Ecography

E7361
Wiegand, T., Fanglinag, H. and Hubbell, S. P. 2012. A systematic comparison of summary characteristics for quantifying point patterns in ecology. – Ecography 35: xxx–xxx.

Supplementary material

Supplementary material

Appendix 1. Details on example patterns

Recruit patterns

The recruits of the two shade tolerant understorey species *Faramea occidentalis* (Fig. A1a) and *Desmopsis panamensis* (Fig A1b) showed little evidence for clustering except some weaker small scale clustering and some weak larger scale structures. The recruit pattern of the canopy and gap species *Cecropia insignis* (Fig. A1e) has been analyzed in detail in Wiegand et al. (2009). It shows a complex superposition pattern in which a random pattern of mostly isolated recruits was independently superimposed by a strongly clustered pattern in which small clusters of recruits are nested within larger clusters. Recruits of the understorey gap species *Croton billbergianus* shows similar extreme level of small-scale clustering, but also many isolated recruits (Fig. A1g). The pattern of the shade tolerant canopy species *Trichilia tuberculata* appears somewhat non-stationary with more larger-scale clustering on the right half of the plot and more isolated recruits on the left part of the plot (Fig. A1c). The canopy species *Calophyllum longifolium* with intermediate light sensitivity is an example of a non-stationary pattern (Fig. A1f). The shade-tolerant canopy species *Alseis blackiana* shows also a complex pattern in which isolated recruits and recruits with clustering at two scales are mixed (Fig. A1d). Finally, the shade tolerant understorey species *Coussarea curvigemmia* shows a sparse pattern with many isolated recruits and some clumps (Fig. A1f). Comita et al. (2007) analyzed all eight species and found in some cases a significant association of small trees to topographically defined habitat types which however did not translate into obvious spatial association patterns (compare patterns with map of habitat types in Fig. A1f with the distribution maps).

Simulated patterns

Because we do not know to which extent possible non-stationarity in the real-world patterns (Fig. A1) influences the ability of the summary statistics to characterize the spatial structure of

point patterns we repeated our analyses with eight simulated patterns that resemble the observed patterns, but were stationary by construction. The observation window was for all patterns a 1000 × 500m plot, the same as for the BCI data. The first pattern (Fig. A2a) was generated by a homogeneous Poisson process (Wiegand and Moloney 2004) by randomly and independently placing 400 points within the observation window. The second pattern is a regular "soft core" pattern (Fig. A2b) with an interaction range of 15m. The pattern was simulated by randomly and independently placing points within the observation window, but a new point within the interaction range of an already placed point (= 15m) was rejected with probability $(1 - r/15)^{0.5}$ where r is the distance between the two points. Thus, if a new point was very close to an existing point the rejection probability was high, and if it was placed at the border of the zone of influence of an existing point the rejection probability was low.

The third pattern was a clustered pattern generated by a Thomas process (Wiegand et al. 2009). The construction principle behind the Thomas process is simple; it consists of a number of randomly and independently distributed "clusters". The position of the cluster centers follows a homogeneous Poisson process with intensity ρ, the number of points per cluster follows a Poisson process, and the location of the points in a given cluster, relative to the cluster centre, have a bivariate Gaussian distribution $h(r, \sigma)$ with variance σ^2 (Stoyan and Stoyan 1994). The Thomas process has thus three parameters: the intensity λ of the process which determines the number of points of the pattern, the intensity ρ of the cluster centers, and the parameter σ that determines the cluster size. The third pattern consisted of 400 points located in 50 clusters with parameter σ = 5m. The typical diameter of the clusters is therefore 4σ = 20m and the mean number of points per cluster is 8. These parameters create patterns with strong smaller-scale clustering (Fig. A2c). The parameter σ = 5m was typical for small-scale clustering of recruits at BCI (Wiegand et al. 2009).

The fourth pattern was similar (400 points, 50 clusters), but now the clustering was governed by parameter $\sigma = 20$ which yields typical diameters of $4\sigma = 80$m. The parameter $\sigma =$ 20 was typical for the larger-scale clustering at BCI and the Sinharaja plot (Wiegand et al. 2007, 2009). This creates patterns with more diffuse clustering (Fig. A2d). The fifth pattern is the (independent) superposition of the two cluster patterns shown in Fig. A2c and Fig. A2d. We selected this superposition pattern because it resembles patterns of recruits and seedlings/saplings at the BCI and Sinharaja plot that often show two critical scales of clustering.

We also simulated a pattern with two nested critical scales of clustering using the same cluster parameters ($\sigma = 5, 20$). That means that the small clusters were not placed independently from the large clusters (as in Fig. A2f), but centered on the points of the pattern with larger-scale clustering (Wiegand et al. 2007, 2009). The pattern comprised also 800 points, and 200 small clusters with an average of 4 points were placed at the 200 points of the pattern with the larger-scale clustering. This creates a pattern (Fig. A2f) which is similar to the superposition pattern shown in Fig. A2e. However, because of the nested construction most points have their nearest neighbor within 10m whereas the points in Fig. A2e stemming from larger clusters may have their nearest neighbor only within 40m.

Finally, the pattern shown in Fig. A2g is the independent superposition of the random pattern (Fig. A2a) and the superposition of the two cluster patterns (Fig. A2e), and the pattern shown in Fig. A2f is the independent superposition of the small-scale cluster pattern (Fig. A2c) with the regular pattern (Fig. A2h).

Reference

Harms, K.E. et al. 2001 Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. - J. Ecol. 89: 947–959.

