Ecological Informatics 49 (2019) 13–21

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Mark-mark scatterplots improve pattern analysis in spatial plant ecology

Felix Ballani[a], Arne Pommerening[b,*], Dietrich Stoyan[a]

[a] Institut für Stochastik, Bergakademie Freiberg, Prüferstraße 9, Freiberg D-09596, Germany
[b] Department of Forest Ecology and Management, Faculty of Forest Sciences, Swedish University of Agricultural Sciences SLU, Skogsmarksgränd 17, Umeå SE-901 83, Sweden

ARTICLE INFO

Keywords:
Spatial ecological patterns
Longleaf pine
Shorea congestiflora
Scots pine
Plant interaction
Fire ecology
Mark variogram
Mark correlation function
Spatial scales

ABSTRACT

Point process statistics provides valuable tools for many ecological studies, where 'points' are commonly determined to represent the locations of plants or animals and 'marks' are additional items such as species or size. In the statistical analysis of marked point patterns, various correlation functions are used such as the mark variogram or the mark correlation function. Often the interpretation of these functions is not easy and the non-spatial ecologist is in need of support. In order to make the analysis of spatial point patterns more accessible to ecologists, we introduced and tested a new graphical method, the mark-mark scatterplot. This plot visualises the marks of point pairs of inter-point distances r smaller than some small distance $r_{\max}$. We tested the application of the mark-mark scatterplot by reconsidering three quite different tree patterns: a pattern of longleaf pine trees from the southern US which was strongly influenced by fires, a tropical tree pattern of the species *Shorea congestiflora* from Sri Lanka and a Scots pine pattern from Siberia (Russia). The new method yielded previously undetected cause-effect information on mark behaviour at short inter-point distances and thus improved the analysis with mark correlation functions as well as complemented the information they provided. We discovered important new correlations in clusters of trees at close proximity. The application of the mark-mark scatterplot will facilitate the interpretation of point process summary statistics and will make point process analysis more accessible to ecologists not specialized in point process statistics.

1. Introduction

Point process statistics plays an important role in the ecological analysis of marked point patterns that are often analysed in spatial ecology. Including marks in the analysis enables the ecologist to consider not only the patterns of plant, animal, den or nest locations, but also important characteristics of these objects such as their sizes, or other attributes, see textbooks such as Illian et al. (2008), Wiegand and Moloney (2014), Baddeley et al. (2016) and papers, e.g. Stoyan and Penttinen (2000), Suzuki et al. (2008) and Pommerening and Särkkä (2013) published in this field. Of particular interest are quantitative marks, particularly size, height, number or weight of fruits and biomass.

In ecological-statistical analyses, summary functions from point process statistics are commonly used, in particular the mark correlation function $k_{mm}(r)$ and the mark variogram $\gamma_m(r)$ as described in the books referred to above as well as in the Methods Section below. These functions yield valuable information about the correlations between the marks of point pairs in dependence on the inter-point distance r, about mutual inhibition or mutual stimulation and on the degree of mark similarity. However, despite all their advantages it is the nature of statistical summary functions to average and thus to smooth away details that may be important for ecological interpretation.

Therefore, there is a need for an additional statistical tool with a nature between the original marked point pattern and the data-compressing summary functions that helps to identify the concrete mark-point configurations causing the shapes of certain summary functions.

The *mark-mark scatterplot*, a 2D-scatterplot of the marks of all point pairs at an inter-point distance r smaller than some $r_{\max}$, is such a tool. Since also the mark correlation function and the mark variogram both depend on point pairs at given distances, such a plot is expected to be a strong support to these summary functions. The idea of the mark-mark scatterplot was inspired by geostatistics, where spatial data are often analysed with so-called h-scatterplots (Pannatier 1996).

In this paper, we introduced the mark-mark scatterplot and reconsidered three classical datasets from spatial ecology to demonstrate the possibilities of this graphical tool. The patterns under consideration were from a longleaf pine fire ecosystem in Georgia (USA), where the trees were highly clustered, mainly as the result of uncontrolled fires (caused by lightning or man), a tropical tree pattern of the species

* Corresponding author.
E-mail addresses: ballani@math.tu-freiberg.de (F. Ballani), arne.pommerening@slu.se (A. Pommerening), stoyan@math.tu-freiberg.de (D. Stoyan).

https://doi.org/10.1016/j.ecoinf.2018.11.002
Received 11 September 2018; Received in revised form 7 November 2018; Accepted 14 November 2018
Available online 16 November 2018
1574-9541/

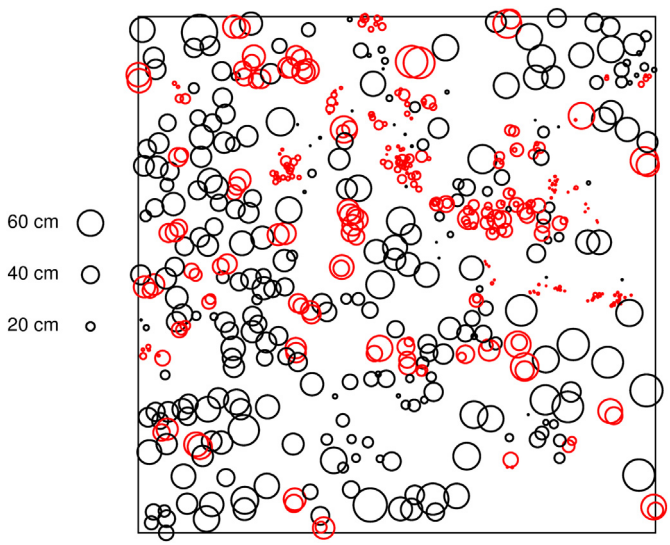

Fig. 1. 584 longleaf pine trees in a 200 m × 200 m observation window using dbh values (in cm) as marks. The trees contributing to the mark-mark scatterplot with $r_{\max} = 3$ m in Fig. 8 are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

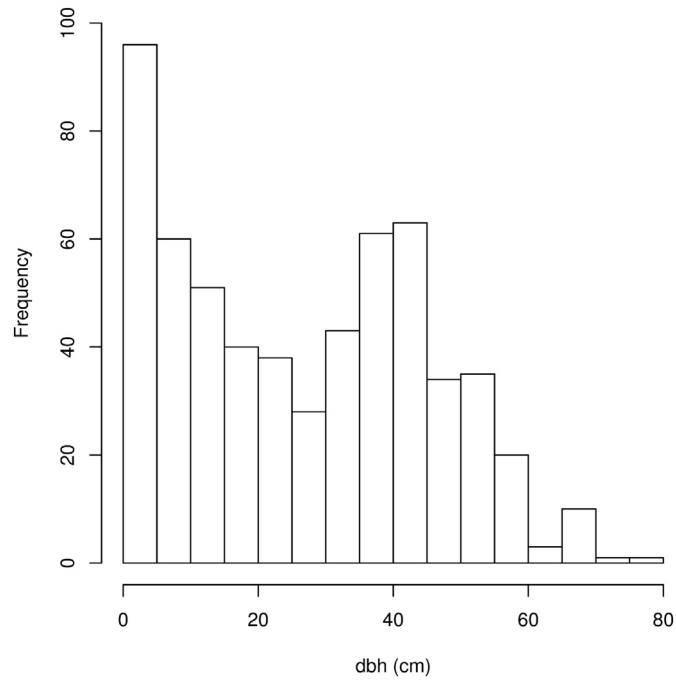

Fig. 2. Histogram showing absolute frequencies of the stem diameters (dbh) of the longleaf pine trees in the observation window.

Shorea congestiflora from Sri Lanka and a very dense even-aged stand of Scots pine in Siberia (Russia). Our hypothesis is that the newly introduced statistical tool can markedly improve the interpretation of the ecological relationships in spatial point patterns by preparing the data analysis and by identifying new cause-effect information.

