Transactions in GIS

WILEY

Transactions in GIS

RESEARCH ARTICLE OPEN ACCESS

Exploration-Based Statistical Learning for Selecting Kernel Density Estimates of Spatial Point Patterns

Michael Govorov[1] | Giedrė Beconytė[2] | Gennady Gienko[3]

[1]Vancouver Island University, Nanaimo, British Columbia, Canada | [2]Vilnius University, Vilnius, Lithuania | [3]University of Alaska Anchorage, Anchorage, Alaska, USA

Correspondence: Michael Govorov (michael.govorov@viu.ca)

Received: 15 January 2025 | **Revised:** 7 April 2025 | **Accepted:** 9 April 2025

Keywords: bandwidth selectors | crime events | cross-validation | kernel density estimation | residuals | spatial point pattern events | validation measures

ABSTRACT

This paper addresses the use of nonparametric kernel density estimation (KDE) to estimate point-based data density in spatial modeling using Geographic Information Systems (GIS). The paper highlights challenges in selecting the appropriate settings for generating the best fitting KDE surfaces and validating their accuracy, as many GIS packages lack sufficient tools for this purpose. The paper focuses on providing guidelines for choosing the best bivariate KDE surface to approximate point patterns, using principles of machine learning for evaluation of the accuracy of KDE using internal and external metrics. Performance evaluation is based on the mass-preservation property of spatial point processes with the introduction of metrics such as residuals, cross-validation errors, and out-of-sample errors. These approaches are demonstrated on statistical data for violent crime in Lithuania but can be applied to other datasets with spatial point patterns.

1 | Introduction

Kernel density estimation (KDE) is a multipurpose nonparametric technique. It is used to estimate the probability density function (PDF) and probability mass function (PMF) of a random variable, intensity function of a point process, relative risk function, spatial regression function, and other quantitative measures. It can be used for exploratory and confirmatory analysis of spatial and temporal data, as well as cartographic visualization.

While the statistical distribution of a dataset can be assumed, in most cases, there is no parametric estimation of dataset parameters, so the KDE employs different nonparametric functions (estimators) to estimate the PDF/PMF directly from the source data. Kernel smoothing with continuous PDF is a well-known approach for estimation of surface density and intensity for spatially distributed point-like sampled datasets (Davies et al. 2018). The work of Davies et al. (2018) offers a comprehensive and practical guide to using kernel estimation techniques for analyzing spatial point patterns, with a case study in the field of epidemiology. In the context of KDE, density (the number of registered events per area) and intensity are closely related concepts and are discussed in detail in Section 3.2.2. Discrete kernel estimations of PMF have been far less investigated, especially for bivariate spatial surfaces (Kiessé 2017). Kiessé (2017) provides a thorough examination of the finite sample properties of nonparametric discrete asymmetric kernel estimators.

While there is no universal classification of KD estimators, they can be grouped by the type of the kernel function (Gaussian, Uniform, Epanechnikov, etc.), the function used to generate the multidimensional KDE (spherical multivariate kernel or product kernel), the type of a vector norm used in the spherical kernel, the size of the KDE bandwidth, which is its free tunable parameter that controls the degree to which density/intensity variations are smoothed out, bandwidth variability (fixed or adaptive/variable), the parameterization class of the multivariate bandwidth matrix (constrained (fixed or diagonal) or unconstrained (full)), and the edge correction factor.

Due to the large number of proposed KD estimators, choosing one KD estimator over another is far from a trivial task, and it is almost impossible to design a universal estimator that generates the best kernel density estimation (Gramacki 2017). There is no single best rule of thumb for choosing the optimal bandwidth, which is a primary driver in the accuracy of KDE. At first, this task may seem trivial, but it turns out to be extremely challenging. As illustrated below, it is not uncommon that different variants, extensions, and modifications of bandwidth selectors produce different sizes of bandwidth, and they can all be considered "optimal" according to the data-driven criteria defined for each selector.

