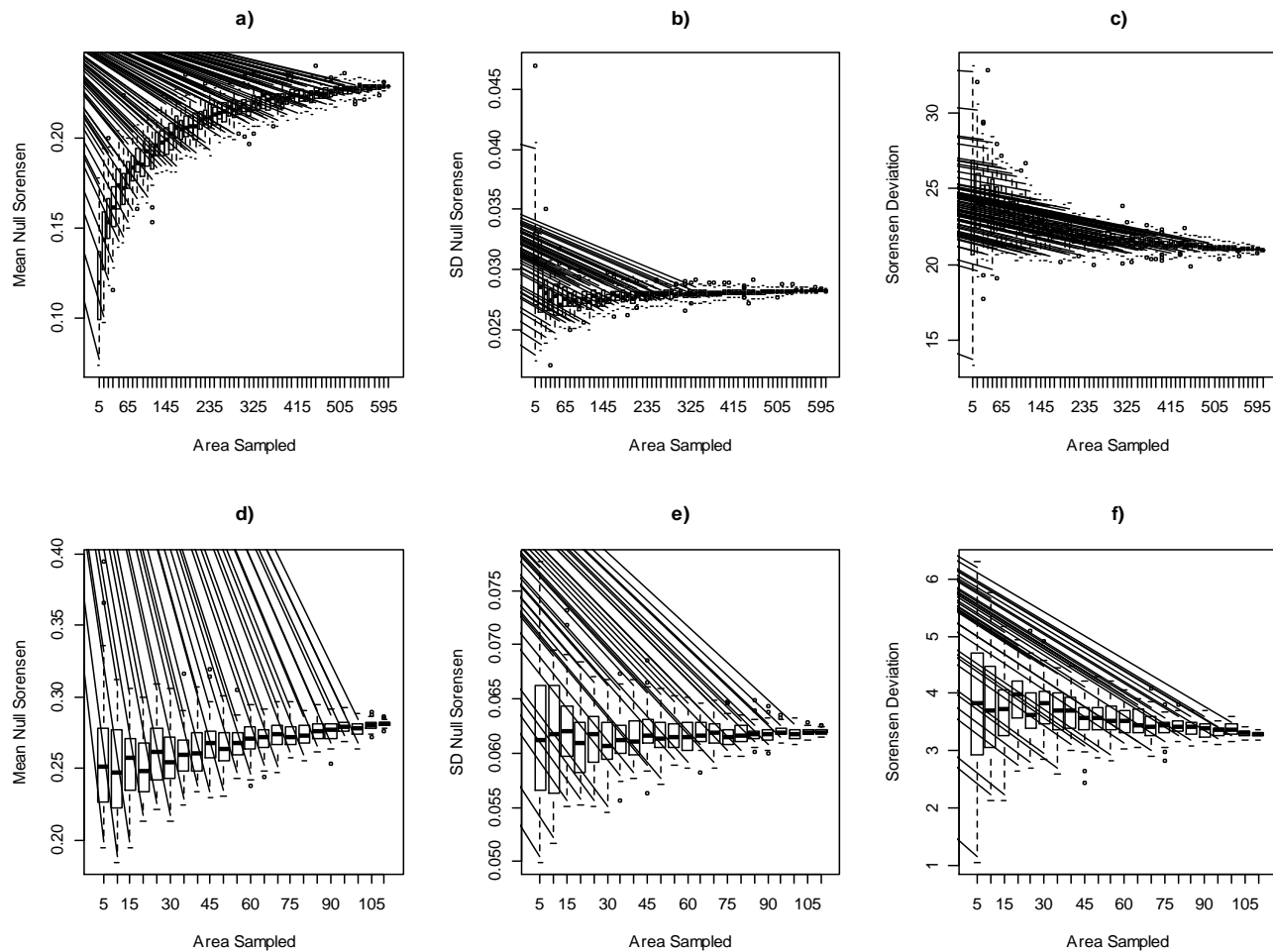
Fig. S16: Components of Sørensen deviations (Sørensen distances generated using the null model approach used to generate β_{dev}) across area sampled (i.e., number of 1m^2 plots) for meadow ecosystem and early successional ecosystem. Mean null Sørensen distances (a), standard deviation of mean null Sørensen distances (b), and Sørensen deviation (c) for meadow ecosystem; mean null Sørensen distances (d), standard deviation of mean null Sørensen distances (e), and Sørensen deviation (f) for early successional ecosystem. Correlations between Sørensen deviation and area sampled: r = -0.92, P>0.0001 (meadow ecosystem); r = -0.95, P>0.0001 (early successional ecosystem).
