Spatial analyses of ecological count data: a density map comparison approach

Claire Lavigne, Benoit Ricci, Pierre Franck, Rachid R. Senoussi

Version définitive du manuscrit publié dans / Final version of the manuscript published in : Basic and Applied Ecology, 2010, In Press, DOI: 10.1016/j.baae.2010.08.011

Spatial analyses of ecological count data: a density map comparison approach

Claire Lavigne[a*], Benoît Ricci[a], Pierre Franck[a] and Rachid Senoussi[b]

[a] INRA, UR 1115, Plantes et Systèmes de culture Horticoles, F-84000 Avignon

[b] INRA, UR546 Biostatistique et Processus Spatiaux, INRA, F-84000 Avignon

Running title: Map comparison for spatial analysis

Number of words: 4562 (see letter to the editor)

* Corresponding author: Tel.: 33 (0)4 32 72 26 66; fax: 33 (0)4 32 72 24 32

E-mail address: claire.lavigne@avignon.inra.fr

Abstract

Analysing spatial patterns of population distributions may help to infer the decisive underlying ecological processes. Here we propose a method adapted to the spatial analysis of count data. Named MAPCOMP (MAP COMParison), it is based on the calculation of a formal distance, the Hellinger distance, between the density map of counts and the density map of sampling effort. Statistical tests of spatial homogeneity are based on count permutations across sampling sites and on valuable properties of the Hellinger distance. We assessed the efficiency of MAPCOMP by simulating different types and locations of clusters of individuals and compared its performance to the classical red-blue SADIE method, used as a reference. The two methods were also compared with respect to counts of codling moth larvae in orchards. Thanks to its better theoretical properties than SADIE, MAPCOMP was efficient in detecting spatial inhomogeneity when clusters were located on square or elongated spatial domains and more or less close to the edges, even for small sample sizes. It also appeared not very sensitive to edge effects. Another advantage of MAPCOMP is a bandwidth parameter that allows assessing the spatial extent of heterogeneity, if any.

Zusammenfassung

Die Analyse der räumlichen Muster in den Verteilungen von Populationen kann dazu beitragen, entscheidende grundlegende ökologische Prozesse abzuleiten. Hier stellen wir eine Methode zur räumlichen Analyse von Zähldaten mit dem Namen MAPCOMP (MAP COMParison) vor. Sie basiert auf der Berechnung der Hellinger-Distanz zwischen der Dichteverteilung der Beobachtungen und der Dichteverteilung der Untersuchungsintensität.

Die statistischen Tests auf räumliche Homogenität basieren auf Permutationen der Beobachtungszahlen über die Probestellen und auf nützlichen Eigenschaften der Hellinger-Distanz. Wir ermittelten die Effizienz von MAPCOMP, indem wir verschiedene Typen und Anordnungen der Cluster von Individuen simulierten und die Leistung von MAPCOMP mit der

der klassischen red-blue SADIE-Methode verglichen. Die beiden Methoden wurden auch auf Fänge von Apfelwicklerlarven in Obstplantagen angewendet.

Dank seiner, verglichen mit SADIE, besseren theoretischen Eigenschaften konnte MAPCOMP effektiv räumliche Inhomogenität selbst bei geringen Individuenzahlen aufdecken, wenn die Cluster auf quadratischen oder rechteckigen Gittern verteilt und mehr oder weniger nah am Rand positioniert waren. MAPCOMP erschien nicht sehr empfindlich gegenüber Randeffekten zu sein. Ein weiterer Vorteil von MAPCOMP ist ein Bandbreitenparameter, der es erlaubt, die räumlichen Ausdehnung der Heterogenität, so vorhanden, abzuschätzen.

Keywords: *Cydia pomonella,* clustering, Hellinger distance, heterogeneity, MAPCOMP, Monte Carlo permutations, spatial pattern, spatial statistics.

Introduction

Analysing spatial patterns of population distributions at various scales, from local patches to landscapes may help to infer the underlying ecological processes (McIntire & Fajardo 2009). Spatial statistics offer a number of tools for point pattern analysis (reviewed in Dale, Dixon, Fortin, Legendre, Myers et al. 2002; Perry, Liebhold, Rosenberg, Dungan, Miriti et al. 2002). A persistent issue is the statistical detection and characterization of spatial heterogeneity, such as gradients or clustering (Dale et al. 2002; Perry et al. 2002). Numerous methods exist for presence/absence data (e.g. Diggle, Gomez-Rubio, Brown, Chetwynd & Gooding 2007). We shall focus here on more general count data. When the form of the count spatial distribution is known a priori from knowledge of ecological processes (e.g. marked Poisson processes), parametric methods can be used to characterize spatial patterns. We consider more frequent situations where the form of the data distribution is not known. Methods then may follow two approaches. The first one is based on the counts of pairs of sampling points that exhibit similar count values at a given distance (e.g. variograms). This approach makes it possible to test a hypothesis of global clustering and determine some of its characteristics but does not allow the explicit mapping of heterogeneity. The second approach is based on the local departure of the spatial density of observations from an expected density. This approach makes it possible to map particular spatial patterns (*e.g.* geographic clusters of disease cases, Gay, Barnouin & Senoussi 2006). SADIE (Perry 1998; Perry, Winder, Holland & Alston 1999) is its most popular representative despite some drawbacks (edge effects: Xu & Madden 2005). We propose a new method (MAPCOMP) based on the second approach. We know of no other method that is adapted to counts, accounts for heterogeneous sampling effort, has well defined and stable statistical properties and allows mapping of heterogeneity. Note, furthermore, that MAPCOMP would also apply to continuous positive data on continuous or discrete spatial supports.

We first present a brief modelling framework. Using simulated data, we then compare our method to the red-blue analysis of SADIE, used here as a reference. Cluster detection methods

are known to be sensitive to the forms of clusters and the sampling area, as well as to edge effect (Yamada 2003). We thus focus on these issues. Finally, we analyze the spatial distribution of codling moth larvae in eight orchards using both methods and we compare these results.

Material and Method

Statistical methods

Fig. 1 describes the MAPCOMP method. Symbols used in this study are summarized in Table 1.

General statistical framework

The red-blue SADIE and the MAPCOMP methods are based on permutation tests. Permutation test methods consist in random permutations of indices s of an ordered data set $\mathbf{X} = (X_s, s \in S)$, *e.g.* of counts observed at sites s. The index set S is structured via neighbourhood relations in a geographic domain D. To test a specific null hypothesis H_0 (e.g. spatial homogeneity)**,** one has to define an adequate test statistic, T**(X),** and a specific subset, Ω, of all data permutations. The choice of Ω should reflect H_0. In the following, the set of all permutations of indices s was chosen as in SADIE, as we wished to keep the same H_0 hypothesis. By its very definition, T**(X)** is designed to statistically behave differently under H_0 and the alternative hypothesis H_1. The test relies on the comparison between the single statistic value calculated using the observed data T_{obs} and the statistic values calculated using a large number of independently sampled data permutations.

*Test statistic T(***X***)*

The four SADIE statistics are based on the minimal sum of distances that individuals have to move from site to site to reach a distribution such that either (1) the number of individuals (possibly fractional) is exactly the same over all sampling sites (statistics I_A, $\overline{v}_i$, $\overline{v}_j$; associated probabilities P_A, $P\overline{v}_i$, $P\overline{v}_j$,), or (2) all individuals are situated on a single sampling site (statistic J_A, probability Q_A) (Perry 1998, Perry et al. 1999).

The suggested statistical test measures a formal (not spatial) distance between the density map of counts and the density map of sampling effort. It increases with increasing inhomogeneity of count distribution over sample sites. We thus needed (1) an estimation of density maps for sample sites and observations and (2) a statistical test to assess the closeness of spatial densities.