2. Materials and methods

2.1. Data

2.1.1. Longleaf pine from Georgia (USA)

Cressie (1991) studied a pattern of 584 longleaf pine trees (*Pinus palustris* Mill.) (146 trees per hectare) with stem diameter at breast height (dbh; measured at 1.3 m above ground level) as quantitative mark, see Fig. 1. Remarkably, the map in Fig. 1 shows some clusters of large trees. The longleaf-pine data are from the Wade Tract, an old-growth forest in Thomas County, Georgia (USA). Longleaf pine is a fire-adapted species; surface fires frequently occured in these forests, removing most competing hardwoods. The population was uneven-aged and much varied in size (Platt and Rathbun 1993), large trees were only loosely aggregated. By contrast, juvenile trees were highly aggregated and were located in areas of low adult densities. Recruitment within this population thus appeared to occur primarily within open spaces created by the decline of large trees. Platt and Rathbun (1993) suggested that fire facilitation typically results in an extended, but indefinite, increase in the persistence of environmental conditions in which longleaf pine, but no other tree species, can survive and reproduce. The dbh mark distribution is bimodal (Fig. 2), i.e. there are two groups of nearly equal size of small and large trees.

Although marks were included in the dataset, Cressie (1991) considered only the tree locations and their clustering. Various authors have continued this approach by fitting Poisson cluster process models to the point data (Ghorbani 2013; Mecke and Stoyan 2005; Stoyan and Stoyan 1996; Tanaka et al. 2008). Until now, only Platt and Rathbun (1993) considered the marks as well as locations and most likely were the first to apply a mark variogram to marked point pattern data.

2.1.2. Shorea congestiflora in Sri Lanka

Wiegand et al. (2007) and Wiegand and Moloney (2014) analysed a tropical tree pattern of the species *Shorea congestiflora* ((Thw.) P. Ashton) from Sri Lanka with dbh as quantitative mark, see Fig. 3. *Shorea congestiflora* is a dominant species in a rain forest at the Sinharaja World Heritage Site (Sri Lanka). It is a medium-large-sized tree and, in contrast to many other species on this site, shows only minor habitat association (Gunatilleke et al. 2006). *Shorea congestiflora* fruits may be carried a short distance away from the crown by wind, however, they can easily be washed down steep slopes along with surface water runoff. The total number of trees in the observation window is 850 (34 trees per hectare; Wiegand et al. 2007). The empirical distribution based on original dbh marks is unsuitable for our analysis (Fig. 4a), because the majority of observations is concentrated in only a

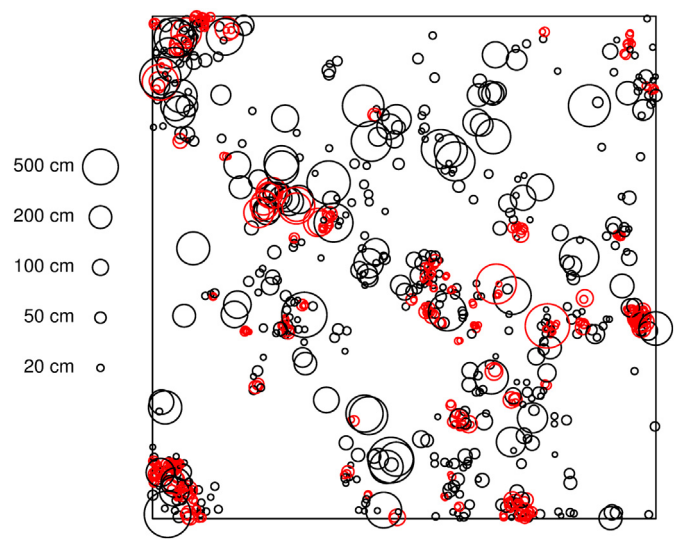

Fig. 3. 850 *Shorea* trees in a 500 m × 500 m observation window using dbh values (in cm) as marks. For better visualization the stem diameters are transformed by the square root transformation $\tau(m) = 0.8\sqrt{m}$. The trees contributing to the mark-mark scatterplot with $r_{\max} = 3$ m in Fig. 11 are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

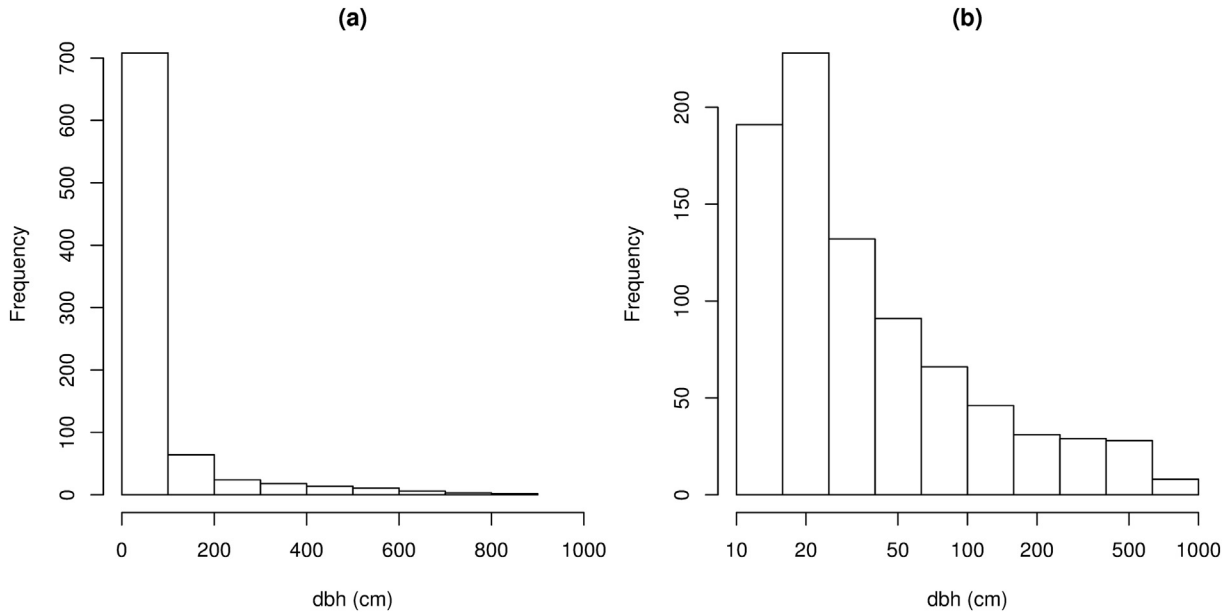

Fig. 4. Histogram showing absolute frequencies of the original stem diameters (dbh) of the *Shorea* trees (a) and after logarithmic transformation (b) in the observation window.

small part of the distribution. Therefore we adopted a logarithmic transformation of marks (Fig. 4b). The pattern was statistically analysed in Wiegand et al. (2007) by means of pair correlation functions for several size and age classes. In Wiegand and Moloney (2014), mark correlation function and mark variogram were estimated and briefly interpreted.

2.1.3. Scots pine in Siberia

This pattern includes 484 (1543 per hectare) Scots pine trees (*Pinus sylvestris* L.) from the Irkutsk region (Siberia, Russia). The climate there is strongly continental with dry spells in spring that often lead to forest fires. Following this disturbance many even-aged forests developed. Growth conditions are generally poor (severe climate and low soil fertility) resulting in low growth rates and many forest stands retained high tree densities even after many years.

The trees of the plot considered are 90 years old. The data were collected and ecologically interpreted by Busykin et al. (1985). As in Gavrikov et al. (1993) and Gavrikov and Stoyan (1995) we analysed the inner region of the originally 60 m × 60 m window here in order to reduce spatial inhomogeneity. The height mark distribution, shown in Fig. 6 is weakly bimodal and negatively skew. Point process methods were applied in the papers by Gavrikov et al. (1993) and Gavrikov and Stoyan (1995). In these papers the pair correlation function was discussed, which was close to a horizontal line through one as for the case of complete spatial randomness. However, the pattern resulted from a strong self-thinning: between age 25 and age 90 only < 10% of the trees survived (Busykin et al., 1985). This reveals strong interaction between the trees. This interaction was studied with the help of a stochastic forest model, the so-called area-saturation model, see Genet et al. (2014). Until now no mark correlation analysis has been carried out for these data.