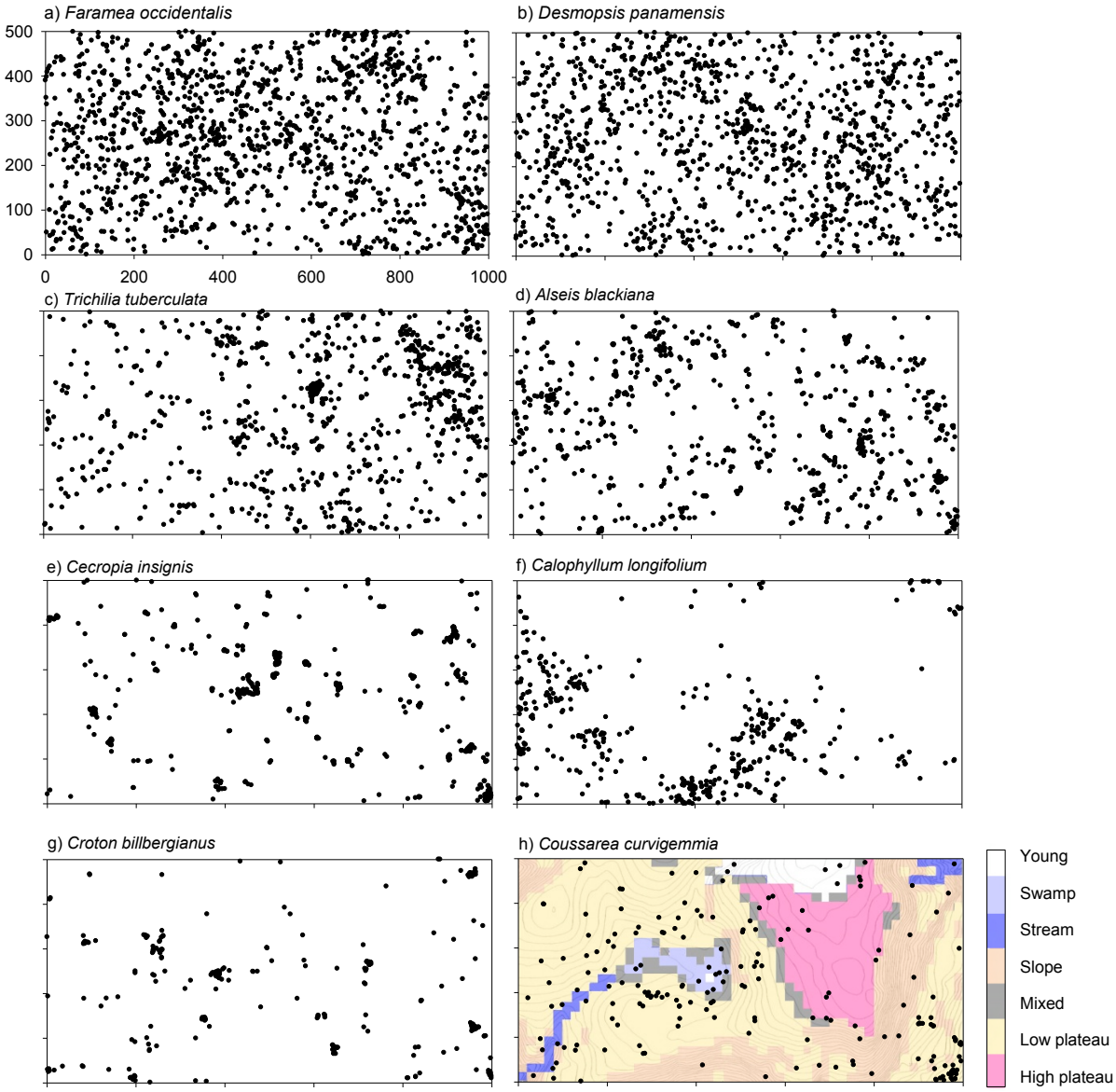

Figure A1. The eight representative patterns of newly recruited trees at the BCI forest used for our analyses. Recruits were all trees not present in the 2000 census but alive in the 2005 census. In panel H) we show additionally the different habitat types identified in Harms et al. (2001).

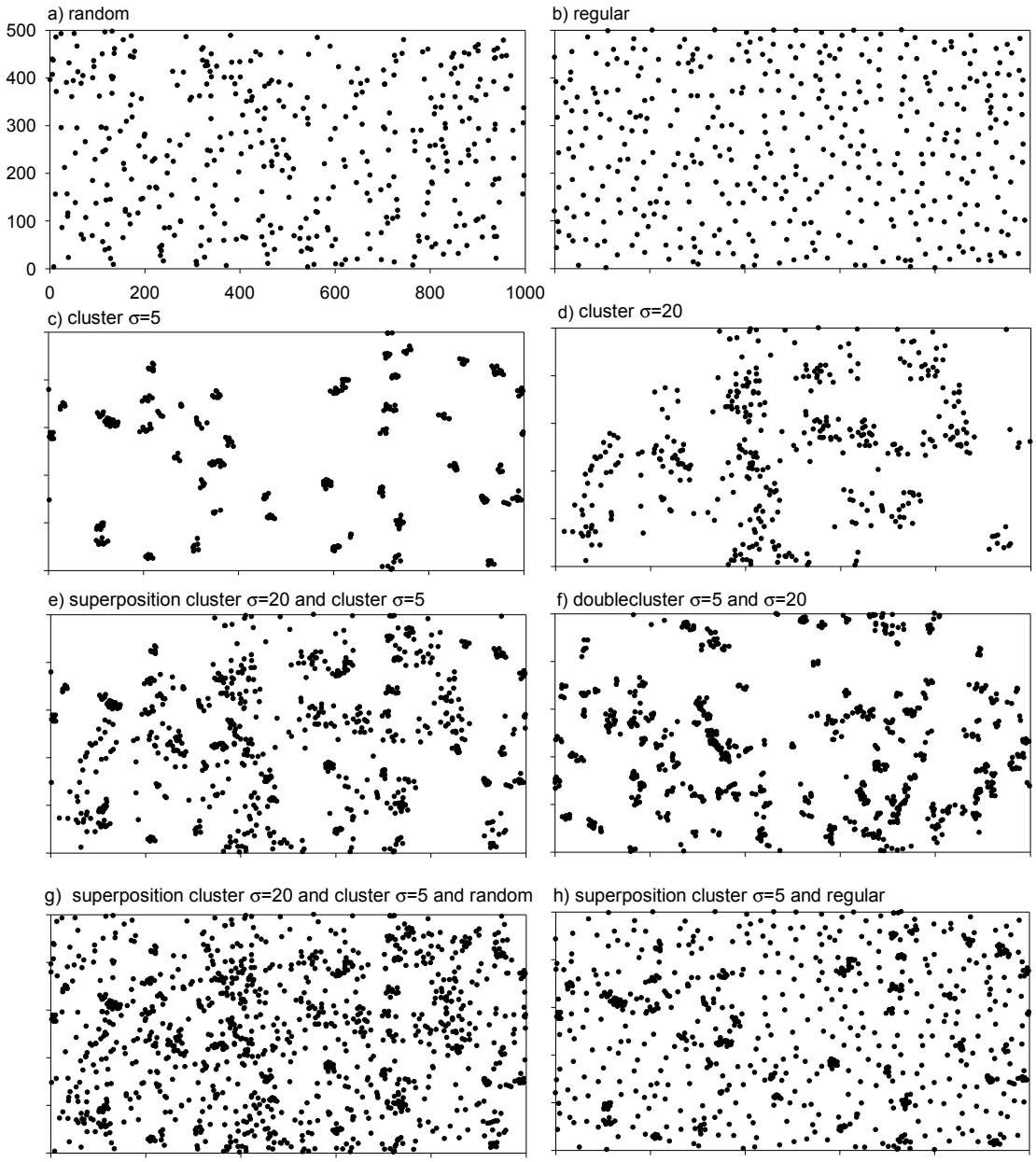

Figure A2. The eight simulated patterns used in our study. The values of σ refer to the parameters of the Thomas processes that determine the critical scale(s) of clustering.

Appendix 2. Estimation of summary statistics

Pair correlation function.—We used the pair correlation function $g(r)$ not directly for pattern reconstruction, but the related quantity $2\pi r dr\, \lambda g(r)$ which gives the mean number of points at rings with radius r and width dr around the points of the pattern. We used this quantity instead of $g(r)$ because it varies over a much reduced range, especially if the pattern shows strong small-scale aggregation (e.g., Fig. A3).

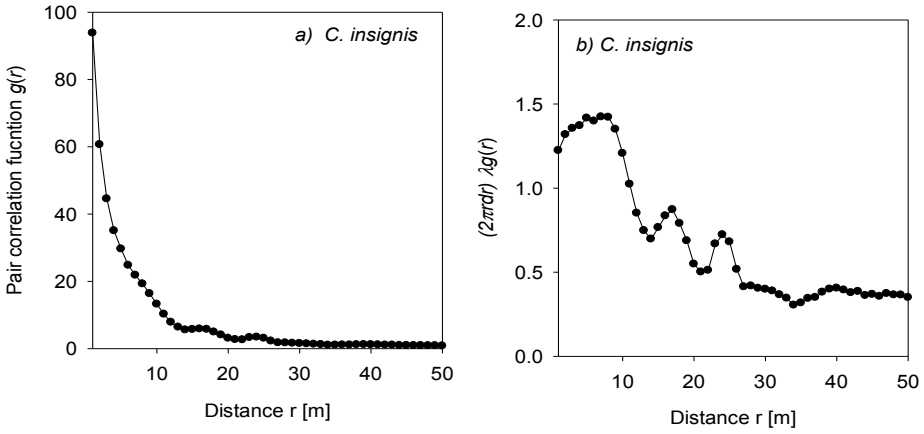

Figure A3. Pair correlation function $g(r)$ and the transformation $2\pi r dr\, \lambda g(r)$ used for pattern reconstruction for the highly clustered species *C. insignis*.