Comparing spatial densities

We relied on the Hellinger distance between two probability densities *p(s)* and *q(s)* over a domain *D* (Gibbs & Su 2002). It is defined as follows:

$$H(p,q) = \left(\frac{1}{2} \int_D \left(\sqrt{p(s)} - \sqrt{q(s)} \right)^2 ds \right)^{1/2} \qquad s = (x,y) \in D \qquad \text{eq. 1}$$

Estimation of density maps

Let $S = \{s_i=(x_i,y_i) \quad i=1,...,\#S\}$ be the set of the geographic coordinates of #*S* sites (Fig.1).

Considering the data **X** as a realization of a random measure whose theoretical normalized intensity is denoted *p(s)***,** we could estimate *p* at any location *s* via a local smoothing using a probability density function, named kernel, *K,* that provides for every location *s* the weights of all observations (Scott 1992):

$$\hat{p}_h(s) = \sum_{s_i \in S} K_h(s,s_i) X_{s_i} \,/ \sum_{s_i \in S} X_{s_i} \qquad \text{eq. 2}$$

where, the function $K_h(s, s_i)$ indexed by the positive parameter *h* (bandwidth parameter) is usually the *h*-scaled and renormalized kernel *K* on the plane as follows:

$$K_h(s,s_i) = K((s-s_i)/h)/h^2. \qquad \text{eq. 3}$$

However, to account for the fact that bordering points intrinsically suffer from a lack of neighbouring observation sites, we introduced the following edge correction:

$$K_h(s,s_i) = K\left(\frac{s-s_i}{h}\right) \Big/ \int_D K\left(\frac{s-s_i}{h}\right) ds \ ,$$

where *D* is the study domain.

We chose the following square supported kernel: for **s**=*(x, y)*,

$$K(s) = C_1 \exp\left(-\left(\frac{1}{1-x^2}+\frac{1}{1-y^2}\right)\right) \text{ if } -1 \le x, y \le 1 \text{ and 0 elsewhere} \quad \text{eq. 4}$$

and C_1 is such that $\int_{-1}^{+1}\int_{-1}^{+1} K(s)\, ds = 1$, *i.e.* $C_1 \simeq 20.3$.

The estimated density $\hat{p}_h(s)$ is thus a local average of counts of neighbouring sites within a $2h$x$2h$ square (Appendix A: Fig. 1).

The normalized sampling intensity *q(s)* is calculated similarly replacing **X** by a vector of sampling effort. In the case of a uniform distribution of sampling sites, one could simply take $\hat{q}_h(g_{kl}) = 1/\#S$.

The bandwidth h: a focusing tool

In functional estimation, the bandwidth h is usually optimized to balance the global bias and variance of $\hat{p}_h(\cdot)$ yielding either $h_{opt} = C(\#S)^{-1/6}$ where C is a constant that can be chosen a priori (depending on domain area) or by a cross-validation method (Hardle, 1989). Here we used h differently, giving it the role of a scale parameter to investigate the range at which sites sharing similar values clustered together. The tested h values should be such that the minimum value is larger than the distance between neighbouring sampling points (otherwise no smoothing occurs) and in the range of values at which dependencies among observations are expected. Exploring a range of values for h may reveal specific patterns at different spatial scales.

Computation procedure

Statistics and maps were computed using R 2.1.1 (R Development Core Team 2005). The estimated densities were approximated over a regular grid **G** of mesh δ with nodes (g_{kl}) covering the study domain D (*e.g.* δ can be chosen as half the minimal distance between sampling sites). The density of counts, p**,** was given values $\hat{p}_h(g_{kl})$ whereas q estimating the sampling density

was given values $\hat{q}_h(g_{kl})$ defined the same way as $\hat{p}_h(g_{kl})$ except that all X_{s_i} were set to one. The test statistic measuring the distance to homogeneity was thus:

$$T_h(X) = \frac{\delta}{\sqrt{2}} \left(\sum_{g_{kl} \in G \cap D} \left(\sqrt{\hat{p}_h(g_{kl})} - \sqrt{\hat{q}_h(g_{kl})} \right)^2 \right)^{1/2} \quad \text{eq. 5}$$

Data sets

Simulated data

Sampling domains *D* were represented by square or rectangular grids with approximately the same size (*i.e.* either 200 x 200 or 284 x 142). Sampling sites were chosen on the domains setting the first site at coordinates (10, 10). Then coordinates of the sampling sites were incremented by value 20 ((10, 30), (30, 10), (30, 30), etc.). With this procedure, the total of sampling sites amounted to 100 for square domains and 98 for rectangular domains.

We then considered spatial distributions of either *N*=20 or 100 sampled individuals, corresponding to species with moderate and high population abundances. To study the effect of the domain shape and the count distribution over this domain, individuals were randomly distributed over sampling sites following five configurations: (1) uniform distribution in both the square and the rectangular domain; (2 & 3) two Gaussian clusters located along the central horizontal axis of the square domain either close to each other at sites (100, 75) and (100,125) or farther away at sites (100, 50) and (100, 150); (4) two Gaussian clusters on one side of the square domain, *i.e.* at sites (100, 25) and (100, 75); and (5) the same two clusters as in (3) but in the middle of the rectangular domain, *i.e.* at sites (70,90) and (70,190) (Appendix A: Fig.2).

For Gaussian clusters, the value of standard deviation σ, hereafter denoted cluster radius, was taken in set {12, 25, 35, 70}. For σ=12 all sampled individuals were in sampling sites neighbouring the centre of the cluster, hereafter denoted cluster focus, whereas for σ=70 some individuals could be distributed along the edge of domains (Fig. 2).

In total, we thus considered 32 clustered patterns (4 configurations x 2 *N* values x 4 σ values).

We then simulated ten replicate distributions for each pattern, resulting in a total of 320 distributions.

Codling moth sampling

We sampled codling moth diapausing larvae in eight apple orchards located in south-eastern France (WGS84 : 43°46'27"N to 43°51'23"N and 4°51'12"E to 4°57'34"E). The mean area of orchards was 0.40 $\pm$0.07 ha with values varying from 0.13 to 0.76 ha. Larvae were caught on 10-cm wide corrugated cardboard strip traps wrapped around tree trunks in July 2008 and collected the following October. Approximately 30 traps were distributed on a regular grid over each orchard.

Testing homogeneity of spatial patterns

We tested the homogeneity of spatial distributions of individuals over sampling points with the help of both the SADIE method (free software SADIEShell 1.22) and of MAPCOMP using R 2.1.1 (R Development Core Team 2005). When using SADIE on simulated data, we used the maximum allowed number of Monte-Carlo simulations for significance test (*k5psimul*=153, *i.e.* 5967 randomisations). We recorded the *P*-value of each of the three permutation tests associated with the three statistics I_A, $\overline{v}_i$, $\overline{v}_j$ (i.e. P_A, $P\overline{v}_i$, $P\overline{v}_j$) for each of the 320 spatial distributions, as Perry (1998) does not recommend considering the fourth statistic, J_A, in case of more than one cluster.

To use the MAPCOMP method, grids with a 2x2 mesh size over domains were chosen to calculate density maps. Then, for each bandwidth *h* taken in {12, 25, 50, 90}, the density maps of both sampling sites and count distributions were calculated. The smallest value *h*=12 corresponds to a smoothing over only the nearest neighbours and the largest value, *h*=90, to a smoothing over about half the sampling sites. For each *h* value the Hellinger distance was calculated between the two density maps and the homogeneity hypothesis was tested using 10,000 permutations of

simulated counts over sampling sites. We recorded the lowest *P*-value among those obtained with the four *h* values, i.e. the result corresponding to the most appropriate *h* value, h_{Sig}.. These *P*-values were used for comparison with SADIE.

We also tested the homogeneity of the spatial pattern in codling moth data using both SADIE and MAPCOMP. For the latter method, a 2 m x 2 m approximating grid was used and five *h* values ranging from a value encompassing only the nearest neighbouring traps to a value encompassing about $1/4^{th}$ of the number of traps per orchard were chosen. The *P*-values for all *h* values in each orchard were then recorded.

