Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Validation of point process predictions with proper scoring rules

Note no SAMBA/17/20

Authors Claudio Heinrich
Max Schneider
Peter Guttorp
Thordis Thorarinsdottir

Date 20th May 2020

The authors

Claudio Heinrich is Senior Research Scientist, Peter Guttorp is Professor II and Thordis L. Thorarinsdottir is Chief Research Scientist at the Norwegian Computing Center. Max Schneider is Ph.D. Student at the University of Washington, Seattle, U.S.A.

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Validation of point process predictions with proper scoring rules**
Authors	**Claudio Heinrich , Max Schneider , Peter Guttorp , Thordis Thorarinsdottir**
Date	20th May 2020
Publication number	SAMBA/17/20

Abstract

We introduce a class of proper scoring rules for evaluating spatial point process forecasts based on summary statistics. These scoring rules rely on Monte-Carlo approximation of an expectation and can therefore easily be evaluated for any point process model that can be simulated. In this regard, they are more flexible than the commonly used logarithmic score; they are also fruitful for evaluating the calibration of a model to specific aspects of a point process, such as its spatial distribution or tendency towards clustering. We show using simulations that our scoring rules are able to discern between competing models better than the logarithmic score. An application on growth in Pacific silver fir trees demonstrates the promise of our scores for scientific model selection.

Keywords	Point processes; Forecast verification; Proper scoring rules
Target group	Scientists
Availability	Open
Project	PointProcess
Project number	220708
Research field	Statistics
Number of pages	26
© Copyright	Norwegian Computing Center og forfatterne

1 Introduction

Point process methodology is applied in diverse scientific fields to model and predict earthquakes (Eberhard et al., 2012), crime rates (Mohler et al., 2011), urban development (Pourtaheri and Vahidi-Asl, 2011), plant and cellular systems (Waller et al., 2011), and animal colonies (Edelman, 2012), to name but a few examples. In prediction settings, model validation methods are needed to assess the predictive performance of competing models. Such methods may further provide goodness-of-fit diagnostics to identify potential shortcomings in either the fitted model or the underlying scientific hypothesis. Previously proposed model validation methods for point processes include the work of Baddeley et al. (2005) and Zhuang (2006), who define a pixel-based residual diagnostic framework similar to that commonly applied to Poisson log-linear regression, pattern transformation methods as reviewed in Clements et al. (2011), and diagnostics based on summary statistics such as the K-function (Baddeley et al., 2011).

There is a variety of applications such as earthquake rate forecasting (Schorlemmer et al., 2018) or distribution of species where spatial point process models are used as probabilistic forecasts for spatial point patterns, and forecast validation techniques are required. Forecast validation differs from model fitting in that it requires a predictive distribution being issued unknowing the observation, whereas in model fitting it is typically assessed how good a parametric model can explain known observations. Probabilistic forecasts are commonly evaluated by means of scoring rules. A scoring rule comprises the information contained in a predictive distribution F and the target observation y into a single number, the score $S(y, F)$. A unified theory of scoring rules has been established in the two papers Gneiting et al. (2007); Gneiting and Raftery (2007). They argue that scoring rules ought to be *proper*, i.e. the expected score should get optimal when the true distribution of the observation is predicted.

In the theory of spatial point processes, the use of scoring rules has so far been limited to the logarithmic score, which is the negative log-likelihood of the predictive model at the observation, $S_{\log}(y, f) := \log(f(y))$ where f is the predictive density, see Daley and Vere-Jones (2004). This is perhaps due to the complexity of the observation space, which makes it challenging to construct proper scoring rules for point processes that are useful in practice.

Motivated by the lack of scoring rules in the point process literature, we introduce a new class of proper scoring rules by combining well-known summary statistics for point processes with the continuous ranked probability score (CRPS). Unlike the logarithmic score, these scores can be approximated by Monte-Carlo methods, and do not require knowledge of the exact density function of the process. This is a substantial advantage in the context of spatial point processes, as for many common models, densities are only known up to untractable normalizing constants.

Our scores rely on summary statistics such as the intensity function or Ripley's K-function that are well-known to practicioners in the field. This makes the scores easily interpretable in the sense that a better score indicates that the summary statistic at hand is in good agreement between the predictive distribution and a given observation. Importantly, different summary statistics can be used in order to target different properties of the point process, such as homogeneity or clustering. We demonstrate the favourable performance of these scoring rules in a simulation study.

The construction of the scores is based on a simple proposition stating essentially that proper scores remain proper under measurable mappings. We believe this proposition to be useful beyond point processes, as it can generally be used to construct proper scoring rules when the observation space is complicated, simply by mapping the observation space into a simpler space, where proper scores do exist.

Proper scoring rules necessarily provide less information than many diagnostic tools, since they reduce observation and predictive distribution to just a single number. On the other hand, this can often be an advantage, as they make it easy for decision-makers to discriminate between competing models that have potentially very different characteristics. Generally, no parametric assumptions are posed on the predictive distributions. Propriety prevents hedgeing and encourages forecasters to report their true beliefs and thus ensures that decision-theoretic principles are obeyed.

There are many diagnostic tools for the validation of point process models available in the literature (see overviews from the perspective of earthquake modeling can be found in Bray and Schoenberg (2013); Gordon et al. (2015). In the following we briefly discuss some of these tools, and explain why we believe that proper scoring rules are a valuable addition to this tool set.

Residual plots were introduced in Baddeley et al. (2005) and are powerful for detecting misspecifications of fitted point process models. Several residuals can be defined which typically have mean 0 when the fitted model is correct. Systematic deviations from 0 therefore indicate miscalibration, and a range of diagnostic plots based on residuals can be considered. While these can unveil systematic model errors, it is often difficult to compare different models based on these plots. According to their diagnostic nature, they do not provide an objective measure allowing directly to infer which of several models is calibrated best.

Other commonly applied evaluation tools are various information criteria (e.g. AIC and BIC). These are closely related to the logarithmic score, but apply an additional penalty to the number of parameters used by the models. Therefore, they suffer from similar limitations as the logarithmic score and cannot be computed for some point process models. Dawid (1984) argues that, in a standard Bayesian setting where the predictive distribution is known up to several parameters, the distribution of which is updated whenever a new observation becomes available, the BIC and the logarithmic score are asymptotically equivalent, in the sense that their quotient converges to 1 as the number of observations goes to ∞. We refer to (Gneiting and Raftery, 2007, Section 7) for more details. Generally, penalizing the number of parameters is uncommon in the context of forecast validation, as the number of parameters might depend on how the competing models are parametrized which might lead to ambiguities. In this setting, overfitting by choosing models with too many parameters is automatically punished, since models are fitted in-sample and then compared to out-of-sample observations.

There are, moreover, several tests whether an observed point pattern is likely to be a random draw of an issued predictive distribution. These tests mostly originating from earthquake forecasting, see Schorlemmer et al. (2018) and references therein. They measure the agreement between a forecast and an observation set, outputting simulation-based p-values, that can indicate a significant disagreement. They do, however, not allow for model comparison, as a higher p-value cannot be directly interpreted as better performance. Some comparative tests and metrics also exist (e.g. Clements et al. (2011) and Baddeley et al. (2011)), some based on residual scores from Baddeley et al. (2005); however, these do not focus on specific properties of interest on point process model performance (e.g spatial distribution or clustering). They also do not provide an objective score by which more than two models may be ranked.

This article is organized as follows. Section 2 contains the theoretical background, including a brief introduction to proper scoring rules and spatial point processes. In Section 3 we derive proper scoring rules for point processes based on summary statistics. Section 4 provides simulation studies analyzing the performance of the introduced scores. In Section 5 we apply the summary statistic scores to an example data set. Section 6 concludes.

2 Proper scoring rules and point processes

Scoring rules assess the accuracy of probabilistic forecasts by assigning a numerical penalty to each forecast-observation pair. Given a measurable observation space $\mathcal{O}$ and a set $\mathcal{P}$ of probability measures on $\mathcal{O}$, a scoring rule is a mapping

$$S : \mathcal{O} \times \mathcal{P} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}, \tag{2.1}$$

such that the mapping $y \mapsto S(y, F)$ is integrable with respect to the measure G for every $F, G \in \mathcal{P}$. We generally assume scoring rules to be negatively oriented, interpreting the score as a penalty, such that smaller scores indicate better predictions.