We estimated the average number of points within distances (r-dr/2, r+ dr/2) from the points of the pattern with an estimator based on Ohser and Mucklich (2000):

$$\frac{A}{\bar{\gamma}_W(r)} \frac{1}{n} \sum_{a=1}^{n} \sum_{b=1}^{n,\neq} k(\|\mathbf{x}_a - \mathbf{x}_b\| - r) \qquad \text{(A1)}$$

The function $k(\|\mathbf{x}_a - \mathbf{x}_b\| - r)$ is an indicator function which yields 1 if the distance r is within the interval (r-dr/2, r+ dr/2) and zero otherwise. The term $\sum_{b=1}^{n,\neq} k(\|\mathbf{x}_a - \mathbf{x}_b\| - r)$ therefore

calculates the number of points $\boldsymbol{x}_b$ which are located within (r-dr/2, r+ dr/2) from the focal points $\boldsymbol{x}_a$, and the term $\frac{1}{n}\sum_{a=1}^{n}\sum_{b=1}^{n,\neq} k(\|\mathbf{x}_a - \mathbf{x}_b\| - r)$ is the average over all focal points $\boldsymbol{x}_a$. Clearly, if the focal point $\boldsymbol{x}_a$ is located close to the edge of the window W the ring is only in completely located inside W and the unobserved points outside W cause a bias. This bias is corrected by the factor $\frac{A}{\bar{\gamma}_W(r)}$ which approximates a value of 1 for small distances r and increases with increasing r to compensate for the bias due to edge effects. The isotropized set covariance $\bar{\gamma}_W(r)$ (Stoyan and Stoyan 1994, p. 123) is proportional to the expected boundary length of a circular line inside the observational window W with radius r and random center location in W. Note that the estimator given in equation A1 can be effectively calculated because it does not individually correct the bias for every point pair as done for example by the Ripley estimator (Illian et al. 2008).

K-function.—We estimated the quantity $\lambda K(r)$ with an estimator analogous to that of the pair correlation function:

$$\left(\frac{\pi r^2}{2\pi\int_0^r t\bar{\gamma}_W(t)dt}\right)\frac{A}{n}\sum_{a=1}^{n}\sum_{b=1}^{n,\neq}\mathbf{1}(\|\mathbf{x}_a - \mathbf{x}_b\|, r) \qquad \text{(A2)}$$

The function $\mathbf{1}(\|\mathbf{x}_a - \mathbf{x}_b\|, r)$ is an indicator function which yields 1 if the distance between points $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ is smaller or equal to r and zero otherwise. The $\frac{1}{n}\sum_{a=1}^{n}\sum_{b=1}^{n,\neq}\mathbf{1}(\|\mathbf{x}_a - \mathbf{x}_b\|, r)$ thus estimates the mean number of points $\boldsymbol{x}_b$ within distance r of focal points $\boldsymbol{x}_a$. The weighting factor (the first factor in Eq. A2) that corrects edge effects is the area of the full circle (numerator) divided by the expected area (within W) of disks with radius r and random center location in W (denominator).

Nearest neighbor statistics.— Estimation of the nearest neighbor statistics may also require edge correction methods (Illian et al. 2008) because here the kth nearest neighbor may be located outside the observation window. This is likely if the observed distance to the kth nearest neighbor is larger than the distance to the border. Illian et al. (2008) recommend the Hanisch method for edge correction (Hanisch 1984, Stoyan and Stoyan 1994: p. 296). We implemented this method following Stoyan (2006) for the spherical contact distribution $H_S(r)$. However, we did not use edge correction for estimation of $D_k(r)$ because the Hanisch method needs to decide for each point individually if it is used for estimation or not. This makes this estimator slow when only one point of the pattern is exchanged. Nevertheless, we implemented the Hanisch method in our software also for $D_k(r)$ and found no substantial differences if the pattern had a larger number of points (say > 200). Note that not using edge correction for $D_k(r)$ is less severe than for $g(r)$ of $K(r)$. This is because in the worst case where a point is located in the corner of W there is still a ¼ probability that the kth neighbor is within W, and if this is not the case there is a high probability that the k+1, k+2, or k+n'th neighbor is within W which provides a good approximation to the distance to the kth neighbor.

Similarly, we did not use edge correction for estimation of the $E(r)$ and $P(k, r)$.

References

Baddeley, A. J., J. Møller and R. Waagepetersen. 2000. Non- and semiparametric estimation of interaction in inhomogeneous point patterns. Statistica Neerlandica 54: 329–350.

Hanisch, K.-H. 1984. Some remarks on estimators of the distribution function of nearest neighbour distance in stationary spatial point patterns. Mathematische Operationsforschung und Statistik, Series Statistics 15, 409-412.

Ohser, J., and Mucklich, F. 2000. Statistical analysis of microstructures in materials science. John Wiley & Sons, Ltd, Chichester.

Appendix 3. Estimation of intensity function

We estimated the intensity function from the recruit patterns using standard procedures of non-parametric kernel estimation based on the Epanechnikov kernel with bandwidth h (Stoyan and Stoyan 1994). We used a bandwidth of h = 50 m. Thus, the intensity estimate removes spatial structures below 50 m. However, for patterns with strong small-scale clustering a direct non-parametric intensity estimate is problematic because the strong clustering is a second-order phenomenon independent on the first-order structure we aim to capture. Biologically, the small clusters can be due to some inherent clustering mechanisms (e.g., a smaller treefall gap, limited seed dispersal, clumped seed deposition; Wiegand et al. 2009) and not caused by locally elevated habitat suitability. We therefore removed the strong small-scale clustering before estimating the intensity function (= "de-clustering").

We checked first if strong small-scale clustering was present by means of the pair correlation function (Fig. A4). If this was the case we removed the small-scale clustering by using a cluster detection algorithm based on an amalgamation rule called "unweighted pair-group centroid" (Sneath and Sokal 1973). Briefly, this algorithm determines the pair of points which has the smallest interpoint distance and amalgams this pair into a new point (termed "cluster") which is the mean location of the two original points. The algorithm proceeds by searching the pair of points of the updated pattern which has the smallest interpoint distance and again amalgams the two points but uses all original points contained in a given "cluster" for determining the new mean location. This is repeated as long as point pairs can be found that have an interpoint distance smaller than a predefined maximal distance r_c. Finally we replaced the amalgam points of the pattern of the final step by the point of the respective amalgam cluster that is located closest to its center of gravity. In this way all small clusters are represented by only one point. The maximal distance r_c is based on a plot of the pair correlation function of the original pattern that allows to determine the scale of strong small-scale clustering (Fig. A4).

Reference

Sneath, P.H.A. and R.R., Sokal. 1973. Numerical Taxonomy. Freeman, San Francisco.

