Finding exceedance locations in a large spatial database using nonparametric regression

Gabrielle Elaine Moser, Sucharita Ghosh

Statistics Lab, Swiss Federal Research Institute WSL, Switzerland

ARTICLE INFO

Keywords:
Big data
Environmental monitoring
Kernel regression
Spatial data
Surface estimation

ABSTRACT

In the era of big data analysis, it is of interest to develop diagnostic tools for preliminary scanning of large spatial databases. One problem is identification of locations where certain characteristics exceed a given norm, e.g. timber volume or mean tree diameter exceeding a user-defined threshold. Some of the challenges are, large size of the database, randomness, complex shape of the spatial mean surface, heterogeneity and others. In a step-by-step procedure, we propose a method for achieving this for large spatial data sets. For illustration, we work through a simulated spatial data set as well as a forest inventory data set from Alaska (source: USDA Forest Services). Working within the framework of nonparametric regression modeling, the proposed method can attain a high degree of flexibility regarding the shape of the spatial mean surface. Taking advantage of the large sample size, we also provide asymptotic formulas that are easy to implement in any statistical software.

1. Introduction

In the context of analyzing complex systems involving analysis of spatial observations, one issue is to uncover spatial domains or areas where the observations display certain characteristics. In forestry or ecology, there are ample examples where this may be relevant. In general, understanding structural properties may be of interest; see for instance (Pommerening, 2002). In forestry, knowledge of spatial locations where the data show specific characteristics can lead to better forest management. In concrete terms, it is of interest to know at or around which geographical coordinates, the expected value (mean) of the random variable of interest substantially deviates from a given threshold. When this happens, we call this an exceedance location. Of particular interest are large spatial databases, such as data collected in nationwide forest monitoring programs. Due to the presence of randomness, it is not enough to perform simple searches in the databases. Appropriate estimation and testing procedures are needed that account for statistical properties of the data, so that local variations in the mean can be detected. It is also of interest to keep the entire procedure sufficiently flexible for applicability to various different data sets.

We consider a nonparametric method which is particularly suitable when the mean surface, in spite of being smooth, may have a complex shape over the entire range of the spatial scale. This is often the case when one is dealing with large scales such as nationwide inventory data. In particular, for instance, linear models or other simple parametric models may not be an adequate description of the spatial patterns. Therefore, instead of assuming a parametric form for the mean, we use a nonparametric function for the purpose. It turns out that given a large sample size, as long as the mean surface is smooth, i.e. satisfies some simple differentiability conditions, no matter which shape it has, an appropriate smoothing procedure can successfully estimate the unknown mean function. We use kernel smoothing for estimation and hypothesis testing for exceedance location detection, which uses distributional properties of the test statistic that rely on large sample theory of nonparametric curve estimators. This avoids the necessity of a normal distribution for the errors in the regression model. The asymptotic normal distribution of the test statistic is a result of the so-called central limit theorem, which is applicable when one has a large data set. The proposed method has excellent statistical properties, in particular since kernel smoothing can be carried out with high precision when one has a large sample. For details, see (Ghosh, 2018).

To start with, however, a definition of what is meant by exceedance is needed. A definition of threshold exceedance may differ depending on the problem and the area of application at hand. In particular, there can be various types of threshold exceedances. A spatial peak may be one idea where, the local mean is higher than the others in the neighborhood. However, not all local peaks may be of practical relevance. Some authors have defined exceedances in terms of derivatives, e.g. speed of...
change (first derivative) above a threshold. Such an example can be found in the palaeo sciences (time series) literature, where, past rapid climate changes define important climate epochs and, biotic responses to rapid climatic changes are of interest (Ammann et al., 2000). In this context, the idea of threshold exceedance is being used to define a rapid change point (Menendez et al., 2010). Threshold exceedance may also be related to the idea of risk assessment defined in terms of a probability of exceedance. The probability that a time series of annual precipitation means will exceed a given threshold is considered in Ghosh and Dra-ghiscescu (2002). In other sciences, such as in medical imaging, the problem of threshold exceedance translates to identifying areas of brain activity in response to different stimuli. Areas in the brain with active voxels can be thought of as regions with a high response value. Marchini and Presanis (2004) discuss numerous techniques for detecting active voxels, but the most common is to announce a voxel active if some associated test statistic is above a certain threshold. In forestry, in a different context, Innes et al. (1996) look at identification of trees with unusual foliage. While the problem there is not finding locations where the population mean exceeds a fixed threshold, i.e. exceedances in the mean, these authors apply the idea of exceedances to Mahalanobis distances from the population mean. They identify 3-dimensional vector of observations (color data) which are farther from the center of a nation-wide forest inventory data cloud, thus shifting the task of identifying true or forest stand characteristics from assessments in the field to a data analysis based approach. Other areas of potential applications may include studies of critical loads in soil science and their implications for forest growth, see e.g. Ouimet et al. (2001) and McNulty et al. (2013) among others. In this note, we consider a direct approach where the expected value locally exceeds a given threshold. In particular we emphasize the idea of ‘local estimation’, so that the search procedure can be adapted to local variations of the mean surface across the landscape. For further concepts, see Zhang et al. (2004), among others.

For a linear model, French and Sain (2013) consider finding exceedance regions for a random quantity g + e where g is a linear regression function and e is a normally distributed error term. The authors use simulations for this purpose. In another interesting work, Bolin and Lindgren (2017) consider excursion sets for Gaussian processes. In both articles, the assumption of an underlying normal distribution is relevant and authors focus on exceedances or excursions of a random quantity, whereas in the present manuscript, the focus is on exceedances of the regression function; in particular in our work, no assumption of Gaussianity is required and the regression function can have any arbitrary smooth shape, without the restriction of a linear model. The kernel estimator that we use is consistent (Ghosh, 2018; Wand and Jones, 1995), i.e. it will be asymptotically precise, no matter which shape our regression function has. If however, the regression part is non-smooth, instead of using a normal distribution based approach, a wavelet based approach may be considered (Daubechies, 1992). This will be addressed elsewhere. We detect exceedences by means of Z-tests, without having to use simulations, the validity of which follows from a central limit theorem and unlike in the works mentioned above, in our case the exceedence errors need not be normal. From a practical point of view, if the set of locations where observations are available span a large area, then various types heterogeneity may be expected. In such a case, the assumption of linearity or Gaussianity may be too restrictive. Our method is proposed in a broader context having a non-Gaussian and non-linear framework, where detecting exceedances of the mean (expected value) of interest, hence having different applicability compared to the works mentioned above.

The spatial mean surface is the expected value of the response variable, considered to be a function of the coordinates. We will use the notation μ(x, y) to denote the expected value of some response, W(x, y), such as plot mean of dbh at location with coordinate (x, y). It is clear that identification of the locations or the regions where this mean surface μ(x, y) is higher or lower than a user-specified threshold, say, η, can be highly relevant not just in forestry or ecology, but also in other areas such as medicine, sociology, meteorology etc., to name just a few. The regions where exceedances occur, may in particular point to episodes of ‘unusual’ events, hence requiring further investigation once these exceedence locations have been identified. In the context of forest economics, silviculture or forest resource management, data driven software for detecting exceedence locations can lead to faster decisions about possible management. However, difficulties may arise when the spatial scale is large, or if the signal is weak in the sense that the spatial mean surface is topographically somewhat flat. This, for instance, is the case for the tree diameter data set from Alaska (data source: USDA Forest Services) considered in this note. The data set contains geographic locations of forest stands and the sample mean of the observed tree diameter at breast height (dbh) at those locations. In the spatial scatter plot (Fig. 3.6a) however, no obvious clustering of large or small values are detectable. In cases like this, statistical procedures specifically designed for the purpose can reveal interesting structures.

