Computational Statistics (2025) 40:1583–1595
https://doi.org/10.1007/s00180-024-01540-y

ORIGINAL PAPER

A precise and efficient exceedance-set algorithm for detecting environmental extremes

Thomas Suesse[1,2,3] · **Alexander Brenning**[2,4]

Received: 25 January 2024 / Accepted: 12 August 2024 / Published online: 6 September 2024

Abstract
Inference for predicted exceedance sets is important for various environmental issues such as detecting environmental anomalies and emergencies with high confidence. A critical part is to construct inner and outer predicted exceedance sets using an algorithm that samples from the predictive distribution. The simple currently used sampling procedure can lead to misleading conclusions for some locations due to relatively large standard errors when proportions are estimated from independent observations. Instead we propose an algorithm that calculates probabilities numerically using the Genz–Bretz algorithm, which is based on quasi-random numbers leading to more accurate inner and outer sets, as illustrated on rainfall data in the state of Paraná, Brazil.

Keywords Geospatial models · Predicted exceedance sets · Kriging

1 Introduction

In environmental and health sciences, researcher often try to determine an exceedance region where the environmental variable of interest is exceeding a certain safety threshold, using point measurements on a set of locations. For example public and scientific interest in nitrate in groundwater and in particulate matter in the air is high.

✉ Thomas Suesse
thomas.suesse@uk-halle.de

Alexander Brenning
alexander.brenning@uni-jena.de

[1] National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522, Australia

[2] Friedrich Schiller University Jena, Department of Geography, Jena, Germany

[3] Deanery, Medical Faculty, Martin-Luther University Halle-Wittenberg, Halle, Germany

[4] Michael Stifel Center Jena for Data-Driven and Simulation Science, Jena, Germany

The legal limit for nitrate in groundwater in the European Union (EU) is 50 mg/l (Ohlert et al. 2023). For particulate matter in the air there are several limits, the European Air Quality Directive has thresholds of 40 $\mu g/m^3$ for PM10 and 20 $\mu g/m^3$ for PM2.5 concentrations, while the WHO air quality guideline has thresholds of 20 $\mu g/m^3$ and 10 $\mu g/m^3$ (Beloconi et al. 2018). The main public interest is in determining regions where these thresholds are exceeded. Then in these regions measures can be taken to reduce exposure levels.

Consider the univariate stochastic process

$$\{Y(\mathbf{s}) : \mathbf{s} \in D\}, \tag{1}$$

in the spatial domain D with $D \subset \mathbb{R}^d$. One of the major objectives in spatial statistics is prediction of $Y(\cdot)$ from observed data,

$$\mathbf{Z}(\mathcal{S}) \equiv (Z(\mathbf{s}_1^o), \ldots, Z(\mathbf{s}_n^o))^\top,$$

at known locations $\mathcal{S} \equiv \{\mathbf{s}_1^o, \ldots, \mathbf{s}_n^o\}$.

The observed variables $Z(\mathbf{s}_i^o)$ are considered to be a noisy version of $Y(\mathbf{s}_i^o)$ by adding a Gaussian measurement error

$$Z(\mathbf{s}_i^o) = Y(\mathbf{s}_i^o) + \varepsilon(\mathbf{s}_i^o);\ i = 1, \ldots, n, \tag{2}$$

where $\varepsilon(\cdot)$ is a Gaussian white-noise process with mean zero and variance σ_ε^2.

An exceedance region is a set of locations in the spatial domain D where the univariate stochastic process $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ exceeds some fixed threshold u, that can also vary with location $\mathbf{s}$, i.e. $u(\mathbf{s})$.

The greater exceedance region is defined as follows

$$G_Y \equiv \{\mathbf{s} \in D : Y(\mathbf{s}) > u(\mathbf{s})\}. \tag{3}$$

French and Sain (2013), French (2014), and French and Hoeting (2016) considered the estimation of outer and inner predicted exceedance sets, the former contains and the latter is contained in the true unknown greater exceedance set G_Y with a certain high pre-specified probability.

Formally, a $(1-\alpha)100\%$ outer predicted exceedance set G_O^α satisfies

$$Pr\big(G_Y \subset G_O^\alpha | \mathbf{Z}(\mathcal{S})\big) = 1 - \alpha; \tag{4}$$

and a $(1-\alpha')100\%$ inner predicted exceedance set $G_I^{\alpha'}$ satisfies

$$Pr\big(G_I^{\alpha'} \subset G_Y | \mathbf{Z}(\mathcal{S})\big) = 1 - \alpha'. \tag{5}$$

These sets and the probabilities in Eqs. (4) and (5) that must equal $1-\alpha$ and $1-\alpha'$, respectively, are obtained from the conditional distribution of $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ given the data $\mathbf{Z}(\mathcal{S})$. This distribution is obtained by Bayes' theorem and is often referred to as empirical Bayes since first the distribution of $Z(\cdot)$ (could be called prior) is estimated from the data $\mathbf{Z}(\mathcal{S})$, Cressie (1992).