A scoring rule is *proper* relative to $\mathcal{P}$ if

$$\mathbb{E}_G S(Y, G) \leq \mathbb{E}_G S(Y, F) \quad \text{for all } F, G \in \mathcal{P}, \tag{2.2}$$

that is, if the expected score for a random observation Y with distribution G is optimized if the true distribution is issued as the forecast, F. The scoring rule is *strictly proper* relative to the class $\mathcal{P}$ if (2.2) holds, with equality only if $F = G$. Forecast evaluation based on proper scoring rules encourages an honest forecast and prevents hedging to how the forecast is evaluated. That is, the perceived performance cannot be improved by a willful divergence of the forecast from the true distribution; see e.g. the discussion in Section 1 of Gneiting (2011).

Competing forecasting methods can be compared by evaluating their mean scores over an out-of-sample test set, and the method with the smallest mean score is preferred. For a small set of forecast-observation pairs, the mean score is commonly associated with a large uncertainty, see Thorarinsdottir and Schuhen (2018). Formal tests of the null hypothesis of equal predictive performance can also be employed, such as the Diebold-Mariano test (Diebold and Mariano, 1995) or permutation tests (Good, 2013).

We consider scoring rules for spatial point processes on $\mathbb{R}^d$ with $d = 2, 3, \ldots$, with a bounded observation window $W \subset \mathbb{R}^d$. The observation space $\mathcal{O}$ is the space of countable subsets of W, which we will denote by $W^{\cup}$. A *spatial point process* $\mathbf{X}$ is a random variable taking values in $W^{\cup}$ with almost surely finitely many points. We use $F, G, \ldots$ to denote distributions of point processes. Lowercase bold letters denote non-random point patterns, and in particular $\mathbf{y}$ is used to denote the observation. The definition (2.2) relies on the construct of a random observation, which we will denote $\mathbf{Y}$. For some function f, the notation $\mathbb{E}_F[f(\mathbf{X})]$ is used to denote the expectation of $f(\mathbf{X})$ when $\mathbf{X}$ is distributed according to F. For a comprehensive overview on spatial point processes, we refer to Daley and Vere-Jones (2007); Møller and Waagepetersen (2003).

Summary statistics of point processes are powerful tools for exploratory data analysis and model selection. We introduce two important examples.

Example 2.1 (Intensity function)**.** *The intensity function $\lambda : W \to \mathbb{R}$ of a point process model F is defined by the property*

$$\int_B \lambda(w)\, dw = \mathbb{E}_F[n(\mathbf{X} \cap B)],$$

for all measurable sets $B \subset W$. Here, $n(\mathbf{X} \cap B)$ denotes the number of points of $\mathbf{X}$ that fall into the set B.

The intensity measures the spatial distribution of points in the sense that a high intensity highlights areas where more points are expected. Whereas Poisson point processes are fully defined by their intensity, the intensity contains no information about interaction of points, i.e. whether the points repel each other or tend to cluster. This interaction behaviour is analyzed by Ripley's K-function, see Baddeley et al. (2000).

Example 2.2 (Ripley's K-function)**.** *For a point process F with intensity λ, Ripley's K-function is defined as*

$$K(r) = \frac{1}{|W|}\mathbb{E}_F\left[\sum_{\substack{x_1, x_2 \in \mathbf{X}, \\ x_1 \neq x_2}} \frac{\mathbb{1}\{\|x_1 - x_2\| < r\}}{\lambda(x_1)\lambda(x_2)}\right],$$

for $r > 0$.

Roughly speaking, $K(r)$ indicates clustering at distances up to r. The K-function of a Poisson process is $K(r) = \frac{2\pi^{d/2}}{\Gamma(d/2)d} r^d$. If, for a point process model, $K(r)$ is larger than this value for small r, the model has more expected point pairs with distance less than r than a Poisson model, and the process exhibits clustering. Other examps of popular summary statistics include the F-, G-, and J-functions as well as the second order intensity. For more details we refer to Møller and Waagepetersen (2003, Chapter 4).

Bearing the two examples above in mind, we make the following definition. Note that in both examples the summary statistic is function-valued, taking values from a space $\mathcal{R}$. For the intensity function, we have $\mathcal{R} = W$ and for the K-function, we have $\mathcal{R} = (0, \infty)$.

Definition 2.1. *Consider a class of predictive distributions $\mathcal{P}$ on $W^{\cup}$ and a measurable space $\mathcal{R}$. A summary statistic is a mapping $T : \mathcal{P} \times \mathcal{R} \to \mathbb{R}$. We sometimes denote $T_F(r)$ instead of $T(F, r)$. A summary statistic estimator is a mapping $\widehat{T} : W^{\cup} \times \mathcal{R} \to \mathbb{R}$.*

In particular, we assume estimators for summary statistics to be based on a single point pattern, which is the case for all standard estimators for the summary statistics mentioned above.

Not all summary statistics are well-defined for all point process models. For example, the K-function is only well-defined for second order reweighted stationary processes, see Baddeley et al. (2000). In view of Definition 2.1, let us remark that throughout this paper we assume all mappings between measurable spaces to be measurable. Products of measurable spaces are equipped with the product σ-algebra. For mappings $\phi : \mathcal{P} \times \mathcal{M} \to \mathcal{M}'$, where $\mathcal{M}, \mathcal{M}'$ are measurable spaces and $\mathcal{P}$ is the space of predictive distributions, we assume that $\phi(F, \cdot) : \mathcal{M} \to \mathcal{M}'$ is measurable for all $F \in \mathcal{P}$.

3 Proper scoring rules based on summary statistics

When dealing with forecasts taking values in a complex observation space $\mathcal{O}$, we can more effectively validate and compare models by focusing on a certain property of interest. This approach is not new; in the context of multivariate forecasts, the Dawid-Sebastiani score (Dawid and Sebastiani, 1999) that focuses on mean and covariance of a multivariate forecast, and the variogram score (Scheuerer and Hamill, 2015) focuses on how the spatial autocorrelation of a spatial prediction decays with distance. We adapt this principle and show how it can be applied to validate point process forecasts.

In point process forecasts, researchers often focus on the number and spatial distribution of points or clustering behavior of the process. We can use summary statistics for such specific properties to construct proper scoring rules sensitive to the same property. This approach has several advantages. It is easily applicable and does not impose any conditions on the predictive distribution. Thus, it can be used to directly compare predictive performance of any collection of point process models. Secondly, the derived scoring rules are always proper, and therefore allow for easy comparison of predictive perfomance following decision-theoretic principles.

We now provide several results that can be used to construct proper scoring rules from summary statistics. We denote in the following by $\mathcal{P}$ an arbitrary but fixed class of predictive distributions, and speak of propriety rather than propriety relative to $\mathcal{P}$.

Proposition 3.1. *Let $r \in \mathcal{R}$ be fixed. Assume that $\widehat{T}$ is an unbiased estimator for T in the sense that $\mathbb{E}_F[\widehat{T}(\mathbf{Y}, r)] = T(F, r)$ for all $F \in \mathcal{P}$. Then the scoring rule*

$$S_T(\mathbf{y}, F, r) := (\widehat{T}(\mathbf{y}, r) - T(F, r))^2$$

is proper.

Proof. This follows directly from the fact that for any random variable Y, the function $c \mapsto \mathbb{E}[(Y-c)^2]$ gets minimal in $c = \mathbb{E}[Y]$. □

The score S_T is usually not strictly proper as we may have $T(F, r) = T(G, r)$ for distributions $F \neq G$, see e.g. Baddeley and Silverman (1984).

In this proposition, both $\widehat{T}$ and T get evaluated at a specific point $r \in \mathcal{R}$, whereas in practice we will be more interested in an overall fit across all points. To this end, we can use the following result, which is an immediate consequence of Tonelli's theorem.

Proposition 3.2. *Let $A \subset \mathcal{R}$ be measurable. If $S(\mathbf{y}, F, r)$ is a non-negative proper scoring rule for all $r \in A$, then*

$$S_A(\mathbf{y}, F) := \int_A S(\mathbf{y}, F, r) dr \tag{3.1}$$

is a proper scoring rule.

Note that non-negativity is not required in the case that the integral in (3.1) exists (possibly taking the value $+\infty$) for all $\mathbf{y}$ and F. These two propositions readily allow the construction of proper scoring rules based on summary statistics when their estimators are unbiased.