The proposed method is a two-step procedure. In the first step, surface estimation is carried out using nonparametric regression. In the second step, hypothesis testing is used for identifying the exceedence locations. The estimation step needs optimal bandwidth selection. The hypothesis testing step needs a standardization test statistic which must be computable from the given data. This requires estimation of the asymptotic variance of the estimated mean. All formulas are explained and given; see Tables 2.1 and 2.4.

To set the technical context, we consider our spatial observations to be realizations of some spatial random field. The aim is statistical identification of locations or spatial coordinates where the local mean μ(x,y) at location (x, y) is above (or below) a given threshold η. In statistical terms, this corresponds to testing of hypothesis about μ(x, y). For the dbh data, considering E(dbh(x,y) = μ(x, y) to be the mean dbh at location (x, y), setting η at a user-specified numerical value, e.g. 10 inches, we may be interested in identifying those (x, y) coordinates where the null hypothesis that μ(x, y) does not exceed η = 10 is rejected at a given level of significance α.

The hypothesis to be tested for exceedance can be formulated in a straightforward manner, see e.g. (2.6); to test for non-exceedance, the directions of H0 and H1 would have to be reversed. The challenge is to estimate μ at location (x, y) ∈ (0,1)2 in a nonparametric setting and to construct a suitable test statistic, which requires the standard error of estimate. In this contribution, we have provided the complete estimation and testing procedure and proposed an algorithm for optimal bandwidth selection. The same method can also be used to interpolate exceedence locations because the location (x, y) to be tested for exceedance appears in all formulas only through the kernel and no observation at this location need be available. This is advantageous, when the interest lies in spatial prediction of locations with high timber volume, high species richness etc, where no data are available. Using available data, not only μ(x, y) can be interpolated, one can also test if (x, y) is an exceedence location.

Detailed information about spatial data analysis can be found in Cressie (2015), whereas nonparametric surface estimation using spatial data can be found in Ghosh (2018). For further information on kernel based estimation, see Wand and Jones (1995) and Silverman (1986) among others. The smoothness of the regression surface may be

| Table 2.1 |
| --- |
| **Theoretical expressions for bias, variance and asymptotic mean squared error (AMSE).** |
| Bias | \[ \frac{\sigma^2 K(du) \left[ m_{\text{dbh}}(x, y) \frac{b_1}{n} + m_{\text{dbh}}(x, y) \frac{b_2}{n} \right]}{\hat{\mu}_{\text{dbh}, b_1} \cdot \int K^2(du) + \tau_{2\alpha}} \] |
| Variance | \[ \sigma^2(\mu(x, y)) \cdot \frac{\hat{\mu}_{\text{dbh}, b_1} \cdot \int K^2(du) + \tau_{2\alpha}}{\text{remainder terms: can be ignored in large samples}} \] |
| AMSE | \[ \frac{\sigma^2(\mu(x, y)) \cdot \frac{\hat{\mu}_{\text{dbh}, b_1} \cdot \int K^2(du)}}{\text{remainder terms: can be ignored in large samples}} \] |
characterized via some mathematical properties such as finiteness of derivatives; for details see Ghosh (2015) and Rudin (1965). Also, the spatial data at hand may be heterogeneous, i.e., have non-constant variance. As mentioned earlier, while analyzing large spatial data sets, e.g., nationwide forest monitoring data, we wish to relax strict parametric assumptions about the underlying structures. In contrast to parametric regression, such as linear models, this nonparametric approach helps to maintain a certain amount of flexibility. Having a large number of observations is advantageous for nonparametric surface estimation methods and any subsequent statistical inference. In particular, under some general mathematical assumptions about the underlying structure, for a sufficiently large sample size, the variance of the surface estimate is inversely proportional to the effective number of observations used to estimate the surface mean at a given location. All else remaining unchanged, this effective sample size increases with the sample size.

Although no assumption is made about the parametric form of the underlying probability distribution function of the set of observations, the large sample theory of statistics allows us to construct an approximate $Z$-test (Casella and Berger, 2002). The theoretical background for this case, to achieve the same level of precision of statistical estimates, e.g., compared to exponentially decaying correlations, larger sample sizes are required. In particular, the effective sample size is then a function of the so-called long-memory parameter. Long-memory correlations can cause substantial differences in the quality of the inference, unless taken into account in the modeling and analysis stage. Additional information may be found in Beran et al. (2013) and Cressie (2015). Note that, no matter which type of correlations there are in the data, the formula for bias in nonparametric estimates of the regression surface will remain the same, as this will not be affected by correlations. For this aspect of the inference, the second derivatives $\mu_{xy}(x,y)$ and $\mu_{yy}(x,y)$ of the regression function become relevant. Thus, even for independent observations, estimation of these spatially varying derivative functions plays an important role for detecting exceedance locations. Moreover, for all correlation types, the error variance $\sigma^2(x,y)$ may also be spatially varying, being functionally dependent on locations $(x,y)$. How to accommodate this heterogeneity in the analysis is addressed in the Methods section.

The rest of the paper is organized as follows. Model specifications, methods of estimation and testing, and complete algorithmic details with R-codes appear in Section 2 under ‘Methods’. Applications to a data set (source: USDA) from Alaska is given in Section 3 under ‘Results’. Also included in Section 3 is a simulation of a data set (a toy example), with the purpose of further explaining how the method works. The Alaska data set can be downloaded from the homepage of the USDA. Discussions and conclusions appear in Section 4, followed by a list of references. Appendix contains further technical details.

### 2. Methods

The main steps to statistically identify an exceedance location can be described as follows: Let $W(x,y)$ denote an observation recorded at the spatial coordinate $(x,y)$ with expected value $\mu(x,y)$. For instance, the spatial coordinate $(x,y)$ may be the center of a forest plot on a forest monitoring network, whereas, $W(x,y)$ may be the average of diameters of trees sampled within that plot. One starts with nonparametrically estimating the smooth but unknown mean function $\mu(x,y)$, and subsequently, testing (2.6) at some level of significance $\alpha$.

For nonparametric surface estimation, we use kernel smoothing involving a kernel $K$ and bandwidths $b_1$ and $b_2$. The bandwidths are small positive numbers, which need to be chosen with care. It turns out that with increasing sample size, the bandwidths get smaller, but only at a certain rate. The kernel is a symmetric probability density function (pdf). In the numerical examples considered in this paper, we use a truncated Gaussian pdf with support $[-1,1]$. Up to some mild mathematical constraints, the choice of the specific kernel is less relevant for estimation (Silverman, 1986) than that for the bandwidths, the latter having stronger consequences for bias and variance of the estimator. Usually optimal bandwidths are calculated by minimizing an estimate of the mean squared error (MSE). The MSE is the sum of squared bias and variance. To use the same bandwidth-pair for all locations, global optimal bandwidths are computed. While a popular option is to minimize an estimate of the mean integrated squared error, we use spatial averages of the AMSE. A final quantity to be computed is the critical value that is needed to define the ‘rejection bounds’ for the hypothesis testing procedure. This involves estimation of bias and variance of the regression estimator.