Results

Power comparison of the two tests

As expected, the two methods provided non-significant results when individuals were randomly uniformly distributed over sites, whatever the shape of the domain or the sample size (not shown), meaning that neither method detected false positives. As results are generally very close in terms of P_A, $P\overline{v}_i$, $P\overline{v}_j$ in SADIE, only results obtained with P_A are used for comparison.

Impact of the sample size on detection of spatial heterogeneity

On a square plot, both methods detected the two central clusters efficiently, whether they were separated by 50 (configuration 2) or 100 (configuration 3) distance units when samples were large (*N*=100) except the clusters with the smallest radius (σ=12) that were not detected with SADIE (Figs 3A, 3C). For both methods, detection was also marginally worse in configuration 3 than in configuration 2 when σ=70.

Decreasing the number of observed individuals to *N*=20 resulted in less detection of clustering by both methods (Figs 3A, 3C). However, because MAPCOMP provided significant values in 64/80 simulated patterns and only poorly detected clusters of the largest radius (σ=70), its detection capacity can be deemed good. In contrast, SADIE provided significant results only in 7/80

patterns that corresponded to clusters with radius σ=25 and σ=35.

Obviously, not all *h* values were equally efficient with the MAPCOMP method whatever *N*. h_{Sig}, *i.e.* the *h* value providing more significant results for the ten replicates, tended to increase with increasing radius width σ (not shown).

Impact of the domain shape on detection of spatial heterogeneity

The elongated shape of the sampling domain (configuration 5) slightly modified the behaviour of MAPCOMP (Figures 3C vs 3D): results were not modified for *N*=100, but for *N*=20, MAPCOMP somewhat less efficiently detected clusters with either small or large radii (from 30/40 detections in square domains to 26/40). In contrast, the detection ability of SADIE largely decreased for *N*=100 (from 26/40 to 4/40) and remained close to 0 for *N*=20.

Impact of cluster distance to border

Setting the foci of the clusters closer to a domain edge (configuration 4) had a marginal effect on inhomogeneity detection with MAPCOMP (Figure 3A vs 3B). Thanks to the border effect correction, all clustered patterns were still detected with certainty for *N*=100. Detection probability only increased for σ=70 and *N*=20 as compared with the reference situation of centred clusters (configuration 2). On the contrary, the SADIE method was very sensitive to border effects: in this case, the method detected inhomogeneity for σ=12 and *N*=100, and the probability of detecting a cluster rose from 6/40 to 37/40 for *N*=20.

Codling moth data

Clustering was detected by MAPCOMP in 3 out of the 8 tested orchards (Table 2). These orchards were also pinpointed using SADIE, testing for regularity (for two orchards) and crowding (for the third one, orchard #G). We used this latter test as data suggested the presence of a single cluster. The other five orchards showed no significant departure from homogeneity

using both methods. Interestingly, in the three heterogeneous orchards, the density of codling moth larvae was higher at the edges (Figure 4).

Discussion

We have presented a new method, MAPCOMP, for detecting the inhomogeneity of spatial patterns in count data. We compared it to SADIE, a reference method for ecologists to detect the spatial heterogeneity in count data (*e.g.* Thomas, Parkinson, Griffiths, Garcia & Marshall 2001; Schellhorn, Bellati, Paull & Maratos 2008). Our results confirm that SADIE is efficient in many standard situations (*e.g.* regularly shaped domains, large sample size). However, MAPCOMP is as efficient as SADIE for large sample sizes and wide enough cluster patterns. Moreover, it better detected clustered patterns for small sample sizes and clusters with a small radius. MAPCOMP also appeared less sensitive than SADIE to the shape of the sampled domain and to edge effects.

These differences could be explained by the different nature of the test statistic used. In SADIE, the test statistic measures the total distance that individuals have to move to reach a regular distribution of individuals. Consequently, the SADIE method has the drawback that the value of the statistic associated with observations strongly depends on large distances between sampling points. For instance, its value for a cluster at the short side of an elongated domain will be much larger than the value assessed for this same cluster located in the central part of this same domain. The associated observed *P*-value will thus be smaller since the distribution of the distance under H_0 is the same in both cases. This sensitivity of SADIE to localization of clusters has already been pointed out (Xu & Madden 2005). This explains the difficulty of SADIE to detect small clusters in elongated areas. However, it must not be forgotten that, in case of a single cluster, the SADIE method proposes another test for detection of clustering based on the minimum distance that individuals would have to move to be all clustered on a single sampling site. This test should be very efficient at detecting clusters of small radii. Conversely, the better efficacy of MAPCOMP for detecting heterogeneity in small samples is probably due to the use of a

normalized version of smoothing technique for densities. Its better efficacy with respect to edge effects and shape of the sampled area is certainly due to the insensitivity of the Hellinger distance to size and shape of the observation domain.

Further, MAPCOMP, as SADIE, not only detects spatial heterogeneity but also provides maps of lower and higher than expected individual densities, i.e. patches and gaps, based on the difference $\hat{p}(g_{kl}) - \hat{q}(g_{kl})$. These maps help to gain insight into the factors underlying the perceived inhomogeneity.

Finally, from a completely different perspective, SADIE provides an association statistical test (Perry & Dixon, 2002) that allows the comparison of the spatial distributions of a single species at different dates (*e.g.* Blackshaw & Vernon 2006) or of two species at the same date (*e.g.* Pearce & Zalucki 2006). As MAPCOMP is based on a formal distance between spatial densities, it could readily be used in these contexts by replacing the sampling density with the density of the second species or previous observations of the species under study (for a first application, see Debras, Senoussi, Rieux, Buisson & Dutoit 2008).

From a more general point of view, MAPCOMP has further desirable properties for the statistical analysis of count data because of its use of the Hellinger distance: (1) $H(p,q)$ is a true metric, that is non-negative, null only if $p=q$, symmetric in $p(s)$ and $q(s)$ (so that there is no need to define a reference population) and satisfies the triangular inequality $H(p_1,p_2) \leq H(p_1,q) + H(p_2,q)$, allowing for instance for the comparison (and partial ordering) between different species distributions. (2) H has bounded values in [0,1], it thus avoids formally large and asymmetric values and can be considered as well scaled for value comparisons. (3) H behaves smoothly with respect to data aggregation. For example, if several distinct study regions D_j with respective densities p_j and q_j were aggregated together and given weights $\alpha_j \geq 0, \sum_j \alpha_j = 1$ (*e.g.* $\alpha_j = area(D_j) / \sum_k area(D_k)$), then the squared Hellinger distance of the resultant densities would simply equal the mean of squared Hellinger distances over the areas, *i.e.*

$H^2\left(\sum_j \alpha_j p_j, \sum_j \alpha_j q_j\right) = \sum_j \alpha_j H^2(p_j, p_j)$. (4) H has a robust and non-erratic behaviour when adding new observations in the given domain (temporal sampling or species grouping) (Beran, 1977): adding new data within a given domain can only lead to a distance decrease that will converge to the 'true' H distance between densities as the number of data points increases. For example, the grouping of distinct subpopulations p_j which are given weights $\alpha_j \geq 0, \sum_j \alpha_j = 1$ (*e.g.* α_j = proportion of biomass of population j) observed within a given domain gives $H^2\left(\sum_j \alpha_j p_j, q\right) \leq \sum_j \alpha_j H^2(p_j, q)$. Another advantage of MAPCOMP is the fact that continuous variables can be used instead of count data without the model undergoing any modifications. These variables could be either abiotic factors (*e.g.* temperature, altitude, etc.) or community or population parameters (*e.g.* species diversity, biomass, growth, etc.). Finally, the last but not the least asset rests on the possible use of the bandwidth parameter as a scale analysis tool, specifically to test the spatial extent of heterogeneity, if any.