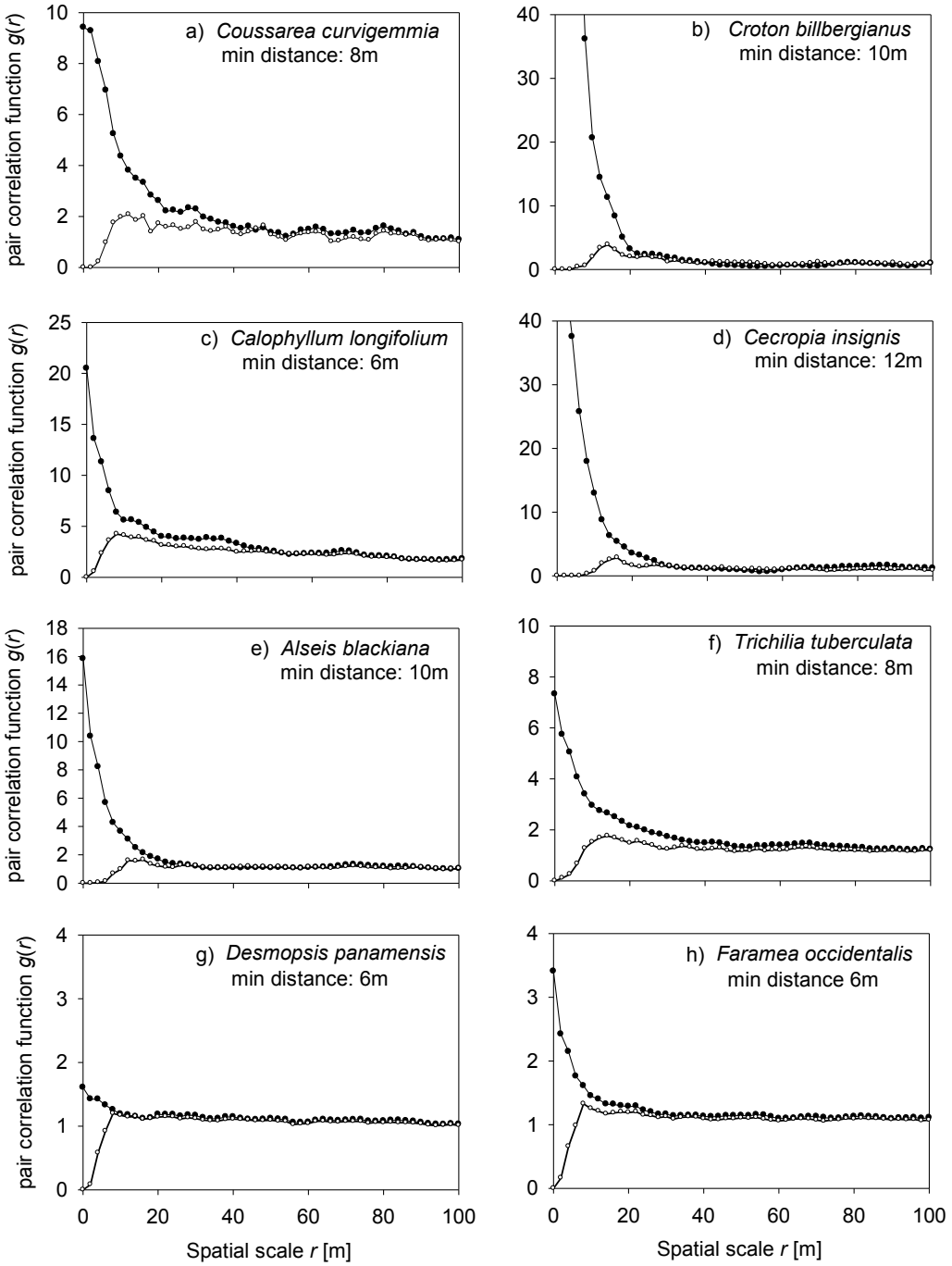

Figure A4. De-clustering. The figures show for each species the pair correlation function of the observed pattern (closed disks) and the pair correlation function of the de-clustered pattern (open disks). The figures show that de-clustering indeed removed the strong small-scale clustering, but in many cases larger scale clustering (heterogeneities) remain.

Appendix 4: thresholds for match of partial energies

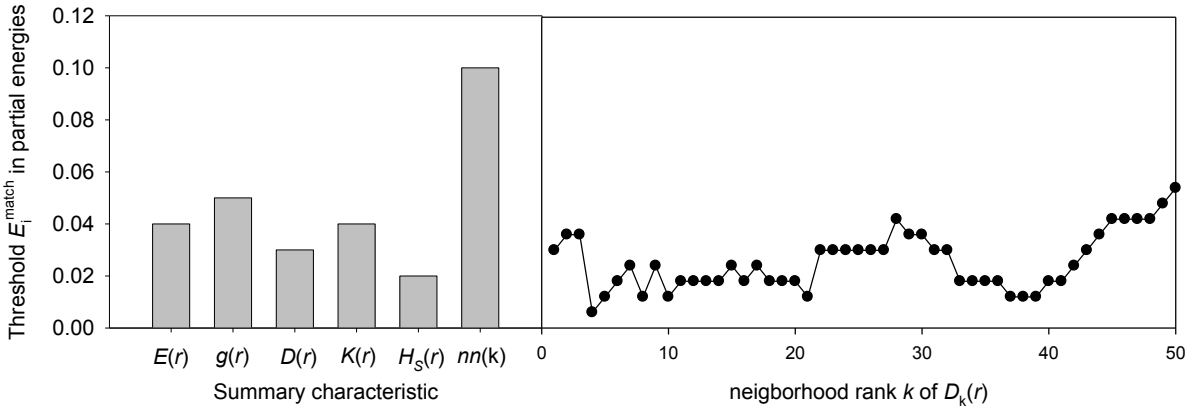

Figure A5. Threshold values E_i^{match} of the partial energies that define match of the corresponding summary characteristic i. As threshold value we used the largest value of the partial energies taken from all reconstructed patterns of simulation experiment 2 in which the corresponding summary characteristic was used. However, for the $nn(k)$ we visually selected a value of 0.1.

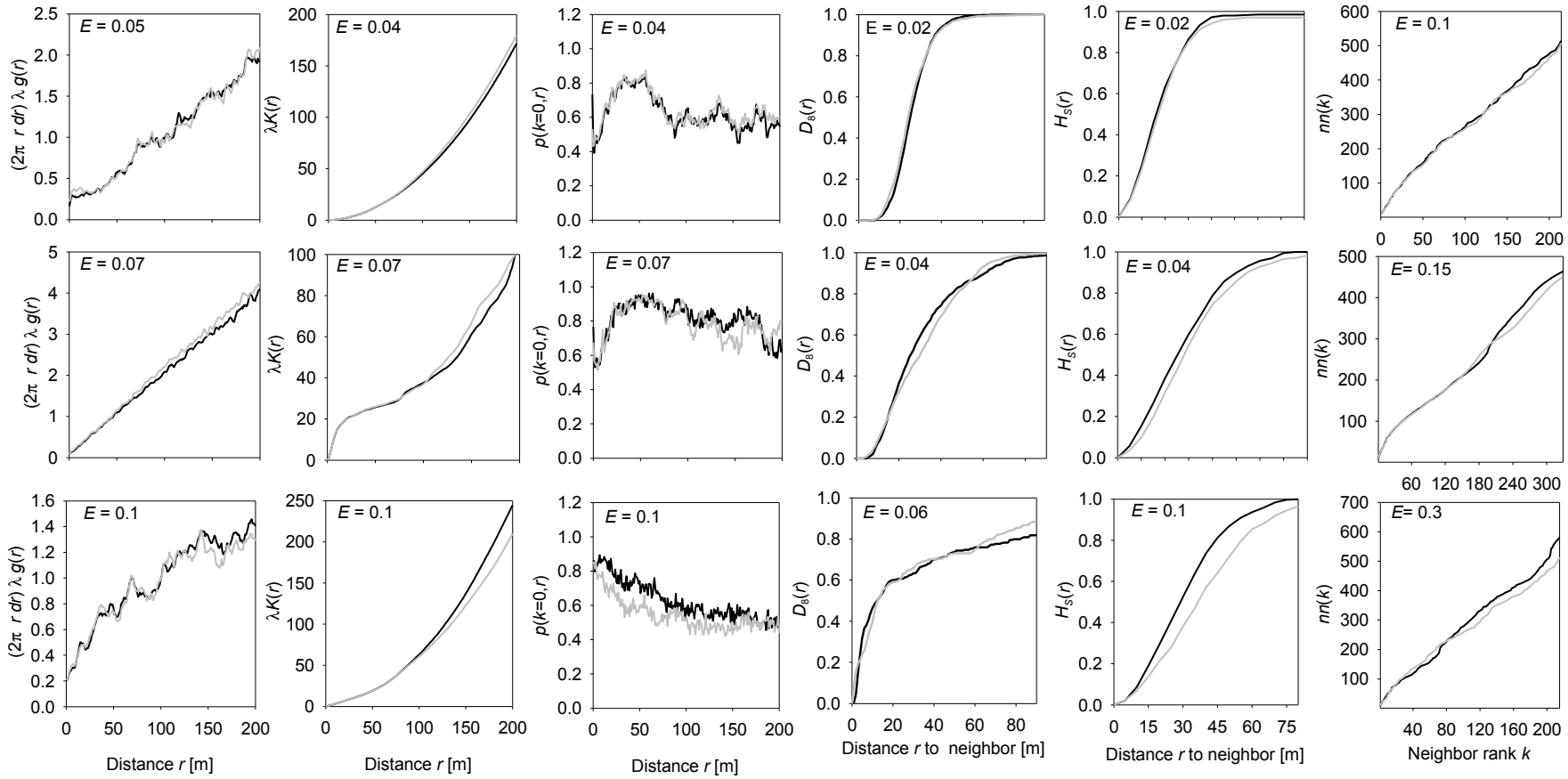

Figure A6. Examples for partial energies. The figures show the summary statistics of the observed pattern (gray), the reconstructed pattern (black), and the corresponding value of the partial energies E_i is given at the left upper corner of the panels. The graphs in the upper row show summary characteristics that just fulfill the threshold criterion, the graphs in the middle row show partial energies that just fail to match and the lower row shows partial energies that clearly fail to match.