Example 3.1. *$\mathcal{F}$-function: The $\mathcal{F}$- or empty-space-function is defined for stationary point processes as the distribution function of the distance from the origin to the nearest point in $\mathbf{X}$. It has the following unbiased estimator:*

$$\widehat{\mathcal{F}}(\mathbf{y}, r) := \sum_{x \in I_r} \frac{\mathbb{1}\{d(x, \mathbf{y}) \leq r\}}{\# I_r},$$

where I is any finite regular grid of points, $I_r := I \cap W_{\ominus r}$, and $W_{\ominus r} = \{w \in W : b(w, r) \subset W\}$, see (Møller and Waagepetersen, 2003, section 4.3). We obtain a scoring rule of predictive distribution F based on the empty-space-function by

$$S_{\mathcal{F}}(\mathbf{y}, F) := \int_0^R (\widehat{\mathcal{F}}(\mathbf{y}, r) - \mathcal{F}_F(r))^2 \, dr,$$

where R is an upper limit that should be chosen to be small relative to the diameter of W. By Propositions 3.1 and 3.2, this scoring rule is proper with respect to the class $\mathcal{P}$ of all stationary point process models.

Proposition 3.1 is quite intuitive, as it compares the estimator $\widehat{T}$ to the true value T_F under the predictive distribution. It comes with two serious restrictions, though. The first is that for many summary statistics, such as for example the K-function and the intensity function, the standard estimators are not unbiased, see Møller and Waagepetersen (2003, Chapter 4). Secondly, even if an unbiased estimator exists, closed form expressions for a given summary statistic T_F are usually only available for selected point process models.

Both of these weaknesses can be overcome by replacing T_F by $\widehat{T}(F)$, the pushforward probability measure of F under the estimator $\widehat{T}$.

Proposition 3.3. *Let $r \in \mathcal{R}$ be fixed. Denote by $\widehat{T}(F, r)$ the pushforward distribution of F under $\widehat{T}(\cdot, r)$. Consider a non-negative scoring rule S on $\mathbb{R}$ that is proper relative to $\widehat{T}(\mathcal{P}) := \{\widehat{T}(F, r)\,,\ F \in \mathcal{P}, r \in \mathcal{R}\}$. Then, the scoring rule*

$$S_{\widehat{T}}(\mathbf{y}, F, r) := S(\widehat{T}(\mathbf{y}, r), \widehat{T}(F, r))$$

is proper.

Proof. This is a direct consequence of the change-of-variables formula. □

Note that $S_{\widehat{T}}$ is usually not strictly proper, even if S is, since we might have $\widehat{T}(F, r) = \widehat{T}(G, r)$ for distributions $F \neq G$. The key for making this result useful is the choice of the proper scoring rule S on the real line. Note that we recover Proposition 3.1 if we choose S to be the mean square error, $S(y, F) = (y - \mathbb{E}_F[X])^2$. However, a preferable choice is the continuous ranked probability score (CRPS) as it is *strictly* proper with respect to all distributions with finite first moment, see Gneiting and Raftery (2007). Moreover, choosing the CRPS allows to approximate $S_{\widehat{T}}$ without requiring detailed knowledge of the pushforward measure $\widehat{T}(F)$. The CRPS, introduced in X, is defined by the formula

$$\mathrm{CRPS}(y, F) := \mathbb{E}_F[|y - X|] - \frac{1}{2}\mathbb{E}_F[|X' - X|],$$

where in the second summand X and X' are independent random variables distributed according to F. When applying Proposition 3.3 with the CRPS, we obtain by the change-of-variables formula, supressing r for brevity,

$$\begin{aligned} S_{\widehat{T}}(\mathbf{y}, F) &= \mathbb{E}_{\widehat{T}(F)}[|\widehat{T}(\mathbf{y}) - X|] - \frac{1}{2}\mathbb{E}_{\widehat{T}(F)}[|X' - X|] \\ &= \mathbb{E}_F[|\widehat{T}(\mathbf{y}) - \widehat{T}(\mathbf{X})|] - \frac{1}{2}\mathbb{E}_F[|\widehat{T}(\mathbf{X}') - \widehat{T}(\mathbf{X})|], \end{aligned}$$

where in the last line, $\mathbf{X}'$ and $\mathbf{X}$ are independent point processes with distribution F. This expression can easily be approximated by Monte-Carlo sampling from the point process distribution F. Therefore, we obtain a scoring rule that is proper and can be computed for any point process distribution by sampling.

Another, somewhat surprising, advantage of this approach is that by using $\widehat{T}(F)$ rather than T_F, the score often can discriminate better between distributions. The reason is that for different predictive models F_1 and F_2 we may have $T_{F_1} = T_{F_2}$ but $\widehat{T}(F_1) \neq \widehat{T}(F_2)$. In this case, since the CRPS is strictly proper, the true distribution will be preferred. This effect can be observed in our simulation study in the next section, where we apply the scoring rule based on the K-function estimator to different Poisson models. These models have identical theoretical K-functions, but, nevertheless, the score gets minimized when the correct model is predicted, since $\widehat{K}$ follows different distributions under the different models.

We sum up the main result of this section in the following corollary of Propositions 3.2 and 3.3.

Corollary 3.1 (summary statistic score)**.** *Consider an estimator for a summary statistic $\widehat{T}$ that is integrable with respect to $F \otimes dr$ for all F in $\mathcal{P}$. The scoring rule defined by*

$$S_{\widehat{T}}(\mathbf{y}, F) := \mathbb{E}_F\left[\int_{\mathcal{R}} |\widehat{T}(\mathbf{y}, r) - \widehat{T}(\mathbf{X}, r)|\, dr\right] - \frac{1}{2}\mathbb{E}_F\left[\int_{\mathcal{R}} |\widehat{T}(\mathbf{X}', r) - \widehat{T}(\mathbf{X}, r)|\, dr\right]$$

is proper.

An advantage of this score are the weak assumptions that are required, namely just the integrability of $\widehat{T}$, which is satisfied for most point process models and summary statistic estimates. Note that $\widehat{T}$ can be any real- or function-valued mapping satisfying these conditions, and no connection to an underlying summary statistic T is required. As a consequence, the constructed proper scoring rule can be considered even for predictive distributions for which the underlying summary statistic T does not exist. An example is the scoring rule $S_{\widehat{K}}$ considered in Example 3.3 below, which may be computed (and remains proper) even for point processes that are not second order intensity reweighted stationary, which is a necessary condition for the K-function to exist. In such a scenario, the score will nevertheless be sensitive to misspecification of the clustering behaviour, since $\widehat{K}$ is.

The following statistics will be used to show the efficacy of our approach for point process model evaluation.

Example 3.2 (Intensity score)**.** *The intensity function λ of a point process is typically estimated by kernel estimators. These estimators are generally biased, making it impossible to apply Proposition 3.1. For a kernel k (i.e. a density on W) and a bandwidth $b > 0$, the kernel intensity estimator is based on the rescaled kernel $k_b(w) := b^{-2}k(w/b)$. It is defined as*

$$\widehat{\lambda}(\mathbf{y}, w) = \sum_{y \in \mathbf{y}} k_b(w - y)/c_{W,b}(y),$$

where $c_{W,b}$ are edge correction factors defined as $c_{W,b}(y) = \int_W k_b(w - y)\, dw$. By Corollary 3.1, the intensity score defined as

$$S_{\widehat{\lambda}}(\mathbf{y}, F) := \mathbb{E}_F\left[\int_W |\widehat{\lambda}(\mathbf{y}, w) - \widehat{\lambda}(\mathbf{X}, w)|\, dw\right] - \frac{1}{2}\mathbb{E}_F\left[\int_W |\widehat{\lambda}(\mathbf{X}', w) - \widehat{\lambda}(\mathbf{X}, w)| dw\right]$$

constitutes a proper scoring rule.

Since this score targets the intensity function, it assesses, roughly speaking, whether the predictive distribution has the correct spatial distribution and number of points, but neglects point interactions. Let us stress again that, unlike the logarithmic score, this scoring rule can be computed for any point process model, in particular also for models defined by a density with an untractable normalizing constant. On the other hand, if we are more interested whether a predictive model reflects point interaction correctly, we can score it using an estimator for the K-function.

Example 3.3. *[Ripley's K-function] The standard estimator for Ripley's K-function is defined as*

$$\widehat{K}(\mathbf{y}, r) := \sum_{y_1 \neq y_2 \in \mathbf{y}} \frac{\mathbb{1}\{|y_1 - y_2| < r\}}{\widehat{\lambda}(y_1)\widehat{\lambda}(y_2)|W \cap W_{y_1 - y_2}|},$$

where $W_{y_1-y_2}$ denotes the shifted set $W + y_1 - y_2$, and $\widehat{\lambda}$ is a kernel estimator for the intensity. Thus, we obtain the proper K-function score

$$S_{\widehat{K}}(\mathbf{y}, F) := \int_0^R \mathbb{E}_F[|\widehat{K}(\mathbf{y}, r) - \widehat{K}(\mathbf{X}, r)|]\, dr - \frac{1}{2}\int_0^R \mathbb{E}_F[|\widehat{K}(\mathbf{X}', r) - \widehat{K}(\mathbf{X}, r)|] dr,$$

where R is an upper limit that should be chosen small relative to the diameter of W.