The algorithms to obtain $G_I^{\alpha'}$ and G_O^α require resampling from the underlying geostatistical model and uses a pseudo-statistic

$$T_{kr}(\mathbf{s}) \equiv \frac{Y_{kr}(\mathbf{s}) - u}{\sigma_{kr}(\mathbf{s})}, \tag{6}$$

where $Y_{kr}(\mathbf{s})$ is the kriging predictor and $\sigma_{kr}(\mathbf{s})$ the kriging standard deviation for a given location $\mathbf{s}$. French and Hoeting (2016) considered a fully Bayesian approach to obtain $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ given the data $\mathbf{Z}(\wp)$, however for the construction of exceedance regions, which they called credible regions, they considered a multiple testing framework that has also been established by French and Sain (2013) and French (2014). This framework was used, because conditions (4) and (5) are equivalent to controlling the family-wise-error rate (FWER) (French and Hoeting 2016).

Since French and Hoeting (2016) considered a Bayesian approach they replaced in Eq. (6) the kriging mean and variance by Bayesian mean and variance estimators. They also considered two other test statistics by using slightly different denominators. Cressie and Suesse (2020) investigated more robust methods by using further adjusted test statistics.

The numerical algorithms to obtain G_O^{α} and $G_I^{\alpha'}$ require simulating a large number of realisations from the conditional distribution of $Y(\mathbf{s})|\mathbf{Z}(\wp)$ to obtain estimates of quantiles or alternatively probabilities. This estimation can lead to different numbers of rejected hypotheses depending on the particular random sample from this conditional distribution and can be very inaccurate by falsely rejecting or failing to reject several hypotheses due to the sampling error, resulting in incorrect sets G_O^{α} and $G_I^{\alpha'}$. In this paper we propose a different algorithm that directly calculates probabilities from $Y(\mathbf{s})|\mathbf{Z}(\wp)$, one for each hypothesis, using the numerical algorithms designed to accurately calculate probabilities from a multivariate normal distribution implemented in the `R` package `mvtnorm` (Genz et al. 2021) leading to more accurate sets G_O^{α} and $G_I^{\alpha'}$.

We review the exceedance set methodology in Sect. 2 including the main algorithm to determine exceedance sets. In Sect. 3 we present an equivalent formulation of this algorithm before we present in Sect. 4 a new algorithm. In Sect. 5 the new algorithm is illustrated using the Paraná rainfall data set and compared with the standard algorithm in terms of sizes of the inner and outer predicted exceedance sets. Finally the paper discusses the findings and future research.

2 Exceedance-set inference

In this section we outline the existing methodolgy and algorithm to obtain inner and outer predicted exceedance sets.

In the following the notation of Cressie and Suesse (2020) is used, unless otherwise stated. We make the assumption that $Y(\cdot)$ is a Gaussian process on $D \subset \mathbb{R}^d$ with mean function $\mu_Y(\cdot) \equiv \{\mu_Y(\mathbf{s}) : \mathbf{s} \in D\}$ and covariance function $C_Y(\cdot,\cdot) \equiv \{C_Y(\mathbf{s},\mathbf{v}) : \mathbf{s},\mathbf{v} \in D\}$.

For convenience we assume that μ_Y and C_Y are known; but realistically they need to be estimated. For example the simple kriging predictor, $Y_{kr}(\cdot)$, see (e.g., Cressie 1993, Ch. 3),

$$Y_{kr}(\mathbf{s}) \equiv E(Y(\mathbf{s})|\mathbf{Z}(\mathcal{S})) = \mu_Y(\mathbf{s}) + \mathbf{c}_Y(\mathbf{s})^\top \boldsymbol{\Sigma}_Z^{-1}(\mathbf{Z}(\mathcal{S}) - \boldsymbol{\mu}_Z), \tag{7}$$

can be used, where $\mathbf{c}_Y(\mathbf{s}) \equiv (C_Y(\mathbf{s}, \mathbf{s}_1^o), \ldots, C_Y(\mathbf{s}, \mathbf{s}_n^o))^\top$ and $\boldsymbol{\mu}_Z \equiv E(\mathbf{Z}(\mathcal{S})) = (\mu_Y(\mathbf{s}_1^o), \ldots, \mu_Y(\mathbf{s}_n^o))^\top$. The corresponding simple kriging covariance is given by

$$\text{cov}(Y(\mathbf{s}), Y(\mathbf{v})|\mathbf{Z}(\mathcal{S})) = C_Y(\mathbf{s}, \mathbf{v}) - \mathbf{c}_Y(\mathbf{s})^\top \boldsymbol{\Sigma}_Z^{-1} \mathbf{c}_Y(\mathbf{v}), \tag{8}$$

with the special case of the simple kriging variance $\sigma_{kr}(\mathbf{s})^2 = \text{cov}(Y(\mathbf{s}), Y(\mathbf{s})|\mathbf{Z}(\mathcal{S}))$. For other kriging predictors, such as universal kriging predictor, see (Cressie (1993), Ch. 3).

Instead of defining the greater exceedance set G_Y, we may also define the lower exceedance set L_Y

$$L_Y \equiv \{\mathbf{s} \in D : Y(\mathbf{s}) < u(\mathbf{s})\}. \tag{9}$$

Similarly to Eqs. (4) and (5), we could define inner and outer lower predicted exceedance sets L_I^α and $L_O^{\alpha'}$. Superscripts α and α' are in the following suppressed, unless needed. Here A^c denotes the complement of the set A. Notice that setting $L_I = G_O^c$ and $L_O = G_I^c$ will satisfy the conditions $L_I \subset L_Y$ and $L_Y \subset L_O$ with certain pre-specified probabilities $1 - \alpha$ and $1 - \alpha'$, respectively.