Appendix 5: Supplementary results

Figure A7 shows several examples of reconstructions for spatial patter of the recruits of the species *Cecropia insignis*. This pattern contained most spatial structure of all eight real-world example patterns. Reconstructions with one summary characteristic (Fig. A7g, i, k) clearly show that one summary statistic alone is not able to capture the spatial structure of the example patterns. Especially poor performed the nearest neighbor statistic $D(r)$ (Fig. A7i) which failed especially in reproducing the typical small clusters. This was somewhat expected because $D(y)$ evaluates only the immediate neighborhood of the points and cannot describe high neighborhood densities. The sequence of 50 distribution functions $D_k(y)$ to the k^{th} neighbor, however, yielded a substantial improvement over using only the first nearest neighbor (cf. Figs. A7i, j). Figure A7j shows that the $D_k(r)$'s were able to produce small-scale clusters because they constrain the aspects of the spatial structure of the first 50 neighbors. However, the $D_k(r)$'s together were still not enough to yield good accordance in the pair correlation function and the *K*-function. This is because we included only $D_k(r)$'s up to a neighbor rank of k=50 which does not cover the entire 0-200 m ranges of scales r used for calculation of the pair correlation function; the mean distance to the 50^{th} neighbor was 85m for the species *C. insignis*.

Figure A7g shows that the *K*-function produced reconstructions with reasonable small-scale clusters but too many scattered isolated points which produced only poor matches in the nearest neighbor statistics. This is because the *K*-function constrains only mean neighborhood densities but not the distribution of the number of neighbors. This is a well known phenomenon that reflects the theoretical finding that substantially different (stationary) cluster processes can have identical second-order statistics (Tscheschel and Stoyan 2006, Wiegand et al. 2007a, 2009). As a consequence, the reconstruction based only on the *K*- or pair correlation function worked best for the pattern of the species *D. panamensis* (Fig. A1b) which shows little clustering and has a stationary appearance. Reconstruction worked worse

for the highly clustered pattern of the species *C. insignis* (Fig. A1e). Overall, these results show that complex point patterns can in general not be reconstructed based on a single summary statistic. Our analysis also showed that the mean distances $nn(k)$ to the kth neighbor captured the small- and intermediate scale spatial structure of the observed pattern poorly (Fig. A7k). This illustrates that the $nn(k)$ is largely independent on the summary characteristics $g(r)$, $K(r)$, $E(r)$, and $D_k(r)$ that capture small- and intermediate scale spatial structure.

When we combined two summary statistics we found that the pair correlation function in combination with the distance distribution functions to the k^{th} neighbors $D_k(r)$ produced the best overall description of the example patterns (Figure 3). This is because second-order statistics [i.e., $g(r)$ and $K(r)$] and k^{th} nearest neighbor characteristics [i.e., $D_k(r)$] characterize the spatial patterns from two fundamentally different points of view (i.e., neighborhood density vs. distribution of nearest neighbor distances). Visual inspection (Fig. A7a vs. Fig. A7e) confirms the improvement in the reconstruction. However, the typical "gaps" in the observed pattern (Fig. A7a) were poorly matched by the combination $g(r)$ and $D_k(r)$ (Fig. A7e), but the reconstruction shown in Fig. A7c that uses also the spherical contact distribution $H_s(r)$ that estimates the "gaps" in a pattern improves the fit of this characteristic (Fig. A7c). Visual comparison of the observed and reconstructed pattern confirms the good reconstruction (cf. Fig. A7a and Fig, A7c).

As noted before, the large-scale properties of the observed pattern can only be recovered by pattern reconstruction if the intensity function is used. This is also shown in Figure A7, although it is not clear if the observed pattern shows heterogeneity or if the larger scale properties (some gaps in the distribution as shown by the intensity function in Fig. A7b) are fluctuations possible under homogeneity.

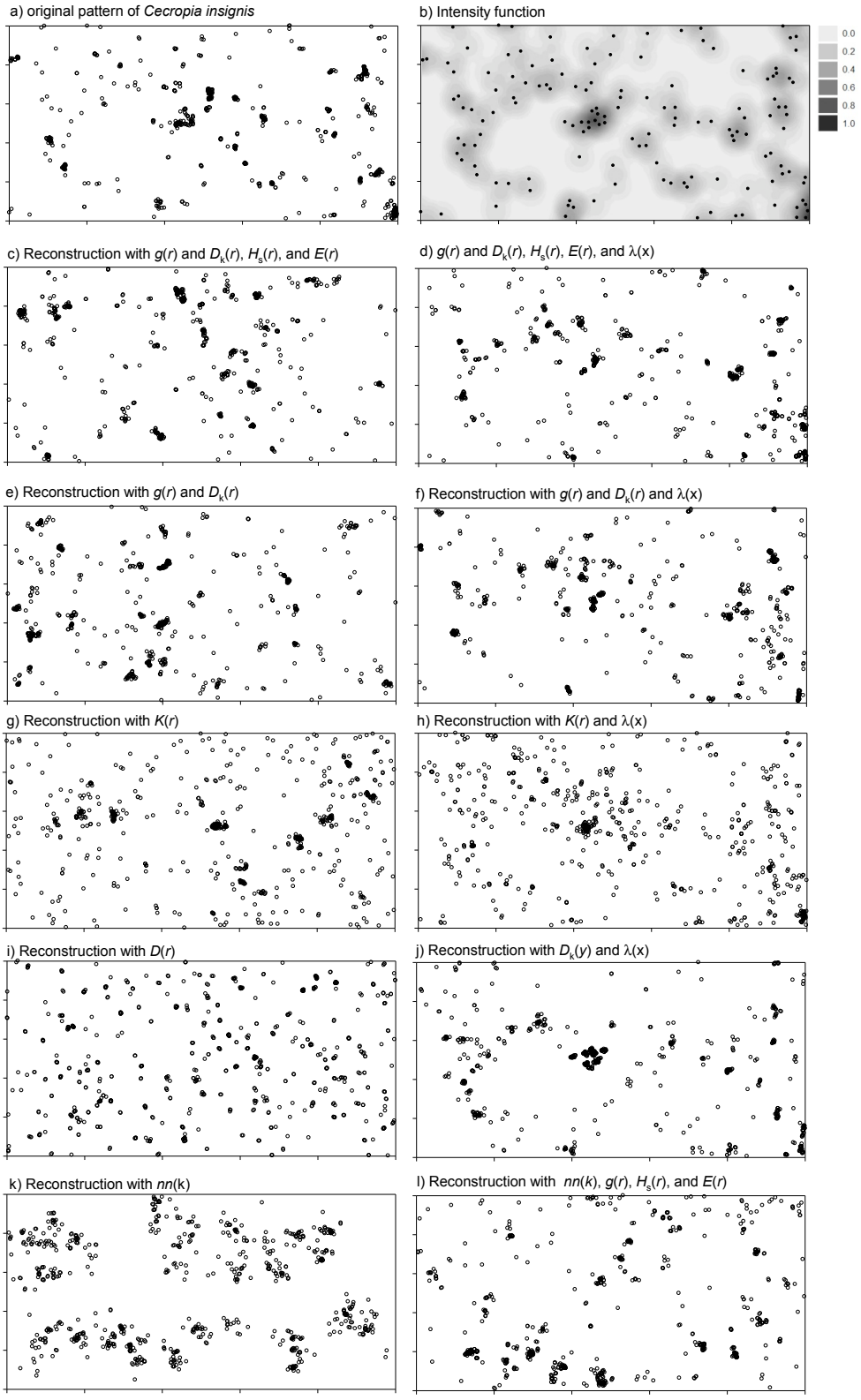

Figure A7. Examples of reconstructions of the pattern of the species *Cecropia insignis*. a) the observed pattern; b) the de-clustered pattern and the intensity estimate with a bandwidth of 50m; c) – j) reconstructions with different combinations of summary characteristics without intensity function $\lambda(\mathbf{x})$ (left column) and with intensity function $\lambda(\mathbf{x})$ (right column).