A number of statistical packages exist which carry out nonparametric regression estimation. To our knowledge, exceedance location detection for spatial data using the method presented here in a non-linear and non-Gaussian framework has not been addressed in the literature; see however French and Sain (2013) for related work. The numerical calculations for this paper were carried out using S-Plus and the R statistical software. However, for convenience of the reader, all formulas are...
explicitly given in the paper. The entire procedure presented here is a stand-alone technique and can be implemented using any statistical software.

2.1. Model

Our method is based on a nonparametric regression model given in (2.2). Consider the response variable \( W \) on which spatial observations \( W_1, W_2, \ldots, W_n \) are available from a total of \( n \) locations with coordinates \((x_i, y_i)\), \( i = 1, 2, \ldots, n \). Explicitly

\[
W_i = W(x_i, y_i), \quad i = 1, 2, \ldots, n. \tag{2.1}
\]

The coordinates \((x_i, y_i)\) are assumed to be on a regular spatial monitoring grid. For unevenly spaced data, minor adjustments can be made to the formula for the regression surface estimator. We consider the situation when the plot centers are far apart and the centered observations (mean subtracted) are uncorrelated. Note however that like the situation when the plot centers are far apart and the centered observations are available from a total of \( n \) locations is an exceedance location with respect to this threshold. Obviously, the directions of the two hypotheses can be reversed, i.e., we may also consider non-exceedance locations, or two-sided exceedance locations, etc., and the ideas presented here generalize to all these cases.

To do so, we first define the null hypothesis and the alternative hypothesis as follows:

\[
H_0 : \mu(x, y) \leq \eta, \quad H_1 : \mu(x, y) > \eta. \tag{2.6}
\]

If \( H_0 \) is rejected at a level of significance \( \alpha \), \( 0 < \alpha < 1 \), then \((x, y)\) is an exceedance location with respect to threshold \( \eta \) and level \( \alpha \).

Let \( \mu(x, y) \) be an estimate of the unknown mean \( \mu(x, y) \). Due to the rather flexible conditions of \( \mu \), we will use a nonparametric kernel estimator defined earlier. As previously mentioned, the nonparametric estimator \( \hat{\mu}(x, y) \) will be asymptotically unbiased, i.e., its bias will converge to zero with increasing sample size. Also, because the number of observations \( n \) is large, we will be using approximate \( Z \)-tests, so that the rejection rule for the hypothesis testing problem can be based on the quantiles of the standard normal distribution. Specifically, our test statistic will be,

\[
Z = \frac{\hat{\mu}(x, y) - \text{bias}(\hat{\mu}(x, y)) - \eta}{\sqrt{\text{var}(\hat{\mu}(x, y))}} \tag{2.7}
\]

where the "bias" notation denotes an estimate based on data, and \( \text{bias}(\hat{\mu}(x, y)) \) and \( \text{var}(\hat{\mu}(x, y)) \) respectively denote the bias & the variance of the estimator \( \hat{\mu}(x, y) \); see Table 2.1.

We can reject the null hypothesis \( H_0 \) at the \( \alpha \) level if for the given data set, the observed value \( Z_{\text{obs}} \) of \( Z \) satisfies: \( Z_{\text{obs}} > z_{\alpha} \), where \( z_{\alpha} \) is the \( 100(1 - \alpha) \) quantile of the \( \text{N}(0, 1) \) distribution. We are using a one-sided critical value here, because we have defined our exceedance detection problem as finding locations where the expected value is likely to be higher than the threshold \( \eta \). If rejected, we call this location an exceedance location.

There is a vast literature on derivation of bias and variance of nonparametric curve estimators under a wide variety of technical assumptions. The specific formulas depend on the type of observations that are available for analysis. For the technical assumptions mentioned before, the necessary formulas (theoretical expressions) appear in Table 2.1. An outline of the derivations (proofs) of these formulas is given in the Appendix.

These formulas play crucial roles for optimum bandwidth selection as well as for defining the exceedance locations. We also need the error variance \( \sigma^2(x, y) \) to be estimated at each target location \((x, y)\). This is done by smoothing the squared residuals \( \hat{e}_i^2 \) where \( \hat{e}_i = y_i - \hat{\mu}(x_i, y_i) \). Theoretical properties of this type of variance estimators have been established elsewhere; see for example Ghosh (2015) and Herrmann (2000) and references therein. Moreover, estimation of bias needs second derivatives of \( \mu(x, y) \) to be estimated; see Gasser and Müller (1984) and Gasser et al. (1991). Additional details are described in the sequel.

2.2. Exceedance locations

Let \( \eta \) be previously specified and we define one-sided exceedance locations with respect to this threshold. Obviously, the directions of the two hypotheses can be reversed, i.e., we may also consider non-exceedance locations, or two-sided exceedance locations, etc., and the ideas presented here generalize to all these cases.

Before we identify exceedance locations, we need to estimate our regression function \( \mu \) nonparametrically. Recall that our observations are \( W_1, W_2, \ldots, W_n \) where \( n \) denotes the sample size. For instance, in the Alaska dbh data example, \( W_i \) is the sample mean of observed tree diameters in plot \( i \) having plot center coordinates \((x_i, y_i)\).

2.3. Estimation and testing

2.3.1. Estimation of the mean surface

We use the Priestley-Chao kernel regression estimator (Priestley and
Chao, 1972) to estimate the mean surface $\mu(x,y)$ at location $(x,y)$. This is given by

$$\hat{\mu}(x,y) = \sum_{i=1}^{n} a_i(x,y) W_i$$

(2.8)

The weights $a_i(x,y)$ are to be recalculated at each target location $(x,y)$ where surface estimation is needed and the location is being tested for exceedance. In a Priestley–Chao regression estimator, the formula for the weights would be given as:

$$a_i(x,y) = \frac{1}{h b_1} K\left(\frac{x-x_i}{h}, \frac{y-y_i}{b_1}\right)$$

(2.9)

If we use the Nadaraya–Watson regression estimator, then the formula for the weights becomes

$$a_i(x,y) = \frac{K\left(\frac{x-x_i}{h}, \frac{y-y_i}{b_2}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}, \frac{y-y_i}{b_2}\right)}$$

(2.10)

In (2.10), $\sum_{i} a_i(x,y) = 1$, whereas, when we have a large sample size, in (2.9), $\sum_{i} a_i(x,y) \sim 1$. In other words, because we have large sample size and fixed design (evenly spaced coordinates), (2.9) and (2.10) are equivalent. Now, these formulas involve a kernel $K$ and bandwidths $b_1$ and $b_2$. The estimator in (2.8) is therefore a weighted average of the observations $W_1, ..., W_n$ on moving windows. The window-size is decided by the bandwidths $(b_1, b_2)$.

An example of a kernel function is the truncated Gaussian kernel:

$$K(u) = \begin{cases} \frac{e^{-u^2/2}}{0.6827\sqrt{2\pi}} & |u| < 1 \\ 0 & \text{otherwise.} \end{cases}$$

(2.11)

This kernel satisfies all conditions listed in Table 2.2. Using this kernel, the weights to be substituted in (2.8) are (also see (2.10)):

$$a_i(x,y) = \frac{-0.5 \left(\frac{x-x_i}{h}, \frac{y-y_i}{b_2}\right)^2}{\sum_{i=1}^{n} e^{\frac{-0.5 \left(\frac{x-x_i}{h}, \frac{y-y_i}{b_2}\right)^2}} }$$

(2.12)

### 2.3.2. Optimal bandwidth selection

The bandwidths $b_1$ and $b_2$ appearing in formula (2.9) must be chosen with care; see Table 2.3. This step is crucial for optimal smoothing and the strategy for doing so is dependent on the property of the estimator (2.8) in large samples.