French and Sain (2013) formulated for each location $\mathbf{s}$ the following hypotheses test

$$H_0^G(\mathbf{s}) : Y(\mathbf{s}) > u \;\text{ versus }\; H_1^G(\mathbf{s}) : Y(\mathbf{s}) \leq u. \tag{10}$$

to determine L_I (and $G_O = L_I^c$, see Eq. (4)).

Similarly to determine G_I, see Eq. (5), (and $L_O = G_I^c$) the following hypotheses test is used

$$H_0^L(\mathbf{s}) : Y(\mathbf{s}) \leq u \;\text{ versus }\; H_1^L(\mathbf{s}) : Y(\mathbf{s}) > u. \tag{11}$$

The algorithms presented by French (2014) and French and Hoeting (2016) to determine G_I begin with a discretisation of the spatial domain into finite-grid representation of the original continuous index set. For convenience, we use the same notation D for this finite-grid representation. The nodes of the grid are $D \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$.

Standard G_I-Algorithm

(1) Conditional on $\mathbf{Z}(\mathcal{S})$, simulate B realizations of $\{Y(\mathbf{s}_i) : i = 1, \ldots, m\}$

$\{Y^{(b)}(\mathbf{s}_i) : i = 1, \ldots, m; \quad b = 1, \ldots, B\}$.

(2) For each realisation b we determine $S_I^{(b)} \equiv \{\mathbf{s}_i : Y^{(b)}(\mathbf{s}_i) < u; \quad i = 1 \ldots, m\}$
(3) Calculate $\Psi_b \equiv \max\{T(\mathbf{s}) : \mathbf{s} \in S_I^{(b)}\}$
(4) Determine $C_I^{\alpha'}$ the $(1 - \alpha')$-th quantile of $\Psi_1, \ldots, \Psi_B$
(5) Determine G_I by

$$G_I^{\alpha'} = \{\mathbf{s}_i : T(\mathbf{s}_i) \geq C_I^{\alpha'}; i = 1, \ldots, m\}. \tag{12}$$

Obtaining G_O can be achieved by applying the same algorithm and obtaining L_I first, by using the following properties: (i) $Pr(Y < u|\mu, \sigma^2) = Pr(Y > -u| - \mu, \sigma^2)$ and (ii) $L_I = G_O^c$ and (iii) $G_I = L_O^c$. Property (i) states that we can obtain lower exceedance sets from greater exceedance sets by changing the sign of u and μ, and properties (ii) and (iii) imply that the sets G_O can be obtained from L_I and L_O from G_I. So in general, the above algorithm can be used to obtain G_I and L_I and then the remaining sets G_O and L_O can be obtained, depending on whether G_I and G_O, or L_I and L_O are required. Due to this equivalence, w.l.o.g. we only consider the construction of G_I.

Obtaining the predictive distribution $\{Y(\mathbf{s}_i) : i = 1, \ldots, m\}|\mathbf{Z}(\wp)$ can be computationally demanding. A general algorithm to simulate realisations from a Gaussian random field has computational complexity $O((n+m)^3)$ due to calculating the Cholesky factorisation of the $(n+m) \times (n+m)$ covariance matrix $\mathbf{\Sigma}$ of observed and predicted locations (Givens and Hoeting 2012). The decomposition only has to be done once, before the simulaton process starts. There are other fine-tuned algorithms to simulate from the predictive destribution, for example turning bands method (TBM) for Gaussian random fields, see Chevalier et al. (2015), that can be faster for particular situations.

In the above G_I algorithm, quantiles are calculated to obtain critical values, or equivalently p-values and more generally probabilities are estimated by simulating a large number B of gridded data sets. Storing and processing matrices of size $m \times B$ is required. To be very accurate in the estimation of probabilities, the value of B is commonly a multiple of thousands, often 1000, 5000 or 10,000 (but could be much larger). To have a certain numerical accuracy in the probability estimates that are needed in the numerical algorithm (later defined as q_k) to determine the exceedance sets, or in other words to be within a certain error tolerance most of the time, say at least $1 - \alpha$, we can calculate the sample size B needed to have a margin of error of say $M = 10^{-3}$. Using the well-known formula for the margin of error (e.g. Tanur 2011)

$$B = \left(\frac{z_{\alpha/2}}{M}\right)^2 p(1-p). \tag{13}$$

Since p is often unknown and $p(1-p) \leq 0.25$, a conservative formula is

$$B = \left(\frac{z_{\alpha/2}}{M}\right)^2 0.25,$$

obtaining $B = 1{,}658{,}724$ for $\alpha = 1\%$ and $M = 10^{-3}$. Even storing a matrix of size $m \times B$ can be problematic, when m or B are very large. When decreasing M by a

factor of 10 (or 100), then B increases 100 (or 10,000) times to a value of approximately 166 million (or 16,6 billion).

In the following we propose a different algorithm that still requires the calculation of the conditional distribution, an m-dimensional Gaussian distribution of $\{Y(\mathbf{s}_i) : i = 1, \dots, m\}|\mathbf{Z}(\mathcal{S})$ with mean vector $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$, but calculates probabilities using a Quasi-Monte-Carlo (QMC) algorithm with high numerical accuracy ϵ without simulating a large number of realisations B from the conditional distribution.