A very large number of statistical methods are available to ecologists for the detection of spatial inhomogeneity. Using the Dale et al. (2002) classification for these methods indicates that MAPCOMP is not redundant with any of those taken into consideration in their study. Being based on continuous density functions and varying bandwidths, MAPCOMP bears relationships to the wavelets methods, but is not based on many fitting parameter estimation. It also shares similarities with the block and quadrat variance methods as it attempts to detect larger than expected variance in adjacent data using windows of varying sizes, but it is not based on a direct measure of local variances.

While the comparison of methods on simulated data tends to favour the MAPCOMP method, it should be noted that MAPCOMP and SADIE behaved rather similarly in the case of our field study dataset. However, the cluster in orchard *G* was detected by SADIE only when using the crowding test which is not the most recommended one (Perry et al. 1999). Interestingly, the

detection of spatial heterogeneity did not depend on sample size or orchard area. From the ecological point of view, the heterogeneity of codling moth distribution in all three cases was due to higher than expected numbers of larvae on orchard edges not bordered by a hedgerow or another orchard (Fig. 4).

To conclude, both methods proved to be efficient at detecting spatial heterogeneity for large sample sizes with regular domains. Moreover, MAPCOMP efficiently detects inhomogeneity on small sample sizes and less regular domains. This could be particularly useful in the case of conservation biology of rare species as well as in the case of agricultural pests where population densities are expected to be low and population habitats may be geometrically intricate.

Acknowledgments

We are grateful to Jean-François Toubon for contacts with farmers, to members of EPI for counting codling moth larvae, to farmers for access to their orchards and to regional experimental station La Pugère. B.R. was funded by INRA and PACA district. This research was partly funded by the ECOGER project “Ecco des vergers”. Tanis Plant and D.G. Biron edited a draft of this paper in English.

Appendix A: Supplementary Material

The online version of this article contains additional supplementary data. Please visit XXXX.

References

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *The annals of Statistics, 5*, 445-463.

Blackshaw, R.P., & Vernon, R.S. (2006). Spatiotemporal stability of two beetle populations in non-farmed habitats in an agricultural landscape. *Journal of Applied Ecology, 43*, 680-689.

Dale, M.R.T., Dixon, P., Fortin, M.J., Legendre, P., Myers, D.E., & Rosenberg, M.S. (2002). Conceptual and mathematical relationships among methods for spatial analysis. *Ecography, 25*, 558-577.

Debras, J.F., Senoussi, R., Rieux, R., Buisson, E., & Dutoit, T. (2008). Spatial distribution of an arthropod community in a pear orchard (southern France) - Identification of a hedge effect. *Agriculture Ecosystems & Environment, 127*, 166-176.

Diggle, P.J., Gomez-Rubio, V., Brown, P.E., Chetwynd, A.G., & Gooding, S. (2007). Second-order analysis of inhomogeneous spatial point processes using case-control data. *Biometrics*, 63, 550-557.

Gay, E., Barnouin, J., & Senoussi, R. (2006). Spatial and Temporal Patterns of Herd Somatic Cell Score in France. *Journal Dairy Science, 89*, 2487-2498.

Gibbs, A.L., & Su., F.E. (2002). On choosing and bounding probability metrics. *International Statistical Review, 70*, 419-435.

Härdle, W. (1989) *Applied non parametric regression*, Cambridge University Press.

McIntire, E.J.B., & Fajardo, A. (2009). Beyond description: the active and effective way to infer processes from spatial patterns. *Ecology,90*,46-56.

Pearce, S., & Zalucki, M.P. (2006). Do predators aggregate in response to pest density in agroecosystems? Assessing within-field spatial patterns. *Journal of Applied Ecology, 43*, 128-140.

Perry, J.N. (1998). Measures of spatial pattern for counts. *Ecology, 79*, 1008-1017.

Perry, J.N., & Dixon, P.M. (2002). A new method to measure spatial association for ecological

count data. *Ecoscience, 9*, 133-141.

Perry, J.N., Liebhold, A.M., Rosenberg, M.S., Dungan, J., Miriti, M., Jakomulska, A., & Citron-Pousty, S. (2002). Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. *Ecography, 25*, 578-600.

Perry, J.N., Winder, L., Holland, J.M., & Alston, R.D. (1999). Red-blue plots for detecting clusters in count data. *Ecology Letters, 2*, 106-113.

Schellhorn, N.A., Bellatib, J., Paullb, C.A., & Maratos, L. (2008). Parasitoid and moth movement from refuge to crop. *Basic and Applied Ecology, 9*, 691-700.

Scott, D.W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization.* New York: Wiley.

Thomas, C.F.G., Parkinson, L., Griffiths, G.J.K., Garcia, A.F., & Marshall, E.J.P. (2001). Aggregation and temporal stability of carabid beetle distributions in field and hedgerow habitats. *Journal of Applied Ecology, 38*, 100-116.

Xu, X.M., & Madden, L.V. (2003). Considerations for the use of SADIE statistics to quantify spatial patterns. *Ecography, 26*, 821-830.

Yamada, I., & Rogerson, P. (2003). An Empirical Comparison of Edge Effect Correction Methods Applied to K -function Analysis. *Geographical Analysis, 35*, 97-109.

Table 1: Symbols used in this study.

Symbol	Meaning
$\mathbf{X} = (X_s, s \in S)$	A set of ordered observations (Data)
X_s	Observation at site *s*
S	Set of observation sites
$\#S$	Number of sites in *S*
$s= (x,y)$	Observation site s identified by its coordinates *(x,y)*
D	Geographical domain over which observations are performed
Ω	A subset of permutations of sites used for testing the null hypothesis H_0 among all possible permutations
T(**X**)	The test statistic calculated from the ordered data set **X**
T_{obs}	The test statistic calculated from the actual observations
I_A, $\overline{v}_i$, $\overline{v}_j$, J_A	The four test statistics from SADIE
P_A, $P\overline{v}_i$, $P\overline{v}_j$,Q_A	The four P-values associated to the four above statistics in SADIE
p	Probability density function of observations
q	Probability density function of sampling effort
$H(p,q)$	Hellinger distance between *p* and *q* (integral over domain *D)*
K	The basic kernel smoothing function *(* any *p*robability density function)
h	Bandwidth parameter for smoothing
K_h	The *h*-scaled and renormalized smoothing kernel *K* (K_1=*K)*
G	Grid over *D* used to compute the integral *H*
δ	Mesh size of **G**
g_{kl}	Node of coordinates (k,l) on **G**
$\hat{p}_h(g_{kl})$	Estimation of *p* using K_h at node g_{kl}
$\hat{q}_h(g_{kl})$	Estimation of *q* using K_h at node g_{kl}
$T_h(X)$	Computed test statistic in MAPCOMP for banwidth *h* and data set X
N	Number of sampled individuals in simulated data sets
σ	Radius of simulated clusters

Table 2: Characteristics of sampled orchards and tests of codling moth spatial homogeneity.

h_i: ith bandwidth parameter used for test and *P*: P-value using MAPCOMP; $P_A, P\overline{v}_i, P\overline{v}_j$, Q_A: P-values for tests in SADIE. Orchards exhibiting significant heterogeneity are in bold.