An algorithm for bandwidth selection is included in the proposed exceedance location detection method. All else remaining fixed, large bandwidths tend to increase bias and small bandwidths tend to increase variance. Thus, there is a trade-off between bias and variance. One popular approach is to minimize the MSE with respect to the bandwidths until an optimum solution is found.

For a data-driven solution, estimation of the MSE is needed. We refer to Table 2.4, where formulas for estimating $\text{Bias}(\hat{\mu}(x,y))$, $\text{Var}(\hat{\mu}(x,y))$, and the Asymptotic Mean Squared Error (AMSE) of the regression function estimator are given. The theoretical formulas for these quantities appear in Table 2.1. The estimates of these quantities require several steps and involve smoothing using a chosen kernel. Estimation of bias requires that the second derivatives of the regression functions be estimated. We use the Priestley–Chao estimators for estimating the derivatives. These are also consistent estimators further simplifying derivation of the formulas; see Priestley and Chao (1972). On the other hand, estimation of the error variance requires smoothing of the squared residuals. For all these smoothing operations, we have used the truncated Gaussian kernel defined in (2.11), but other kernel can be also used. Once the AMSE is estimated at each location $(x_i, y_i)$ $i = 1, 2, ..., n$, we compute a global value $\text{AMSE}_\text{global}$ of the ‘estimated asymptotic mean squared error’ (AMSE) by averaging over all $n$ locations:

$$\text{AMSE}_\text{global} = \frac{1}{n} \sum_{i=1}^{n} \text{AMSE}(\hat{\mu}(x_i, y_i)).$$

(2.13)

At the next step, $\text{AMSE}_\text{global}$ is minimized with respect to $b_1$ and $b_2$. As an exploratory effort, we set $b_1 = b_2 = b$ and let this be equal to $b$. A search for the optimum $b$ is then performed on a grid. For instance, for the Alaska data set, a grid of length 500, ranging from $b = 0.01$ to $b = 0.5$ is used. The optimum is found at $b = 0.155$ as shown in Fig. 3.7. The fact that all else remaining fixed, the theoretical AMSE (see Table 2.1) is a convex function of $b$, can be established by differentiation, leading to a unique value for the theoretical optimum bandwidth. Indeed, while setting $b_1 = b_2$ is an option for a preliminary search in an exploratory stage, a refined analysis will also need to include unequal bandwidths, involving bandwidth pairs $(b_1, b_2)$, where $b_1$ and $b_2$ may not be equal.

### 2.3.3. Identifying exceedance locations: testing

It turns out that due to the large sample size, a $Z$ – test can be carried out, the required theory for which is addressed in Ghosh and Moser.

### Table 2.4

| $n$ Total number of observations in the database | $W_i$ Observation at location $i = 1, 2, ..., n$ | $(x,y)$ Geographical coordinates scaled to the unit square $[0,1]^2$ | $(x,y)$ Coordinates (scaled) being tested for exceedance | $b_1, b_2$ Bandwidths used for estimation | $I_i$ Indicator function: $I_i = 1$ if $[x_i - x_i < b_1$ and $y_i - y_i < b_2]; I_i = 0$ otherwise. | $K(u)$ Truncated Gaussian kernel on $[-1, 1]: K(u) = \frac{1}{0.6827\sqrt{2\pi}}e^{u^2/2}; |u| < 1, K(u) = 0$ otherwise. | $U_2(K)$ | $R(K)$ | $\hat{\mu}(x,y)$ As in (2.8) | $\hat{\gamma}$ Residuals: $W_i - \hat{\mu}(x_i, y_i)$ | $\hat{\sigma}_y^2(x,y)$ Smoothed squared residuals | $\text{AMSE}_i$ An example of a kernel function is the truncated Gaussian kernel:

$$\begin{align*}
\text{Bias}(\hat{\mu}(x,y)) &= 0.5[\text{Bias}(\hat{\mu}(x,y)|b_1^2 + \text{Bias}(\hat{\mu}(x,y)|b_2^2)] \\
\text{Var}(\hat{\mu}(x,y)) &= \text{Var}(\hat{\mu}(x,y)|b_1^2 + \text{Var}(\hat{\mu}(x,y)|b_2^2)] \\
\text{AMSE}_i &= \text{Bias}(\hat{\mu}(x,y)|b_1^2 + \text{Bias}(\hat{\mu}(x,y)|b_2^2)]^2 + \hat{\sigma}_y^2(x,y)|b_1b_2) \\
\text{AMSE}_\text{global} &= \frac{1}{n} \sum_{i=1}^{n} \text{AMSE}(\hat{\mu}(x_i, y_i))
\end{align*}$$

(2.13)
(2019). Formula (2.7) gives the test statistic for a general kernel, whereas the calculations in Section 3 are done for a truncated Gaussian kernel. A location \((x,y)\) at which the null hypothesis has been rejected is labeled an exceedance location. For the given data set and a threshold \(\eta\), the (bias corrected) test statistic \(Z_{duk}\) is computed as follows:

For every target location \((x,y)\) to be tested for exceedance, define

\[
Z_{duk} = \frac{\hat{\mu}(x,y) - \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}(x_i,y_i)) \left( \frac{h_i}{h_{opt}} \right)^2 + \hat{\mu}_y(x,y) \left( \frac{h_{opt}}{h_i} \right)^2}{\sqrt{\hat{\sigma}^2(x,y)R(K)/\{nh_{opt}^2h_{opt}^2\}}} - \eta
\]  

(2.14)

The definitions of \(\hat{\mu}(x,y), \hat{\mu}_x(x,y), \hat{\mu}_y(x,y)\) and \(\hat{\sigma}^2(x,y)\) are in Table 2.4. Moreover, \(R(K) = \int K^2(u)du\) and \(U_2(K) = \int u^2K(u)du\). For the truncated Gaussian kernel on \((-1,1), U_2(K) \approx 0.29\) and \(R(K) = 0.51\). The null hypothesis is rejected at level \(\alpha\) if \(Z_{duk} > z_\alpha\), where \(z_\alpha\) is the \((1-\alpha)\) quantile of the standard normal distribution.

Thus, we say that the location \((x,y)\) is an exceedance location at threshold \(\eta\) and level of significance \(\alpha\), if

\[
\hat{\mu}(x,y) > C_{\alpha,\eta}(x,y)
\]

(2.15)

where the critical level \(C_{\alpha,\eta}(x,y)\) is given by

\[
C_{\alpha,\eta}(x,y) = \eta + \text{correction factor}
\]

(2.16)

correction factor \(= \frac{\hat{\mu}(x,y) - \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}(x_i,y_i)) \left( \frac{h_i}{h_{opt}} \right)^2 + \hat{\mu}_y(x,y) \left( \frac{h_{opt}}{h_i} \right)^2}{\sqrt{\hat{\sigma}^2(x,y)R(K)/\{nh_{opt}^2h_{opt}^2\}}} + z_\alpha \sqrt{\hat{\sigma}^2(x,y)R(K)/\{nh_{opt}^2h_{opt}^2\}}
\]

(2.17)

Similarly, other alternatives can also be tested. The entire set of calculations are summarized in the Tables 2.4 and 2.5.

2.4. Computational details

To implement the procedure, a number of estimation steps are needed before testing can be carried out. While any statistical software can be used for the purpose, R, the user may write a set of generic functions as indicated below, e.g. for estimating \(\mu\), derivatives of \(\mu\) or \(\sigma^2\).