2.2. Second-order summary functions

In this paper, we concentrate on three summary functions of marked point patterns. The simplest is the *mark distribution*, i.e. the probability density function of the marks, which is empirically represented by histograms. Analysing the mark distribution is the classical starting point of any analysis of marked point patterns (e.g. Illian et al. 2008). The corresponding mean is referred to as *mean mark* and is denoted by $\overline{m}$ and the corresponding variance is ${\sigma_m}^2$. This distribution gives first clues about the processes involved in a marked point pattern and may also be helpful when interpreting spatial characteristics, since it greatly influences them. As is common in statistics, it is sometimes desirable to achieve a better structured, empirical mark distribution through transformation so that the transformed distribution resembles known statistical distributions such as the normal or the gamma distributions.

The spatial characteristics included in our paper are the mark correlation function and the mark variogram. These functions are well explained in the aforementioned textbooks. The value of the *mark correlation function* $k_{mm}(r)$ at r is the mean of the product of marks m_i and m_j of points i and j that are separated by distance r. This value is normalised by the square of the mean mark $\overline{m}$. Similarly, the value of the *mark variogram* function $\gamma_m(r)$ at r is the mean of $0.5\ (m_i - m_j)^2$ for these points.

The mark correlation function characterises the sizes of marks and their dependence on inter-point distance r. Often points at close proximity tend to have small marks, perhaps because they belong to the same cluster of plants, either young, or older and inhibited in growth by mutual competition. This typically causes mark correlation functions with values smaller than 1, as often observed for dbh marks. In ecological applications, $k_{mm}(r) = 0$ is hardly possible at $r = 0$; this would only happen if at least one mark of points at close proximity is zero. Also values larger than 1 are possible: in that case, points at close range tend to have marks larger than the mean mark, as reported for height marks by Suzuki et al. (2008). For large r, $k_{mm}(r)$ always tends towards the limit of 1, although with empirical mark correlation functions the limit may differ from 1 because of statistical fluctuations and spatial inhomogeneity.

The mark variogram characterises the similarity (or continuity) of marks in dependence on inter-point distance r. If the marks of points at close proximity are similar, $\gamma_m(r)$ has small values for small r. However, in ecological applications the case $\gamma_m(r) = 0$ at $r = 0$ cannot occur, since this would imply that points very close together have exactly the same mark, which is hardly possible because of biological variability. Therefore in ecological analyses mark variograms always show a so-called nugget effect, i.e. a positive value of $\gamma_m(r)$ at $r = 0$. With increasing r a (theoretical) mark variogram tends towards a limit, which is the mark variance ${\sigma_m}^2$. Similarly as for the mark correlation function, for an empirical mark variogram the corresponding limit (often called

the "sill") may differ from the empirical mark variance, because of statistical fluctuations and spatial inhomogeneity.

Both functions provide valuable information about the correlation range, i.e. the maximum distance up to which statistically measurable interactions between marks exist. The range is given by that value of distance r from which onward the functions tend to be constant or fluctuate around limits. (If points appear in clusters the range may be related to cluster diameters.) Quite often the corresponding ranges of the *empirical* functions $k_{mm}(r)$ and $\gamma_m(r)$ differ.

Furthermore, the functions provide valuable information on the nature of statistical interactions. The mark variogram allows to distinguish between negative and positive association, also referred to as negative and positive autocorrelation. 'Negative' denotes a situation where two marks under consideration are of significantly different size and 'positive' where they are of similar size. Small values of the mark variogram indicate positive association. The mark correlation function provides information on the absolute values of the marks of point pairs. For example, the function may have large values if the marks of points with a distance r tend to be large (perhaps one point with a large and the other with a medium-sized mark), but do not need to be similar.

However, both functions tell comparatively little about the variability of the marks for small inter-point distances r. The numerical values of $k_{mm}(r)$ and $\gamma_m(r)$ at $r = 0$ characterize the variability of the marks at short distances only in a rather condensed form.

2.3. Mark-mark scatterplot

The mark-mark scatterplot is obtained by plotting the marks of all point pairs with an inter-point distance r smaller than some $r_{\max}$. The mark pairs of these points are arranged in a mark-mark coordinate system, where the abscissa is related to the mark of the first point i and the ordinate to that of the second point j.

For ecological applications intervals $[0, r_{\max}]$ for suitable $r_{\max}$ are of main interest, since interactions of ecological importance mainly occur at short distances. In analogy, also all point pairs with an inter-point distance in an interval $[r_{\min}, r_{\max}]$ can be considered, where the interval includes distances that attract particular interest when studying the aforementioned summary functions for a given ecological effect that is likely to occur in this interval.

When plotting pairs of marks, it is not obvious which is the first and which the second data point. Therefore, we recommend assigning two points to each point pair related to (i, j) and (j, i). The resulting graph is symmetric with respect to the line $m_i = m_j$.

The mark-mark scatterplot can be refined by plotting the points in grey tones. According to the inter-point distance of a point pair in the interval $[0, r_{\max}]$, the points are then plotted in the corresponding grey tone between white (= distance 0) and black (= distance $r_{\max}$). To reduce overlapping effects, we recommend plotting these points in descending order of the corresponding inter-point distances, i.e. from largest to smallest. Our R package 'mmsc' does this automatically and this helps identifying the inter-point distances of the point pairs included in the mark-mark scatterplot. All points with a grey tone below some limit that corresponds to a distance $r_{\max}'$ form a nested mark-mark scatterplot for inter-point distances smaller than $r_{\max}'$.

The choice of the limiting distance $r_{\max}$ is crucial. We recommend values of $r_{\max}$ clearly smaller than the range of correlation. For this we have three reasons: (i) The ecologically most interesting interaction occurs at short distances corresponding to the distances to near neighbours, (ii) the shape of the summary functions, which are aimed to be interpreted by the scatterplot, is strongly determined by the values for small $r_{\max}$, and (iii) for large $r_{\max}$ the scatterplot includes too many points making interpretation difficult.

In order to facilitate the interpretation the mark-mark scatterplot should be complemented by two contour lines. We recommend using the (curved) contour line corresponding to mark pairs (m_i, m_j) which satisfy $m_i m_j = \overline{m}^2$, thus dividing the mark-mark scatterplot into two regions of mark pairs (m_i, m_j) resulting in contributions $m_i m_j / \overline{m}^2$ less or larger than one. This is helpful when discussing certain values of the mark correlation function $k_{mm}(r)$. The second contour line (a pair of straight lines) corresponds to mark pairs (m_i, m_j) which satisfy $0.5(m_i - m_j)^2 = \sigma_m{}^2$, where $\sigma_m{}^2$ denotes the variance of the marks. This contour line is a boundary between comparatively similar and comparatively dissimilar mark pairs, likewise separating small and large contributions of $0.5(m_i - m_j)^2$ to the mark variogram $\gamma_m(r)$. Note that these contour lines are devised to match the test functions of $k_{mm}(r)$ and $\gamma_m(r)$. If other correlation functions were used, different contour lines may apply.

The mark-mark scatterplot magnifies the information on interactions between the marks of point pairs at short inter-point distances: It uncovers more crucial ecological information on the association of marks than the summary functions $k_{mm}(r)$ and $\gamma_m(r)$, since the mark-mark scatterplot displays the original marks before averaging. This provides an opportunity to identify extreme pairs, with very big or very small mark differences, which leads to a better understanding of the mark variogram $\gamma_m(r)$ for small r and the nugget effect. Furthermore, the mark-mark scatterplot yields detailed information on the type of association between marks for short inter-point distances r. In the case of positive association, the scatterplot points (m_i, m_j) tend to have locations close to the line $m_i = m_j$, while in the case of negative association, the scatterplot points tend to be far from this line, i.e. for large m_i there are small m_j and vice versa. Also, the mark-mark scatterplot may reveal that positive and negative association depends on the size of marks, for example, in a way that positive association in a given pattern holds preferably for small marks.