As $\widehat{K}$ is sensitive to point interaction, this scoring rule specifically targets correct representation of point interaction in the predictive model. On the other hand, it will be relatively insensitive to misspecification of the intensity function, and for example, be inadequate for differentiating between different Poisson processes, which have the same K-function.

4 Simulation study

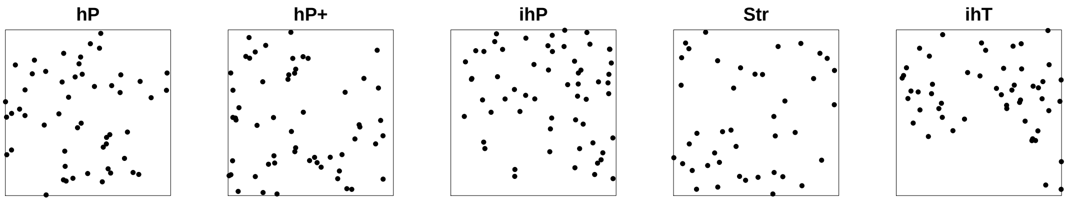
In order to demonstrate the usefulness of our scores, we present a simulation study where we consider five different point process models. Their characteristics are summarized in Figure 1, which also shows an example realization of each model. The spatial window considered is $W = [0, 10] \times [0, 10]$. The first two models are homogeneous Poisson processes with 50 and 60 expected points in the considered area, respectively. Model 3 is an inhomogeneous Poisson process with 50 expected points, and with an intensity that increases linearly in the distance from the lower left corner of the window. Model 4 is a homogeneous Strauss process, i.e. the points repel each other and the typical point pattern is more regular than for the homogeneous Poisson process. The Strauss process is defined by its density

$$f(\mathbf{x}) = c\beta^{n(\mathbf{x})}\gamma^{s_R(\mathbf{x})},$$

where c is a normalizing constant, $\beta > 0, R > 0$, and $\gamma \in (0, 1)$ are parameters, $n(\mathbf{x})$ denotes the number of points in $\mathbf{x}$ and $s_R(\mathbf{x})$ is the number of pairs of points in the pattern $\mathbf{x}$ with distance less than R. The value R is the range of interaction between points, and γ determines the strength of the interaction, with smaller γ leading to stronger inhibition between close points. We choose $\gamma = 0.5$ and $R = 1$. By letting $\beta = 1.15$, we obtain an expected number of points of approximately 50, the same as for models 1 and 3.

As fifth model we consider an inhomogeneous Thomas process, which is constructed by generate an (invisible) Poisson process of parent points, and then letting each parent generating a random number of offsprings that are spatially distributed according to a Gaussian kernel centered at the parent point. We choose an inhomogeneous parent process, with the same intensity as model 3, divided by 2. The number of offsprings per parent is Poisson distributed with mean 2 and the standard deviation for the Gaussian kernel is set to 0.5. By these choices, the Thomas process has an intensity similar to model 3, and the number of expected points in the observation window is again approximately 50.

We consider each of the models both as true distribution G and as predictive distribution F, for a total of 25 combinations of true and predictive distributions (see Figure 1). For



Name	Model	Intensity	Point interaction	$\mathbb{E}[n(\mathbf{X})]$
hP	homogeneous Poisson	$\sim c$	none	50
hP+	homogeneous Poisson	$\sim \frac{6}{5}c$	none	60
ihP	inhomogeneous Poisson	$\sim \sqrt{x^2+y^2}$	none	50
Str	homogeneous Strauss	$\sim c$	inhibition	≈ 50
ihT	inhomogeneous Thomas	$\sim \sqrt{x^2+y^2}$	clustering	≈ 50

Figure 1. Example plots and characteristics of the considered models. The considered spatial window is $[0,10] \times [0,10]$, i.e. for the inhomogeneous processes, the intensity increases with distance from the lower left corner.

each combination we compute $\mathbb{E}_G[S_{\widehat{T}}(\mathbf{Y},F)]$ by simulating 100 i.i.d. copies of $\mathbf{Y} \sim G$ and averaging $S_{\widehat{T}}(\mathbf{Y},F)$. For the computation of $S_{\widehat{T}}(\mathbf{Y},F)$ the expectations are approximated by simulating 100 i.i.d. copies of $\mathbf{X} \sim F$. As summary statistic estimator $\widehat{T}$ we consider both the kernel estimator $\widehat{\lambda}$ and the K-function estimator $\widehat{K}$. The computations are carried out using the R-package `spatstat` (Baddeley et al., 2015). For the kernel density estimator an isotropic Gaussian kernel is used with standard deviation 1.25 and bandwidth 1, which are the default values of the `spatstat` function `density.ppp`.

The results are presented in Figure 2. For both scoring rules the expected score is minimized under all distributions when the true model is predicted. Not surprisingly, the scores are sensitive to the underlying summary statistic. For example the score $S_{\widehat{\lambda}}$ has difficulties to differentiate between the homogeneous Poisson model hP and the Strauss model which have the same intensity, but can clearly differentiate between homogeneous and inhomogeneous models.

On the other hand, the score $S_{\widehat{K}}$ is capable of detecting mismatches in the point interaction between predictive and true distribution, and can in particular differentiate between the Strauss and the hP model. It is therefore important to be aware of which property of the process is targeted by the scoring rule, and, in practice, to use multiple scoring rules to assess predictive skill. Figure 1 shows the results of permutation tests assessing the significance of the difference of the mean scores. The differences between the K-function scores for the different Poisson models is not significant at a 5% level, and especially the difference between the two homogeneous models hP and hP+ is not reliably picked up. However, given that these models have the same theoretical K-function, it is still noteworthy that the mean score of $S_{\widehat{K}}$ is minimized when the true distribution is predicted (see Figure 2). This is due to the fact that the distribution of $\widehat{K}$ still varies between these models, which is detected by the CRPS. At the same time, the score $S_{\widehat{\lambda}}$ reliably tells the Poisson models apart. In addition, by basing inference on both scoring rules, the true

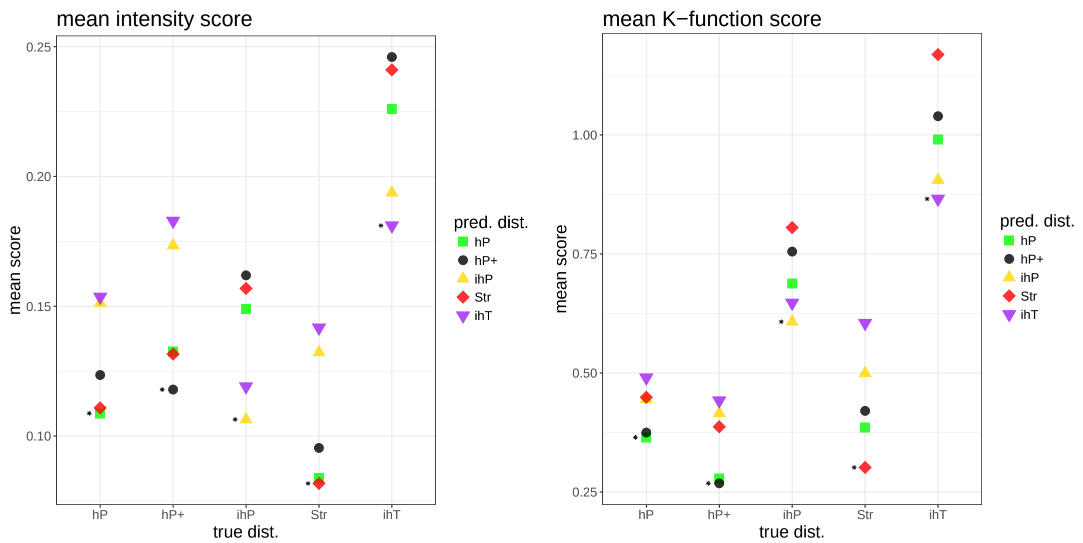


Figure 2. The mean scores $\mathbb{E}_G[S_{\widehat{\lambda}}(\mathbf{Y},F)]$ (left hand side) and $\mathbb{E}_G[S_{\widehat{K}}(\mathbf{Y},F)]$ (right hand side) for each combination of the 5 considered models. The x-axis shows the true distribution G of the data, and the score for the correct distribution is additionally marked by a star.