3 An equivalent algorithm

In this section, we define adjusted p-values termed q_k that can be used to define a new but equivalent algorithm.

For notational convenience, $T_1, \dots, T_m$ denote the values of the test statistic $T_{kr}(\cdot)$ of the m grid locations $\mathbf{s}_1, \dots, \mathbf{s}_m$. The values $T_1, \dots, T_m$ are assumed to be in descending order $T_1 \geq T_2 \geq \cdots \geq T_m$ (i.e. $T_i = T_{(n-i+1)}$ is the $(n-i+1)$th order statistic), and the corresponding locations are $\mathbf{s}_1, \dots, \mathbf{s}_m$. Similarly, the null hypothesis $H_0^L(\mathbf{s}_i)$ is denoted by H_0^i and the alternative $H_1^L(\mathbf{s}_i)$ by H_1^i, $i = 1, \dots, m$. Large values of the test statistic should indicate deviations from the null hypothesis.

Consider the following equation

$$\begin{aligned} Pr(Y(\mathbf{s}) > u|\mathbf{Z}(\mathcal{S})) &= Pr\left(\frac{Y(\mathbf{s}) - Y_{kr}(\mathbf{s})}{\sigma_{kr}(\mathbf{s})} > \frac{u - Y_{kr}(\mathbf{s})}{\sigma_{kr}(\mathbf{s})}\middle|\mathbf{Z}(\mathcal{S})\right) \\ &= \Phi\left(\frac{Y_{kr}(\mathbf{s}) - u}{\sigma_{kr}(\mathbf{s})}\right) = \Phi(T_{kr}(\mathbf{s})). \end{aligned}$$

It shows a monotonic relationship between T_{kr} and $Pr(Y(\mathbf{s}) > u)$, suppressing conditioning on $\mathbf{Z}(\mathcal{S})$ for convenience. Under the null hypothesis, see Eq. (10), T_{kr} and $Pr(Y(\mathbf{s}) > u)$ are small and $Pr(Y(\mathbf{s}) < u)$ is large.

In steps 2 and 3 of the G_I-algorithm, the maximum Ψ_b is calculated. Next we consider its distribution under the null denoted by $\max_M T_i$, where M refers to the set of grid cells to be tested. Let us consider the distribution of $\max_M T_i$ under H_0^L of the set M. For example suppose $M = \{1\}$. Then $\max T_i = T_1$ when $Y_1 \leq u$ (null) or $\max T_i = -\infty$ when $Y_1 > u$ (alternative) with $P(\max_{i\in\{1\}} T_i = T_1) = P(Y_1 \leq u)$ and $P(\max_{i\in\{1\}} T_i = -\infty) = P(Y_1 > u)$, a discrete variable with two outcomes $-\infty$ and T_1.

More generally let $M \equiv \{1, \dots, l\}$, then $\max_{i\in M} T_i$ has a domain with $l+1$ values $T_1, T_2, \dots, T_l, -\infty$ with $T_1 > T_2 > \dots T_l > -\infty$, a discrete variable with $l+1$ outcomes.

The algorithm by French and Sain (2013) is equivalent to rejecting the first l hypotheses $H_0^1, \dots, H_0^l$ when $P(\max_{i\in M} T_i \geq T_l) \leq \alpha$. Define $q_k^I \equiv P(\max_{i\in I} T_i \geq T_k)$ for any $I \subset M \equiv \{1, \dots, m\}$. We reject $H_0^1, \dots, H_0^k$ if $q_k^M \leq \alpha$. The value q_k^M can be considered as an adjusted p-value for the hypothesis H_0^k. In the following the notation q_k is used without M, unless required.

4 Proposed new algorithm

As the calculation of q_k can be computationally demanding, we consider some simplifications, avoiding the calculation of q_k in many steps of equivalent algorithm. Then we formulate the newly proposed algorithm to obtain inner (and outer) predicted exceedance sets incorporating these simplifications.

Define $A_k \equiv \{Y_1 > u, \ldots, Y_k > u\}$ and $B_k \equiv \{Y_1 > u, \ldots, Y_{k-1} > u, Y_k \le u\}$. The following relationship holds $q_k^M = P(\max T_i \ge T_k) = \sum_{l=1}^{k} P(\max T_i = T_l) = \sum_{l=1}^{k} P(B_k)$.

We can express q_k^M as

$$q_k^M = P\left(\bigcup_{i=1}^{k}\left\{Y_i \le u\right\}\right) = 1 - P\left(\overline{\bigcup_{i=1}^{k}\left\{Y_i \le u\right\}}\right) = 1 - P\left(\bigcap_{i=1}^{k}\left\{Y_i > u\right\}\right)$$
$$= 1 - P\Big(Y_1 > u, Y_2 > u, \ldots, Y_k > u\Big) = 1 - P(A_k). \tag{14}$$

We also have $A_{k-1} = A_k \cup B_k$ with $A_k \cap B_k = \emptyset$ leading to

$$P(A_{k-1}) = P(A_k) + P(B_k) \tag{15}$$

and

$$q_k^M = 1 - P(A_k) = 1 - (P(A_{k-1}) - P(B_k)) = 1 - P(A_{k-1}) + P(B_k). \tag{16}$$

So we may calculate q_k in different ways, either directly $P(A_k)$ to obtain q_k, or calculate $P(B_k)$ to obtain $P(A_k)$ and finally q_k, but this requires using the previous $P(A_{k-1})$.