2.4. Relabellings

It is well known that in the case of independent marks (no spatial correlation) $k_{mm}(r)$ and $\gamma_m(r)$ are theoretically constant, i.e. the curves approach 1 and $\sigma_m{}^2$ (for $r > 0$), respectively; for the corresponding empirical curves $\widehat{k}_{mm}(r)$ and $\widehat{\gamma}_m(r)$ this is satisfied at least approximately. This trait is used in statistical tests to identify significant deviations from independence. The corresponding envelopes of these functions we also use in this paper. In analogy, the point cloud in mark-mark scatterplot has a simple shape in the case of independent marks: a pattern of uniformly distributed points (m_i, m_j). Therefore, we recommend independent relabelling of the points of the marked pattern based on the empirical mark distribution and determining the corresponding mark-mark scatterplot. The original mark-mark scatterplot and the new one may show interesting differences that can lead to a better ecological interpretation.

Finally, there is another way of relabelling points, i.e. the so-called *uniform rank transformation* (Skibba et al. 2013). This transformation is performed by replacing each mark m_i by the value $u_i = \text{rank}(m_i)/n$, where $\text{rank}(m_i)$ is the rank of the mark m_i in the whole set $m_1, m_2, \ldots, m_n$ of the n marks after ordering them ascendingly. (In the case of ties, i.e. marks with equal values, it is common either to assign to them the rank given by the arithmetic mean of the ranks calculated for them or to assign the ranks randomly.) Since this transformation is order preserving, mutual relationships among the marks are maintained, i.e. after transformation originally small marks continue to be comparatively small and originally large marks continue to be comparatively large.

The uniform rank transformation yields new marks u_i the mark distribution of which is the uniform distribution between 0 and 1 and therefore the corresponding mark-mark scatterplot is restricted to the unit square. Since this method reduces the influence of the original mark distribution on the shape of the mark-mark scatterplot it may serve as a kind of standardisation (Skibba et al. 2013) and should be considered a valuable tool for the comparison of marked point patterns with different mark distributions. For a thorough analysis it is instructive to compare the two mark-mark scatterplots (original and

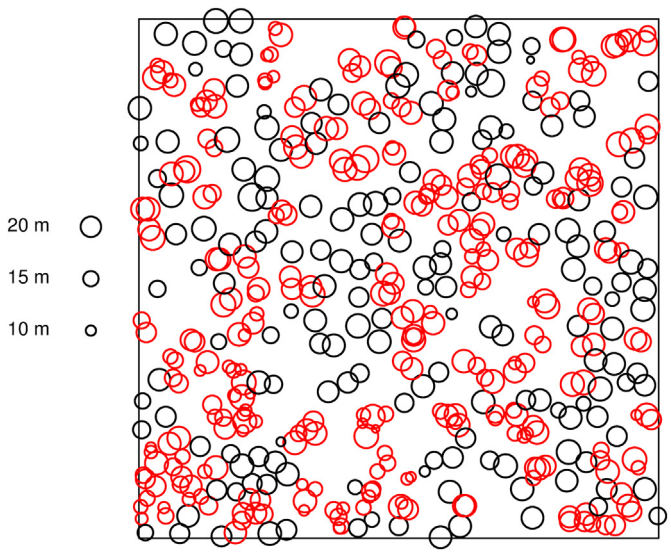

Fig. 5. 484 Scots pine trees in a 56 m × 56 m observation window using total height values (in m) as marks. The trees contributing to the mark-mark scatterplot with $r_{\max} = 1.5$ m in Fig. 13 are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

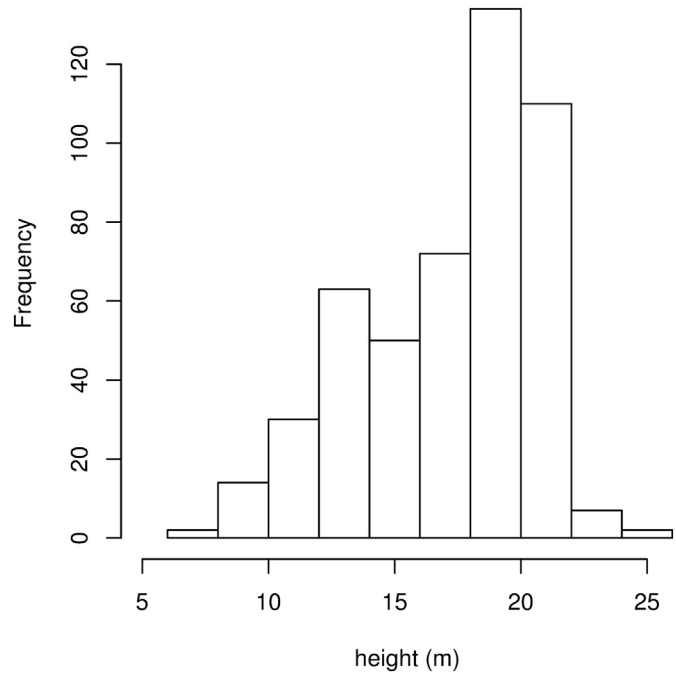

Fig. 6. Histogram showing absolute frequencies of the total heights of the Scots pine trees in the observation window.

transformed) and the corresponding mark correlation functions.

3. Results

3.1. Longleaf pine from Georgia (USA)

The mark correlation function related to the longleaf pine trees in Georgia is presented and discussed in Chiu et al. (2013, p. 124), see our Fig. 7a: for small r the function $k_{mm}(r)$ increases and for $r > 15$ m it fluctuates around a value of 0.9, which here replaces the theoretical limit 1. This result indicates that trees close together tend to have smaller diameters than the mean dbh, which is 26.8 cm. In Chiu et al. (2013) this was explained "as the price trees have to pay for being close together". A more comprehensive ecological interpretation would state that this part of the curve describes clusters of trees.

The mark variogram shown in Fig. 7b has the typical shape of a geostatistical variogram and Platt and Rathburn (1993) estimated the range of correlation as 28 m, since this is the value of r where the empirical mark variogram begins to fluctuate around 280. Based on this information they determined the size of patches with similarly sized trees. (Note that the value of 280 is smaller than the empirical mark variance, which is 336.) The variogram clearly has a nugget effect, the empirical value of $\gamma_m(r)$ at $r = 0$ is 34.9.

The corresponding mark-mark scatterplot for $r_{\max} = 3$ m, presented in Fig. 8, shows the detailed mark structure of point pairs at short distances. It is immediately clear that the mark-mark scatterplot supports the statements given by the correlation functions: there is great variability of marks, which explains the nugget effect and there are many mark pairs where both marks are smaller than $\overline{m} = 26.8$ cm, eventually leading to values of $k_{mm}(r)$ smaller than 0.9 (the analogue to the theoretical value 1 for the longleaf data). The latter fact is even more clearly indicated by the two large groups of mark pairs below and above the contour line $m_i m_j = \overline{m}^2$ (blue) leading to relatively small and, respectively, large contributions $m_i m_j / \overline{m}^2$ to $k_{mm}(r)$ for $r \leq 3$. Likewise, the majority of mark pairs lies between the two contour lines $0.5\ (m_i - m_j)^2 = \sigma_m{}^2$ (red), i.e. most values $0.5\ (m_i - m_j)^2$ contributing to the mark variogram are small and, hence, the result is a relatively small value of $\gamma_m(r)$ for $r \leq 3$.