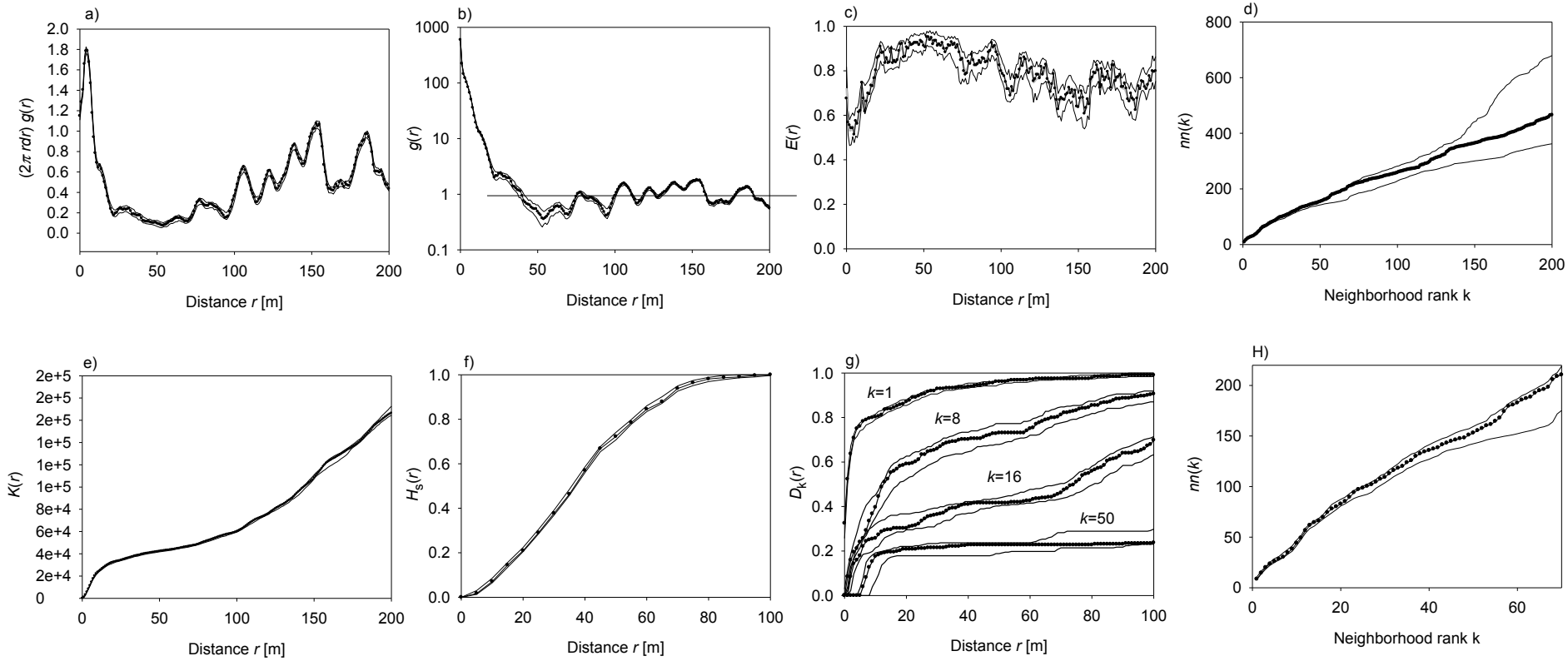

Figure A8. Assessment of the reconstruction quality for the different summary characteristics measured by means of simulation envelopes. Shown are results for the pattern of the species *Croton billbergianus*. The simulation envelopes for a given summary characteristic are the maximal and minimal values among all reconstructions (conducted under simulation experiment for the pattern of *C. billbergianus*) that used that summary characteristic.

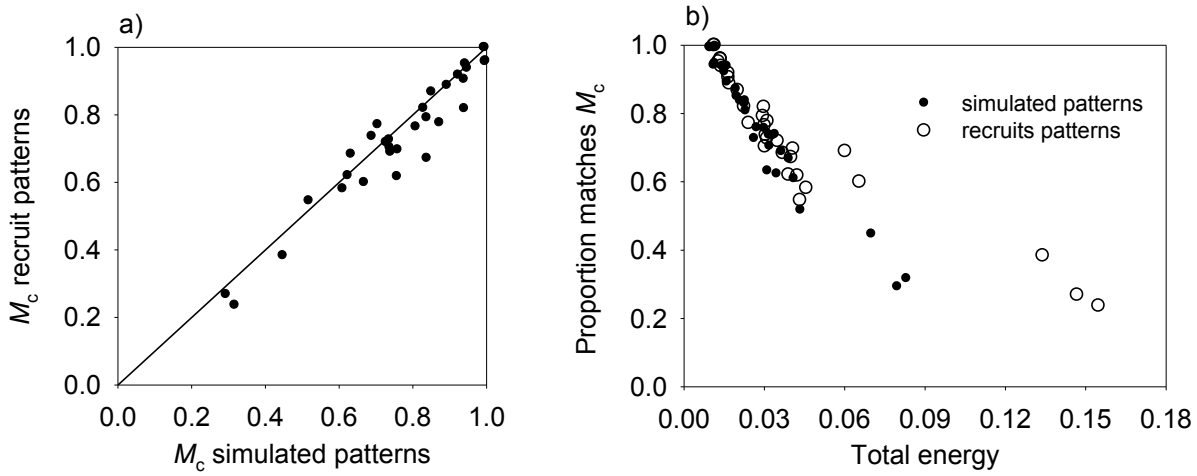

Figure A9. Comparison of performance of different combinations *c* of summary characteristics between simulated (stationary) patterns and (partly non-stationary) recruit patterns. a) the index M_c that measures the information contained in a given combination *c* of summary characteristics for simulated patterns (x-axis) and real-world patterns (y-axis). b) relationship between the index M_c and the total energy E_{total} averaged over the six summary characteristics $g(r)$, $K(r)$, $D(r)$, $D_k(r)$, $H_s(r)$, and $E(r)$, the eight patterns and 10 replicates for the simulated patterns (closed circles) and the recruit patterns (open circles).