For instance, to estimate \(\mu(x_0,y_0)\) at a given location \((x_0,y_0)\) using a bandwidth pair \((b_1,b_2)\) and kernel \(K\), a generic R function may look like

\[
\text{mu.hat} = \text{function}(x_0,y_0, X, Y, W, b_1, b_2, K) \text{ \text{where X and Y are the vectors of the spatial coordinates where raw data are available (e.g. plot center coordinates in the Alaska data set) and W = (W_1, W_2, ..., W_n) is the vector of all n observations (e.g. plot means of dbh) at locations (X, Y). The function mu.hat should implement the formulas (2.8)-(2.10). Next, to estimate the error variance \(\sigma^2(x_0,y_0)\), to be used to compute the standard error, one would replace \(\hat{\nu}\) in the above code by the squared residual vector \((\epsilon_1^2, \epsilon_2^2, ..., \epsilon_n^2)\) where \(\epsilon_i^2 = W_i - \hat{\nu}(x_i,y_i)\). Computation of the bias term needs estimation of the second derivatives of \(\mu\). In order to estimate a second derivative of \(\mu(x_0,y_0)\) for instance \(\frac{d^2}{dx^2}\mu(x_0,y_0)\), in mu.hat, one would plug in a different kernel, such as the second derivative \(K''(u) = \frac{d^2}{du^2}K(u)\) of the kernel used to estimate \(\mu(x_0,y_0)\). For further details, see Gasser et al. (1991), Fan and Gijbels (1996) and Herrmann (2000). These quantities are then substituted in the formulas for bias and variance, and finally a test statistic is computed at a given location. The entire set of formulas are summarized in Table 2.4.

2.4.1. R-codes

The calculations are done for a large data set. Thus, various large sample approximations are already integrated in the codes. The procedure for detecting exceedance locations involves several steps or modules. Here we demonstrate how some of these modules work. We use R to write our codes, however, using the formulas given in Tables 2.4 and 2.5, the method can be implemented using any statistical software.

1. Data scanning: The spatial coordinates are in the unit square: \((x,y) \in (0,1)^2\). \(W\) denotes observations at rescaled coordinates \((x,y)\).

- \(n = \text{number of triplets}: (W[i], x[i], y[i]), i = 1,2, ..., n\).
- \(W = (W[1], W[2], ..., W[n]) = \text{vector of all (diameter) observations to be smoothed}\)
- \(x = (x[1], x[2], ..., x[n])' = \text{vector of all rescaled x-coordinates}\)
- \(y = (y[1], y[2], ..., y[n])' = \text{vector of all rescaled y-coordinates}\)

\textbf{R code - data:}

\[
\text{mydata = read.table("myfile", header=T)}
\]
\[
\text{names(mydata)}
\]
\[
[1] "W" "x" "y"
\]

2. Kernel: The truncated Gaussian kernel and its second derivative have been used for estimation. Below, \(\alpha\) is assumed to be between –1 and 1 : \(u \in (-1,1)\).

\textbf{R code - kernel:}

\[
\text{kernel = function(uall) } \{ 
\text{K0.all=c(); K2.all=c()} 
\text{for(u in uall)} \{ 
\text{if(abs(u) > 1) } \{ \text{K0=0; K2=0} \} \text{ else } \{ \text{K0=dnorm(u)/(2*pnorm(1)-1)} \}
\text{K2=(u**2-1)*K0} \}
\text{K0.all=c(K0.all,K0)} \}
\text{K2.all=c(K2.all,K2)} 
\text{drop(list(u=u.all,K0=K0.all, K2=K2.all))} 
\}
\]

\text{Integrability: Here, we test that the above kernel satisfies the integrability condition: i.e. it integrates to 1.}

\textbf{R code - integrate:}

\[
\text{K0=functiuon(u) } \{ \text{kernel(u)@K0} \}
\text{integrate(K0, lower=-1, upper=1)}
\]

3. Estimation: The function below estimates the regression surface and its second derivatives at spatial (rescaled) location \((x_1,x_2) \in (0,1)^2\) using vectors \(x, y \text{ and W, bandwidth pair } (b_1, b_2) \text{ and a kernel. Here, we use the kernel that is defined by the kernel function above. Ideally, the following should be written up using matrices. However, for ease of understanding, here we provide a more direct description of the different steps.}

\textbf{R code - regression estimation:}
4. Error variance: To estimate the error variance function $\sigma^2$ we use the \texttt{mu.hat.spatial} function where we smooth the squared residuals.

**R code - error variance:**
```
mu0 = kernel(u01)K0
wv0 = w0 * v0 / (1/(n*b1*b2))
wv0 = wv0/sum(wv0)
mu0 = sum(wv0)*W
n0 = length(W[abs(u01) <= 1 & (abs(u02)<1)])]
b1.2 = b1*sqrt(n)***(1/10)
b2.2 = b2*sqrt(n)***(1/10)
u01 = (x-s1)/b1.2
u02 = (y-s2)/b2.2
w0 = kernel(u01)K0
w0 = w0 * v0 / (1/(n*b1.2*b2.2))
c0 = sum(wv0)
w2 = kernel(u01)SK2 v2 = kernel(u02)SK2
wv.20 = w2 * v0 / (n * b1.2**3 * b2.2)
wv.20 = wv.20/c0
wv.02 = w0 * v2 / (n * b1.2 * b2.2**3)
wv.02 = wv.02/c0
mu.xx = sum(wv.20 * W)
mu.yy = sum(wv.02 * W)
drop(list(x1=x1, x2=x2, x=x, y=y,
b1=b1, b2=b2, mu.hat=mu0,
mu.xx=mu.xx, mu.yy=mu.yy))
```

5. Bias, Variance, AMSE: At this step we obtain estimates of Bias, Variance and AMSE as well as the average AMSE over all locations.

**R code - amse:**
```
res=c(); mu.xx=c()
mu.yy=c(); mu.hat=c()
for(i in 1:n){
s1=x[i]; s2=y[i]
r0=mu.hat.spatial(x[i], y[i], x, y, W, b1, b2)
mu.hat=c(mu.hat, r0$mu.hat)
mu.xx=c(mu.xx, r0$mu.xx)
mu.yy=c(mu.yy, r0$mu.yy)
res=c(res, W[i]-r0$mu.hat)
}
sigma=sq(c)
for(i in 1:n){
sigma.sq.=mu.hat.spatial(x[i], y[i], x, y, res**2, b1, b2)$mu.hat
sigma.sq=c(sigma.sq, sigma.sq.i)
```

6. Optimum Bandwidth Selection: Consider a grid of bandwidth pairs \((b_1, b_2)\) and evaluate the average AMSE over all locations (AMSE, mean, see above) for each pair \((b_1, b_2)\) on the grid. The optimum bandwidth is defined to be the bandwidth pair \((b_1, b_2)\) for which the minimum of the average AMSE occurs.

7. Testing: For this step, the optimum bandwidth pair defined in step-6 is used. Let $\eta$ denote a threshold and let the level of significance be denoted by $\alpha$. If we fix the threshold at $\eta$ and level at $\alpha$, the (large sample) approximate $Z$-test can be carried out as follows:

**R code - testing:**
```
eta= eta0
alpha = alpha0
stat = (mu.hat - Bias - eta)/sqrt(Var)
pval = 1-pnorm(stat)
index = 1:n
index.ex=index[stat < alpha]
y.ex = x[index.ex]
y.ex = y[index.ex]
plot(x,y, pch="*"")
points(x.ex, y.ex, pch="o")
title("A toy example using simulated data")
```

3. Results

In this section, we analyze two data sets. The first data set is a toy example, a simulated data set, with standard normal errors and a known regression function. The second data set consists of diameter data from Alaska (data source: USDA).