				MAPCOMP					SADIE			
Orchard Id.	Number of traps	Area (ha)	Codling moths/trap mean ±se	h_1 P	h_2 P	h_3 P	h_4 P	h_5 P	P_A	$P\overline{v}_i$	$P\overline{v}_j$	Q_A
A	32	0.32	0.59 ±0.20	*11* 0.94	*13* 0.95	*15* 0.96	*18* 0.96	*20* 0.95	0.58	0.61	0.54	0.77
B	33	0.76	12.48±1.64	*19* 0.32	*21* 0.28	*23* 0.27	*25* 0.25	*28* 0.20	0.21	0.21	0.39	0.36
C	34	0.13	4.30±0.58	*7* 0.76	*9* 0.62	*11* 0.61	*13* 0.53	*15* 0.46	0.65	0.83	0.83	0.93
D	32	0.19	1.06±0.18	*8* 0.29	*11* 0.14	*14* 0.18	*17* 0.37	*20* 0.57	0.53	0.50	0.56	0.11
E	31	0.45	9.35±1.15	*16* 0.95	*20* 0.93	*24* 0.92	*28* 0.92	*32* 0.92	0.97	0.945	0.905	0.43
F	**30**	**0.33**	**5.56±1.11**	***12*** **8.10^{-4}**	***15*** **4.10^{-4}**	***18*** **4.10^{-4}**	***21*** **$<10^{-4}$**	***23*** **$<10^{-4}$**	**2.10^{-4}**	**0.003**	**5.10^{-4}**	0.29
G	**33**	**0.70**	**2.06±0.42**	***19*** **0.012**	***21*** **0.008**	***24*** **0.009**	***26*** **0.008**	***28*** **0.009**	0.105	0.159	0.115	**0.021**
H	**30**	**0.27**	**0.87±0.25**	***10*** **0.015**	***12*** **0.009**	***13*** **0.007**	***14*** **0.006**	***15*** **0.008**	**0.016**	**0.019**	**0.048**	**0.005**

Figure legends:

Fig. 1: Schematic representation of the MAPCOMP method to analyze spatial patterns of population distributions for a given species. The density map of observed counts p^h_{obs} is derived from the smoothing of the observed counts using kernel K_h on grid G. The density map of sampling q_h is derived from locations of sampling sites using the same kernel and grid. The test statistic $T_{obs}(X_s)$ based on the Hellinger distance between these two maps is compared to each of those obtained between the same sampling density map q_h and each map calculated from permutated counts $T(X_{permut})$.

Fig. 2: Examples of simulated cluster: A) N=100, σ=25, configuration 2, B) N=100, σ=25, configuration 3:, C) N=100, σ=25, configuration 4, D) N=100, σ=25, configuration 2 (See Appendix A: Fig. 2 for other examples).

Fig. 3: Detection probability of clustered patterns in simulated data, i.e. number of cases (out of 10) where the P-value is below 0.05 as a function of σ, the radius of the cluster. White bars: MAPCOMP, grey bars: SADIE.

Fig. 4: Density maps of the three orchards on which we detected clustering of codling moths larvae. h: bandwidth used for MAPCOMP, dots= traps, thick lines=hedgerows or neighbouring orchard or forest edge.

Figure 1

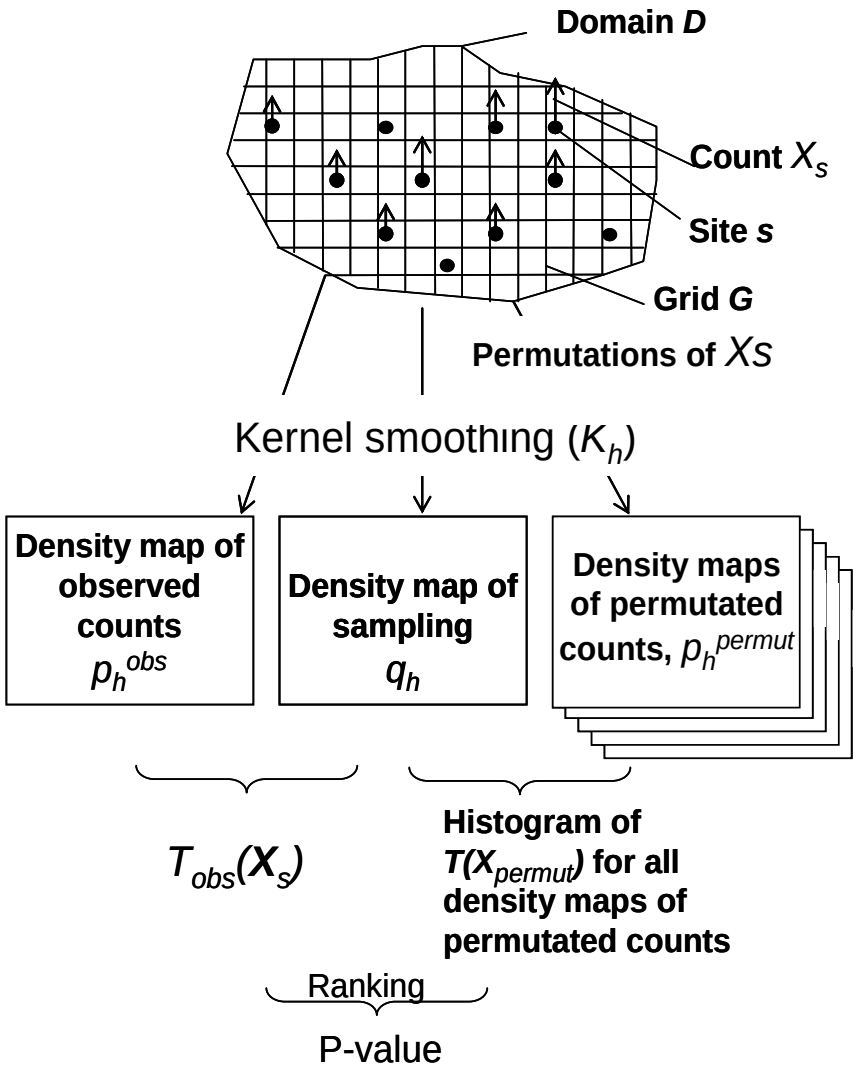
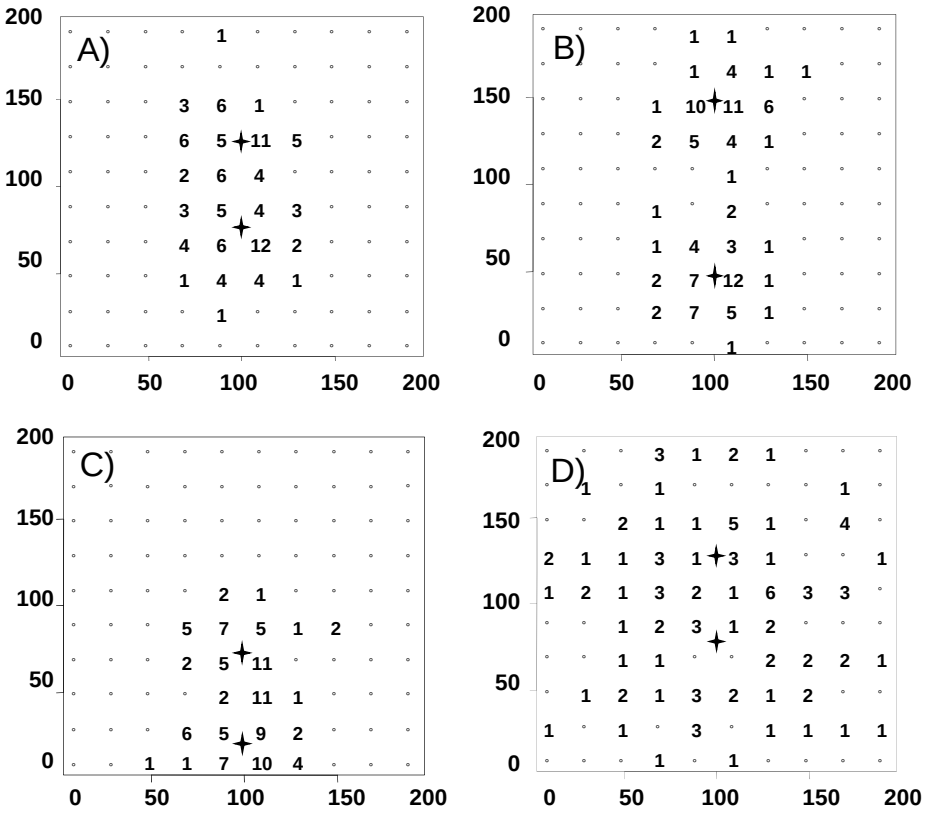