However, the mark-mark scatterplot provides more information. Obviously, the scattering around the line $m_i = m_j$ depends on the values of the marks. There are three regions of particular interest corresponding to different ecological interaction strata: small marks (0…20), medium-sized marks (20…35), and large marks (35…). For *small marks* the square [0, 20] × [0, 20] of the plot is nearly uniformly filled with points. This suggests that the small marks of trees at close proximity are nearly independent, i.e. for these trees there is no spatial association and therefore no interaction. For *medium-sized marks* the points in the plot tend to be comparatively far from the line $m_i = m_j$, i.e. medium-sized trees tend to have neighbours that are clearly smaller. This qualifies for negative association. For *large marks* spatial interactions vary in type: there are trees that have a large or medium-sized neighbour of similar size (positive association) but others have large neighbours of varying size. Particularly interesting are trees with a dbh of (around) 36 cm: they have neighbours at a distance smaller than 3 m with dbhs between 5 and 65 cm. In addition the nested mark-mark scatterplot formed by all points with white to light grey colour shows that the observed structure of three distinctive regions is also visible at interpoint distances smaller than one metre; however, naturally large-large mark pairs are then rare.

Based on the information given by the mark-mark scatterplot it is now possible to highlight in red all trees in the forest map which have at least one neighbouring tree less than $r_{\max} = 3$ m apart (Fig. 1): Now we see clearly several clusters of small trees and particularly the interesting, rare clusters of large trees, perhaps resulting from fires long ago.

Finally, we compared the original mark-mark scatterplot in Fig. 8 with the mark-mark scatterplot corresponding to independent relabelling, shown in Fig. 9a. The structure of both mark-mark scatterplot differs greatly: The latter tends to a uniform distribution of the points corresponding to independent marking. By contrast, the original mark-mark scatterplot shows a clear concentration of the points along the line $m_i = m_j$ and a large cluster of points corresponding to pairs of small trees. This comparison is another way of proving that the size development in the tree pattern under consideration is not independent.

The mark-mark scatterplot for uniformly ranked marks (Fig. 9b) is similar to the plot in Fig. 8, indicating that the original marks and the rank marks are similar in their correlation behaviour. This observation is confirmed by the correlation functions, since $k_{mm}(r)$ and $\gamma_m(r)$ for the uniformly ranked marks (not shown here) are similar to the original

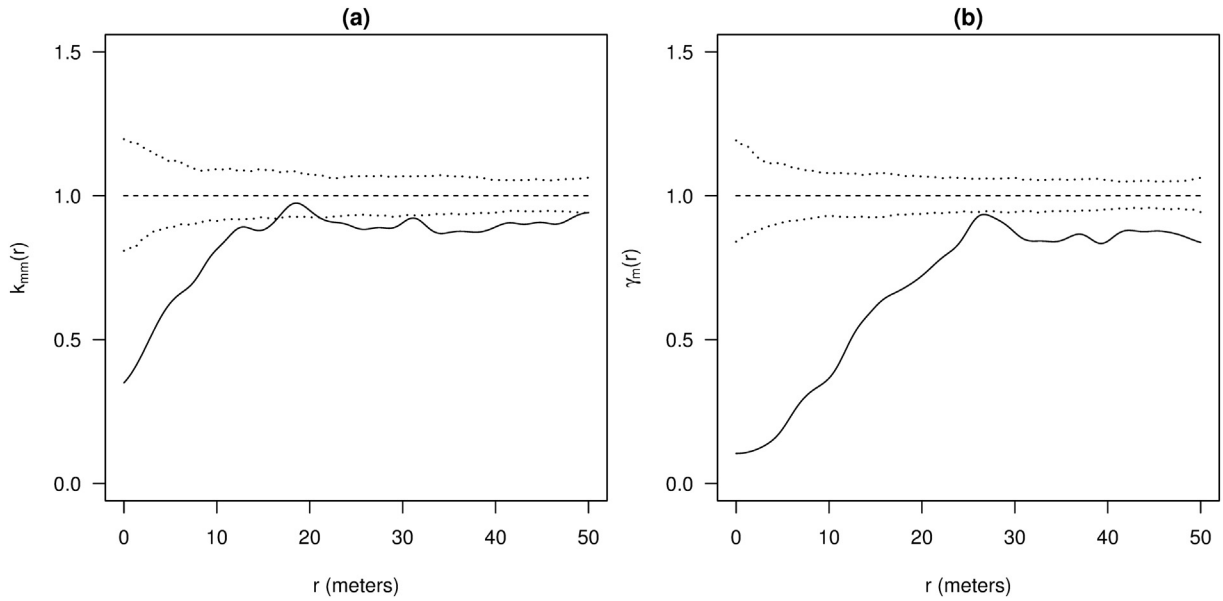

Fig. 7. Mark correlation function $k_{mm}(r)$ (a) and mark variogram $\gamma_m(r)$ (b) for the longleaf pine trees (solid lines); $\gamma_m(r)$ is normalised with the empirical mark variance. Additionally the 95% pointwise envelopes (dotted lines) from 999 random relabellings of the marks of the longleaf-pine data are given.

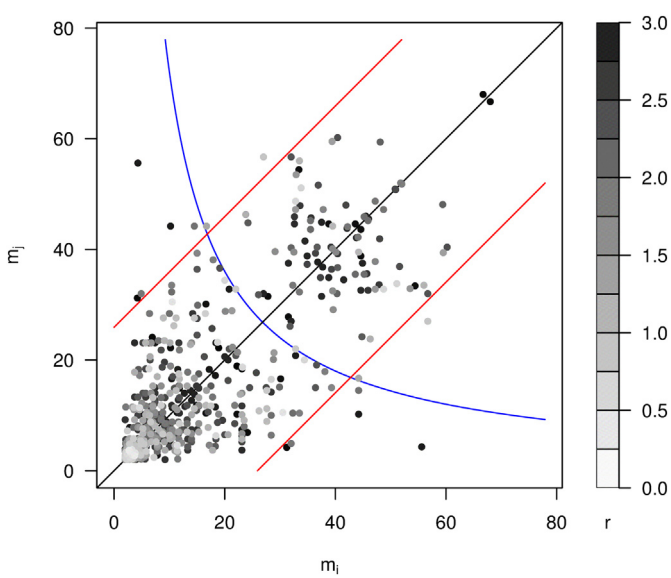

Fig. 8. Mark-mark scatterplot for the longleaf pine trees with $r_{\max} = 3$ m including additional contour lines at $m_i m_j = \overline{m}^2$ (blue) and 0.5 $(m_i - m_j)^2 = \sigma_m{}^2$ (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

functions shown in Fig. 8. Generally we note large variability in the dbh of tree pairs at distances smaller than 3 m.

3.2. Shorea congestiflora in Sri Lanka

The mark correlation function and the mark variogram for the *Shorea* trees in Sri Lanka (Fig. 10) were briefly explained in Wiegand and Moloney (2014, p. 67). (The more detailed paper Wiegand et al. (2007) did not consider quantitative marks.) Wiegand and Moloney (2014) concluded that many small trees were grouped in clusters. They also mentioned the presence of some larger trees close to clusters of small trees, which produced the larger values in the mark products leading to $\hat{k}_{mm}(r)$ of nearby trees. Accordingly, the standardized mark variogram started with values of 0.2 at very small distances.

This is a typical behaviour that can be observed for many tree species around the world: large trees often represent parent trees while the clusters of small trees are their offspring. However, in some situations the species of large individual trees are even different from that of clusters of small trees close to. This is then due to natural processes of maintaining biodiversity as described by the Janzen-Connell and herd immunity hypotheses (Pommerening and Uria-Diez 2017). The birth and self-thinning processes involved lead to the development of size hierarchies, i.e. differences in size at close proximity.

We agree with Wiegand and Moloney's general interpretation and, based on Illian et al. (2008), add: the mark correlation function $k_{mm}(r)$ has a shape that mainly results from large clusters of small trees. The curve increases with increasing inter-tree distance r and is always smaller than 1 for $r < 35$ m. Thus the typical, most frequent case in this pattern is that both individuals of a pair of trees at close proximity occur in the same cluster of small trees and have a dbh smaller than the mean dbh, which is 72.8 cm. The range of correlation indicated by $k_{mm}(r)$ is at about 35 m.