3.1. A toy example

The purpose here is to demonstrate how estimation & testing works for a given data set. We simulate \((x_i, y_i, W)\), $i = 1, 2, ..., 100$ from the regression model (2.2) where we set the regression function to be:

$$
\mu(x, y) = 5x^2 + 5y^2 - 1.5x^3 - 1.5y^3, \quad (x, y) \in (0, 1)^2
$$

and the regression errors $\epsilon_i$ are pseudo random numbers (independent standard normal variables), with zero mean and unit variance, the $x_i$ and $y_i$ having been independently sampled from the uniform distribution on (0,1). The R-code for the data generating mechanism is shown below. Histogram of the simulated $W_i$ observations is in Fig. 3.1. The threshold to be tested for exceedance is set at $\eta = 3.5$.

**R code - simulation:**
```
set.seed(12345)
n=100
x=runif(n)
y=runif(n)
W=norm(n)
eta0=3.5
alpha_bonferroni = alpha0/n
```

To find the optimum bandwidth, we need to conduct a search on a grid of bandwidth pairs. Considering the grid $b = 0.05, 0.1, 0.15, ..., 1.40, 1.45, 1.50$, and setting $b_1 = b_2 = b$ for every $b$. AMSE is estimated at each $(x_i, y_i)$, $i = 1, 2, ..., n$. Finally, averaging the AMSE over all $n$ spatial locations, global AMSE is found. The summary statistics in Table 3.1 shows the resulting global (average) AMSE values for given bandwidth pair. The minimum of the average AMSE occurs at $b_1 = b_2 = 0.55$; also see Fig. 3.2. As discussed elsewhere in this article, high resolution grids with uneven combinations of $b_1, b_2$ should also be included to conduct refined searches of optimum bandwidths. The exceedance locations are shown in Fig. 3.3.
3.2. Alaska diameter data

In this sub-section we will go through a numerical example of this method using the 2014 forest inventory data set from the US state of Alaska. This data set is publicly available and can be found on the USDA Forest Services website http://www.fs.fed.us U.S. Department of Agriculture Forest Service (2015). This data set consists of 148,503 sampled trees from 2368 plots in the southern region of Alaska, (see Fig. 3.4), of which 2366 plot means were used for this analysis. This map was created in R 3.0.3 using the library ggmap, see Kahle and Wickham (2013). The point cloud is in Fig. 3.5; see for instance the cloud functions in R (Sarkar, 2008) as well as in S-plus (TIBCO Software Inc.).

To identify geographical coordinates where exceedance takes place, surface estimation is called for. The spatial distribution and the histogram of the diameter values are shown below in Fig. 3.6 (a) and (b). The histogram for the plot mean of the tree dbh is nicely bell-shaped. Note that, the procedure used in this contribution does not require the data to be so well-behaved, however, this property does make the prescribed methods ‘behave better’, without the need to transform the data in any manner. Roughly speaking, nicely tapering tail of the histogram has implications for any assumptions about the errors in the regression model, such as finiteness of the moments of the error distribution (see Appendix).

In Fig. 3.6, the spatial distribution of the plot averages lacks any obvious clustering, the darker dots indicating plot means higher than the threshold 10.65 inches. Thus a finer search based on local estimation and testing would be needed to identify significant spatial patterns, and no simple search in the database would suffice as such a search would be affected by randomness in the data.

In order that a hypothesis testing can be carried out, first of all, the local mean needs to be estimated. In the nonparametric smoothing procedure that we use, while the choice of the kernel can be done from a list of several kernels that are available, bandwidth selection needs to be carried out because of its effect on bias and variance of the estimator. For this, using the formulas in Table 2.4, we plotted the estimated global AMSE (see Eq. (2.13)) for different values of the bandwidth $b$ where, for
simplicity, we set \( b_1 = b_2 = b \). Then we choose that bandwidth which gives rise to the minimum global AMSE (see Fig. 3.7). In particular, we have for this data set, \( b_{opt} = 0.155 \), since this gave the smallest estimated global AMSE. The general steps for writing a generic code for implementing the above formulas is described in the Methods section.

Now that we have chosen an estimate for the regression function, derived its expected value, bias and variance, constructed a test statistic and chosen our bandwidths, we can begin to test for exceedance locations for our spatial data. The average mean diameter at breast height for the Alaska data set is 10.07 inches. The summary statistics are given in Table 3.2. Note that being based on a smoothing operation, the location \((x,y)\) appears in all formulas through the kernel. Thus, not only the mean \( \mu(x,y) \) can be estimated (interpolated) at any \((x,y)\) location via smoothing, one can also test if \((x,y)\) is an exceedance location. In particular, as in a theorem in Bradley (1983), it is possible to establish that nonparametric regression estimates at two distinct and fixed locations will be asymptotically independent, having implications for hypothesis testing.

Based on the estimated mean values at each plot computed with a global optimal bandwidth of \( b_1 = b_2 = 0.155 \), exceedance locations were detected using the methods described in the previous section. Fig. 3.8 shows individually tested locations of forest sites where the mean dbh significantly exceeds \( \eta = 10.65 \) inches \( (\alpha = 0.05) \). Out of a total of 2366 sites, 1091 sites are detected as exceedance locations. Lifting the threshold decreases the number of exceedance locations substantially for this data set. For instance, for \( \eta = 10.75 \) at \( \alpha = 0.05 \), a total of 197 plots can be detected to be exceedance locations (Fig. 3.9).

While the spatial scatter plot of the raw plot means in Fig. 3.6 does not show any obvious spatial clustering of the data, as a first step, individual plots are tested for exceedance. The exceedance locations in Figs. 3.8 and 3.9, do show clearer spatial tendencies. Using the proposed method, it is possible to identify regions of the map where the 'thicker' trees may be expected to occur. Changing the threshold \( \eta \) and the level of

![Fig. 3.4. A map of the sampled plots in Alaska.](image-url)

![Fig. 3.5. Alaska diameter data - raw observations: The data cloud for the 2366 observations from the Alaska dbh data set.](image-url)
significance $\alpha$ would change the locations of these clusters, which can be adjusted by the user. However, for a large spatial data set, instead of testing individual plots, one may be interested in regional patterns, in particular if there are larger regions with tendencies of high mean tree diameter.

**Regional testing** We define a region to be an area consisting of groups of plots. To test if a regional expected value will exceed the threshold $\eta$ or not, one would start with stratifying the large landscape into regions or groups of plots. Each such region can then be tested for exceedance. We superimpose a regular grid on the map, thus stratifying the entire landscape consisting of $n = 2366$ plots into regions or groups. The grid

![Spatial distribution of dbh](image1)

![Histogram of dbh](image2)

![Alaska diameter data - raw observations](image3)

![Alaska diameter data summary statistics](image4)

![Alaska diameter data - individually tested locations](image5)

![Regional optimal bandwidth](image6)

![Global optimal bandwidth](image7)

![Each location tested](image8)

**Example** We define a region to be an area consisting of groups of plots. To test if a regional expected value will exceed the threshold $\eta$ or not, one would start with stratifying the large landscape into regions or groups of plots. Each such region can then be tested for exceedance. We superimpose a regular grid on the map, thus stratifying the entire landscape consisting of $n = 2366$ plots into regions or groups. The grid

**Table 3.2** Alaska diameter data summary statistics (data source: USDA).

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max.  |
|------|--------|--------|------|--------|-------|
| 1.100000 | 7.582400 | 9.749138 | 10.068560 | 12.165541 | 27.627273 |

The bandwidth that gives the smallest global AMSE (see (2.13)) is used in subsequent analysis.