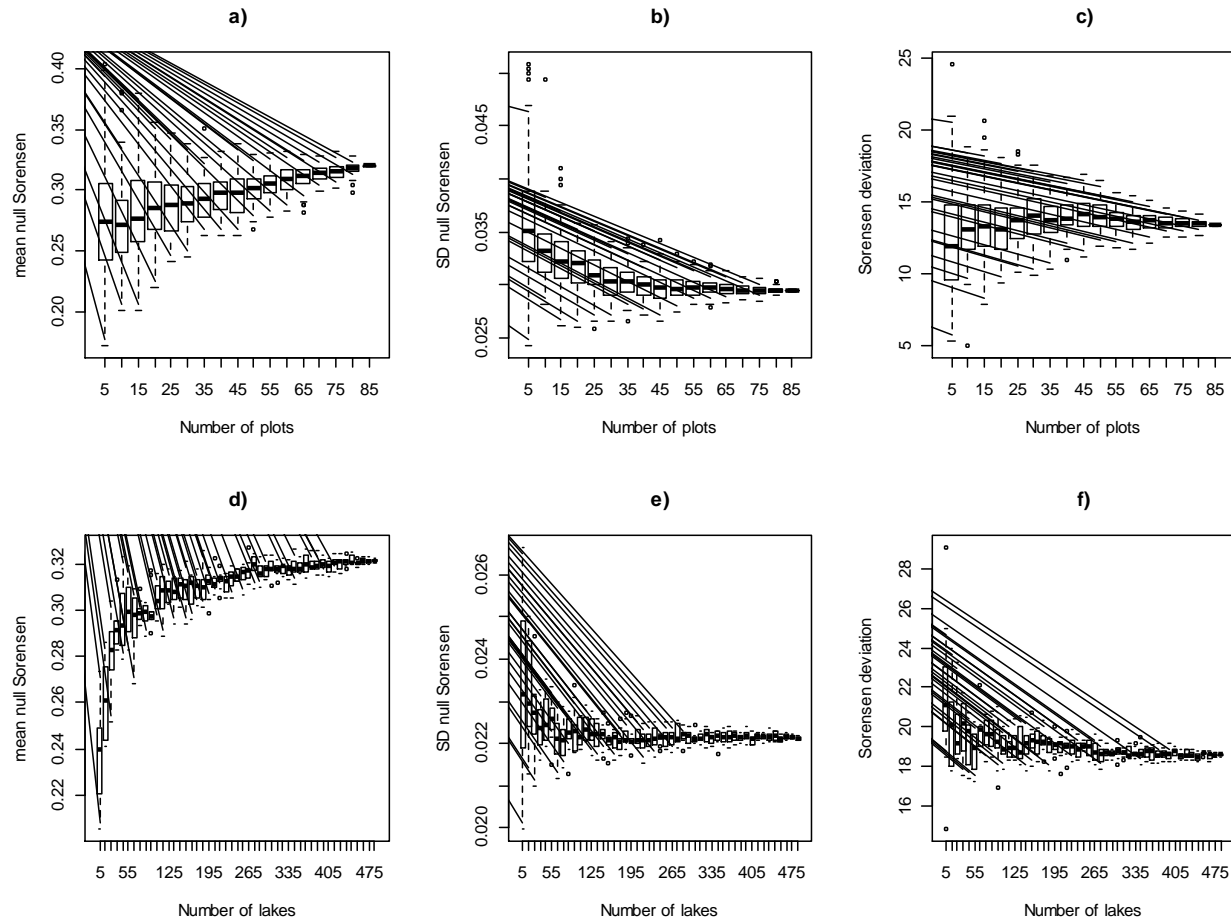
Fig. S17: Components of Sørensen deviations (Sørensen distances generated using the null model approach used to generate β_{dev}) across number of 50m^2 plots for forest dataset and number of lakes for diatom surveys. (a) mean null Sørensen distances for forest plots; (b) standard deviation (SD) of mean null Sørensen distances for forest plots; (c) Sørensen deviation for forest plots; (d) mean null Sørensen distances for diatoms; (e) standard deviation (SD) of mean null Sørensen distances for diatoms; (f) Sørensen deviation for diatoms. Correlations between Sørensen deviation and number of plots/lakes: r = 0.24, P=0.35 (forest plots); r = -0.42, P>0.0001 (diatoms).