We can also construct bounds based on subsets of A_k and B_k. Define $A^{i_1,\ldots,i_l} \equiv P(Y_{i_1} > u, Y_{i_2} > u, \ldots, Y_{i_l} > u)$ and similarly $B^{i_1,\ldots,i_l} \equiv P(Y_{i_1} > u, Y_{i_2} > u, \ldots, Y_{i_{l-1}} > u, Y_{i_l} \le u)$ $B^{i_1,\ldots,i_l} \equiv P(Y_{i_1} > u, Y_{i_2} > u, \ldots, Y_{i_{l-1}} > u, Y_{i_l} \le u)$.

Suppose $\{i_1, \ldots, i_l\} \subset \{1, \ldots, k\}$, then $P(A_k) \le P(A^{i_1,\ldots,i_l})$ and this leads to a lower bound q_k^L for q_k, i.e.

$$q_k \ge 1 - P(A_{i_1,\ldots,i_l}) = q_k^L.$$

Similarly $B_k \subset B_{i_1,\ldots,i_l}$ if $i_1, \ldots, i_l \subset \{1, \ldots, k\}$, hence

$$P(B_k) \le P(B_{i_1,\ldots,i_l})$$

leading to an upper bound for

$$q_k = \sum_{l=1}^{k} P(B_k) \le \sum_{l=1}^{k} P(B_{i_1,\ldots,i_l}) = q_k^U.$$

In particular using sets with one index only leads to upper and lower bounds based on marginal probabilities

$$q_k^L = 1 - \min(p_1, p_2, \ldots, p_k), q_k^U = \sum_{l=1}^{k} \min(p_1, p_2, \ldots, p_{l-1}, 1 - p_l) \tag{17}$$

with $p_k = P(Y_k \leq u)$, which are easy to calculate with standard software. Notice that the marginal probability p_k equals $q_k^{k,\ldots,m}$, for example $q_1^M = 1 - P(Y_1 > u)$ is identical to $p_1 = P(Y \leq u)$. We could also term these marginal or unadjusted p-values to contrast them with the adjusted p-values q_k. We could also use different sets with several indices, for example two indices leading to pairwise probabilities. But since there are $O(k^2)$ of such pairwise probabilities, the benefit of using these lower dimensional probabilities is limited (due to the increase in these probabilities that need to be computed).

The calculation of q_k is only necessary to compare this value with α. If $q_k^M < \alpha$, then H_0^k is rejected, otherwise we fail to reject H_0^k and the algorithm stops. Hence knowing lower q_k^L and upper bounds q_k^U can help in making the decision, i.e. if $q_k^L > \alpha$, then we can stop, as we failed to reject H_0^k, likewise $q_k^U \leq \alpha$ then H_0^k is rejected. Using lower and upper bounds based on lower-dimensional integrals, such as marginal probabilities $P(Y_k > u)$ and $P(Y_k \leq u)$ can avoid more complex calculations.

We can also use the Holm procedure (Holm 1979) to accomplish this, i.e. when $p_k = P(Y_k \leq u) \leq \frac{\alpha}{m-k+1}$, then we can reject H_0^k without calculating an integral. To determine G_I we propose the following algorithm

New G_I-Algorithm

(1) $k = 1$
(2) If $p_k \leq \frac{\alpha}{m-k+1}$, then reject H_0^k and go to Step 8.
(3) If $q_k^L > \alpha$, then we fail to reject $H_0^k, \ldots, H_0^m$ and stop.
(4) Calculate q_k^U, if $q_k^U < \alpha$, then reject H_0^k go to Step 8.
(5) Calculate $\hat{q}_k$ and $\hat{\epsilon}_k$ using Genz–Bretz algorithm with tolerance $\epsilon = 0.01$, if $\hat{q}_k + \hat{\epsilon}_k < \alpha$, then reject H_0^k and go to Step 8.
(6) Calculate $\hat{q}_k$ and $\hat{\epsilon}_k$ using Genz–Bretz algorithm with lower tolerance, e.g. $\epsilon = 10^{-6}$, if $\hat{q}_k + \hat{\epsilon}_k < \alpha$, then reject H_0^k and go to Step 8.
(7) Retain $H_0^k, \ldots, H_0^m$ and stop.
(8) Set $k = k + 1$, proceed with Step 2.

In Step 2 the Holm algorithm is used. Steps 3 and 4 use the lower and upper bounds, see Eq. (17). In Step 5, the Genz–Bretz algorithm is used (Genz 1992, 1993; Genz and Bretz 2002) to calculate q_k, see Eq. (14)

$$q_k = 1 - \int_u^\infty \int_u^\infty \cdots \int_u^\infty f(\mathbf{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) dy_k \ldots dy_1, \tag{18}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of $\mathbf{Y} = (Y_1, \ldots, Y_k)^\top$ conditional on $\mathbf{Z}(\wp)$.