![Fig. 3.7. Alaska diameter data - global AMSE](image9)

![Fig. 3.8. Alaska diameter data - individually tested locations](image10)
region may be defined as $T_k = \sqrt{Z_{\text{obs,}A_k}} \text{,}$ where, $Z_{\text{obs,}A_k}$ is the test statistic at an individual plot inside $A_k$ and $Z_{\text{obs,}A_k}$ is the sample mean of these $k$ statistics. Recall that $Z_{\text{obs,}A_k}$ (see (2.14)) is asymptotically standard normal, so that $T_k$ is also asymptotically standard normal under the null hypothesis. Let $p_k = P(Z > T_k)$ where $Z \sim N(0,1)$ denote the usual $p$-value. One can then use e.g. the p.adjust function in R for multiple comparisons, to obtain adjusted $p$-values. For the purpose of illustration, we use the default option (holm) but any other option can also be used. The results are shown in Figs. 3.10 and 3.11, where exceedance regions are highlighted, both before and after multiple comparison adjustments. We have discussed tests for both individual plots as well as regions or groups. Thus the purposes differ so that the test statistics and results for these tests are different. However, the general tendencies as to where the exceedances may be found in the entire landscape seem somewhat similar. It is important to note that only after considering multiple comparison adjustments (adjusted $p$-values) the overall level of significance may be set at the desired level, e.g. $\alpha = 0.05$. If multiple comparison adjustments are not made, the level of significance is valid for single plots or groups but not for the overall tests.

4. Discussion

The paper gives a detailed account of a nonparametric procedure for detecting exceedance locations in large spatial data sets. While doing so, in particular, the various formulas have been given explicitly, so that these can be easily implemented using any programming language and used for analysis of other data sets. In this paper, the calculations were done using the R software environment and illustrated using a simulated data set as well as a forestry data set from the state of Alaska, USA. The Alaska diameter data are maintained by the USDA Forest Service.

In contrast to the literature, in our case, no parametric model assumption is imposed on the regression function, and also, the error distribution is allowed to belong to a larger class of distributions, of which the Gaussian is only a special case. While we identify exceedances of the regression function based on asymptotic theory in a nonparametric and non-Gaussian context, the methods and results remain valid also if the regression function can be declared a linear model and the errors are Gaussian. This is so because, due to the consistency property, optimal smoothing will correctly estimate any regression function including linear models. The proposed method can test if a location is an exceedance location without the need of simulations. Smoothness of the regression surface leads to visualizing some clusters of the exceedance
locations. In the examples shown, exceedances can be found where, on the average, values larger than the threshold tend to occur.

It is important to keep in mind that the methods discussed in this paper were tailored specifically to large data sets, and would most likely lead to biased results if the sample size was not sufficiently large. For simplicity, for estimating the various unknown functions such as the derivatives of $\mu$ and the error variance, the same bandwidth pair $b_1$ and $b_2$ have been used. The aim here has been to see which bandwidth pair gives rise to the minimum AMSE. Needless to say, in a second level analysis of these data, faster convergence rates may be possible to achieve by using other bandwidths for estimating these functions. This is not pursued here, since the aim has been to derive a simple tool for an exploratory analysis of a large spatial data set. The large sample size is specifically important for all of the asymptotic assumptions we make throughout this paper and without it, the consistency arguments would not hold. Similarly, our approach would need to be modified if our plot coordinates were randomly spaced rather than on an evenly spaced grid. In this case, one would need to adjust the formulas for the expected value and the variance of a ratio of random variables. In the computations done in this paper, we have used a separable kernel that can be factorized with respect to the coordinates. As a referee points out, more generally, one may consider a two-dimensional kernel. Needless to say, the estimator will remain consistent, i.e., with MSE converging to zero with increasing sample size, as long as the kernel satisfies certain basic conditions, such as total integral is unity and a few other criteria. However, in detailed analysis, using kernels that are not product of two univariate kernels would be an interesting topic to look into, especially, as the referee points out, in cases of anisotropic behavior.

One particular advantage of our approach is that, in addition to testing if a location in the data set, where observations are available, is an exceedance location or not, it also can test if a location where there are no observations, has also the potential to be an exceedance location. This is possible since the test statistic is based on the idea of interpolation of the mean surface $\mu(x,y)$ as well as interpolation of the variance $\sigma^2(x,y)$. We use nonparametric estimation methods, because this type of problems typically occur where the surfaces encountered are often complex. But in principle, this can also be done using parametric regression models, where however, one would have to find models that fit the entire range of the spatial observations. For large spatial scales, e. g. for nation-wide data, this task may be formidable. Therefore, use of nonparametric approach may be easier when one is dealing with very large spatial scales, and expects that local features may not be the same everywhere.

Needless to say, the proposed method is also easily transferable to answer the question, if there has been a change in the landscape over two time periods. One would simply take the difference of surface mean estimates at two time points and carry out the hypothesis testing by setting the threshold $\eta = 0$. A rejection of the null hypothesis in that context would imply that a change has occurred. We consider a general threshold $\eta$ and answer the question if the unknown population mean $\mu(x,y)$ at location $(x,y)$ exceeds the threshold $\eta$.

Among various areas of applications, finding hotspots is a popular aim and the ideas presented here can also be related to uncovering hotspots. While we use the idea of threshold exceedance and statistical significance in a spatial context, some other authors have also considered use of spatial statistics to identify hotspots; see for instance (Harris et al., 2019), or locations with high species diversity etc. To account for heterogeneity, it is recommended that standardized test statistics be used to identify exceedances and their locations. The proposed method takes into account randomness (probabilistic structure) in the data and uses the large sample size to its advantage (leading to simpler formulas). The exceedance locations are functions of the threshold $\eta$ and the level of significance $\alpha$, both of which can be adjusted depending on the user’s need.

We have analyzed a forest inventory data set from the state of Alaska, USA (data source: USDA Forest Services), as well as a simulated data set, in order to work through numerical examples. The proposed steps are simple to program and can be implemented in standard software. The fact that nationwide forestry data sets are so large leads us to the conclusion that a method tailored specifically to large data should be established. The goal here has been to propose a statistical procedure for detecting locations in a large spatial data set given a previously specified threshold, i.e., statistically identify exceedance locations. The modern theoretical know-how from nonparametric curve estimation procedures have been specially adapted in the spatial context to identify exceedance locations & to provide with a user-friendly tool.

**Declaration of Competing Interest**

We declare no conflict of interest.

**Acknowledgments**

We would like to thank the Associate Editor and two referees for detailed comments that led to a substantial improvement of the manuscript. Part of this research is based on the Master thesis of the first author at the Department of Mathematics, University of Zürich. The tree diameter data analyzed here is available from the USDA Forest Services website [http://www.fs.fed.us](http://www.fs.fed.us). Permission to use these data is gratefully acknowledged. Data analysis was carried out using S-PLUS (TIBCO Software Inc.) and R (The R Foundation for Statistical Computing).