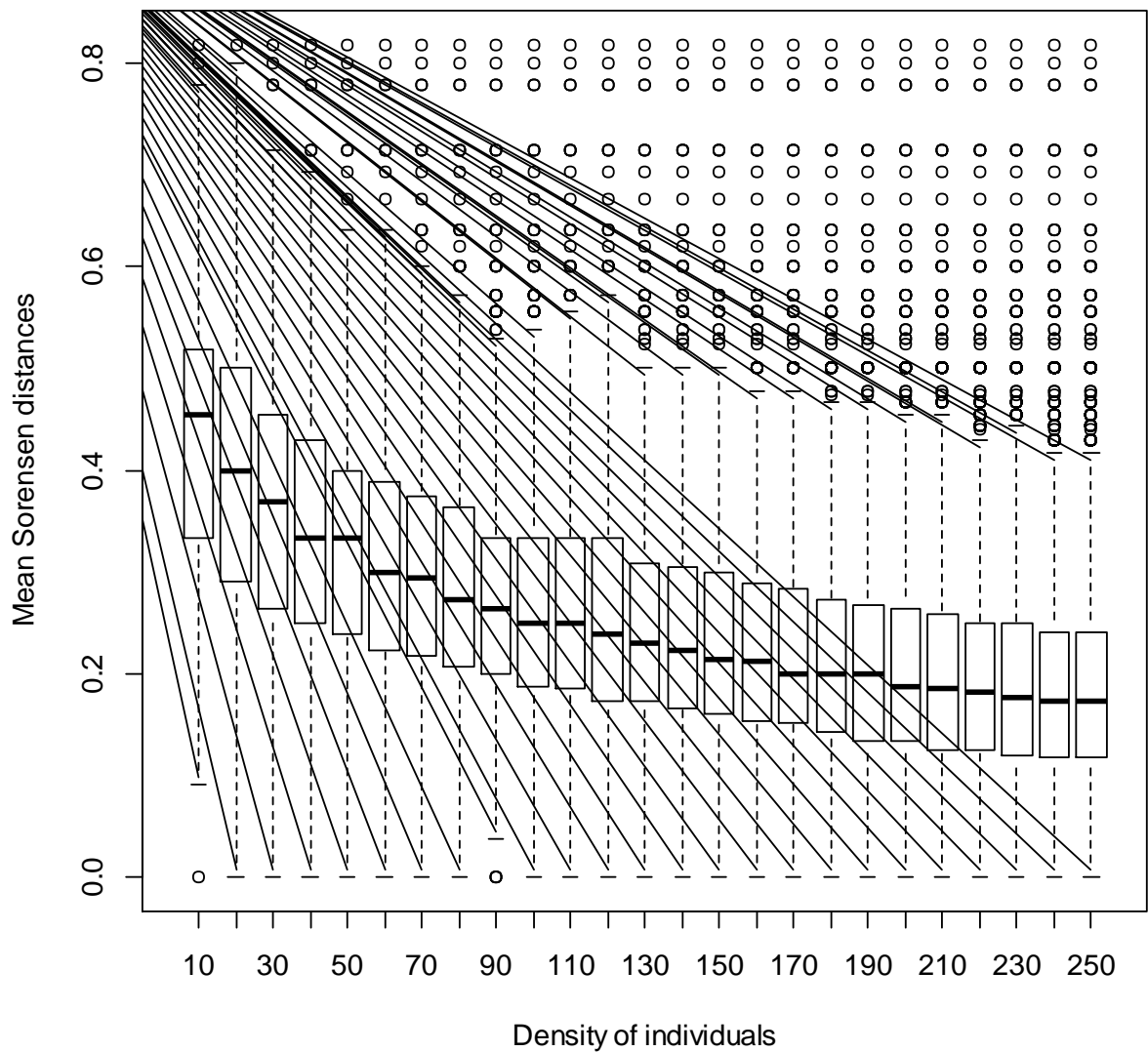
Fig. S18: Effect of density of individuals per sample on mean Sørensen distance between samples, using 100 pairs of samples for 20 simulated species drawn from abundance distributions simulated using the Fisher et al. (1943) log-series. The abundance of each species is calculated as follows: $Y = (-1/\log(1-c)) \times c^X/X$, where Y is the relative abundance of each species of rank X (in our case 1:20), and c is a coefficient determining the evenness of the abundance distribution. We used