The Monte-Carlo algorithms by Genz (1992, 1993) use a transformation to calculate the value of the integral q_k and use a random sample of size N from the uniform distribution. Genz and Bretz (2002) considered instead a Quasi-Monte-Carlo (QMC) method by obtaining a series of quasi random numbers to achieve a higher convergence rate of $O(N^{-1})$ instead of $O(N^{-1/2})$. We refer to this QMC algorithm as Genz–Bretz algorithm, following the convention in the `R` package `mvtnorm` Genz et al. (2021). This algorithm also estimates the standard error of the probability estimate $\hat{q}_k$ denoted by $\sigma^2_{N,k}$. Since the interval $\hat{q}_k \pm 3\sigma_{N,k}$ based on a normal approximation contains most of the values (99.73% for a truly normal distribution), the term $\hat{\epsilon}_k = 3\sigma_{N,k}$ was considered by these authors as an error estimate. The user of the algorithm can set a pre-specified value of ϵ_k (error tolerance) and the algorithm increases N when $\hat{\epsilon}_k > \epsilon_k$ until $\hat{\epsilon}_k \leq \epsilon_k$. Often the estimated error $\hat{\epsilon}_k$ is very small, for example $\hat{\epsilon}_k < 10^{-6}$, even when ϵ_k is set to much larger values, e.g. $\epsilon_k = 0.01$, as in Step 5 of the algorithm. The criterion $\hat{q}_k + \hat{\epsilon}_k < \alpha$ uses the upper bound for the estimated q_k to ensure a correct decision is made with high probability. If the hypothesis in Step 5 could not be rejected we lower ϵ_k to a very small value and base inference on this new q_k - estimate. It appears that the implemented Genz–Bretz algorithm now uses $\epsilon = 3.5 \times \sigma_N$ as reported by Genz and Bretz (2002), which indicates a further improvement from 99.73% to 99.95% confidence.

5 Example

The computations are now illustrated on a spatial data set of long-term mean May-to-June precipitation in the state of Paraná, Brazil (Diggle and Ribeiro 2002). Figure 1 shows the state along with the 143 weather stations. The data set is freely available

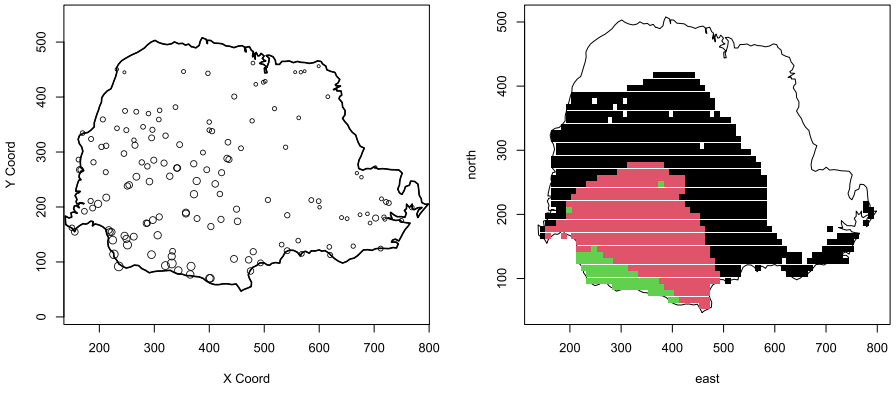

Fig. 1 Rainfall data (left) and predicted greater exceedance sets (right) for the Paraná data set. Left: Weather stations as circles scaled proportionally to May-to-June precipitation amount. Right: Inner and outer predicted exceedance sets with $\alpha = 0.10$ for a threshold value of $u = 300$mm, as well as plugin-predictor: G_I (green), plugin-predictor (green and red), G_O (green, red and black). Inner and outer sets obtained by new algorithm (color figure online)

through the R package geoR (Ribeiro Jr and Diggle 2018). We use the same fitted model as presented by Cressie and Suesse (2020) and choose a distance of 10 km between neighbouring grid points resulting in 1,953 grid points. Then we apply Universal Kriging using two predictors, the spatial coordinates, and obtain the corresponding mean vector $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$ from the predictive Gaussian distribution $\{Y(\mathbf{s}_i) : i = 1, \ldots, m\}|\mathbf{Z}(\mathcal{S})$ (Stein and Corsten 1991; Le Riche 2014).

We calculate G_I and G_O with both algorithms, the standard and the newly proposed with $\alpha = \alpha' = 0.10$. All computations are done on one core of a HPC, the Dell R750 PowerEdge.

First we discuss the results of the new algorithm. In total, 65 hypotheses were rejected to obtain G_I, and to obtain G_O, 711 hypotheses were rejected, yielding $|G_I| = 65$ and $|G_O| = 1953 - 711 = 1242$ grid points. Figure 1 shows the results of these inner and outer predicted exceedance sets. To obtain G_I, first 3 hypotheses were rejected in Step 2 (Holm) and then the next 61 in Step 4 (upper bound). To make decisions about hypotheses 65 and 66 the Genz–Bretz algorithm was applied. For hypothesis 65, $\hat{q}_{65} = 0.094992$ with an upper bound of $\hat{q}_{65} + \hat{\epsilon}_{65} = 0.095025 < \alpha$. Then $\hat{q}_{66} = 0.100426$ with lower bound $\hat{q}_{66} - \hat{\epsilon}_{66} = 0.10038 > \alpha$, hence we can confidently conclude that H_0^{65} must be rejected but H_0^{66} cannot be rejected. The average time to calculate $\hat{q}_{65}$ and $\hat{q}_{66}$ was 0.38 s with an average error of $\hat{\epsilon} = 0.000041$.

