Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Validation of point process forecasts

Note no SAMBA/20/19

Authors Claudio Heinrich
Max Schneider
Peter Guttorp
Thordis Thorarinsdottir

Date 31st July 2019

The authors

Claudio Heinrich is 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 forecasts**
Authors	**Claudio Heinrich , Max Schneider , Peter Guttorp , Thordis Thorarinsdottir**
Date	31st July 2019
Publication number	SAMBA/20/19

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 which cannot be evaluated for many point process models, as their density is only known up to an untractable constant. In simulation studies we demonstrate the usefulness of our scores. Furthermore we consider a scoring rule, the quantile score, that is commonly used to validate earthquake rate predictions, and show that it lacks propriety. As a consequence, several tests that are commonly applied in this context are biased and systematically favour predictive distributions that are too uniform. We suggest to remedy this issue by replacing the commonly used one-sided by two-sided tests.

Keywords	Point processes; Forecast verification; Proper scoring rules
Target group	Scientists
Availability	Open
Project	PointProcess
Project number	220708
Research field	Statistics
Number of pages	28
© Copyright	Norwegian Computing Center

1 Introduction

Motivated by the lack of a unified theory for validation of meteorological forecasts, Gneiting et al. (2007) attempted to formulate fundamental principles of probabilistic forecasting and to provide a mathematical framework for validating to what degree a forecaster satisfies these principles. They introduced the paradigm that the main goal of a probabilistic forecast is

> *'maximizing the sharpness of the predictive distributions subject to calibration.'*

The term *calibration* means that the relative frequencies of an event manifesting in the observations should match the probability of this event happening under the predictive distribution. Intuitively, this means that the observation should look like a random draw from the predictive distribution. For a more detailed introduction of calibration we refer to the original article by Gneiting et al. (2007). *Sharpness* refers to the spread in the predictive distribution. Maximizing the sharpness therefore refers to issuing predictions that are as precise as possible. The most common way of assessing calibration and sharpness is by using *scoring rules*, which comprise the information contained in a predictive distribution F and the target observation y into a single number, the score $S(y, F)$. Scoring rules ought to be proper in order to be useful in practice, meaning that the expected score gets optimal when the true distribution of the observation is predicted, see Gneiting and Raftery (2007). Proper scoring rules are valuable tools for decision making, since they allow to easily compare different forecast models based on a single number, and have therefore become widely successful in forecast validation throughout various sciences.

In the context of point process forecasts, the use of scoring rules has so far mostly been limited to the logarithmic score, i.e. the negative log-likelihood of the observation under the predictive model, see Daley and Vere-Jones (2004). The main reason for this is presumably 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 this lack of scoring rules in the literature, we introduce a new class of proper scoring rules by combining well-known summary statistics for point process 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 density of the process. This is a tremendous advantage in the context of spatial point processes, as for many common point process models, such as Markov processes and many Cox processes, 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, making the scores easily interpretable. 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 very 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 involved, simply by mapping the observation space into a simpler space, where proper scores exist.

One of the most prominent example of point-process valued forecasts are earthquake predictions. We consider a scoring rule, sometimes called the quantile-score, that is often used in this context and show that it is improper. We demonstrate both theoretically and in simulation studies that a forecaster evaluated by the quantile score achieves a better expected score by issuing predictions that are too uniformly distributed.

The quantile score is used as test-statistic for three tests that are commonly applied in earthquake rate prediction and are typically referred to as the L-, M-, and S-test. We demonstrate that, by impropriety of the quantile score, these tests are biased towards too uniform distributions. Our findings explain several questions raised in the corresponding literature: Schorlemmer et al. (2007) remarked that it is difficult to explain too high values of the scoring rule used in the L-test, a gap that is filled by our results. Gerstenberger et al. (2009) note that models significantly underestimating the occurance of earthquakes in regions of high activity can pass the L-test. Indeed, such models issue predictive distributions that are too uniform and are preferred by the test. We propose a simple alteration to these tests that resolves this issue.

This article is organized as follows. Section 2 contains the theoretical background, including a brief summary 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. In Section 6 we introduce the quantile score and show that it is improper. Section 7 relates this result to the corresponding tests used in the literature and provides a simulation study demonstrating the bias of the tests. Section 8 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. The scoring rule is *strictly proper* relative to the class $\mathcal{P}$ if (2.2) holds, with equality only if $F = G$. Evaluating a forecaster based on proper scoring rules encourages honesty and prevents hedging. That is, the preceived 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 (ref. to T. and Schuhen, 2019). 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 $d = 2, 3, ...,$ with a bounded observation window $W \subset \mathbb{R}^d$. The observation space $\mathcal{O}$ is then the space of countable subsets of W, which we will denote by $W^\cup$. A *spatial point process*, usually denoted $\mathbf{X}$ or $\mathbf{Y}$, is a random variable taking values in $W^\cup$ with almost surely finitely many points. We use $F, G, ...$ to denote distributions of point processes and use the notation $\mathbb{E}_F[f(\mathbf{X})]$ for the expectation of $f(\mathbf{X})$ for some function f, when $\mathbf{X}$ is distributed according to F. For a overviews on the topic we refer to Møller and Waagepetersen (2003) and Daley and Vere-Jones (2007).

Summary statistics are powerful tools for exploratory data analysis and model selection for point processes. We introduce two important examples that are useful to bear in mind.
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 many 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 repell 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*

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

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 clusters. Other examples of popular summary statistics include the F-, G-, and J-function as well as the second order intensity. For more details we refer to Møller and Waagepetersen (2003, Chapter 4). Bearing these examples in mind, we make the following definition suitable for our purposes. Note that in both examples the summary statistic is function-valued, taking values from a space $\mathcal{R}$. For the intensity function we have $\mathcal{R} = W$, 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 a summary statistic to be based on a single point pattern, which is the case for all standard estimators for the summary statistics mentioned above. Let us remark that not all summary statistics are well-defined for all point process models. For example is the K-function 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 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 $\mathcal{O}$ it is quite natural not to attempt validating the full predictive distribution, but rather focus on a certain property of interest. This approach is not new, examples in the context of multivariate forecasts being the Dawid-Sebastiani score (Dawid and Sebastiani, 1999) that focusses on mean and covariance of a multivariate forecast, and the variogram score (Scheuerer and Hamill, 2015) that focusses on the variogram of a spatial prediction. We adapt this principle and show how it can be applied to validate point process forecasts. Examples for properties of interest in the context of point process forecasts are number and spatial distribution of points or clustering behavior of the process. We demonstrate in the following how proper scoring rules can be obtained that are sensitive with respect to these properties. The key idea is to utilize well-known summary statistics that target said property. Then, scoring rules can be constructed by comparing estimators of these summary statistics to the corresponding summary statistic of the predictive distribution. 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. These tools do not impose any restrictions on the class $\mathcal{P}$ of predictive distributions considered. We therefore 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 for example Baddeley and Silverman (1984).

In this proposition, both $\widehat{T}$ and T get evaluated at a specific $r \in \mathcal{R}$, whereas in practice we will be more interested in an overall fit. 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, as long as the integral in (3.1) exists (possibly taking the value $+\infty$) for all $\mathbf{y}, F$. These two proposition readily allow the construction of proper scoring rules based on summary statistics in some cases.

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 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 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, there are no unbiased estimators. Secondly, even if there are unbiased estimators, closed form expressions for T_F are usually only available for selected point process models.

A bit surprisingly, both of this 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. Moreover, choosing the CRPS allows to approximate $S_{\widehat{T}}$ without requiring detailed knowledge of the pushforward measure $\widehat{T}(F)$. The CRPS 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 easily 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 nevertheless $\widehat{T}(\mathbf{X}_1)$ and $\widehat{T}(\mathbf{X}_2)$ usually have different distributions for $\mathbf{X}_1 \sim F_1, \mathbf{X}_2 \sim F_2$. 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.