Figure 2



1 **Figure 3**

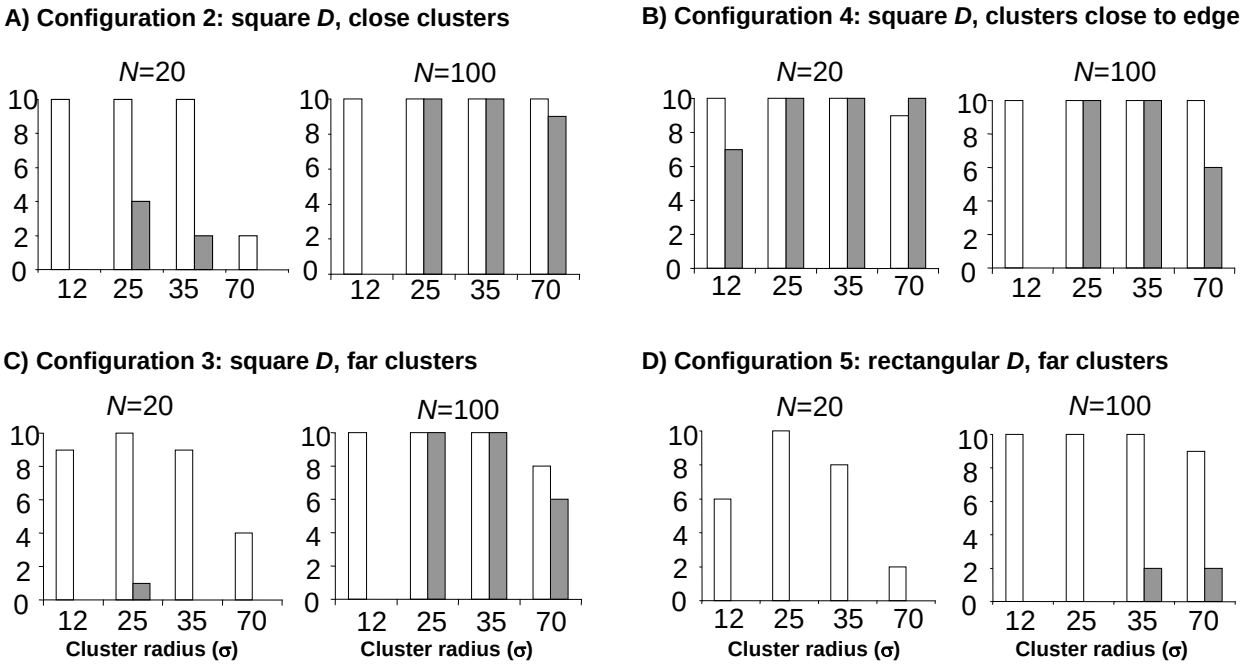
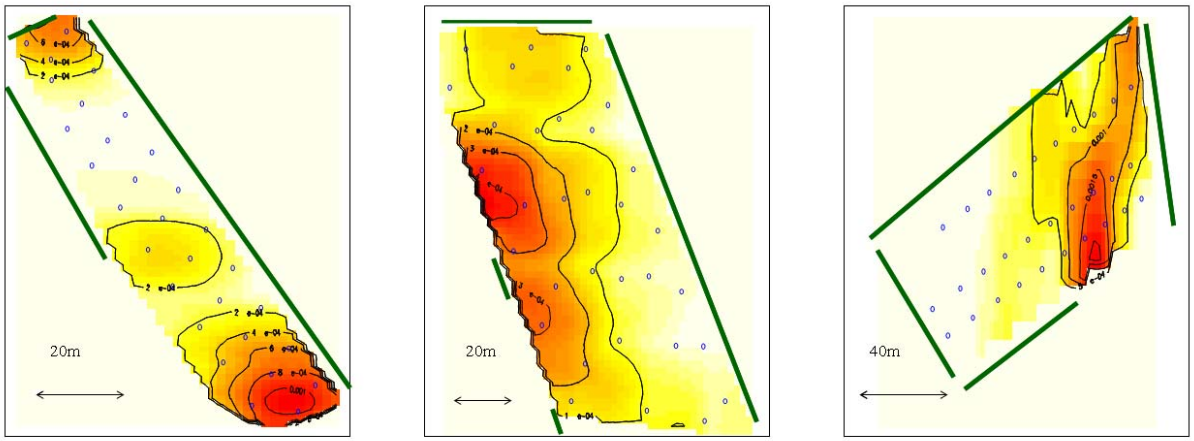


Figure 4



Orchard F, *h*=18 Orchard G, *h*=24 Orchard H, *h*=14

Appendix A:

Figure 1: 2D representation of the kernel used for smoothing: ─ h=1,····· h=2, – – h=0.5.