To obtain G_O, in Step 2 (Holm) 425 hypotheses were rejected. Then in Step 3 (upper bound) a further 274 hypotheses could be rejected. Then hypotheses 700–711 could be rejected with Genz–Bretz (Step 5), for example $\hat{q}_{711} = 0.099525$ with an upper bound of $\hat{q}_{711} + \hat{\epsilon}_{711} = 0.099525 + 0.000350 = 0.099876 < \alpha$. For H_0^{712} the Genz–Bretz algorithm could not provide a clear answer, because the initial estimate of q_{712} was $0.100208 > \alpha$ but its lower bound is $\hat{q}_{712} - \hat{\epsilon}_{712} = 0.100208 - 0.000336 = 0.099873 < \alpha$, making it unclear whether q_{712} is below or above α. Then Step 6 of the algorithm was applied yielding now a lower bound of $\hat{q}_{712} - \hat{\epsilon}_{712} = 0.100208 - 0.000020 = 0.100189 > \alpha$. We conclude 711 hypotheses could be rejected and the remaining could not be rejected, both with high confidence.

The less accurate Genz–Bretz algorithm with $\epsilon = 0.01$ for H_0^{711} has a very small estimated error of 0.000290, which is smaller than the corresponding error of the standard algorithm (3× standard error) given by

$$3\sqrt{\frac{q(1-q)}{B}}. \tag{19}$$

for most reasonable values of B and using $q = 0.10$ (as $\alpha = 0.10$).

Using $B = 1{,}000{,}000$ gives an error of 0.00090 and using $B = 10{,}000{,}000$ gives 0.00028. The more accurate Genz–Bretz with $\epsilon = 10^{-6}$ gave $\hat{\epsilon} = 1.8 \times 10^{-5}$, which cannot be matched for any practically reasonable value of B. The required B, using $z_{\alpha/2} = 3$ in Eq. (13) based on (19), would be $B = 6.24 \times 10^9$. To store even a single vector of this size in R allocates a large chunk of memory, in R approximately 46.5 Gb of memory, making the application of the algorithm to obtain the exceedance sets (on standard computers) impossible.

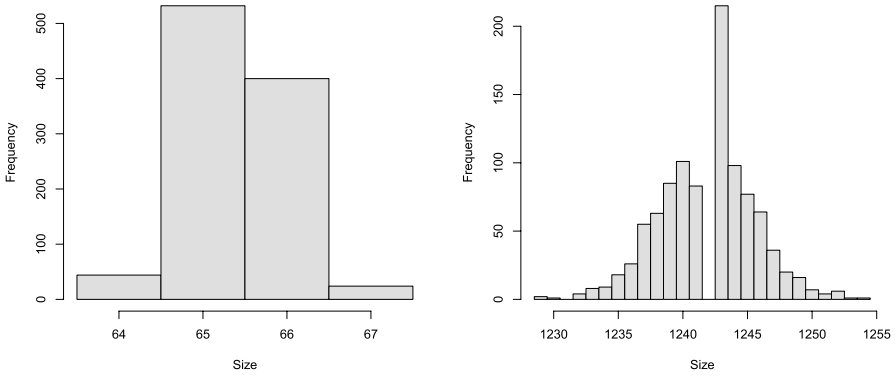

Fig. 2 Histograms of the size of the inner G_I (left) and outer G_O (right) predicted exceedance sets with $\alpha = 0.10$ for the Paraná data set based on 1000 samples each of size 10,000 from the predictive distribution of $\{Y(\mathbf{s}_i) : i = 1, \ldots, m\}|\mathbf{Z}(\mathcal{S})$

To evaluate the results of the old algorithm 1000 samples each of sample size 10,000 from the predictive distribution of $\{Y(\mathbf{s}_i) : i = 1, \ldots, m\}|\mathbf{Z}(\mathcal{S})$ were taken to calculate G_I and G_O for each of the 1,000 samples. Figure 2 shows the histograms of the sizes of the sets G_I and G_O. It can be seen for G_I approximately 50% of the samples rejected the correct number of 65 hypotheses, showing that the old algorithm gave incorrect results in approximately 50% of all samples (with either 64, 66 or 67 rejected hypotheses). For G_O, interestingly the correct size of 1242 was never obtained. The size of $|G_O|$ varied considerably from 1229 to 1254, demonstrating the inacccurate random results of the standard algorithm.

6 Discussion

We considered a new more precise algorithm to obtain inner and outer predicted exceedance sets. A drawback of the algorithm is that currently the Genz–Bretz algorithm in the `R` package `mvtnorm` is limited up to dimension $k = 1000$, even though the algorithms could be easily extended beyond 1000. Furthermore the Genz–Bretz algorithm computes a Cholesky factorisation for the current covariance matrix $\mathbf{\Sigma}_k$ in step $k = 1, \ldots, m$ of dimension $k \times k$, see Eq. (18). The algorithm could be improved by calculating a Cholesky factor once for the full $m \times m$ covariance matrix $\mathbf{\Sigma}$ (or better only for those hypotheses for which $p_k \leq \alpha$ leading to a much smaller submatrix of $\mathbf{\Sigma}$) and then using the upper left submatrix of dimension $k \times k$ in step k as the Cholesky factor that is needed for q_k (this repeated use is possible as the hypotheses are already in correct order).