Let us sum up the main result of this sections 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, the only condition being integrability of $\widehat{T}$ which is satisfied for most point process models and summary statistic estimates. Note that $\widehat{T}$ can be any real 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, by construction of $\widehat{K}$, the score will still be sensitive to a misspecification of the clustering behaviour in the predictive model.

We conclude this section by discussing two important examples that will be used in our simulation studies.

Example 3.2 (Kernel estimator 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 $c_{W,b}(y) = \int_W k_b(w - y)\, dw$. Therefore, by Corollary 3.1, the kernel estimator 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 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 construct a score using the K-function estimator.

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 unsensitive to misspecification of the intensity function, and for example be inadequate for differentiating between different Poisson processes, which have the same K-function.

Finally, let us emphasize that the scoring rules derived in Proposition 3.3 and Corollary 3.1 are proper but not strictly proper, even though they utilize the strictly proper CRPS. Distributions F and G satisfying $\widehat{T}(F) = \widehat{T}(G)$ obtain the same expected score.

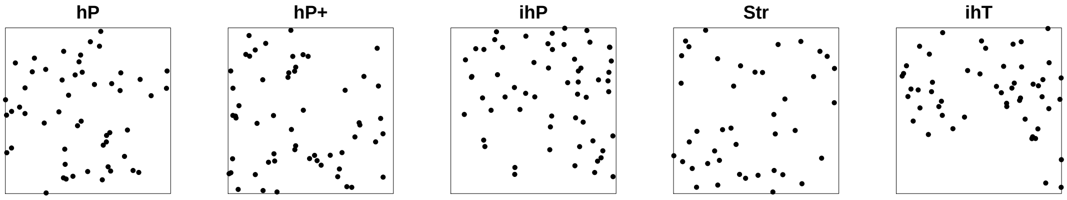
4 Simulation study

In order to demonstrate the usefulness of our scores, we present a simulation study where we consider 5 different point process models. Their characteristics are summarized in Figure 1 where we also show an example plot from each model. We consider the spatial window $W = [0, 10] \times [0, 10]$. 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 repell 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 generating an (invisible) Poisson process of parent points, and then letting each parent



	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.

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. For 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 $\mathbf{X} \sim F$. We do these computations for $\widehat{T} = \widehat{\lambda}$ and $\widehat{T} = \widehat{K}$. The computations are carried out using the `R`-package `spatstat` (Baddeley et al., 2015), and all parameters for the estimators are set to their `spatstat` default values. In particular, we use a Gaussian kernel in the intensity estimator. 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 does 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. The score $S_{\widehat{K}}$, on the other hand, is capable of detecting mismatches in the point interaction between predictive and true distribution. In particular, it can differentiate between the Strauss and the homogeneous Poisson 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 3 shows the results of permutation tests assessing the significance of the difference of the mean scores. The differences

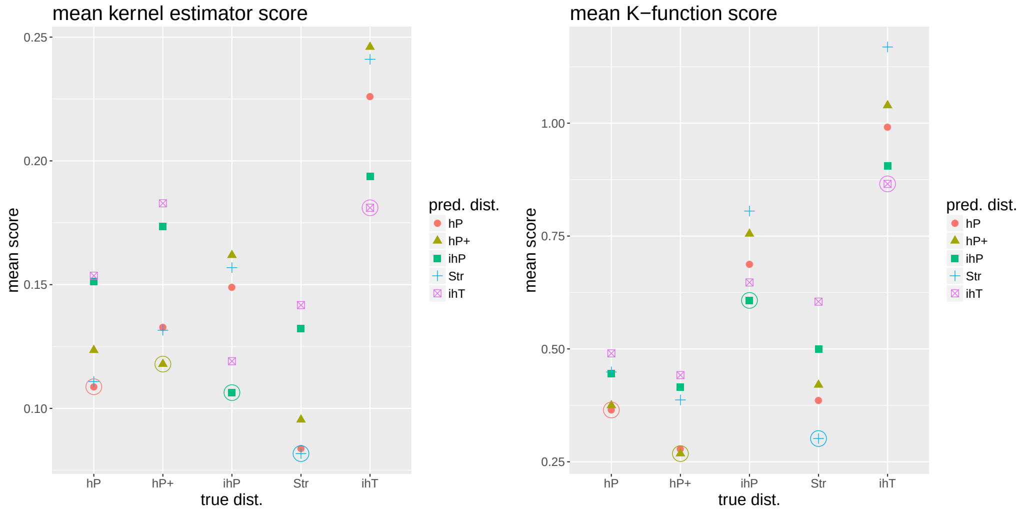


Figure 2. The mean scores $\mathbb{E}_G[S_{\widehat{\lambda}}(\mathbf{Y},F)]$ (left hand side) and $\mathbb{E}_G[S_{\widehat{K}}(\mathbf{Y},F)]$ (right hand side) for each combination of the 5 considered models. The x-axis shows the true distribution G of the data. The score for each of the five predictive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

between the K-function scores for the different Poisson models is not significant at a 5% level. However, given that these models have the same theoretical K-function, it is remarkable that the score $S_{\widehat{K}}$ is able to differentiate between them at all, if also less reliably than between the other models. This is due to the fact that the distribution of $\widehat{K}$ still varies between these models, which is detected by the CRPS. 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.

Nevertheless, the logarithmic score is commonly applied when different Poisson models are compared. We therefore next address the question how the kernel intensity estimator score compares to the logarithmic score, when assessing the skill of different Poisson models. We consider a homogeneous Poisson process with 50 expected points in the window $[0,10]\times[0,10]$ as the true distribution of the data, and compare the performance of four different predictive Poisson models, assessed by both scoring rules. The first predictive model is the true model of the data, model 2 and 3 are homogeneous Poisson processes with 40 and 60 expected points, respectively. Model 4 is an inhomogeneous Poisson process with 50 expected points and intensity function $\lambda(x,y)=\frac{x}{20}+0.25$, i.e. the intensity increases linearly in x from 0.25 to 0.75.

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 kernel estimator score is more sensitive than the logarithmic score and can more reliably identifies the true distribution. This is again likely to be a consequence of applying the strictly proper CRPS, which fully assesses the distribution of $\widehat{\lambda}(\mathbf{Y})$ which is in a sense more sensitive to the full distribution of $\mathbf{Y}$ than the log-likelihood. If, for example, the predictive distribution F is a homogeneous Poisson process, then $S_{\log}(\mathbf{Y},F)$ depends on $\mathbf{Y}$ only via $n(\mathbf{Y})$ and is therefore not sensitive to other parameters such as spatial

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

		hP	hP+	ihP	Str	ihNS
G	hP	-	<0.1%	<0.1%	**8.0%**	<0.1%
	hP+	<0.1%	-	<0.1%	<0.1%	<0.1%
	ihP	<0.1%	<0.1%	-	<0.1%	**8.4%**
	Str	**15.1%**	<0.1%	<0.1%	-	<0.1%
	ihNS	<0.1%	<0.1%	4.4%	<0.1%	-

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

		hP	hP+	ihP	Str	ihNS
G	hP	-	**38.3%**	**17.6%**	<0.1%	<0.1%
	hP+	**38.9%**	-	**16.7%**	<0.1%	<0.1%
	ihP	**8.2%**	**6.2%**	-	<0.1%	0.5%
	Str	<0.1%	<0.1%	<0.1%	-	<0.1%
	ihNS	<0.1%	<0.1%	0.7%	<0.1%	-

Figure 3. p-values 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.

distribution of points or size and shape of the observation window. Let us remark that, for Poisson point processes, evaluating the logarithmic score is less computationally involved than evaluating the kernel estimator score, which relies on Monte-Carlo approximation of a (multiple) integral. However, in typical spatial point process applications there are often only few 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 and ecology). In these cases there is a huge benefit from applying a more robust score, 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. Values of the score itself cannot be interpreted unless scores for different predictive models are compared.