$S_{\widehat{\lambda}}(Y,F)$

F :		hP	hP+	ihP	Str	ihT
	hP	-	<0.1	<0.1	**8.0**	<0.1
	hP+	<0.1	-	<0.1	<0.1	<0.1
G	ihP	<0.1	<0.1	-	<0.1	**8.4**
	Str	**15.1**	<0.1	<0.1	-	<0.1
	ihT	<0.1	<0.1	4.4	<0.1	-

$S_{\widehat{K}}(Y,F)$

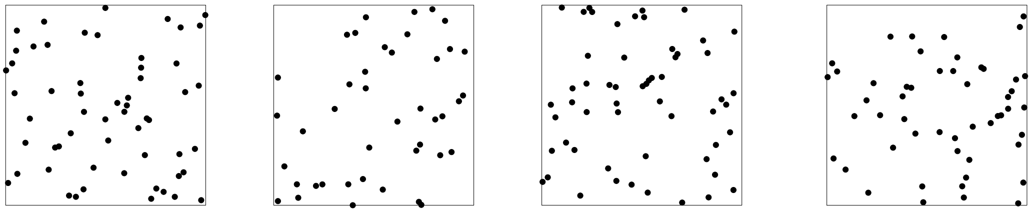
F :		hP	hP+	ihP	Str	ihT
	hP	-	**38.3**	**17.6**	<0.1	<0.1
	hP+	**38.9**	-	**16.7**	<0.1	<0.1
G	ihP	**8.2**	**6.2**	-	<0.1	0.5
	Str	<0.1	<0.1	<0.1	-	<0.1
	ihT	<0.1	<0.1	0.7	<0.1	-

Table 1. p-values in $\%$ of a permutation test assessing the significance of the difference between the score of predictive distribution F and the score of the true distribution G. Values above 5% (in bold) indicate nonsignificance, and the corresponding score cannot reliably distinguish between F and G.

model can easily and clearly be identified in all cases. This is quite remarkable, since the models are chosen to challenge validation tools and are in particular often difficult to tell apart by eye, see Figure 1. Finally, let us note that the logarithmic score could only be computed for the first three models, as the densities for the Strauss and the Thomas process have intractable normalizing constants.

We next address the question how the intensity score compares to the logarithmic score when different predictive Poisson models are compared. The model descriptions and example plots can be found in in Figure 3. Figure 4 shows boxplots for $S_{\log}$ and $S_{\widehat{\lambda}}$, as well as boxplots of bootstrap resamples of the expected score based on 50 and 500 observations, respectively.

The results show that the intensity score is more sensitive than the logarithmic score and can more reliably identify the true distribution, when we use more than one observation. Accordingly, when significance of score differences is assessed by permutation tests, using the intensity score leads to tests with a higher power, more details can be found in the



$F_1: \ \lambda(x,y) \equiv \frac{1}{2}$, $F_2: \ \lambda(x,y) \equiv \frac{2}{5}$, $F_3: \ \lambda(x,y) \equiv \frac{3}{5}$, $F_4: \lambda(x,y) = \frac{x}{20} + 0.25$,

Figure 3. The four different Poisson models considered in the second study. The spatial window is $[0,10] \times [0,10]$. The true distribution of the data, F_1, is a homogeneous process with 50 expected points, F_2 and F_3 are homogeneous processes with 40 and 60 expected points, respectively. F_4 is an inhomogeneous process with intensity that increases linearly in x from 0.25 to 0.75.

appendix. The reason is that $S_{\log}$ is a local proper scoring rule that solely depends on the value of the predictive density at the observation, whereas the CRPS and therefore $S_{\hat{\lambda}}$ depends always on the full predictive distribution. For Poisson point processes, evaluating the logarithmic score is less computationally involved than evaluating the intensity score, which relies on Monte-Carlo approximation of a (multiple) integral. However, in typical spatial point process applications there are often limited observations available, either because generating/collecting observations is involved (e.g. in ecology or epidemiology) or because new observations take several years to materialize (e.g. in earthquake rate prediction). In such cases, there is a substantial benefit from performing model selection with more robust scores, even if it comes at additional computational costs. Let us finally remark that the values of the y-axis in Figure 4 bear no inherent meaning. As typically the case for scoring rules, should only be used for comparing different predictive models.

5 Application to *Abies amabilis* forests

In ecology and forestry spatial point processes have been increasing in popularity over the last decades, as they allow to address many key questions such as local dominance of species or species area relationships, see Velázquez et al. (2016); Wiegand et al. (2017) for overviews. Law et al. (2009) review the use of summary statistics and conclude that the pair correlation function is particularly popular, as it is a natural measure for what ecologists call the 'plant's-eye view'. The authors of Wiegand et al. (2013) carefully analyze the information content of different summary statistics for use in ecology. They conclude that the pair correlation function is the most informative second order summary statistic in this context.

Here, we study location data of *Abies amabilis* (Pacific silver fir) at eight disjoint 6 by 6 meter plots at Findley Lake Reserve in Washington State, USA. See Grier et al. (1981) for a description of the site conditions. Figure 5 shows the location of trees at two of the plots for three different time points over 31 years. The area was clear-cut in 1957; the trees in our dataset were apparently present as seedlings before the clear-cut and there appears

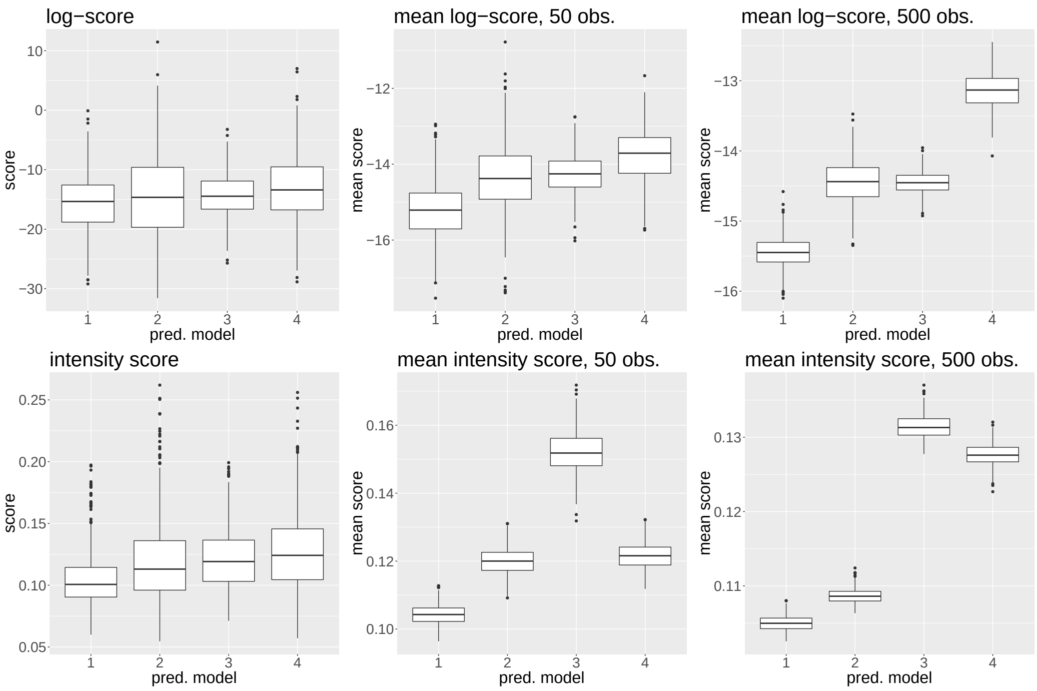


Figure 4. The first column shows boxplots for $S_{\log}(\mathbf{y}, F)$ and $S_{\widehat{\lambda}}(\mathbf{y}, F)$ for the four different predictive Poisson models described in Table 3. Model 1 is the true distribution of the data. Columns two and three show boxplots for bootstrap resamples of the expected scores $\mathbb{E}_G[S_{\log}(\mathbf{y}, F)]$ and $\mathbb{E}_G[S_{\widehat{\lambda}}(\mathbf{y}, F)]$ computed from 50 and 500 observations, respectively.

to have been no reproduction in the stand since then. The first observation was made in 1978, 21 years after the area was clear-cut. On average, roughly 80% of the original trees were still present in the second observation in 1990 and approximately 25% of them were present in the third observation in 2009. The data was previously studied by Sorrensen-Cothern et al. (1993) who investigated the development of tree crown structure under competition for light.