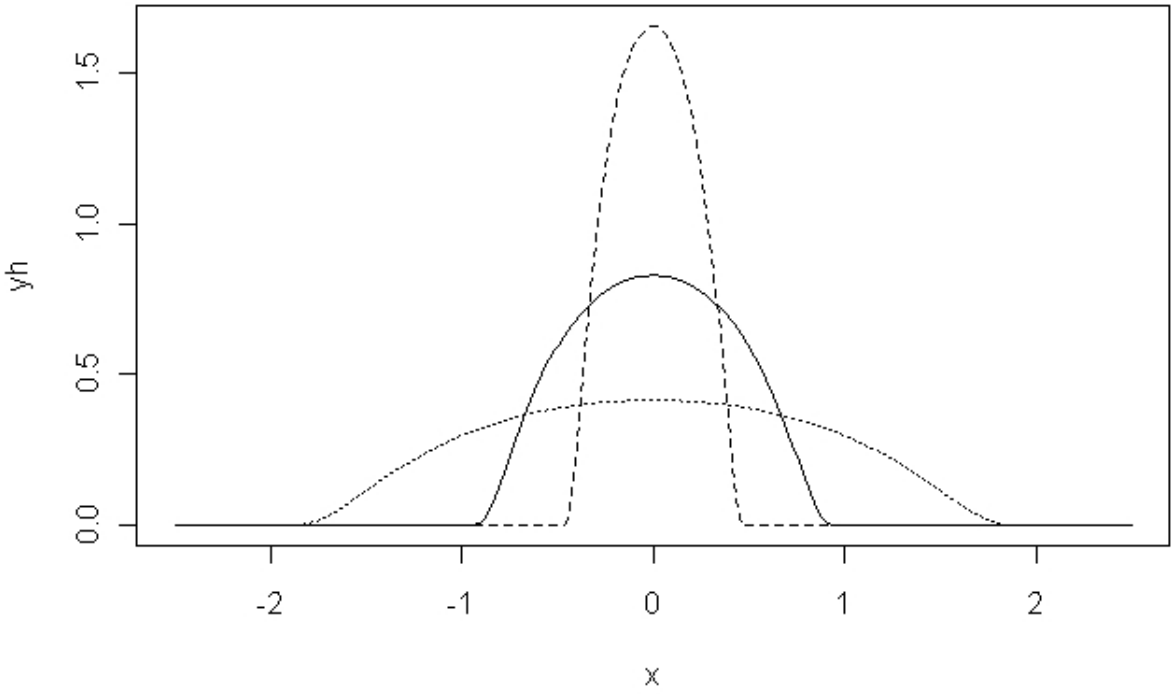
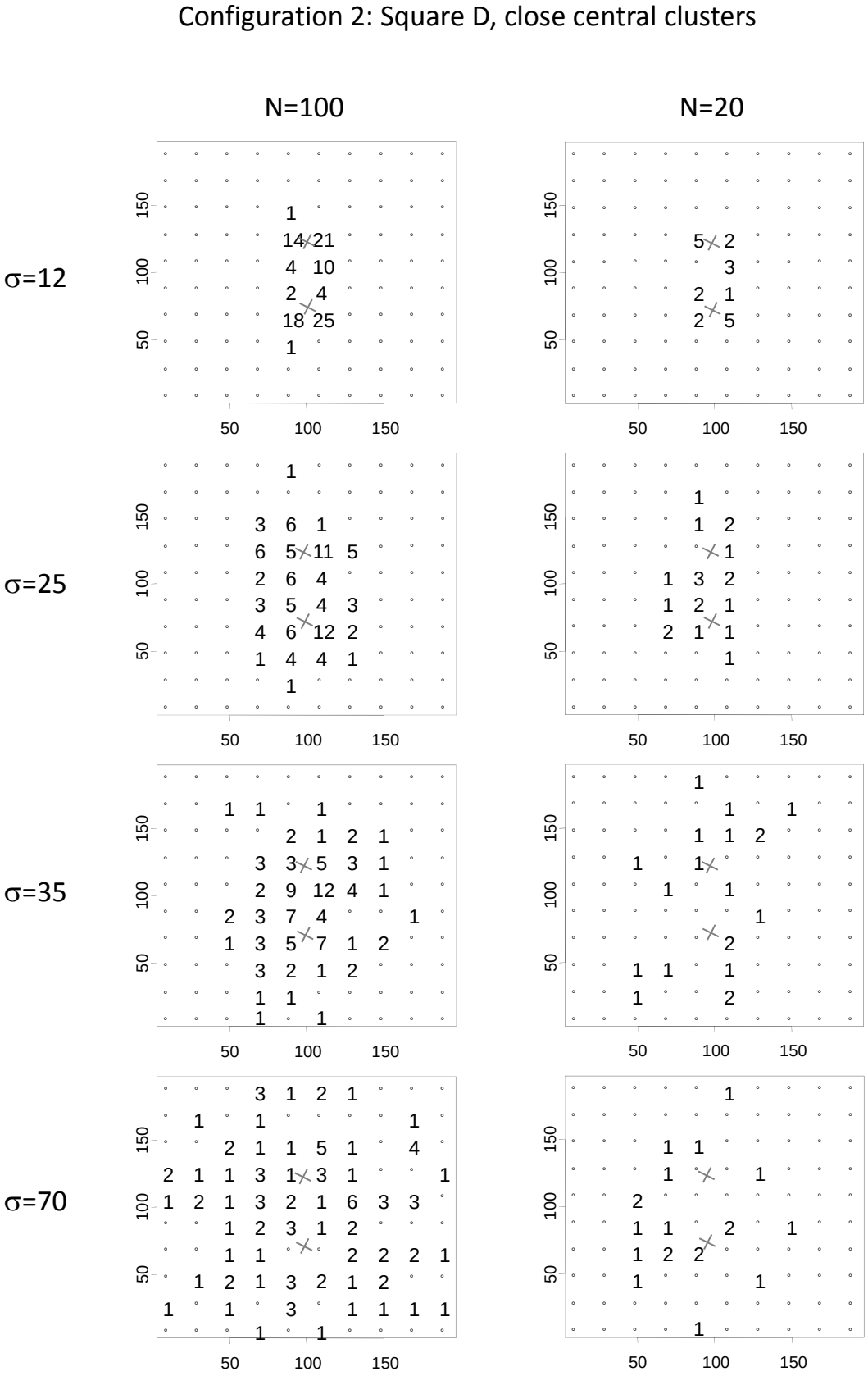
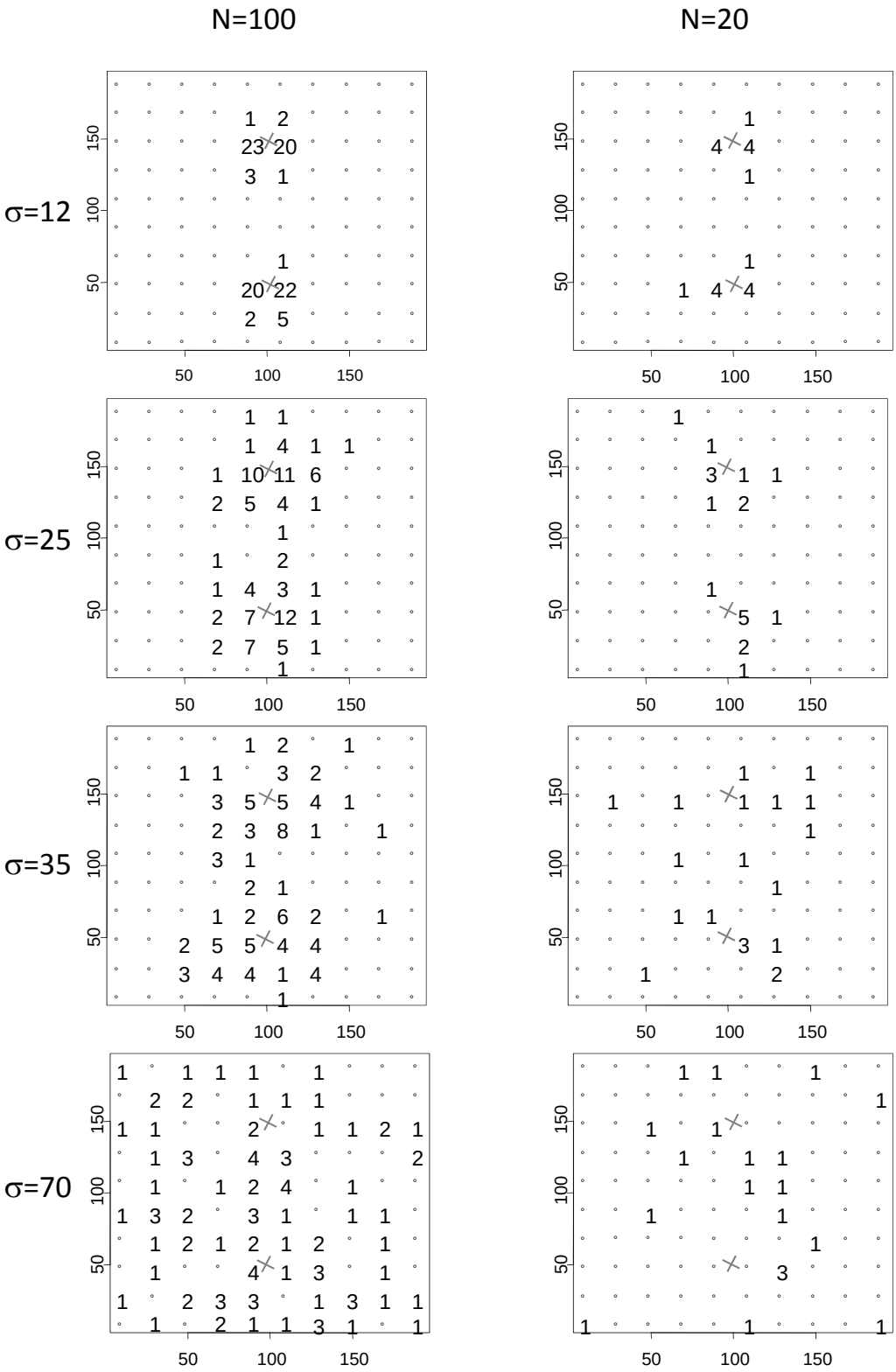


Figure 2: One example of each of the 32 different simulated clustered patterns. *N*=number of sampled individuals, σ=radius of simulated clusters.