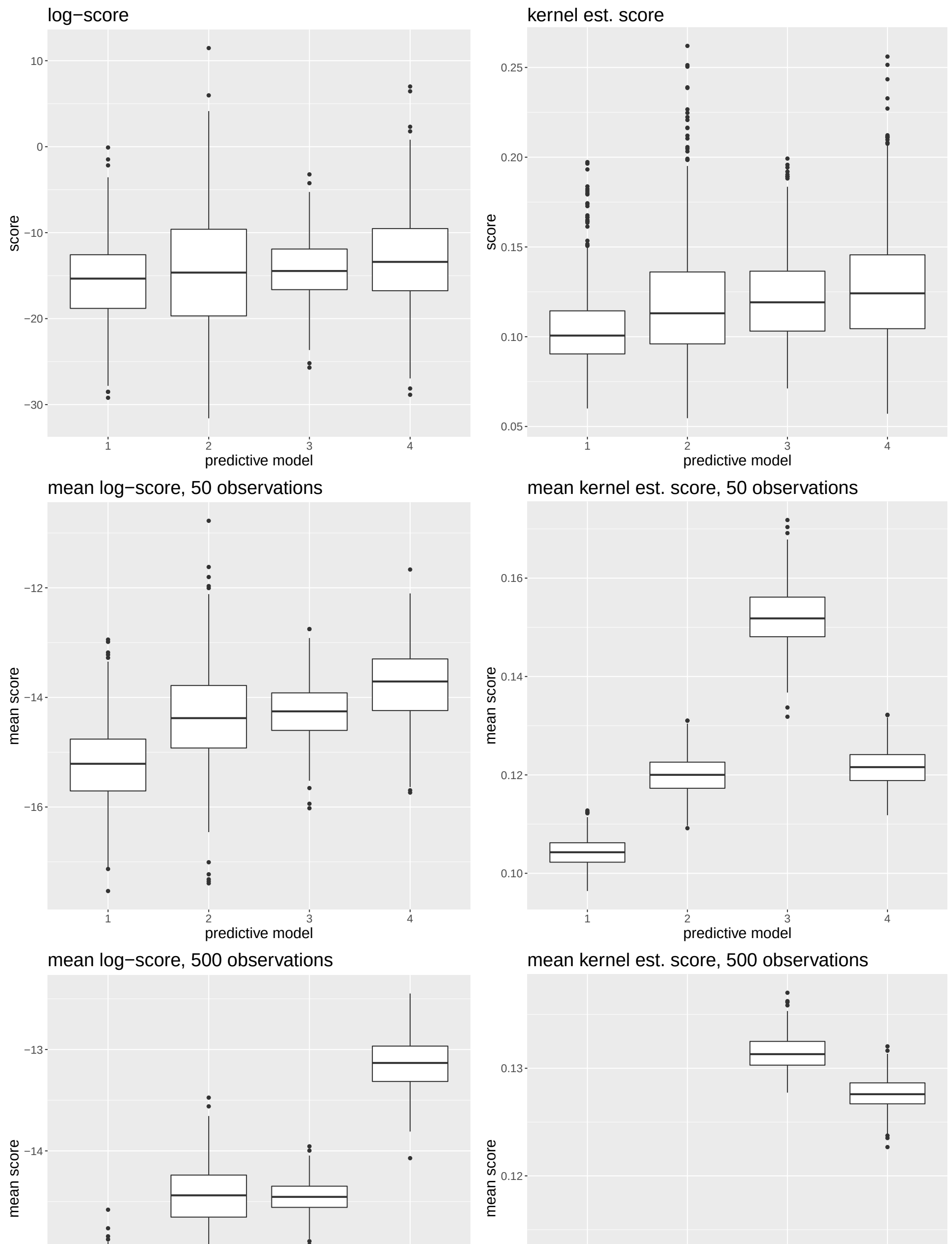


Figure 4. The first row shows boxplots for $S_{\log}(\mathbf{y},F)$ and $S_{\widehat{\lambda}}(\mathbf{y},F)$ for the four different predictive Poisson models described in the text. Model 1 is the true distribution of the data. Rows two and three show boxplots for bootstrap resamples of the expected 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.

5 Application to *Abies amabilis* forests

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, U.S.A., see Grier et al. (1981) for a description of the site conditions. Figure 5 shows the location of trees at four of the plots for three different time points over 31 years. The area was clear-cut in 1957; the trees in our data set were apparently present as seedlings before the clear-cut and there appears 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 investigate the development of tree crown structure under competition for light.

We analyze the observations from each year independently such that in each analysis, we have eight independent realizations of the underlying process. We consider three different predictive models: A Poisson model, a log-Gaussian Cox process and a Matérn cluster process. Each model is fitted to the tree sample in one plot, and then this predictive model is validated against the other 7 plots by using the kernel estimator score and the K-function score. Overall we therefore obtain 56 score values for each model (8 model fits, and 7 validations for each fit), and we can use the mean score as a measure for assessing calibration of the predictive model. We again use bootstrapping to assess the uncertainty associated with the mean scores. The results are presented in Figure 6.

6 The quantile score

In this section we consider the quantile score, a scoring rule that lies at the foundation for several numeric tests typically performed in the validation of earthquake rate predictions. Denote by l_F the log-likelihood function of a predictive distribution F. The quantile score is then defined as

$$\gamma(\mathbf{y}, F) = \mathbb{P}_F[l_F(\mathbf{y}) > l_F(\mathbf{X})]. \tag{6.1}$$

This quantity is then used for testing the hypothesis $\mathbf{y} \sim F$ by rejecting the hypothesis if $\gamma(\mathbf{y}, F) < \alpha$ for some fixed threshold α. The underlying intuition is that, in order for the predictive model to be consistent with the observation, the observation should be as likely under the predictive model as a typical realization of the model. Such tests are popular in the context of earthquake rate predictions and we will discuss some details about these tests in the next section. The typical setting in this context considers a binning $\{b_1, ..., b_n\}$ of the observed spatial window and earthquake magnitudes, each bin resembling a multidimensional interval. Every earthquake is then recorded as an event in one of these bins, depending on its location and magnitude. A rate prediction is then a predictive distribution for how many earthquakes are expected in which bin, over a