The observations for each year are analyzed separately such that in each analysis, eight independent realizations of the underlying process are available (the eight forest stands' data in that year). We consider a spatial prediction scenario where we are given four of the observations in order to fit a predictive distribution, which is then verified against the other four observations using the intensity- and the K-function score. This is repeated for all possible combinations of choosing four training stands out of 8, leading to a total of $4 \times \binom{8}{4} = 280$ score evaluations for each model and each year. Some of the scores are correlated, namely if they are based on the same observation (but use different training sets) or if their training sets overlap.

Visual examination of the point patterns show no sign of anisotropy and indicates that they exhibit clustering, particularly for the years 1978 and 1990 (see Figure 5). Therefore, we consider various isotropic cluster point processes as predictive models, namely a log-Gaussian Cox process and 4 different models of Neyman-Scott type, the Thomas, Matérn cluster, variance Gamma and Cauchy model. Moreover, a homogeneous Poisson model is considered as a benchmark. The log-Gaussian Cox process (LGCP) has a driving field

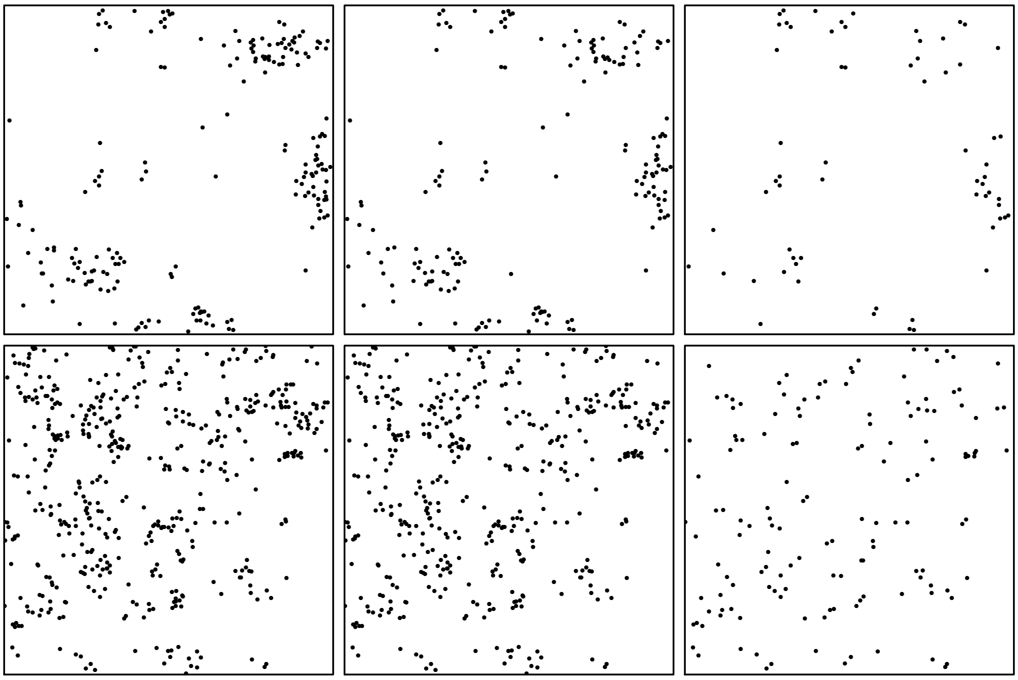


Figure 5. *Abies amabilis* (Pacific silver fir) at two disjoint 6 by 6 meter plots at Findley Lake Reserve in Washington State (rows). The area was clear-cut in 1957; the first column shows trees present in 1978, 21 years after the clear-cut; the second column shows the trees still present in 1990, 33 years after the clear-cut; the third column shows the remaining trees in 2009, 52 years after the clear-cut.

$\{Z(w)\}_{w \in W}$, which is the exponential of a homogeneous Gaussian field with, for our purpose, exponential autocovariance function

$$C(r) = \tau^2 \exp(-r/\sigma)$$

, with variance parameter τ and scale parameter σ. Conditional on Z, the LGCP is then distributed as a Poisson process with intensity Z. Neyman-Scott models have an invisible homogeneous Poisson parent process $\mathbf{U}$ and for each parent point $u \in \mathbf{U}$ an offspring process is considered, which is Poissonian with intensity $\lambda_u(w) = \alpha k(u - w)$, where k is an isotropic kernel. The Neyman-Scott process is then the union of all offspring processes, i.e. conditional on $\mathbf{U}$, it is a Poisson process with intensity

$$\lambda(w) = \alpha \sum_{u \in \mathbf{U}} k(u - w).$$

The Thomas model uses a Gaussian kernel k, which is specified by its standard deviation σ. In the Matérn cluster model, the offsprings are uniformly distributed in a ball of radius r around their parent, see Matérn (2013) for details. In the variance Gamma model, the kernel is a convolution of a Gamma density and a Gaussian density and can be written as

$$k(u) = \frac{1}{\pi 2^{\nu+1} \eta^2 \Gamma(\nu + 1)} (\|u\|/\sigma)^{\nu} K_{\nu}(\|u\|/\sigma),$$

where σ is a scale parameter, ν a shape parameter for the kernel and K_ν denotes the modified Bessel function of the second kind. For more details we refer to Jalilian et al.

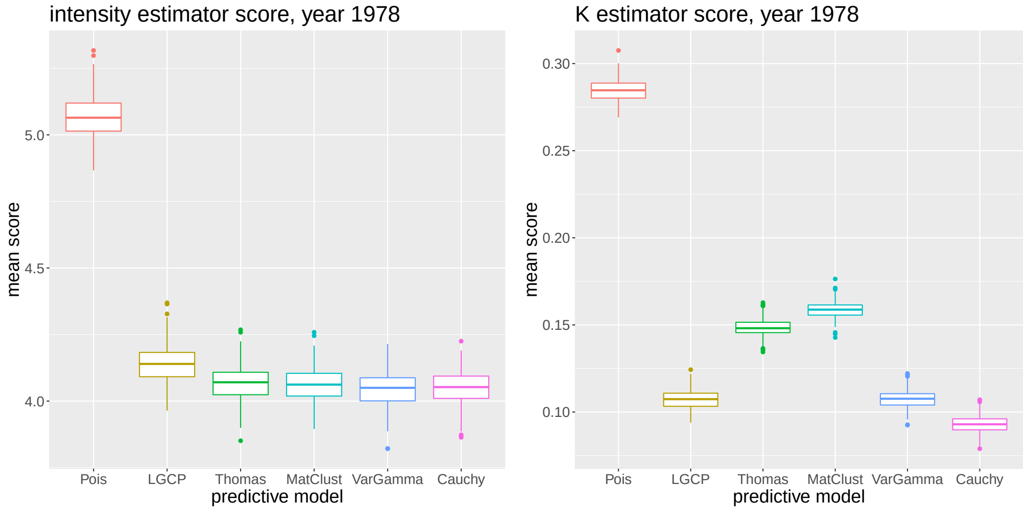


Figure 6. Mean scores for five different models predicting the spatial distribution of trees in 1979 in the described data set.

(2013). The Cauchy cluster model was introduced in Ghorbani (2013). It uses a heavy-tailed Cauchy kernel, with scale parameter σ, i.e.

$$k(u) = \frac{1}{2\pi\sigma^2}\left(1 + \frac{\|u\|^2}{\sigma^2}\right)^{-3/2}.$$

The models differ in their number of parameters: The Poisson process has only one, the LGCP two, the Thomas, Matérn and Cauchy model have 3 (the intensity parameter of the parent process, the expected number of offsprings per parent α and the scale parameter for the kernel), and the variance Gamma model has 4 (an additional shape parameter of the kernel).

All models are fitted by the method of minimum contrast, see Diggle and Gratton (1984); Waagepetersen (2007). To be precise, the K-function is estimated from the training set and the parameters Θ of the model are then fitted to minimize the integrated distance

$$\widehat{\Theta} = \arg\min_{\Theta}\left\{\int_0^{r_{\max}} (\widehat{K}(r)^{1/4} - K(r;\Theta)^{1/4})^2\, dr\right\}, \tag{5.1}$$

where $K(r;\Theta)$ is the theoretical K-function for the model with parameter Θ. The upper limit $r_{\max}$ should be chosen small relative to the diameter of the observation window, we use the default value of `spatstat`, which is one fourth of the diameter. Typically, only one point process observation is used to fit the model, and in this case the estimator $\widehat{K}$ described in Example 3.3 can be used. Our scenario where a single point process model is fitted to multiple observation is somewhat non-standard. However, we can estimate K from multiple observations by computing the estimator from Example 3.3 for all observations, and then taking the mean of these estimators. Thereafter we estimate Θ according to (5.1).