The (standardized) mark variogram has a shape as in geostatistical applications, i.e. it increases with increasing r. This means that with increasing inter-tree distance the variability of dbh differences between tree pairs increases. There is a so-called nugget effect, i.e. $\hat{\gamma}_m(r)$ is not zero for very small r but 0.2. The range of correlation indicated by the variogram is about 40 m, perhaps a bit larger than that of the mark correlation function.

We prepared the corresponding mark-mark scatterplot for $r = 3$ m after performing a logarithmic transformation on the dbh values (Fig. 11). Using the original marks resulted in an amorphous mass of points without any structure. (We also estimated the summary functions for logarithmic marks, but this did not yield any additional information.) The mark-mark scatterplot clearly shows the variability of the mark pairs. The main tendency is positive association: there are many pairs with both small or medium marks and some pairs with both large marks. This explains the shape of the correlation functions well. There are some pairs of exceptionally large trees (dbh > 150 cm) at short inter-tree distances without smaller-sized neighbours. Furthermore, there are particularly three pairs where one tree is small and the other very large (dbh > 500 cm). Since the corresponding inter-tree distances are larger than 2 m, these pairs do only little contribute to $\hat{\gamma}_m(r)$ at $r = 0$; the nugget effect results from (many) pairs of trees of different sizes, one with a medium and one with a small dbh.

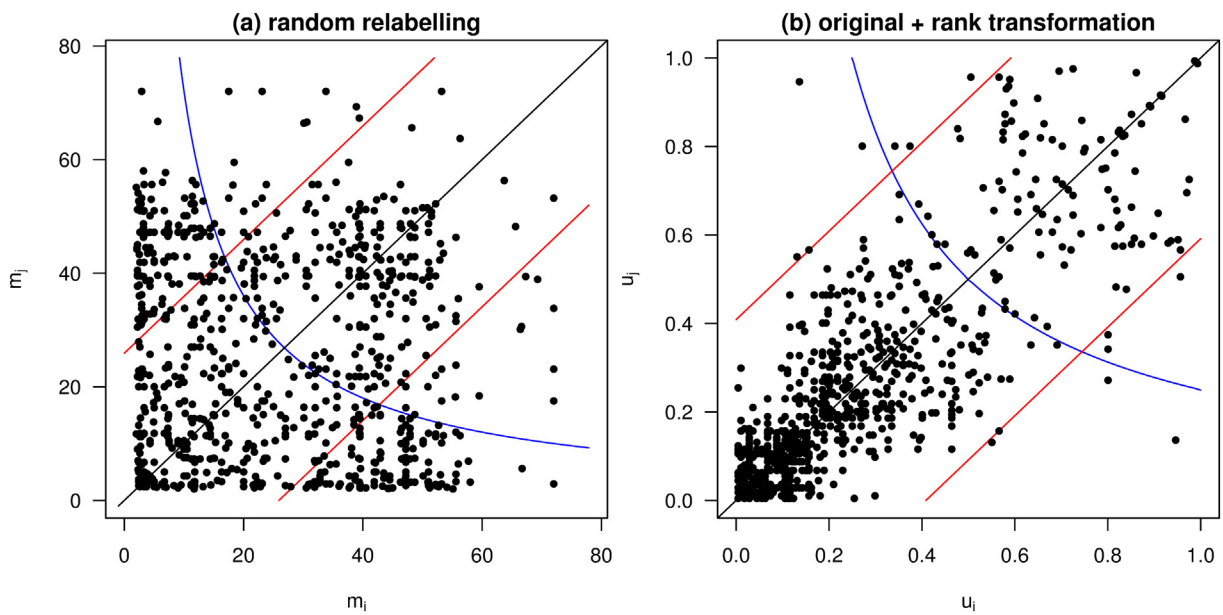

Fig. 9. Mark-mark scatterplots for the longleaf pine trees with randomly relabelled marks (a) and original marks after uniform rank transformation (b) for inter-point distances smaller than $r_{\max} = 3$ m. Additionally the contour lines highlighting $m_i m_j = \overline{m}^2$ (blue) and $0.5\,(m_i - m_j)^2 = \sigma_m{}^2$ (red) are shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

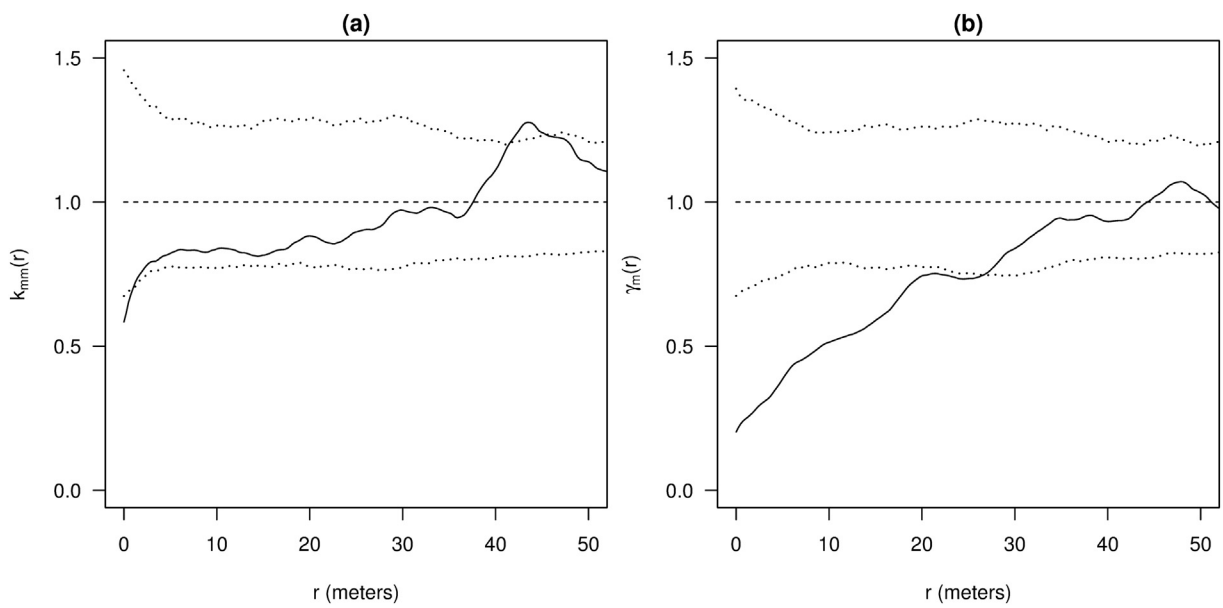

Fig. 10. Mark correlation function $k_{mm}(r)$ (a) and mark variogram $\gamma_m(r)$ (b) for the *Shorea* trees (solid lines); $\gamma_m(r)$ is normalised with the empirical mark variance. Additionally the 95% pointwise envelopes (dotted lines) from 999 random relabellings of the marks of the *Shorea* data are given.

3.3. Siberian Scots pine (Russia)

While the shape of the mark correlation function of the Scots pine trees is similar to those for the longleaf pine trees and the *Shorea* trees, the mark variogram is different (Fig. 12). At short inter-tree distances $r < 4$ m the curve decreases in r. Only for larger values of r the empirical function increases again as with the other data. The ranges of correlation are 6 m for the mark correlation function and 12 m for the mark variogram. Incidentally, for the dbh values the correlation functions are similar; the coefficient of correlation between dbh and height is 0.8.

When interpreting this mark variogram, the experienced statistician expects a considerable number of tree pairs at close proximity with different heights, similar to Illian et al. (2008, p. 419) and Wälder and Stoyan (1996). This can be roughly confirmed when looking at the map of the marked point pattern in Fig. 5 but now, after studying the mark-mark scatterplot, this effect is understood more thoroughly (Fig. 13). We used $r_{\max} = 1.5$ m, where the small value of $r_{\max}$ is an adaptation to the high tree density of the plot. The scatterplot shows two main point clouds. There are many points scattered along the line $m_i = m_j$ between the red lines, indicating positive association. These points belong to tree pairs of nearly equal height. These trees possibly have experienced an extended life history with similar light conditions and densities. There are also point clouds around (20, 12) and (12, 20). The corresponding points belong to tree pairs of different height, where perhaps one tree is dominating and the other is suppressed or germinated slightly later.