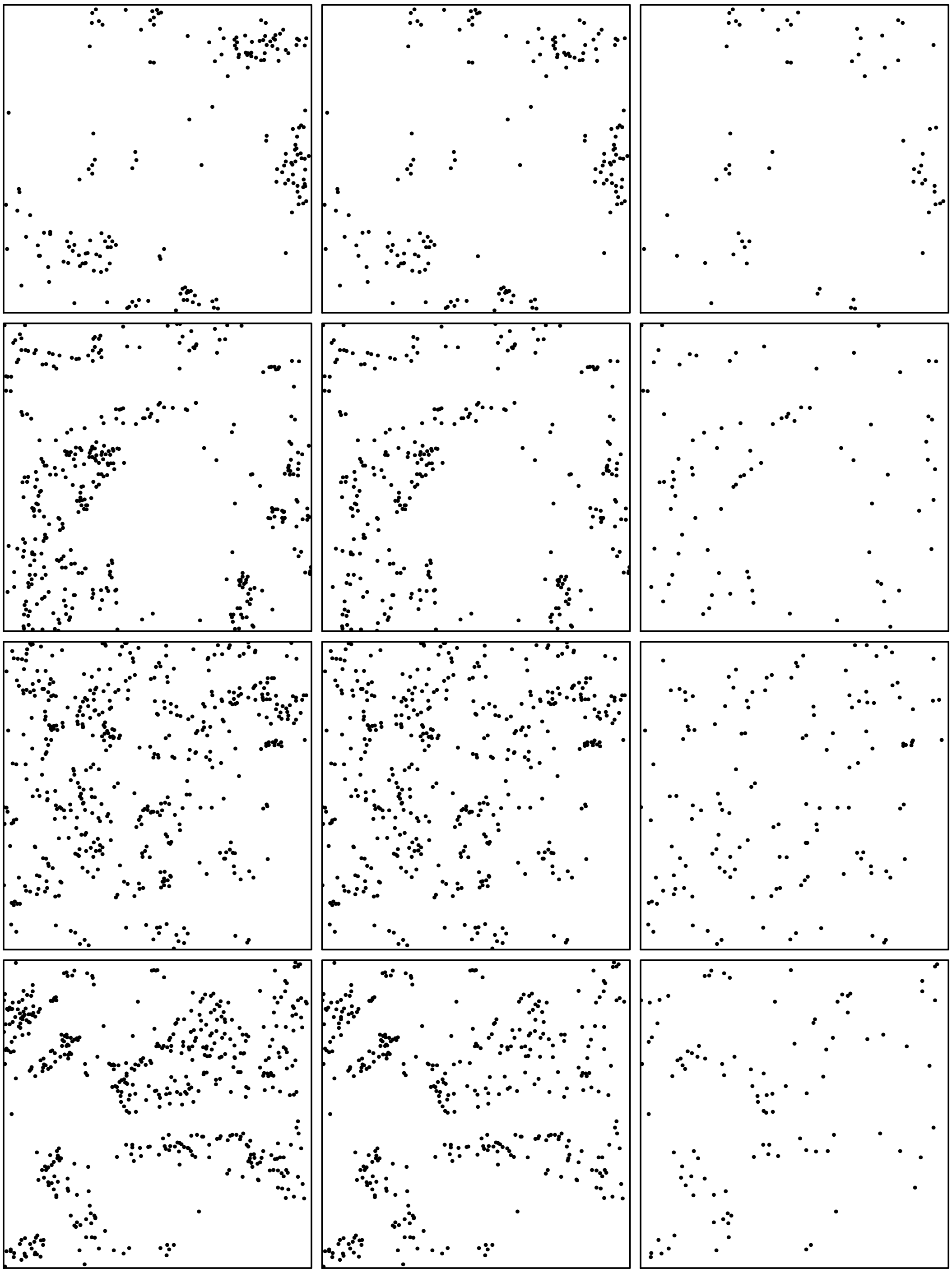


Figure 5. *Abies amabilis* (Pacific silver fir) at four disjoint 6 by 6 meter plots at Findley Lake Reserve in Washington State (rows). The area was clear-cut in 1957; the first column shows trees present in 1978, 21 years after the clear-cut; the second column shows the trees still present in 1990,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

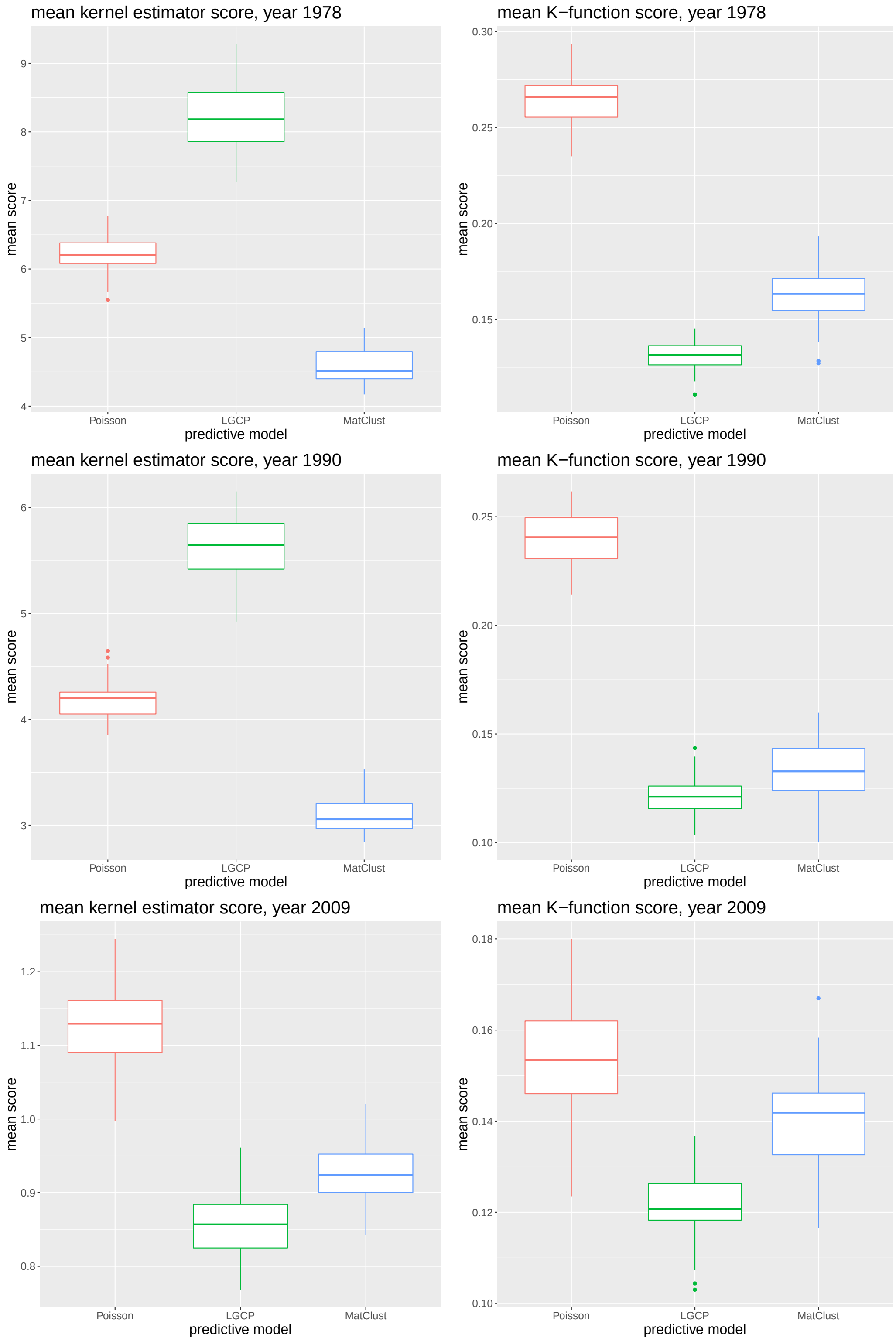


Figure 6. Bootstrap estimates of mean scores for three different models used for predicting the spatial distribution of trees in the described data set.

previously defined timespan, for example 5 years. Accordingly, in this context the observation space $\mathcal{O} = \mathbb{N}_0^n$ contains integer vectors, resembling the amount of earthquakes in each bin. As $\mathbb{N}_0^n$ and $\mathbb{N}_0$ are isomorphic, we may assume for this section that $\mathcal{O} = \mathbb{N}_0$. In its usual definition (6.1) the quantile score is positively oriented in the sense that larger scores indicate better performance. In order to align with our previous sections we therefore consider the equivalent negatively oriented score

$$\widetilde{\gamma}(\mathbf{y}, F) := \mathbb{P}_F[l_F(\mathbf{y}) \leq l_F(\mathbf{X})] = 1 - \gamma(\mathbf{y}, F).$$

Our results show that this score is improper and lower (better) expected scores are obtained by choosing 'more uniform' distributions. The concept of a distribution being more uniform than another is quite intuitive: For example for two Poisson distributions $P_{\lambda_1}, P_{\lambda_2}$ with $\lambda_1 < \lambda_2$, we would say that P_{λ_2} is more uniform than P_{λ_1}, since it spreads the probability weight more evenly on all natural numbers. However, formalizing this concept requires a bit of work, since we cannot define it in terms of distance to the uniform distribution, as the observation space $\mathbb{N}_0$ does not support a uniform distribution. We therefore introduce the following somewhat technical definition.