Figure 6 shows the mean scores and bootstrap confidence intervals for all models for the year 1978. The intensity score can only identify the Poisson model as significantly

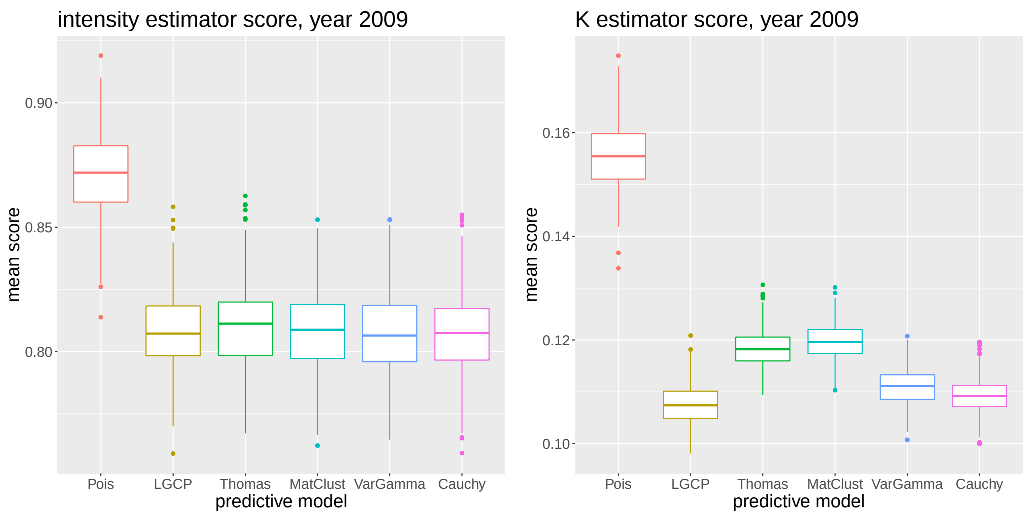


Figure 7. Mean scores for five different models predicting the spatial distribution of trees in 2009 in the described data set.

worse but cannot reliably distinguish between the cluster models. This is not surprising, since all models are homogeneous and their theoretical intensities are identical. The K-function scores indicate that the Cauchy process provides the best fit of the overall clustering behaviour of the trees, closely followed by the LGCP and variance Gamma model. For 1990 we obtained similar results (not shown), whereas for 2009 the Log-Gaussian Cox model performs slightly better than the Cauchy model, see Figure 7. It can also be observed that the difference between the Poisson model and the cluster point process models has grown smaller in 2009. It is quite interesting that the clustering behaviour of the trees changes over the years and that different point process models should be used for different years to achieve the best results. Moreover, the good performance of the Log-Gaussian Cox process is remarkable since it is the only considered cluster process that is not of Neyman-Scott type. Neyman-Scott processes have a great appeal for modelling plant distributions in ecology, since they are constructed by letting a parent process spawning offsprings in close vicinity to the parent points, and therefore mimic the process leading to the observed spatial distributions of the plants.

6 Discussion

We introduced a new class of proper scoring rules by combining estimators for summary statistics with the continuous ranked probability score. Our scoring rules can be computed from simulations of the predictive model. Therefore they can be applied to a wider range of predictive distributions than the commonly used logarithmic score, which requires the density of the predictive model to be known. They constitute, to the best of our knowledge, the first non-trivial and proper scoring rules for spatial point processes applicable to any predictive distribution. Even when comparing different Poisson models, where the logarithmic score is available, the intensity score performed better in our simulation study, in the sense that it expresses more significant differences of the mean scores between different models. The R package spatstat Baddeley et al. (2015) provides manifold tools for simulating various spatial point process models and provides implementations of all common summary statistic estimators. It provides therefore a platform that makes the introduced scoring rules easily applicable and, indeed, we used this package in all our simulation studies.

Our approach is based on the intuitive principle that, when the observation space is complex, the observations and predictions can be mapped into a simpler space for validation. This approach, that we simply call the mapping principle, is not restricted to point processes, and opens a fruitful new perspective on validation of involved forecasts in general. Indeed, when the observation space is involved, finding proper scoring rules can be difficult, even more so when they ought to be sensitive to certain high level properties of the observation-generating process. The mapping principle shifts this task to the much easier task of finding real-(or function-)valued mappings sensitive to these properties. Potential other applications include high-dimensional forecasts and forecasts of spatial fields, as well as function-valued forecasts.

We argue that, in the context of point processes, it is natural to use estimators of summary statistics in the mapping principle, as they are designed to be sensitive to high-level properties of the point process such as clustering or inhomogeneity. The variety of summary statistics available in the literature results in a collection of scoring rules sensitive to different properties of the predictive process. An additional advantage is that many theoretical properties of summary statistics and their estimators have already been investigated in the literature, and practicioners are familiar with them, making the output of the scoring rules easier to interpret.

The mapping principle requires the choice of a scoring rule on the codomain of the mapping. We opted for the CRPS since it is *strictly* proper and can be efficiently approximated by Monte-Carlo methods. When computational time is highly important, an attractive alternative might be the mean square error (MSE), $S(\mathbf{y}, F) := (\mathbf{y} - \mathbb{E}_F[\mathbf{X}])^2$. The expectation of the pushforward measure $\widehat{T}(F)$ can be Monte-Carlo approximated, and the computational costs are lower, since for the CRPS the term $\mathbb{E}_F[|\mathbf{y} - \mathbf{X}|]$ needs to be approximated for each observation $\mathbf{y}$, which is not necessary for the MSE. However, the MSE is proper but not strictly proper, and therefore the resulting score will be less sensitive to miscalibrations.

There are several avenues for potential future research. Generally, assessing the usefulness of the mapping principle in other contexts as mentioned above should be investigated. The good performance of the kernel estimator score when compared to the logarithmic score entails the question, whether scores based on kernel density estimators could outperform the popular logarithmic score in a wider range of settings. An important application for point-process-valued forecasts is earthquake rate forecasting, and a variety of methods for forecast validation has been developped in this community, see Schorlemmer et al. (2018) and references therein. Proper scores should be integrated into the existing validation programs to see how they compare to other methods currently used.

Acknowledgments

We acknowledge the support of The Research Council of Norway through grant 240838 “Model selection and model verification for point processes”.

A Permutation tests

To supplement the comparison study of the logarithmic score and the intensity score in Section 3, we assess the power of permutation tests based on the two scores. To this end we simulate 500 point patterns by each predictive method $F_1, ..., F_4$ and compute both scores. See Figure 3 in the main paper for a description of the models. Permutation tests assess the significance in the difference of mean scores taken over n observations and their power is therefore increasing in n. For a range of n, we subsample n observations and the corresponding scores from the 500 simulated point patterns. Then we compute the mean scores of these n observations and conduct a permutation test. The permutation test assesses significance in score differences of two models by randomly permuting the two models, and we consider 500 permutations in each step. We subsample from a previously simulated catalogue of scores in order to increase computational speed. This allows us to repeat this procedure 500 times for each $n = 5, 10, ..., 50$. For each n, the power of the permutation test is then approximated by the fraction of rejections out of the 500 repetitions, at a significance level of 5%. Recall that the null hypothesis in the permutation test is that the score differences are symmetrically distributed around 0, and a rejection therefore. This experiment is conducted four times, with the true distribution being $F_1, ..., F_4$. Each time we assess the power of the permutation test when the true model is compared with all three alternative models. The results are summarized in Figure A.1.

Note that the values in this figure are random approximations of the true power of the tests, and are intercorrelated since we used subsampling. Nevertheless, they clearly show that permutation tests based on the intensity score are more powerful and therefore the intensity score differentiates more clearly between the predictive models. It is remarkable that, based on the intensity score, the permutation test reaches a power of 0.9 for less than 20 available observations regardless of which models are compared. In contrast, the permutation test based on the logarithmic score has difficulties distinguishing between the relatively similar models F_1 and F_2 as well as F_1 and F_3. Even when 50 observations are available the power is below 0.5.