We compared our findings with the detailed ecological discussion of the data in Genet et al. (2014). That paper considered three size classes

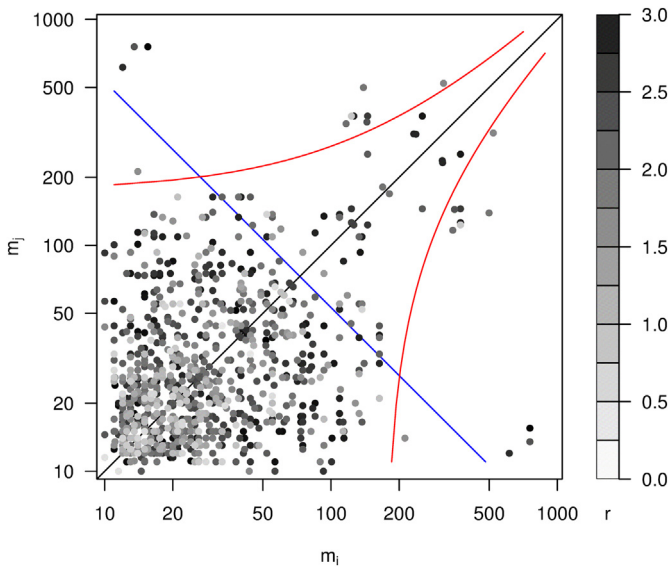

Fig. 11. Mark-mark scatterplot for the *Shorea* trees with $r_{\max} = 3$ m including additional contour lines at $m_i m_j = \overline{m}^2$ (blue) and $0.5\ (m_i - m_j)^2 = \sigma_m{}^2$ (red). Note the logarithmic scaling, which also influences the shape of the contour lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

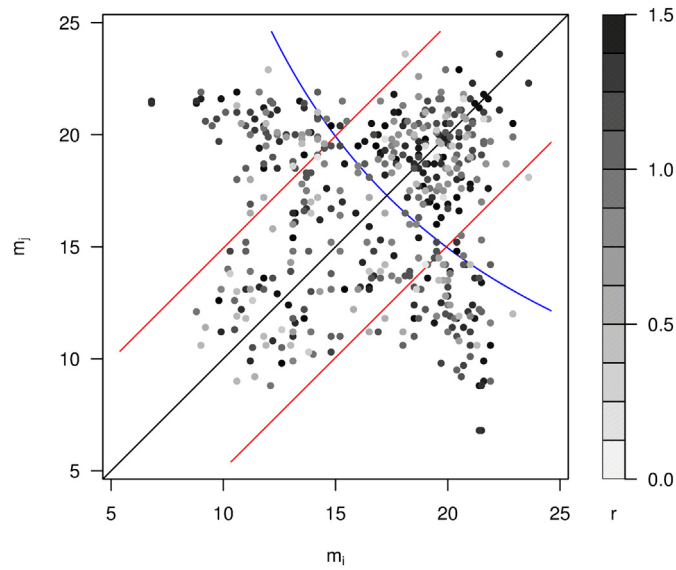

Fig. 13. Mark-mark scatterplot for the Scots pine trees with $r_{\max} = 1.5$ m including additional contour lines at $m_i m_j = \overline{m}^2$ (blue) and $0.5\ (m_i - m_j)^2 = \sigma_m{}^2$ (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(small, medium and large) and comes to two main conclusions: (i) Large and medium-sized trees were both associated with small-scale (< 3 m) repulsion leading to regularly spaced trees, (ii) small-sized trees tended to medium-scale clusters at distances around 2 m and 3 m from large- and medium-sized trees. These clumps were explained by microtopographic variations and not by canopy gaps.

We understand that the mark variogram and the mark-mark scatterplot refined these conclusions: pairs of trees of different size but at close proximity (see the scatterplot in Fig. 13 around (12, 20)) appeared frequently in the stand and large- and medium-sized trees can be close to trees of similar size (see the mark variogram in Fig. 12 for distances of 2–4 m).

4. Discussion

Through applying the mark-mark scatterplot to the longleaf-pine data we have clarified correlations between the trees' dbh values at short inter-tree distances r and we have achieved a better understanding of the particular shapes of the two correlation functions used in the analysis. We now are sure that it is mainly pairs of tree neighbours of the same clusters where both marks are small, that are responsible for the observed shapes and that there is great variability in dbh values at short distances. However, the mark-mark scatterplot helped to detect also some clusters of large trees, see Fig. 1. For the *Shorea* trees we learned from the mark-mark scatterplot that predominantly pairs of both small and medium-sized marks cause the shapes of the summary functions. The plot also highlighted interesting combinations of small and very large trees that did not affect the

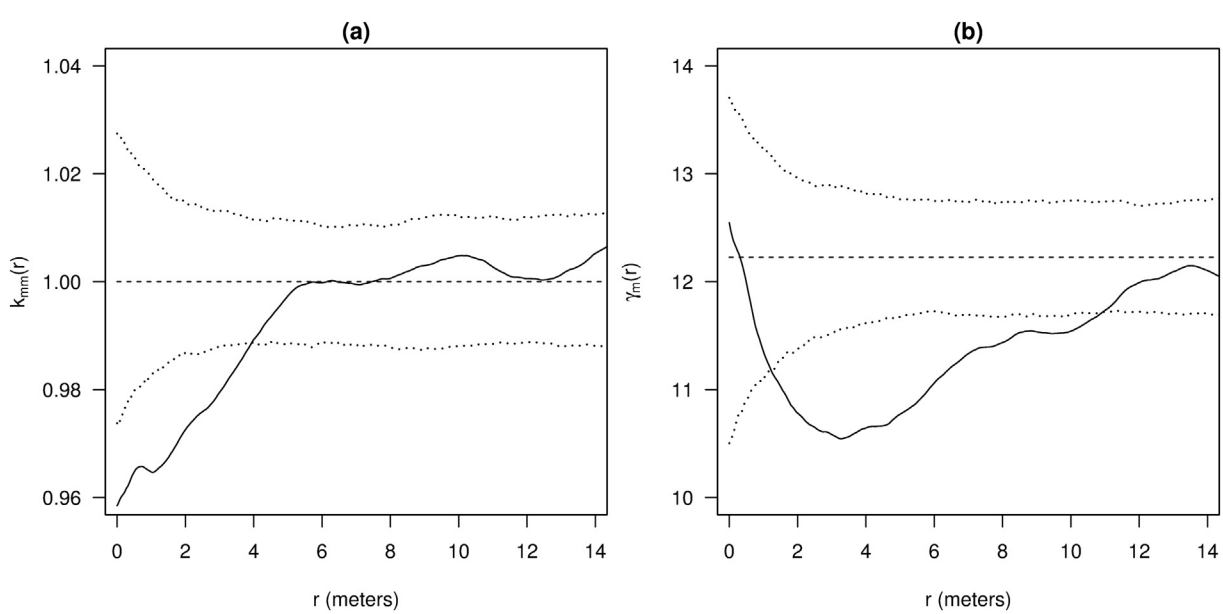

Fig. 12. Mark correlation function $k_{mm}(r)$ (a) and mark variogram $\gamma_m(r)$ (b) for the Scots pine trees (solid lines). Additionally the 95% pointwise envelopes (dotted lines) from 999 random relabellings of the marks of the Siberian pine data are given.

summary functions. For Scots pine the mark-mark scatterplot clarifies the particular shape of the mark variogram: for small values of r the characteristic is essentially governed by tree pairs of different size, while for larger values of r the influence of trees of similar height increases. This removes any doubt that the shape of the mark variogram may be a mere statistical artefact. Instead we understand that the negative mark association is a consequence of local size hierarchies or size inequality (Ford 1975; Suzuki et al. 2008; Weiner and Solbrig 1984).