Definition 6.1. *For two distributions $\mathbb{P} := (p_0, p_1, ...)$ and $\mathbb{Q} := (q_0, q_1, ...)$ on $\mathbb{N}_0$ we denote $\mathbb{P} \succeq \mathbb{Q}$ if the following is satisfied. After reordering the sequences of probabilities $(p_0, p_1, ...)$ and $(q_0, q_1, ...)$ into nonincreasing sequences $\widetilde{q}_0 \geq \widetilde{q}_1 \geq ...$ and $\widetilde{p}_0 \geq \widetilde{p}_1 \geq ...$, it holds that*

$$\sum_{i=0}^{n} \widetilde{q}_i \geq \sum_{i=0}^{n} \widetilde{p}_i, \quad \textit{for all } n \in \mathbb{N}_0. \tag{6.2}$$

If, additionally, the inequality is strict for at least one n, we denote $\mathbb{P} \succ \mathbb{Q}$ and say that $\mathbb{P}$ is more uniform than $\mathbb{Q}$.

This defines a partial order on the set of all probability distributions on $\mathbb{N}_0$. We show in the appendix that for finitely supported distributions, $\mathbb{P} \succeq \mathbb{Q}$ implies that $D(\mathbb{P}, \mathbb{U}) \leq D(\mathbb{Q}, \mathbb{U})$, where D denotes the total variation distance of probabilities, and $\mathbb{U}$ denotes the uniform distribution on the joint support of $\mathbb{P}$ and $\mathbb{Q}$. Note that $\sum_{i=0}^{n} \widetilde{q}_i$ is the probability under $\mathbb{Q}$ of the $n+1$ most likely events. In order for $\mathbb{P}$ to satisfy (6.2) it needs to assigns less mass to its $n+1$ most likely events, and therefore more weight to less probable numbers, making it indeed more uniform.

Moreover, we need the following definition. For distributions $\mathbb{P}_1 = (p_0^1, p_1^1, ...)$ and $\mathbb{P}_2 = (p_0^2, p_1^2, ...)$ on $\mathbb{N}_0$ we say that $\mathbb{P}_1$ and $\mathbb{P}_2$ have *the same probability ranks*, if they satisfy

$$p_i^1 < p_j^1 \Leftrightarrow p_i^2 < p_j^2 \quad \text{and} \quad p_i^1 = p_j^1 \Leftrightarrow p_i^2 = p_j^2 \quad \text{for all } i, j \in \mathbb{N}_0.$$

The main result of this section is the following theorem.

Theorem 6.1. *Consider distributions $\mathbb{P}_1, \mathbb{P}_2, \mathbb{Q}$ on $\mathbb{N}_0$, all with the same probability ranks, and let $\mathbb{P}_1 \succeq \mathbb{P}_2$. It holds that*

$$\mathbb{E}_{\mathbb{Q}}[\widetilde{\gamma}(Y, \mathbb{P}_1)] \leq \mathbb{E}_{\mathbb{Q}}[\widetilde{\gamma}(Y, \mathbb{P}_2)]. \tag{6.3}$$

Moreover, if $\mathbb{P}_1 \succ \mathbb{Q}$, then we obtain the strict inequality

$$\mathbb{E}_{\mathbb{Q}}[\widetilde{\gamma}(Y, \mathbb{P}_1)] < \mathbb{E}_{\mathbb{Q}}[\widetilde{\gamma}(Y, \mathbb{Q})]. \tag{6.4}$$

The proof is given at the end of this section. Inequality (6.4) implies that $\widetilde{\gamma}$ is improper. Inequality (6.3) shows moreover, that, for distributions with the same probability ranks, more uniform distributions always get a lower expected score, regardless of whether they are a better approximation of the true distribution $\mathbb{Q}$. In this sense, the theorem shows much more than just impropriety of $\widetilde{\gamma}$ which would already follow from finding one particular pair of distributions $\mathbb{P}_1, \mathbb{Q}$ satisfying (6.4). It shows that hedging is possible for any true distribution $\mathbb{Q}$, and that more uniform distributions always lead to better expected scores. The technical restriction of having the same probability ranks is used in our proof, but did not seem to be of major importance in our numerical experiments. In particular, the distributions in our simulation study in the next section do not have the same probability ranks.

Some authors consider a slight modification of the quantile score defined as

$$\gamma_2(\mathbf{y}, F) = \mathbb{P}_F[l_F(\mathbf{y}) \geq l_F(\mathbf{X})],$$

i.e. treating the case $l_F(\mathbf{y}) = l_F(\mathbf{X})$ differently. Replacing γ by γ_2 does not change any of our conclusions and requires only minor changes in the proof of Theorem 6.1.

Proof of Theorem 6.1. Let $\mathbb{P}_1 = (p^1_0, p^1_1, ...), \mathbb{P}_2 = (p^2_0, p^2_1, ...)$, and $\mathbb{Q} = (q_0, q_1, ...)$ have the same probability ranks, in particular it holds for any i, j that $\mathbb{1}\{p^1_j \leq p^1_i\} = \mathbb{1}\{p^2_j \leq p^2_i\} = \mathbb{1}\{q_j \leq q_i\}$. It follows that

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[\widetilde{\gamma}(Y, \mathbb{P}_1) - \widetilde{\gamma}(Y, \mathbb{P}_2)] &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p^1_i q_j \mathbb{1}\{p^1_j \leq p^1_i\} - \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p^2_i q_j \mathbb{1}\{p^2_j \leq p^2_i\} \\ &= \sum_{j=0}^{\infty} q_j \sum_{i=0}^{\infty} (p^1_i - p^2_i)\mathbb{1}\{q_j \leq q_i\} \\ &= \sum_{j=0}^{\infty} \widetilde{q}_j \sum_{i=0}^{k(j)} (\widetilde{p}^1_i - \widetilde{p}^2_i), \qquad (6.5)\end{aligned}$$

where in the last line $\widetilde{p}$ and $\widetilde{q}$ denote the probabilities in decreasing order, and $k(j)$ denotes the last index i such that $\widetilde{q}_i \geq \widetilde{q}_j$. In the last equality we used that $\mathbb{P}_1, \mathbb{P}_2$ and $\mathbb{Q}$ can be ordered by applying the same permutation to the indices since they have the same probability ranks. By the assumption $\mathbb{P}_1 \succeq \mathbb{P}_2$ we have

$$\sum_{i=0}^{k(j)} (\widetilde{p}^1_i - \widetilde{p}^2_i) \leq 0, \qquad \text{for all } j, \qquad (6.6)$$

implying (6.3). For (6.4) we assume that $\mathbb{P}_1 \succ \mathbb{Q}$, and denote by j_0 the first index such that $\sum_{i=0}^{j_0} \widetilde{q}_i > \sum_{i=0}^{j_0} \widetilde{p}^1_i$. It holds that $\widetilde{q}_{j_0} > \widetilde{p}^1_{j_0} \geq 0$, as well as $\widetilde{q}_{j_0} = \cdots = \widetilde{q}_{k(j_0)}$ and $\widetilde{p}^1_{j_0} \geq \cdots \geq \widetilde{p}^1_{k(j_0)}$, implying $\sum_{i=j_0+1}^{k(j_0)} \widetilde{q}_i \geq \sum_{i=j_0+1}^{k(j_0)} \widetilde{p}^1_i$. Consequently, it holds that

$$\widetilde{q}_{j_0} \sum_{i=0}^{k(j_0)} (\widetilde{p}^1_i - \widetilde{q}_i) < 0,$$

which together with (6.5) and (6.6) implies (6.4).