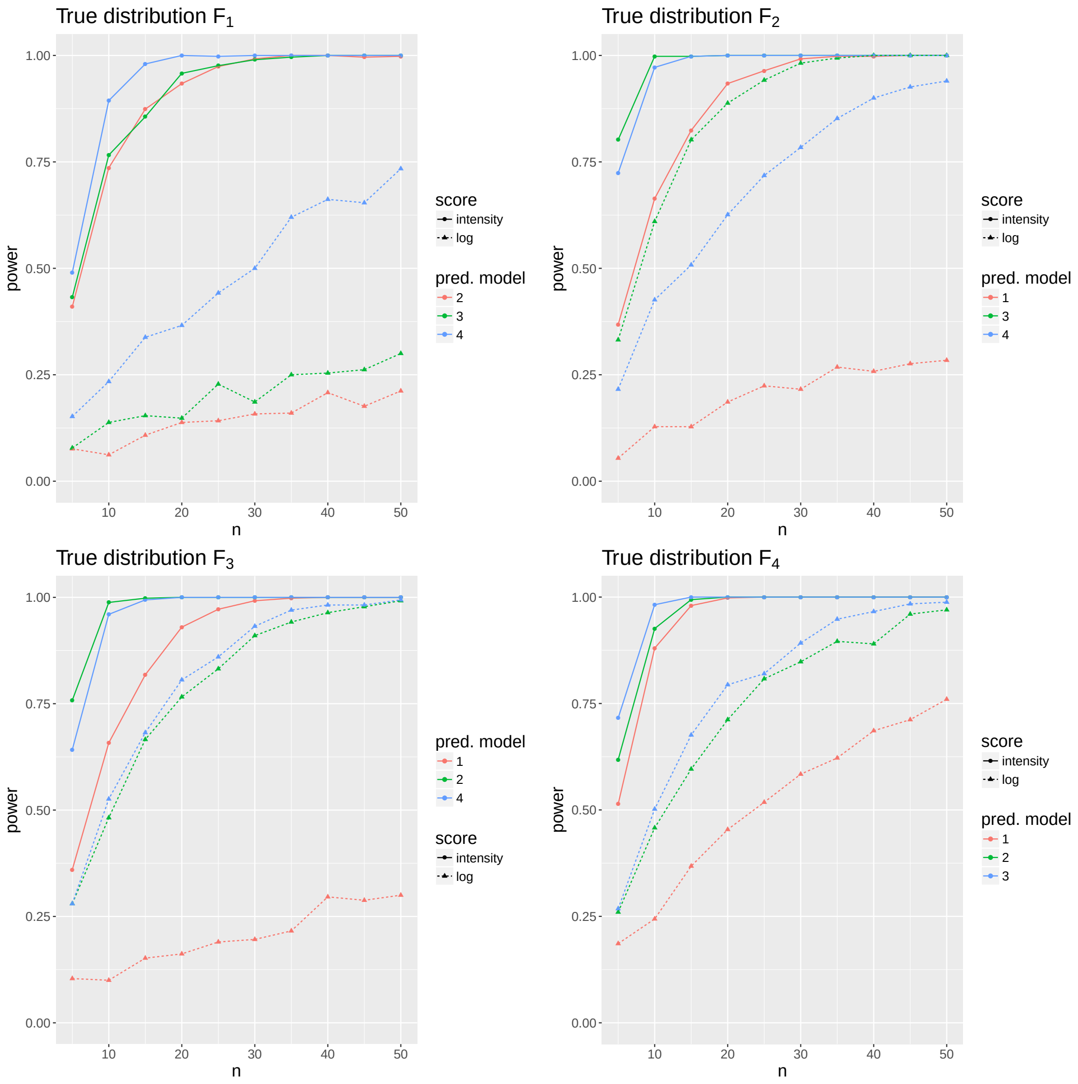


Figure A.1. Power of permutation tests for comparing different predictive models with the true distribution for a range of available observations n. The size of the test is 5%.

References

Baddeley, A., Møller, J., and Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54(3):329–350. 7, 8

Baddeley, A., Rubak, E., and Møller, J. (2011). Score, pseudo-score and residual diagnostics for spatial point process models. *Stat Sci*, 26(4):613–646. 4, 6

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, London. Available from: `http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/`. 13, 20

Baddeley, A. and Silverman, B. (1984). A cautionary example on the use of second-order methods for analyzing point patterns. *Biometrics*, 40(4):1089–1093. 9

Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005). Residual analysis for spatial point processes. *J Roy Stat Soc B*, 67:617–666. 4, 5, 6

Bray, A. and Schoenberg, F. (2013). Assessment of point process models for earthquake forecasting. *Statistical science*, 28(4):510–520. 5

Clements, R., Schoenberg, F., and Schorlemer, D. (2011). Residual analysis methods for space-time point processes with applications to earthquake forecast models in California. *The Annals of Applied Statistics*, 5:2549–2571. 4, 6

Daley, D. and Vere-Jones, D. (2004). Scoring probability forecasts for point processes: The entropy score and information gain. *J. Appl. Probab.*, 41(A):297–312. 4

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes; volume II: general theory and structure*. Springer Science & Business Media. 7

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion and rejoinder). *Journal of the Royal Statistical Society Ser. A*, 147:278–292. 5

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, 27:65–81. 8

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263. 7

Diggle, P. and Gratton, R. (1984). Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):193–212. 18

Eberhard, D. A. J., Zechar, J. D., and Wiemer, S. (2012). A prospective earthquake forecast experiment in the western Pacific. *Geophys J Int*, 190:1579–1592. 4

Edelman, A. J. (2012). Positive interactions between desert granivores: localized facilitation of harvester ants by kangaroo rats. *PloS ONE*, 7(2):e30914. 4

Ghorbani, M. (2013). Cauchy cluster process. *Metrika*, 76(5):697–706. 18

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762. 6

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J Roy Stat Soc B*, 69(2):243–268. 4

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378. 4, 5, 10

Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media. 7

Gordon, J., Clements, R., Schoenberg, F., and Schorlemmer, D. (2015). Voronoi residuals and other residual analyses applied to csep earthquake forecasts. *Spatial Statistics*, 14:133–150. 5

Grier, C. C., Vogt, K. A., Keyes, M. R., and Edmonds, R. L. (1981). Biomass distribution and above- and below-ground production in young mature *abies amabilis* zone ecosystems of the Washington Cascades. *Canadian Journal of Forest Research*, 11:155–167. 15

Jalilian, A., Guan, Y., and Waagepetersen, R. (2013). Decomposition of variance for spatial Cox processes. *Scandinavian Journal of Statistics*, 40(1):119–137. 17

Law, R., Illian, J., Burslem, D. F., Gratzer, G., Gunatilleke, C., and Gunatilleke, I. (2009). Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology*, 97(4):616–628. 15

Matérn, B. (2013). *Spatial variation*, volume 36. Springer. 17

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tite, G. E. (2011). Self-exciting point process modeling of crime. *J Am Stat Assoc*, 106(493):100–108. 4

Møller, J. and Waagepetersen, R. (2003). *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC. 7, 9

Pourtaheri, R. and Vahidi-Asl, M. Q. (2011). Point pattern analysis of regional city distributions. *Qual Quant*, 45:1473–1481. 4

Scheuerer, M. and Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334. 8

Schorlemmer, D., Werner, M., Marzocchi, W., Jordan, T., Ogata, Y., Jackson, D., Mak, S., Rhoades, D., Gerstenberger, M., Hirata, N., et al. (2018). The collaboratory for the study of earthquake predictability: achievements and priorities. *Seismological Research Letters*, 89(4):1305–1313. 4, 6, 20

Sorrensen-Cothern, K. A., Ford, E. D., and Sprugel, D. G. (1993). A model of competition incorporating plasticity through modular foilage and crown development. *Ecological Monographs*, 63(3):277–304. 16

Thorarinsdottir, T. and Schuhen, N. (2018). Verification: assessment of calibration and accuracy. In *Statistical Postprocessing of Ensemble Forecasts*, pages 155–186. Elsevier. 7

Velázquez, E., Martínez, I., Getzin, S., Moloney, K. A., and Wiegand, T. (2016). An evaluation of the state of spatial point pattern analysis in ecology. *Ecography*, 39(11):1042–1055. 15

Waagepetersen, R. (2007). An estimating function approach to inference for inhomogeneous Neyman–Scott processes. *Biometrics*, 63(1):252–258. 18

Waller, L. A., Särkkä, A., Olsbo, V., Myllymäki, M., Panoutsopulou, I. G., Kennedy, W. R., and Wendelschafer-Crabb, G. (2011). Second-order spatial analysis of epidermal nerve fibers. *Stat Med*, 30(23):2827–2841. 4

Wiegand, T., He, F., and Hubbell, S. P. (2013). A systematic comparison of summary characteristics for quantifying point patterns in ecology. *Ecography*, 36(1):92–103. 15

Wiegand, T., Uriarte, M., Kraft, N. J., Shen, G., Wang, X., and He, F. (2017). Spatially explicit metrics of species diversity, functional diversity, and phylogenetic diversity: Insights into plant community assembly processes. *Annual Review of Ecology, Evolution, and Systematics*, 48:329–351. 15

Zhuang, J. (2006). Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *J Roy Stat Soc B*, 68(4):635–653. 4