c = 0.9, as per Chase and Knight (2013), which generates a realistic abundance distribution. Greater density of individuals leads to more shared species among plots (if species are randomly-distributed), which leads to lower mean Sørensen distances.
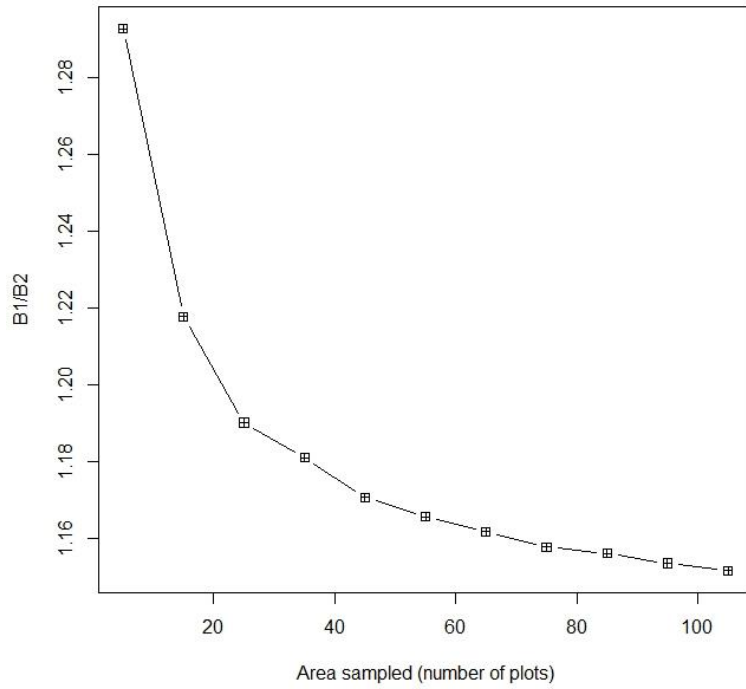
Fig. S19: Scale dependency of differences in classical β between two regions. The y-axis is the ratio of mean classical β in the meadow ecosystem to that in the old field ecosystem at the scale indicated (number of plots, x-axis). As is shown in eqn. S11, the scale dependency of this ratio cannot be expressed through a single index.