□

7 Implications for likelihood testing

There are several tests based on the quantile score that are frequently applied in the context of earthquake rate predictions. The first of these tests is the L- or data-consistency test that has been introduced in Kagan and Jackson (1995). It is a test for the hypothesis $\mathbf{y} \sim F$ that rejects if $\gamma(\mathbf{y}, F) < \alpha$ for some fixed threshold α. The authors of Zechar et al. (2010) developped two tests based on this test, which yield more detailed insight into miscalibrations of predictive models. The key idea is to consider the marginal spatial (S-test) and magnitudial (M-test) distributions separately. They make the underlying assumption that the predictive model is Poissonian, and therefore the marginal distributions are easily obtained by summing the expectations of the corresponding bins. Moreover, they condition the predictive distribution on having the same number of earthquakes as the observation, since the distribution of the number of earthquakes is typically assessed in a separate test, for more details we refer to Zechar et al. (2010). These tests are today still frequently applied to validate earthquake predictions (e.g. Pandey et al., 2019; Taroni et al., 2018).

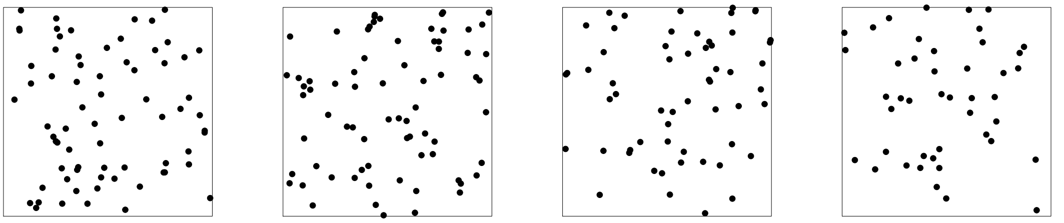
Our results from the last section indicate that it is problematic to apply these one-sided tests, since a higher expected quantile score not necessarily indicates that the predictive model is close to the true distribution, but that more uniform distributions achieve higher expected quantile scores. In particular, while the true model is accepted in these tests with probability $1 - \alpha$, higher acceptance probabilities can be achieved by reporting a too uniform model.

The following simulation study emphasizes that indeed more uniform distributions lead to higher acceptance probabilities in the one-sided tests. We consider bins $b_1, ..., b_{100}$ and assume that the number of occurrences in each bin is Poisson distributed and independent along bins. The predictive distribution is therefore fully specified by a vector $(\lambda_1, ..., \lambda_n)$ containing the expected number of events per bin. For a range of parameters $\mu > 0$ and $a > 0$ we investigate the performance of a Poisson model with $\lambda_1 = \mu - a$ and $\lambda_n = \mu + a$, and linearly interpolated in between, i.e. $\lambda_i := \frac{2(i-1)a}{n-1} + (\mu - a)$. This model gets more uniform when a approaches 0 (where all bins get assigned the same rate) and when μ increases (since a higher expected value yields a Poisson distribution that spreads probability mass more uniformly across $\mathbb{N}_0$). We consider four different models $F_1, ..., F_4$ specified by different choices of μ and a, see Figure 7. The parameters are chosen such that from F_1 to F_4 the distributions are decreasing in uniformity.

We consider each model both as true and as predictive distribution, for a total of 16 combinations, and compute for predictive distribution F_{pr} the expected quantile score $\mathbb{E}_{F_{\text{tr}}}[\gamma(\mathbf{Y}, F_{\text{pr}})]$ under true distribution F_{tr}. As done in applications, we approximate $\gamma(\mathbf{y}, F_{\text{pr}})$ by

$$\gamma_m(\mathbf{y}, F) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{l_F(\mathbf{X}_i) < l_F(\mathbf{y})\} \tag{7.1}$$

where $\mathbf{X}_1, ..., \mathbf{X}_m$ are i.i.d. samples of F_{pr}. We choose $m = 1000$, and approximate the



$F_1 : \mu = 0.6,\ a = 0$ $F_2 : \mu = 0.55,\ a = 0.1$ $F_3 : \mu = 0.5,\ a = 0.2$ $F_4 : \mu = 0.45,\ a = 0.3$

Figure 7. Example plots of the models $F_1, ..., F_4$. The spatial bins are sorted row-wise, with b_1 being in the lower left corner, b_{10} in the lower right, and b_{100} in the upper right corner.

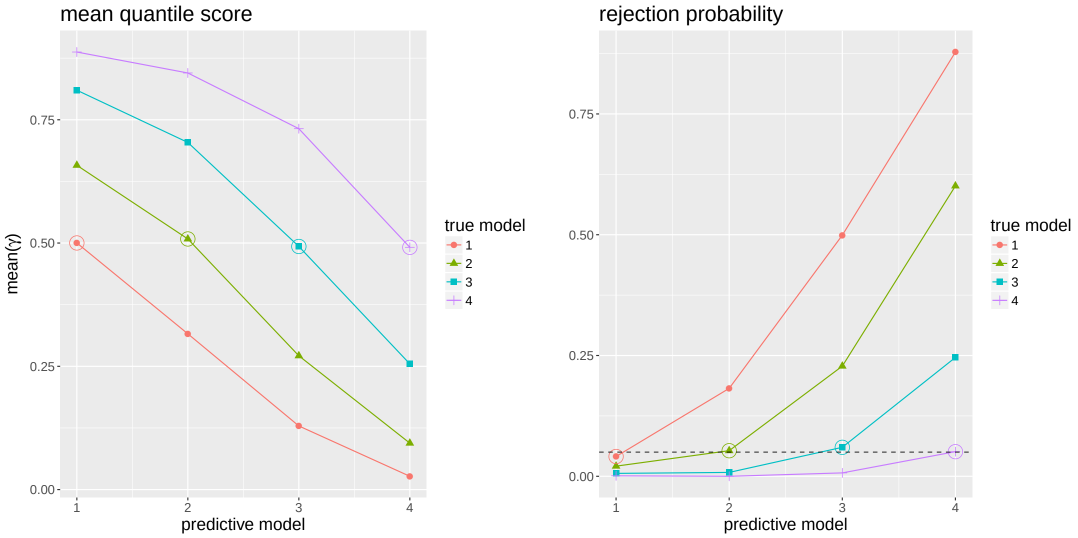


Figure 8. The left plot shows the mean quantile score $\mathbb{E}_{F_{\text{tr}}}[\gamma_m(\mathbf{Y}, F_{\text{pr}})]$ for predictive distribution $F_1, ..., F_4$ and true distribution $F_1, ..., F_4$. The value when the correct distribution is predicted is encircled. The right plot shows the rejection probability for the one-sided test at level 5% (dashed line), approximated by 1000 repetitions of the test.

expectated score by simulating $N = 1000$ observations, i.e. by

$$\mathbb{E}_{F_{\text{tr}}}[\gamma_m(\mathbf{Y}, F_{\text{pr}})] \approx \frac{1}{N}\sum_{k=1}^{N} \gamma_m(\mathbf{Y}_k, F_{\text{pr}}), \tag{7.2}$$

where $\mathbf{Y}_k \sim F_{\text{tr}}$ are independent. Figure 8 shows the expected quantile score of the different predictive models and the rejection probabilities of the one-sided L-test at a level $\alpha = 0.05$. The figure shows that, no matter which model is the true distribution, the mean quantile score decreases and the rejection probability increases with decreasing uniformity of the predictive distribution. It therefore indicates that, regardless of the true distribution, predicting more uniform distributions leads to better results in the L-test.