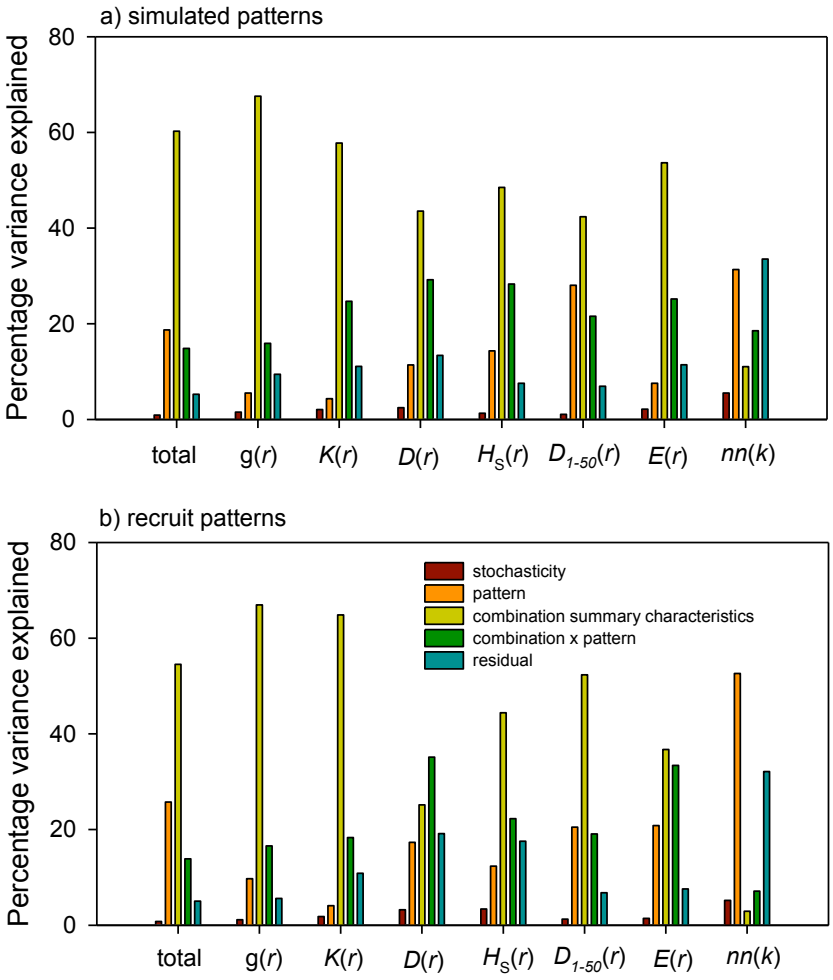

Figure A10. Results of analysis of variance for the index M_c that measures the information contained in a given combination c of summary characteristics the partial indices M_c(i) that measure the information contained in a single summary characteristic i. a) Results for the simulated patterns shown in Fig. A2. b) Results for the recruits patterns shown in Fig. A1. Shown are the percentage of total variance in the index M_c and the partial indices $M_c(i)$ explained by the factors "pattern" (8 different patterns were reconstructed), "summary characteristic" (32 combinations of summary characteristics were used for reconstruction), combination of pattern and combination, and stochasticity (10 replicated were generated for each of the 8 × 32 reconstructions).

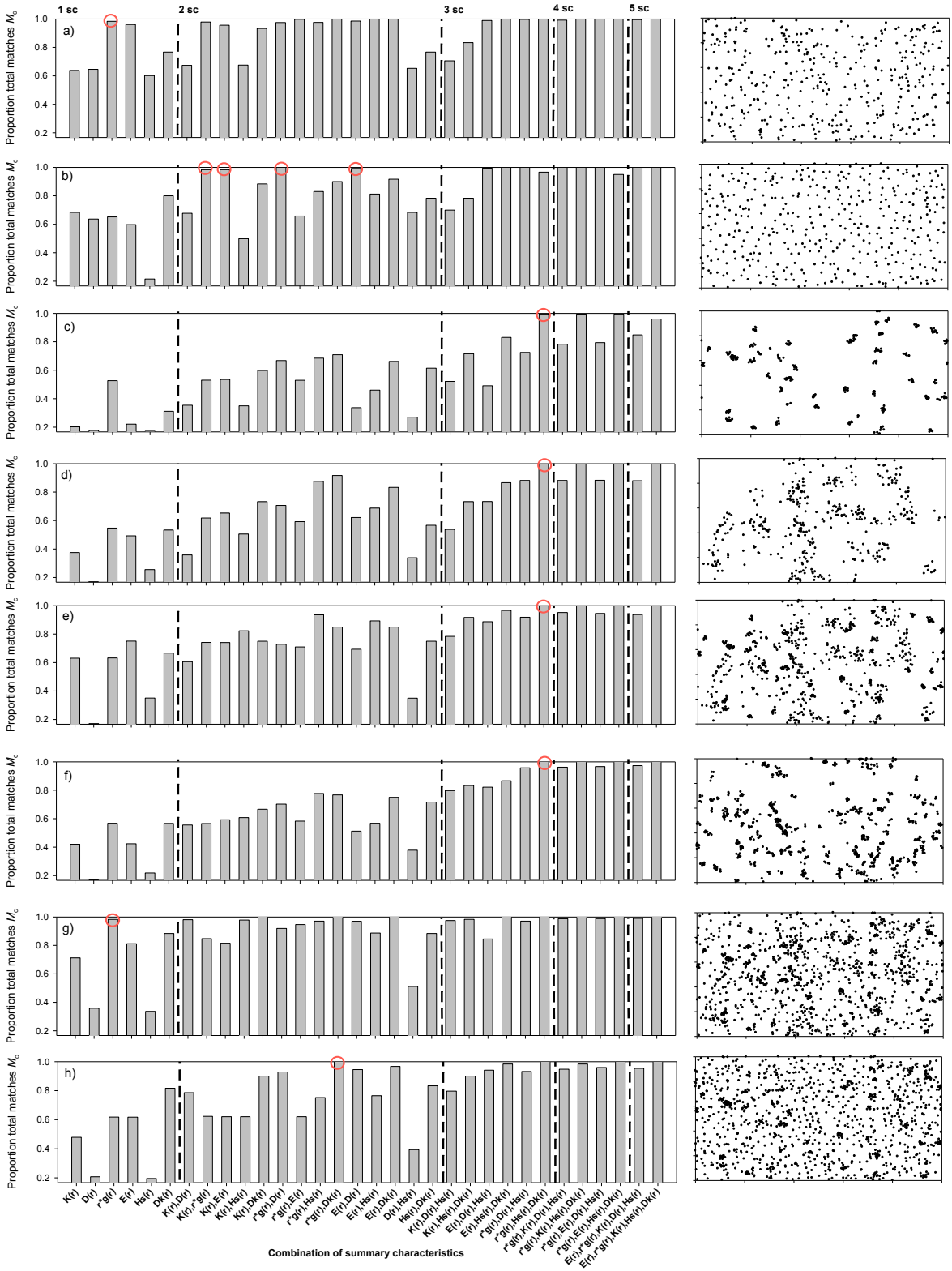

Figure A11. Ranking of combination of summary characteristics (as Figure 3), but individually for the eight simulated patterns. Shown is the total pattern match M_c for each combination *c*. The best combinations are marked with a red circle. The panel letter coincides with that of Fig. A2.