Every ecologist working with spatial data knows that any analysis of such data is a multi-step, iterative process. The map of object locations and marks is first visually assessed, a process which may already lead to working hypotheses and even to tentative ecological theories. The next logical step in the analysis is the estimation and interpretation of the mark distribution. By interpreting this characteristic one learns important facts about the mark structure. A trained ecologist can often link the shape of an observed mark distribution with certain ontogenetic stages in a population of organisms.

Following this, we now recommend constructing mark-mark scatterplots. Here various values of $r_{\max}$ should be tried iteratively. The aim is to find a good balance between a possibly too small value of $r_{\max}$ (resulting in too few data points) and a chaotic plot with too many data points that make interpretation difficult. The visual assessment of mark-mark scatterplots gives a good impression of the type of spatial interactions at short distances, which are of particular ecological value: positive or negative association and the influence of mark size and correlation on these relationships. For answering these questions in a comprehensive way it is instructive to check whether the points of the plot are located close to the diagonal $m_i = m_j$ or far from it.

Mark-mark scatterplots help the ecological analyst to anticipate the shape of correlation functions such as $k_{mm}(r)$ and $\gamma_m(r)$ for small r and to verify them after the estimation. This may be helpful for choosing suitable bandwidth parameters of the kernel functions used in the estimation of these functions.

Once the definite correlation functions have been estimated and plotted, mark-mark scatterplots assist in their interpretation. It is possible that there is a discrepancy between the function shapes anticipated by the analyst on the basis of the mark-mark scatterplots and the real functions. This gives the analyst the opportunity to check whether the initial interpretation of the mark-mark scatterplots was possibly flawed or whether inappropriate parameters such as bandwidth h or $r_{\max}$ have led to misleading function graphs and subsequently to misinterpretations.

One possible outcome of applying uniform rank transformation is that the mark-mark scatterplot and the correlation functions do not significantly change after transformation. This implies that the original marks are good size descriptors and cannot be improved by simple rank ordering. The original marks m_i and the new marks u_i both characterize in some sense the "size" of the objects considered, where rank marks are perhaps closer to an abstract concept of size.

5. Conclusions

The mark-mark scatterplot, a simple graphical description of the variability of the marks of point pairs with an inter-point distance smaller than a certain limit $r_{\max}$, has proved useful for ecological research. This plot has the potential to close the gap between a given marked point pattern and the corresponding mark correlation functions. The mark-mark scatterplot is crucial to the comprehensive understanding and the critical validation of correlation functions and with the help of this new tool these summary statistics become more intelligible and trustworthy tools for ecologists.

Data accessibility

To facilitate the use of mark-mark scatterplots the authors have implemented the required computer routines in the new R package 'mmsc', which can be accessed on https://github.com/fballani/mmsc. The longleaf-pine data are available in the R package 'spatstat' (https://cran.r-project.org/package=spatstat), referred to as dataset 'longleaf'. The *Shorea* data can be accessed on https://forestgeo.si.edu/sites/asia/sinharaja and the Scots pine data are included in the 'mmsc' package.

Author contributions

F.B. and D.S. conceived the fundamental idea and all authors designed the methodology; F.B. implemented the methodology; all authors interpreted the results and wrote substantial parts of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Acknowledgments

The authors thank Thorsten Wiegand for constructive comments on an earlier version of this paper, for giving the *Shorea* data and for advice, which finally led to the contour lines of the mark-mark scatterplot. They are also grateful to Ravi K. Sheth for a discussion on the use of rank-ordered marks and to the referees for their valuable comments and suggestions.

References

Baddeley, A., Rubak, E., Turner, R., 2016. Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC, Boca Raton.
Busykin, A.I., Gavrikov, V.L., Sekretenko, O.P., Hlebopros, R.G., 1985. Analysis of Forest Cenoses. Nauka, Novosibirsk.
Chiu, S.N., Stoyan, D., Kendall, W.S., Mecke, J., 2013. Stochastic Geometry and its Applications, 3rd edn. J. Wiley and Sons, Chichester.
Cressie, N., 1991. Spatial Statistics, 2nd edn. J. Wiley and Sons, New York.
Ford, E.D., 1975. Competition and stand structure in some even-aged plant monocultures. J. Ecol. 63, 311–333.
Gavrikov, V., Stoyan, D., 1995. The use of marked point processes in ecological and environmental forest studies. Environ. Ecol. Stat. 2, 331–344.
Gavrikov, V., Grabarnik, P., Stoyan, D., 1993. Trunk-top relations in a Siberian pine forest. Biom. J. 35, 487–498.
Genet, A., Grabarnik, P., Sekretenko, O., Pothier, D., 2014. Incorporating the mechanisms underlying inter-tree competition into a random point process model to improve spatial tree pattern analysis in forestry. Ecol. Model. 288, 143–154.
Ghorbani, M., 2013. Cauchy cluster process. Metrika 76, 697–706.
Gunatilleke, C.V.S., Gunatilleke, I.A.U.N., Esufali, S., Harms, K.E., Ashton, P.M.S., Burslem, D.F.R.P., Ashton, P.S., 2006. Species-habitat associations in a Sri Lankan dipterocarp forest. J. Trop. Ecol. 22, 371–384.
Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. Statistical Analysis and Modelling of Spatial Point Patterns. J. Wiley and Sons, Chichester.
Mecke, K.R., Stoyan, D., 2005. Morphological characterisation of point patterns. Biom. J. 47, 473–488.
Pannatier, Y., 1996. Variowin. Software for Spatial Data Analysis. Springer, New York.
Platt, W.J., Rathbun, S.L., 1993. Dynamics of an old-growth longleaf pine population. In: Proceedings of the 18th Tall Timbers Fire Ecology Conference, Tall Timbers Research Station, Tallahassee, Florida, pp. 275–297.
Pommerening, A., Särkkä, A., 2013. What mark variograms tell about spatial plant interaction. Ecol. Model. 251, 64–72.
Pommerening, A., Uria-Diez, J., 2017. Do large forest trees tend towards high species mingling? Ecol. Inf. 42, 139–147.
Skibba, R.A., Sheth, R.K., Croton, D.J., Muldrew, S.I., Abbas, U., Pearce, F.R., Shattow, G.M., 2013. Measures of galaxy environment – II. Rank-ordered mark correlations. Mon. Not. R. Astron. Soc. 429, 458–468.
Stoyan, D., Penttinen, A., 2000. Recent applications of point process methods in forestry statistics. Stat. Sci. 15, 61–78.
Stoyan, D., Stoyan, H., 1996. Estimating pair-correlation functions of planar cluster processes. Biom. J. 38, 259–271.
Suzuki, S.N., Kachi, N., Suzuki, J.-I., 2008. Development of a local size hierarchy causes regular spacing of trees in an even-aged *Abies* forest: analyses using spatial autocorrelation and the mark correlation function. Ann. Bot. 102, 435–441.
Tanaka, U., Ogata, Y., Stoyan, D., 2008. Parameter estimation and model selection for Neyman-Scott point processes. Biom. J. 50, 43–57.
Wälder, O., Stoyan, D., 1996. On variograms in point process statistics. Biom. J. 38, 895–905.
Weiner, J., Solbrig, O.T., 1984. The meaning and measurement of size hierarchies in plant populations. Oecologia 61, 334–336.
Wiegand, T., Moloney, K.A., 2014. Handbook of Spatial Point Pattern Analysis. Chapman and Hall, Boca Raton.
Wiegand, T., Gunatilleke, C.V.S., Gunatilleke, I.A.U.N., Okuda, T., 2007. Analyzing the spatial structure of a Sri Lankan tree species with multiple scales of clustering. Ecology 88, 3088–3102.