The Monte-Carlo algorithm proposed by Genz (1992) is designed for arbitrary bounds. The algorithm could be further adapted using the fact that different q_k's require integration within overlapping domains. Suppose we need to calculate q_k and q_{k+1}. Then the integration bounds of the first k integrals are the same, see Eq.

(18). In addition, the main algorithm could be computationally improved further by simultaneously calculating several probabilities at once using that many bounds are identical (avoiding many calculations). We have not implemented and tested these possibly more efficient algorithms, as we are unsure about the exact current implementation of the algorithm. Future research might shed light on this issue.

7 Conclusion

We proposed a more precise algorithm to obtain inner and outer predicted exceedance sets in the sense that the decisions for each hypothesis can often be made with very high confidence, with the current implementation of the error estimate mostly with at least 99.73% probability. This is in contrast to the standard simulation-based algorithm where for some hypotheses no clear decision can be made, as we have demonstrated on the rainfall data set where in particular the sizes of the outer set varied considerably from 1229 to 1254 but never matched the true value of 1242 that was obtained from the new algorithm with very high confidence. In theory, the accuracy of the standard algorithm can be improved by increasing the sample size of the simulated data, but practically the accuracy of the new algorithm cannot be matched by the standard algorithm by simply increasing the sample size.

We have applied the methodology to a rainfall data set and we have shown some erroneous decisions for some regions, the grid points that were not identified to be part of G_O. For these rainfall data the consequences could be incorrect farming investment decisions, however this could have more severe and immediate consequences for other issues, for example when measures to curb nitrate in ground water or particulate matter in the air were incorrectly applied or were mistakenly not considered for some regions.

The methodology is not limited to geospatial data or spatial random fields, but it can be applied to any multivariate predictive distribution where its distribution is finite, here m. In this sense, this algorithm could be applied to other problems, for example calculating exceedance sets based on the predictive distribution of linear mixed models, simultaneous autoregressive models or common time series models.

Acknowledgements Thomas Suesse's research was supported by the Jena Excellence Fellowship Programme.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data and code availability R code to reproduce the results can be found at https://figshare.com/s/ca211f289cc089d06aa1.

References

Beloconi A, Chrysoulakis N, Lyapustin A, Utzinger J, Vounatsou P (2018) Bayesian geostatistical modelling of PM10 and PM2.5 surface level concentrations in Europe using high-resolution satellite-derived products. Environ Int 121:57–70. https://doi.org/10.1016/j.envint.2018.08.041

Chevalier C, Emery X, Ginsbourger D (2015) Fast update of conditional simulation ensembles. Math Geosci 47:771–789

Cressie N (1992) Smoothing regional maps using empirical bayes predictors. Geogr Anal 24:75–95

Cressie N (1993) Statistics for spatial data. Rev. ed, Wiley, New York

Cressie N, Suesse T (2020) Great expectations and even greater exceedances from spatially referenced data. Spatial Statistics **37**, 100420. https://doi.org/10.1016/j.spasta.2020.100420. frontiers in Spatial and Spatio-temporal Research

Diggle PJ, Ribeiro PJ Jr (2002) Bayesian inference in Gaussian model-based geostatistics. Geogr Environ Model 6:129–146

French JP (2014) Confidence regions for the level curves of spatial data. Environmetrics 25:498–512

French JP, Hoeting JA (2016) Credible regions for exceedance sets of geostatistical data. Environmetrics 27:4–14

French JP, Sain SR (2013) Spatio-temporal exceedance locations and confidence regions. Ann Appl Stat 7:1421–1449

Genz A (1992) Numerical computation of multivariate normal probabilities. J Comput Graph Stat 1:141–149

Genz A (1993) Comparison of methods for the computation of multivariate normal probabilities. Comput Sci Stat 25:400–400

Genz A, Bretz F (2002) Comparison of methods for the computation of multivariate t probabilities. J Comput Graph Stat 11:950–971

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2021) mvtnorm: multivariate normal and t distributions. https://CRAN.R-project.org/package=mvtnorm. R package version 1.1-3

Givens GH, Hoeting JA (2012) Computational statistics, vol 703. Wiley, New York

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70

Le Riche R (2014) Introduction to kriging. https://hal.science/cel-01081304/file/kriging_course_mnmuq2014_hal.pdf

Ohlert P, Bach M, Breuer L (2023) Accuracy assessment of inverse distance weighting interpolation of groundwater nitrate concentrations in Bavaria (Germany). Environ Sci Pollut Res 30:9445–9455. https://doi.org/10.1007/s11356-022-22670-0

Ribeiro Jr PJ, Diggle PJ (2018) geoR: analysis of Geostatistical Data. https://CRAN.R-project.org/package=geoR. R package version 1.7-5.2.1

Stein A, Corsten L (1991) Universal kriging and cokriging as a regression procedure. Biometrics 47:575–587

Tanur JM (2011) Margin of error. Springer, Berlin, pp 765–765. https://doi.org/10.1007/978-3-642-04898-2_34