In the M- and S-test the simulation of γ_m is typically carried out conditional on the samples having the same number of earthquakes as the observation, i.e. in (7.1) $\mathbf{X}_1, ..., \mathbf{X}_m$ are sampled from F_{pr} conditional on $n(\mathbf{X}_i) = n(\mathbf{y})$. In order to demonstrate that also in this case too uniform distributions are preferred, we consider a second simulation study. In this study we moreover don't apply a binning to the observation window, and work

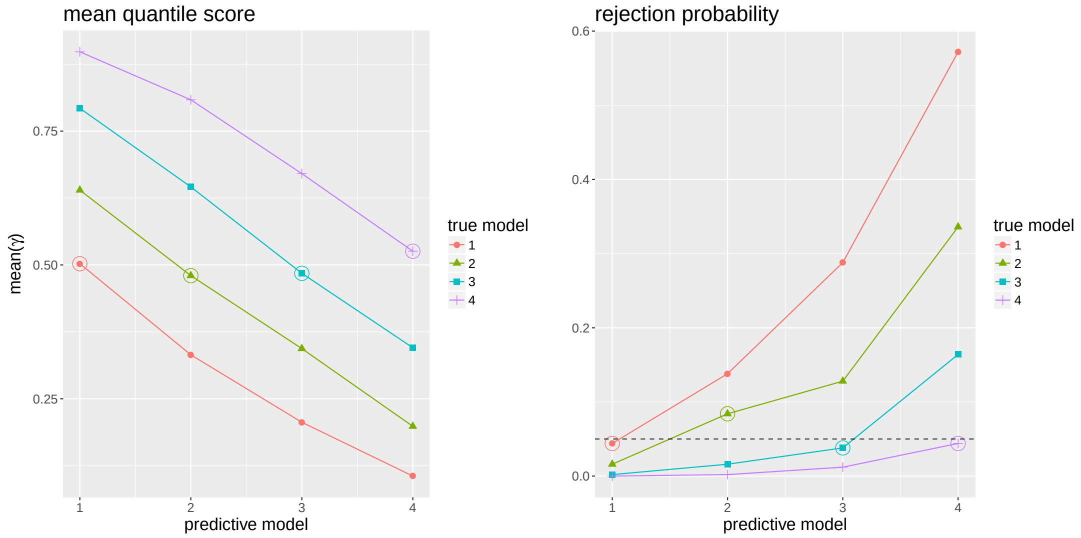


Figure 9. The left plot shows the mean quantile score $\mathbb{E}_{G_{\text{tr}}}[\gamma_m(\mathbf{Y}, G_{\text{pr}})]$ for predictive distribution $G_1, ..., G_4$ and true distribution $G_1, ..., G_4$. The value when the correct distribution is predicted is encircled. The right plot shows the rejection probability for the one-sided test at level 5% (dashed line), approximated by 500 repetitions of the test. For more uniform models the expected quantile score is larger and the rejection probability smaller.

with point processes directly, i.e. $\mathcal{O} = W^{\cup}$. Note that the proof given for Theorem 6.1 does not extend to this more general setting. We consider 4 Poisson processes $G_1, ..., G_4$ on $W = [0, 10] \times [0, 10]$, with intensity linearly increasing in x. The intensity for G_i increases from $1 - a_i$ to $1 + a_i$, i.e. $\lambda_i(x, y) = a_i(\frac{x}{5} - 1) + 1$, with $(a_1, a_2, a_3, a_4) = (0.1, 0.2, 0.3, 0.4)$. In particular, G_i decreases in uniformity for increasing i (intuitively, Definition 6.1 cannot be applied in this setting). We compute expected quantile scores as in (7.2), except that we set $N = m = 500$ and that in the evaluation of γ_m we simulate $\mathbf{X}_1, ..., \mathbf{X}_m$ as i.i.d. samples from the predictive distribution, conditional on $n(\mathbf{X}_i) = n(\mathbf{y})$. Moreover, we compute the likelihood to fail a one-sided test at level 5%. The results are shown in Figure 9, and indicate that our results carry seamlessly over to this more involved framework.

Under the null hypothesis $\mathbf{Y} \sim F$, the quantile score $\gamma(\mathbf{Y}, F)$ is clearly uniformly distributed on $[0, 1]$ and therefore a suitable test statistic for testing consistency of the predictive distribution with the observation. The problem we pointed out for the L-, M-, and S-test originates from the false belief that higher values of γ indicate better agreement between $\mathbf{Y}$ and F. Consequently, these tests reject only if the value of γ is unusually low. Our findings show that also unusually high values of γ also indicate inconsistency between $\mathbf{Y}$ and F. A natural solution is therefore to replace the classical one-sided tests by two-sided tests that reject if $\gamma_m(\mathbf{y}, F) \notin [\alpha/2, 1 - \alpha/2]$. A two-sided test punishes both too volatile and too uniform predictions and therefore seems more appropriate for consistency assessment. High values of γ do occur in practice, see for example Werner et al. (2011). Therefore, using two-sided tests would have major impact on which earthquake rate prediction models are rejected and which pass the data-consistency tests. Figure 10 shows the rejection probabilities for the two simulation studies in this section when two-sided tests are considered. The rejection probability of the two-sided test is minimized in

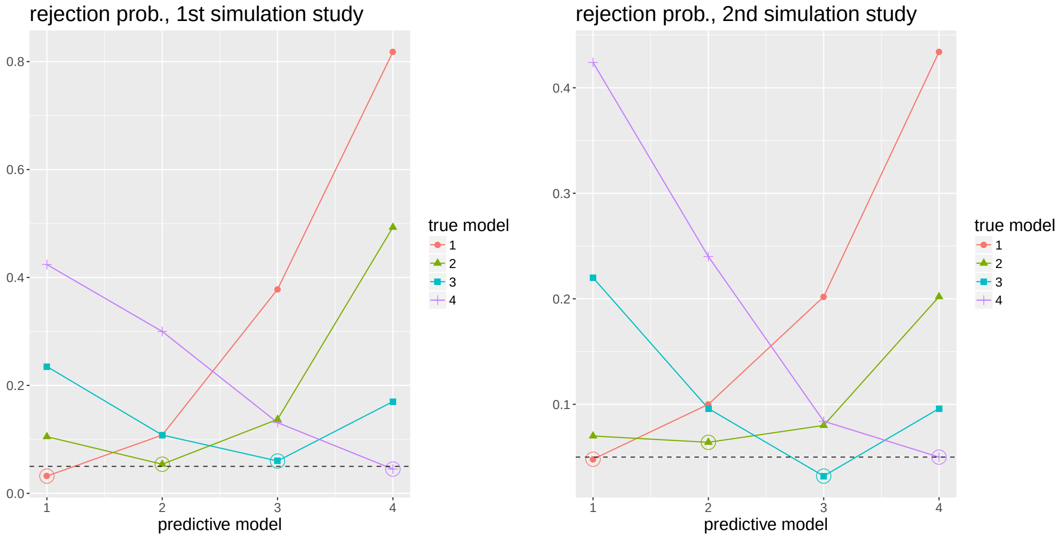


Figure 10. Rejection probabilities for both experiments presented in this section, when two-sided tests are considered. The left plot shows the rejection probabilities for predictive distributions $F_1, ..., F_4$, the right plot for $G_1, ..., G_4$. The true distributions are encirceled. For both tests and all distributions the rejection probability is minimized when the true distribution is predicted. The dashed line highlights the chosen threshold $\alpha = 5\%$.

all cases when the true distribution is predicted.