**Appendix.**

1. **Assumptions on the regression model:**
   
   (a) $\epsilon_i$ are pairwise uncorrelated, so that, $\text{Cov}(\epsilon_i, \epsilon_j) = 0, i \neq j$.
   
   (b) $\epsilon_i$ have zero mean and finite variance, whereas the error variance need not be a constant. Let $\text{var}(\epsilon_i) = \sigma^2 = \sigma^2(x_i, y_i)$.
Bias and Variance

Asymptotic Mean Squared Error

- 

When the sample size is large, the expression for the Asymptotic Mean Squared Error (AMSE) is used for bandwidth selection. This expression is simpler to use in computations and an adequate approximation for (6.5). The AMSE, which will also go to zero as \( n \) goes to infinity can be written as in Eq. (6.6).

\[
\text{AMSE}(\hat{\mu}(x, y)) = \left[ \int u^2 K(u) du \left\{ m_{xx}(x, y) \frac{b_1^2}{2} + m_{yy}(x, y) \frac{b_2^2}{2} \right\} \right]^2 + \frac{1}{nb_1 b_2} \sigma^2(x, y) \int K^2(u) du
\]

If we take \( b = b_1 = b_2 \) and use the notations \( R(K) = \int K^2(u) du \) and \( U_2(K) = \int u^2 K(u) du \) both of which are finite values for our chosen kernel, we have the simplified formula

\[
\text{AMSE}(\hat{\mu}(x, y)) = \left[ U_2(K) \frac{b^2}{2} \left\{ m_{xx}(x, y) + m_{yy}(x, y) \right\} \right]^2 + \frac{1}{nb} \sigma^2(x, y) R(K)
\]

Remark: A formula for the local optimal bandwidth \( b_{opt} \) that minimizes the above expression for AMSE can easily be given by differentiating the above expression and equating to zero. Other approaches include finding global optimal bandwidths. We consider that combination of \( b_1 \) and \( b_2 \) for which \( \sum_{i=1}^{n} \text{AMSE}(\hat{\mu}(x_i, y_i))/n \) is minimized where \( (x_i, y_i) \) are the observed (rescaled) locations in the data set. For the Alaska diameter data set, this
value is $b_{opt} = 0.155$ which is taken to be the optimum bandwidth for either direction.

References

Ammann, B., Birks, H.J.B., Brooks, S.J., Eicher, U., von Grafenstaedt, U., Hofmann, W., Lemdahl, G., Schwander, J., Tobolík, K., Wick, L., 2000. Quantification of biotic responses to rapid climatic changes around the younger dryas—A synthesis. Palaeogeogr. Palaeoclimatol. Palaeoecol. 159, 313–347.

Beren, J., Feng, Y., Ghosh, S., Kulik, R., 2013. Long Memory Processes - Probabilistic Properties and Statistical Models. Springer Verlag, Heidelberg.

Bolin, D., Lindgren, F., 2017. Quantifying the uncertainty of contour maps. J. Comput. Graph. Stat. 26, 513–524.

Bradley, R.C., 1983. Asymptotic normality of some kernel-type estimators of probability density. Stat. Probab. Lett. 1, 295–300.

Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury, Pacific Grove.

Cressie, N., 2015. Statistics for Spatial Data, Revised Edition. Wiley, New York.

Dahlhaus, R., 1989. A flexible and fast method for automatic smoothing. J. Am. Stat. Assoc. 86, 643–652.

Dahlhaus, R., Müller, H.G., 1984. Estimating regression functions and their derivatives by the kernel method. Scand. J. Stat. 11, 171–185.

Ghosh, S., 2015. Surface estimation under local stationarity. J. Nonparam. Stat. 27, 229–240.

Ghosh, S., 2018. Kernel Smoothing: Principles, Methods and Applications. Wiley, Hoboken, New Jersey.

Ghosh, S., Draghiciusc, D., 2002. Predicting the distribution function for long-memory processes. Int. J. Forecast. 18, 283–290.

Ghosh, S., & Moser, G. (2019). On exceedances of a nonparametric spatial regression function under long-memory. Preprint.

Harris, N.L., Goldberg, E., Gabrie, C., Nordling, J., Minnenmeyer, S., Ansari, S., Lippmann, M., Bennett, L., Raad, M., Hansen, M., Potapov, P., 2017. Using spatial statistics to identify emerging hot spots of forest loss. Environ. Res. Lett. 12, 024012.

Herrmann, E., 2000. Variance estimation and bandwidth selection for kernel regression. In: Schimek, M.G. (Ed.), Smoothing and Regression: Approaches, Computation and Application. Wiley, New York.

Innes, J.L., Ghosh, S., Schwzyzer, A., 1996. A method for the identification of trees with unusually colored foliage. Can. J. Forest Res. 26, 1548–1555.

Kahle, D., Wickham, H., 2013. ggplo2: Spatial visualization with ggplot2. R. J. 5 (1), 144–161.http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Kawar, C., Ottenger, M., Wegmann, M., Guo, H., Wang, C., Zhang, J., Dech, S., Wikelski, M., 2014. Earth observation satellite sensors for biodiversity monitoring: potentials and bottlenecks. Int. J. Remote Sens. 35, 6599–6647.

Marchini, J., Pressan, A., 2004. Comparing methods of analyzing IMRI statistical parametric maps. Neuroimage 22.3, 1203–1213.

McNulty, S.G., Cohen, E.C., Moore Myers, J.A., 2013. Climate change impacts on forest soil critical acid loads and exceedances at a national scale. In: Potter, K.M., Conkling, B.L. (Eds.), Forest Health Monitoring: National Status, Trends, and Analysis 2010. Gen. Tech. Rep. SRS-GTR-176. Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station, pp. 95–108.

Menendez, P., Ghosh, S., Beran, J., 2010. On rapid change points under long memory. J. Stat. Plan. Inference 40, 3343–3354.

Ouimet, R., Duchesne, L., Houle, D., et al., 2001. Critical loads and exceedances of acid deposition and associated forest growth in the northern hardwood and boreal coniferous forests in Quebec, Canada. Water Air Soil Pollut. 1, 119–134.

Pardo, L.H., Coombs, J.A., Robin-Abbott, M.J., Pontius, J.H., D’Amato, A.W., 2019. Tree species at risk from nitrogen deposition in the northeastern united states: a geospatial analysis of effects of multiple stressors using exceedence of critical loads. Forest Ecol. Manag. 454, 117528.

Pommerehne, A., 2002. Approaches to quantifying forest structures. Forestry 75, 305–324.

Priestley, M.B., Chao, M.T., 1972. Non-parametric function fitting. J. R. Stat. Soc. Ser. B (Methodological) 34, 385–392.

Rudin, W., 1965. Principles of Mathematical Analysis. McGraw-Hill, New York.

R Development Core Team, 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Sarkan, D., 2008. Lattice: Multivariate Data Visualization with R. Springer, New York.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall.

U.S. Department of Agriculture, Forest Service (2015). Pacific Northwest Research Station, PNW-FIADB: Pacific Northwest Annual Forest Inventory Database (AK, CA, OR, WA). http://www.fs.fed.us/pnw/rma/fia-topics/inventory-data/index.php.

Wand, M.W., Jones, M.C., 1995. Kernel Smoothing. Chapman and Hall.

Zhang, J., Cressie, N., Craigmile, P.F., 2004. Loss function approaches to predict a spatial quantile and its exceedance region. Technometrics 50, 216–227.