Version définitive du manuscrit publié dans / Final version of the manuscript published in : Basic and Applied Ecology, 2010, In Press, DOI: 10.1016/j.baae.2010.08.011

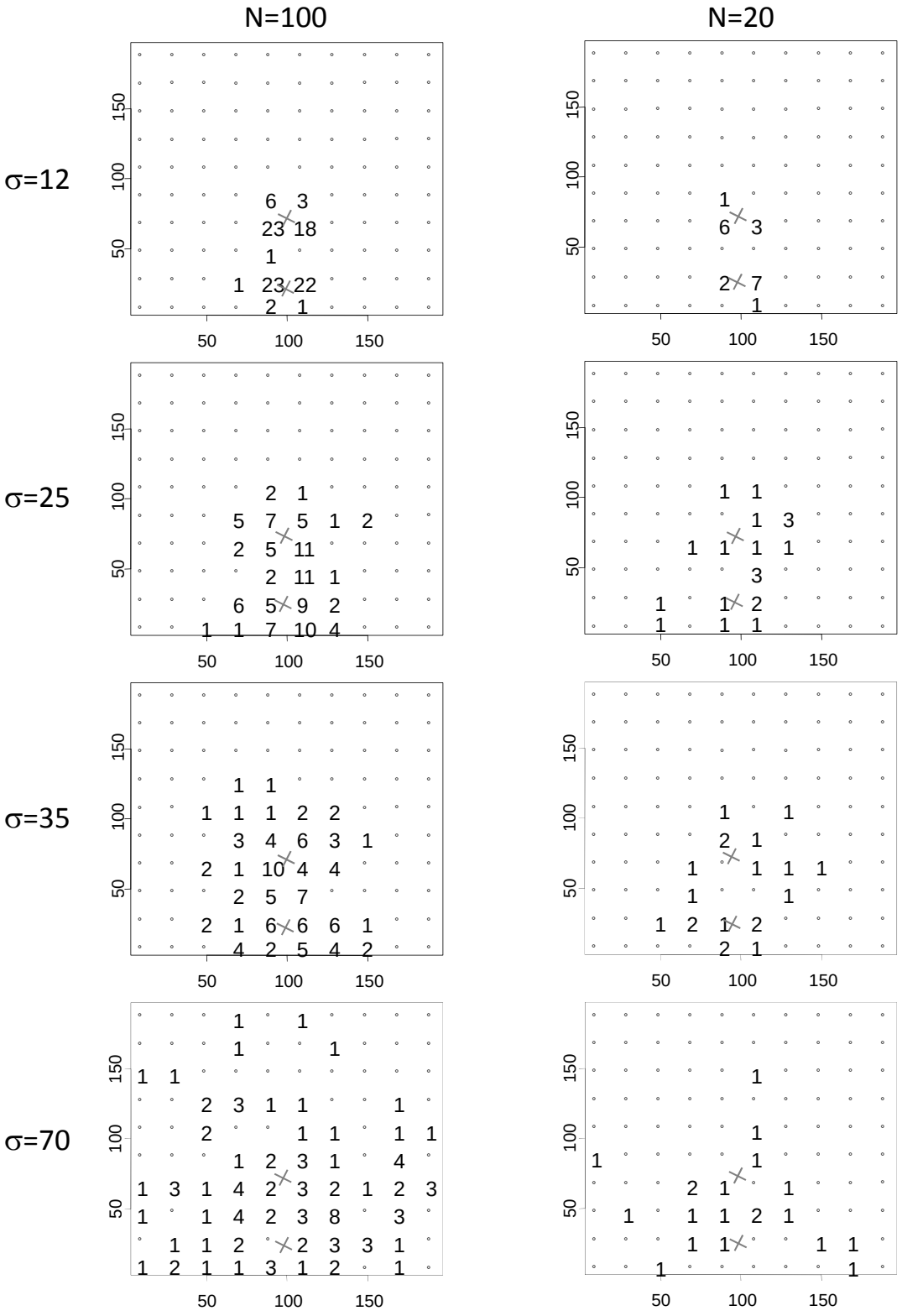
Configuration 3: Square *D*, far central clusters



1
2

Version définitive du manuscrit publié dans / Final version of the manuscript published in : Basic and Applied Ecology, 2010, In Press, DOI: 10.1016/j.baae.2010.08.011

Configuration 4: Square D, clusters close to edge



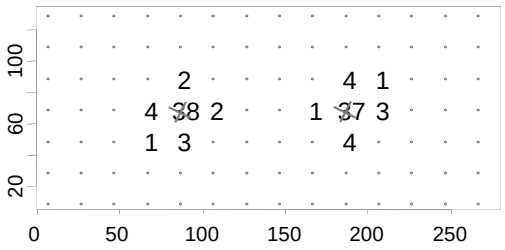
1
2

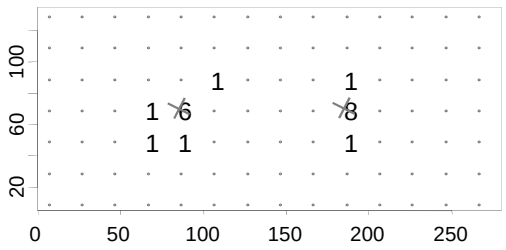
Version définitive du manuscrit publié dans / Final version of the manuscript published in : Basic and Applied Ecology, 2010, In Press, DOI: 10.1016/j.baae.2010.08.011

Configuration 5: Rectangular *D*, central clusters

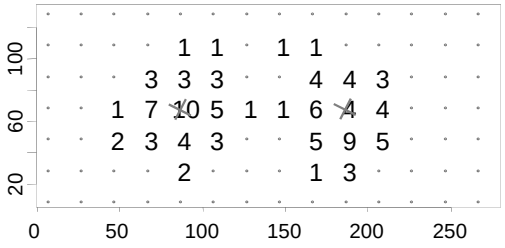
N=100 N=20

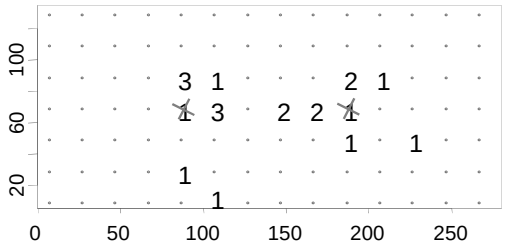
σ=12



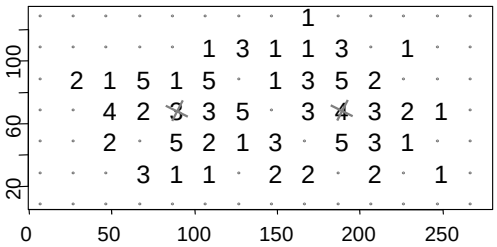


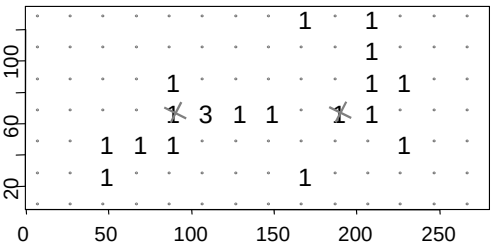
σ=25



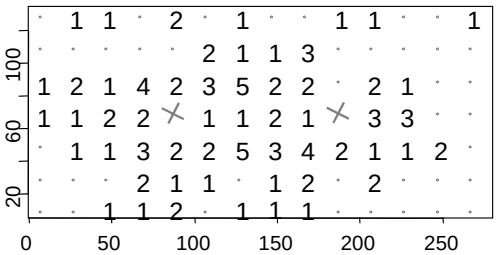


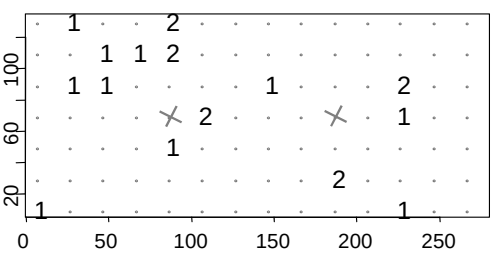
σ=35





σ=70





1