8 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 proper scoring rules for spatial point processes that can be applied to a wide range of predictive distributions. Even when comparing different Poisson models, where the logarithmic score is available, the kernel estimator score performed better in a simulation study, in the sense that its mean score is more robust against outliers. The R package spatstat Baddeley et al. (2015) provides manifold tools for simulating various 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 was used in all our simulation studies.

Our approach is based on the very 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 in itself can be difficult, even more so when they should be sensitive to certain high level properties of the observation-generating process. The mapping principle shifts this to the typically much easier task of finding real-valued functions sensitive to these properties. Possible other applications of this principle include high-dimensional forecasts and forecasts issued as spatial fields, as well as function-valued forecasts of any kind.

We argue that estimator of summary statistics are natural candidates for mappings sensitive to high-level properties of point processes. The resulting score then assesses whether the corresponding summary statistic is in good agreement between the predictive model and the observed data. This comes with the additional advantage that practicioners in the field of point processes are familiar with these summary statistics, making the output of the constructed scores easier to interpret.

We moreover showed that the quantile score constitutes an improper scoring rule. This scoring rule lies at the heart of several tests commonly applied in point process forecast validation, especially for earthquake predictions. We demonstrated that, as a consequence, these tests are systematically biased towards too uniform predictive distributions. Zechar et al. (2010) write about these tests:

> *"We are interested in the question: does* [the observed likelihood] *fall into the lower tail of the distribution of* [simulated likelihoods from the predictive model]*? If it does, this indicates that the observation is not consistent with the forecast..."*

Our results complement this intuition by showing that also falling into the upper tail indicates inconsistency with the forecast. Our simulation studies indicate that this systematic bias can easily be avoided by replacing the L-, S-, and M-test by two-sided tests based on the same test statistics.

Acknowledgments

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

A Appendix

We show the claim made after Definition 6.1.

Proposition A.1. *Let $\mathbb{Q}$ and $\mathbb{P}$ be distributions on $\{0, ..., n\}$ with $\mathbb{P} \succeq \mathbb{Q}$. Denote by $\mathbb{U}$ the uniform distribution on $\{0, ..., n\}$. It holds that*

$$D(\mathbb{P}, \mathbb{U}) \leq D(\mathbb{Q}, \mathbb{U}),$$

where D denotes the total variation distance of probability measures, defined as

$$D(\mathbb{P}, \mathbb{Q}) := \frac{1}{2} \sum_{i=0}^{n} |p_i - q_i|.$$

Proof. Denote by $\widetilde{p}_0, \widetilde{p}_1, ..., \widetilde{q}_0, \widetilde{q}_1, ...$ the ordered probabilities as in Definition 6.1. Further denote by k_p and k_q the smallest indices such that $\widetilde{p}_{k_p} \leq 1/(n+1)$ and $\widetilde{q}_{k_q} \leq 1/(n+1)$, respectively. It holds that

$$\begin{aligned} D(\mathbb{P}, \mathbb{U}) &= \frac{1}{2} \sum_{i=0}^{k_p} \left(\widetilde{p}_i - \frac{1}{n+1} \right) - \frac{1}{2} \sum_{i=k_p+1}^{n} \left(\widetilde{p}_i - \frac{1}{n+1} \right) \\ &= \sum_{i=0}^{k_p} \left(\widetilde{p}_i - \frac{1}{n+1} \right), \end{aligned}$$

where we used that $\sum_{i=0}^{n} \left(\widetilde{p}_i - \frac{1}{n+1} \right) = 0$. It holds now that

$$D(\mathbb{Q}, \mathbb{U}) = \sum_{i=0}^{k_q} \left(\widetilde{q}_i - \frac{1}{n+1} \right) \geq \sum_{i=0}^{k_p} \left(\widetilde{q}_i - \frac{1}{n+1} \right) \geq \sum_{i=0}^{k_p} \left(\widetilde{p}_i - \frac{1}{n+1} \right) = D(\mathbb{P}, \mathbb{U}),$$

where we used in the first inequality that k_q is the index maximizing the sum, and in the second that $\mathbb{P} \succeq \mathbb{Q}$. □

The inverse statement is not true. An example where $D(\mathbb{P}, \mathbb{U}) < D(\mathbb{Q}, \mathbb{U})$ but $\mathbb{P} \not\succeq \mathbb{Q}$ is, for $n = 2$, $(\widetilde{p}_0, \widetilde{p}_1, \widetilde{p}_2) = (\frac{7}{12}, \frac{1}{3}, \frac{1}{12})$ and $(\widetilde{q}_0, \widetilde{q}_1, \widetilde{q}_2) = (\frac{1}{2}, \frac{1}{2}, 0)$. Indeed, it holds that $\widetilde{p}_0 > \widetilde{q}_0$, and therefore $\mathbb{P} \not\succeq \mathbb{Q}$. On the other hand we have $D(\mathbb{P}, \mathbb{U}) = \frac{1}{4} + 0 + \frac{1}{4} = \frac{1}{2}$, whereas $D(\mathbb{Q}, \mathbb{U}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{3} = \frac{2}{3}$. Moreover, $\mathbb{P} \succ \mathbb{Q}$ does *not* imply the strict inequality $D(\mathbb{P}, \mathbb{U}) < D(\mathbb{Q}, \mathbb{U})$. A counterexample satisfying the former but not the latter is $n = 3$, $(\widetilde{p}_0, \widetilde{p}_1, \widetilde{p}_2, \widetilde{p}_3) = (\frac{3}{8}, \frac{3}{8}, \frac{1}{8}, \frac{1}{8})$, and $(\widetilde{q}_0, \widetilde{q}_1, \widetilde{q}_2, \widetilde{q}_3) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0)$.

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. 6, 7

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/. 12, 24

Baddeley, A. and Silverman, B. (1984). A cautionary example on the use of second-order methods for analyzing point patterns. *Biometrics*, pages 1089–1093. 8

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. 6

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, 27:65–81. 7

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263. 6

Gerstenberger, M., Rhoades, D., Stirling, M., Brownrigg, R., and Christophersen, A. (2009). Continued development of the New Zealand earthquake forecast testing centre. *GNS Science Consultancy Report*, 182. 5

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. *Journal of the Royal Statistical Society Ser. B*, 69:243–268. 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

Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media. 6

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. 16

Kagan, Y. Y. and Jackson, D. D. (1995). New seismic gap hypothesis: Five years after. *Journal of Geophysical Research: Solid Earth*, 100(B3):3943–3959. 21

Møller, J. and Waagepetersen, R. (2003). *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC. 6, 7, 8

Pandey, A., Chingtham, P., Prajapati, S., Roy, P., and Gupta, A. (2019). Recent seismicity rate forecast for North East India: An approach based on rate state friction law. *Journal of Asian Earth Sciences*, 174:167–176. 21

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. 7

Schorlemmer, D., Gerstenberger, M., Wiemer, S., Jackson, D., and Rhoades, D. (2007). Earthquake likelihood model testing. *Seismological Research Letters*, 78(1):17–29. 5

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

Taroni, M., Marzocchi, W., Schorlemmer, D., Werner, M., Wiemer, S., Zechar, J., Heiniger, L., and Euchner, F. (2018). Prospective CSEP evaluation of 1-day, 3-month, and 5-yr earthquake forecasts for Italy. *Seismological Research Letters*, 89(4):1251–1261. 21

Werner, M., Helmstetter, A., Jackson, D., and Kagan, Y. (2011). High-resolution long-term and short-term earthquake forecasts for California. *Bulletin of the Seismological Society of America*, 101(4):1630–1648. 23

Zechar, J. D., Gerstenberger, M. C., and Rhoades, D. A. (2010). Likelihood-based tests for evaluating space–rate–magnitude earthquake forecasts. *Bulletin of the Seismological Society of America*, 100(3):1184–1195. 21, 25