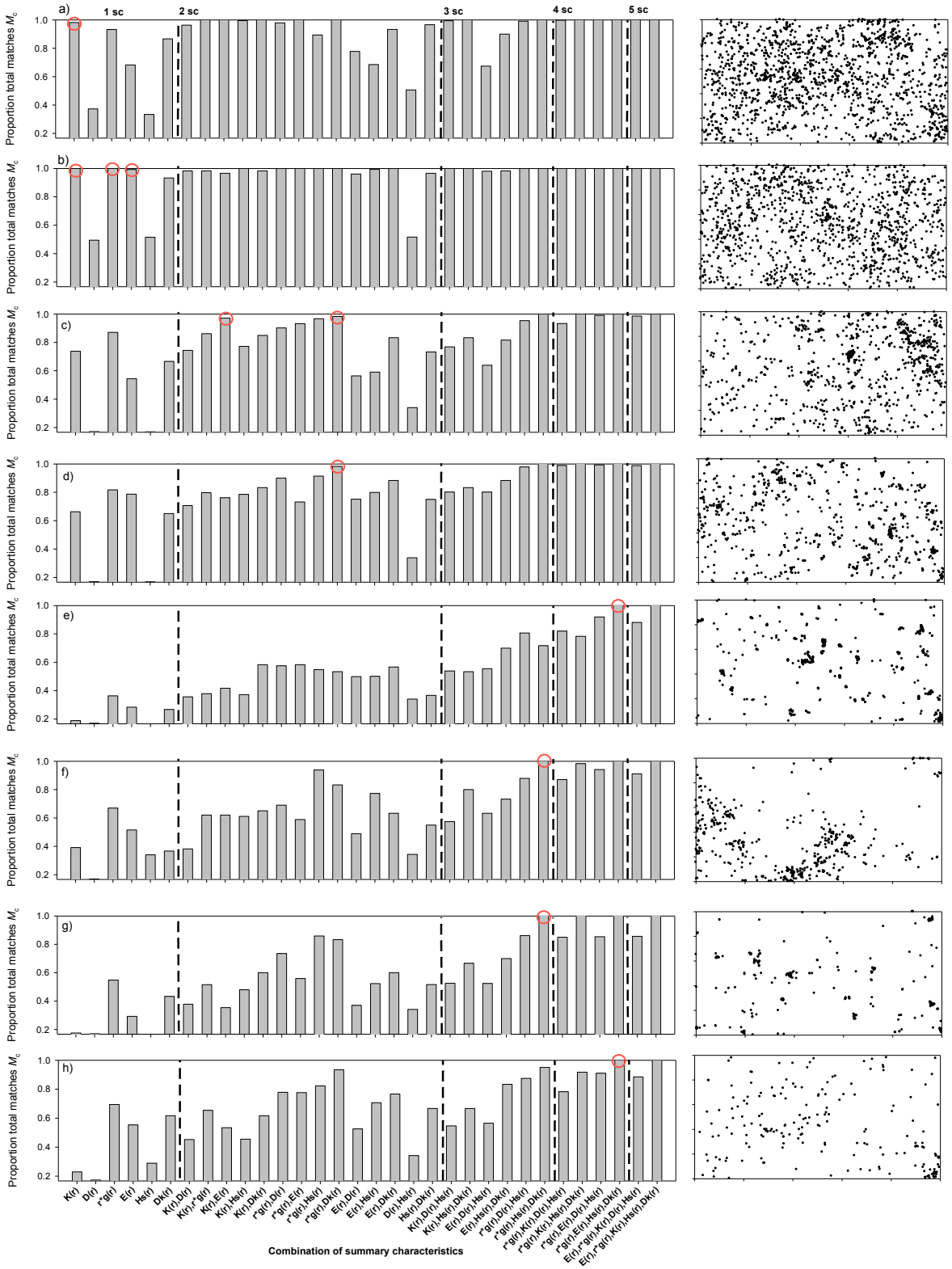

Figure A12. Ranking of combination of summary characteristics (as Figure 3), but individually for the eight observed patterns. Shown is the total pattern match M_c for each combination *c*. The best combinations are marked with a red circle. The panel letter coincides with that of Fig. A1

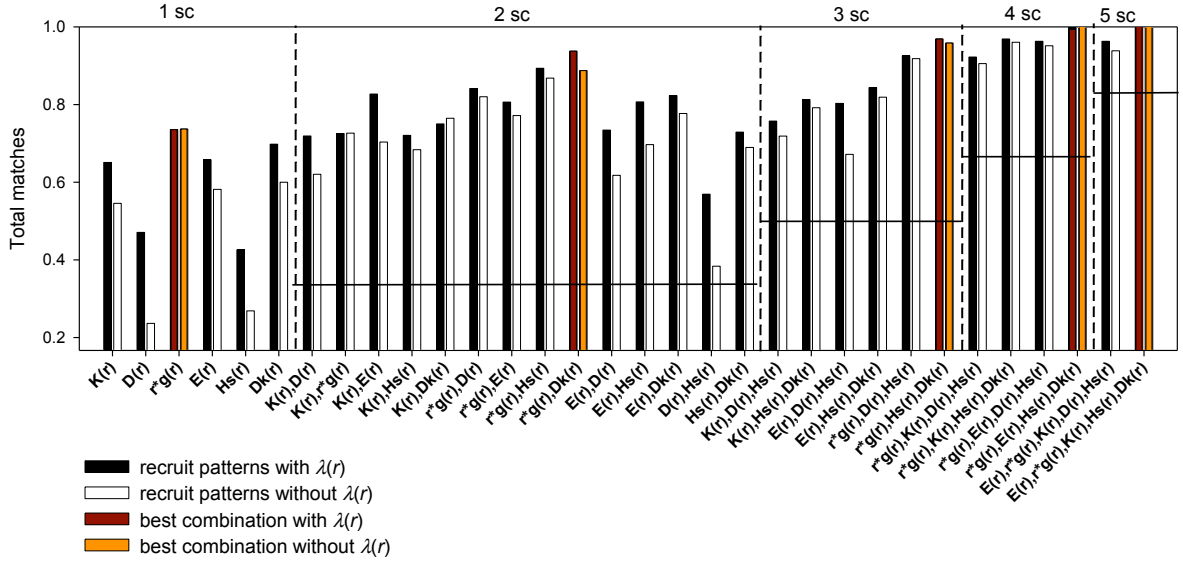

Figure A13. Same as Figure 3, but for simulation experiment 3 which uses the intensity function for pattern reconstruction.

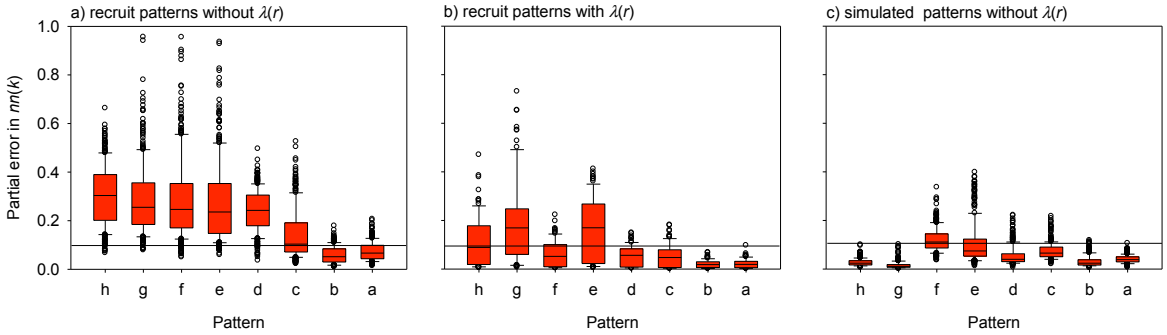

Figure A14. Boxplots of the partial errors of $nn(k)$ from simulation experiment 2 (panels a and b) and simulation experiment 3 (panel c) separately for the different patterns shown in figures A1 and A2. The letters on the x-axis correspond to the panel labels of figures A1 and A2 and the horizontal line indicates the acceptance threshold for the summary characteristic $nn(k)$.

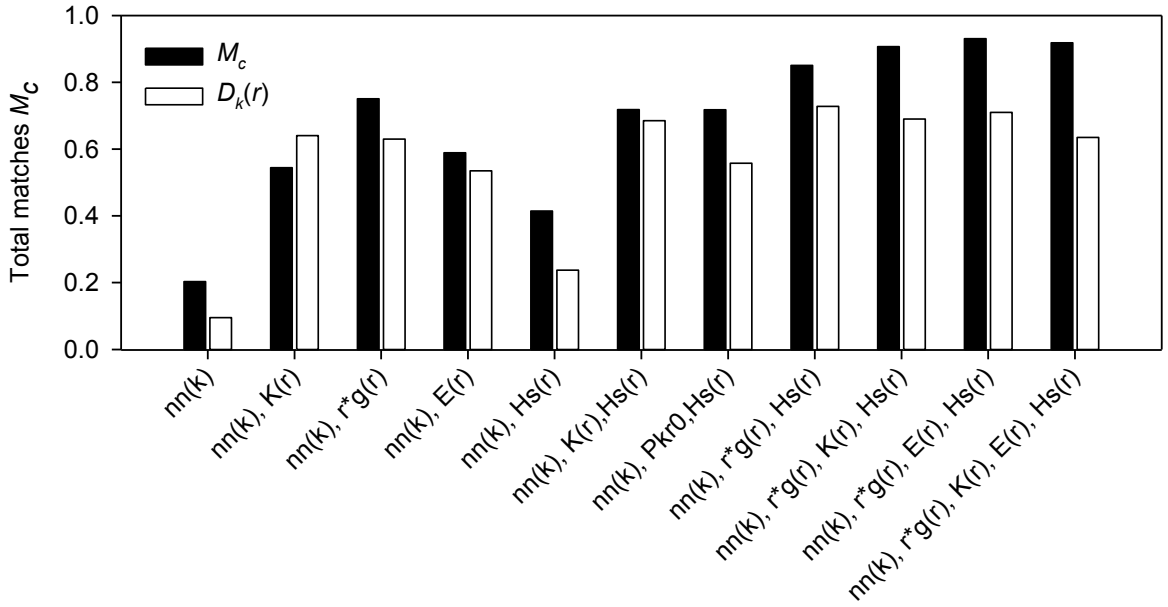

Figure A15. Results of simulation experiment 4. The index M_c that measures the information contained in a given combination *c* of summary characteristics for several combinations *c* that include the mean distance *nn*(*k*) to the kth neighbor (black) and the proportion of cases in which the distribution functions to the kth neighbor $D_k(r)$ were matched (white).