The primary objective of this study is the testing of kernel estimates, including comparison, checking, and selection of the optimal estimator. The selection of the appropriate kernel density estimator is a critical issue in spatial analysis, as it directly affects the accuracy and effectiveness of spatial analyses in various real-world applications, including crime analysis, as demonstrated in this paper. The main aspects and innovations of the study are:

- A machine learning framework is applied to test and compare KD estimators for a set of points using two methodologies: the mass preservation property of spatial point processes governed by intensity, and a cross-validation (CV) approach, which is part of empirical risk minimization and model selection techniques within statistical learning theory.
- Spatial point process residuals are proposed as internal metrics for comparing bivariate KDE surfaces. These residuals have previously been used primarily to validate parametric models of point processes (Baddeley et al. 2016).
- Cross-validation errors, along with several modification techniques, are proposed as internal metrics for estimating and selecting bivariate KDE surfaces. Modifications include spatial cross-validation using space-filling curves and Delaunay triangulation, which group observation points into local neighborhoods forming contiguous blocks for cross-validation folds.
- A point-based deviance residual using cross-validation is proposed as an internal metric for testing and selecting KDE surfaces.
- Inverse lambda point-based residuals are proposed as external extra-sample errors for testing and selecting KDE surfaces.
- Several different KD estimators were tested on a large real-world dataset. The results show that the proposed non-model-based approaches, which use both internal and external error measures of a point process, provide a legitimate way to select the most accurate KDE surface.

The rest of the paper is organized as follows. Section 2 briefly discusses various KD estimators, including the categorization by kernel functions, parameterizations, and variability of bandwidths. The main purpose of the review is to show the diversity of KD estimators and the difficulties associated with the selection of the most suitable one for a particular application. Discussions on the edge correction techniques and joint bandwidth estimation for relative risk functions are limited. Section 3 elaborates on validation measures in the context of statistical learning, and approaches to using internal and external measures to evaluate KDE surfaces. In a case study in Section 4, the most common kernel density estimators and bandwidth selectors were tested on violent crime data for Lithuania. The training dataset, structured as spatial point data, is very large, which introduces computational challenges but also provides a better statistical representation of the studied phenomena. A final discussion and conclusions are presented in Section 5.

2 | Brief Review of Kernel Density Estimators

While the literature about KDE is extensive and a complete list of past and recent developments on the subject cannot be listed even roughly, below is an attempt to give a brief overview of the subject. The most significant foundational principles for bandwidth selection and various kernel functions are derived from Silverman (1986). Multivariate KDE and adaptive KDE techniques are covered in Scott (1992, 2015). Wand and Jones (1995) discuss various KDE methods, including boundary corrections and adaptive bandwidth selection. Chacón and Duong (2018) provide a detailed examination of multivariate and directional KDE techniques.

Univariate KDE, known as the Rosenblatt-Parzen window method, estimates the underlying PDF of a sample dataset with no assumptions on the underlying parametric distribution of the dataset (Silverman 1986; Wand and Jones 1995). KDE considers the contribution of each data point to the density function and can be applied to data drawn from a complex distribution. It has been demonstrated that univariate KDE works well for observation data with inhomogeneous dispersion and can be applied to spatial and spatiotemporal point pattern datasets with high heterogeneity and anisotropy (Davies et al. 2018). Univariate KDE has been extended to estimate multivariate densities based on the same principle: compute an average of densities centered at the events or grid points.

There are several characteristics, or features, of kernel density estimators. The *first feature* is what type of function is used as a kernel. A multivariate kernel can be obtained by two common techniques: by a derivation of univariate kernels or by using spherical or radially-symmetric kernels with l^2 Euclidean vector norm (length of the vector) (Wand and Jones 1995; Härdle and Müller 2000; Scott 2015). In the first case, kernels over multi-dimensional inputs can be constructed by multiplying or averaging different univariate kernels (Li and Racine 2007; Gramacki 2017). For bounded or partially bounded distributions without correlation between the components, a more suitable approach is to use the product of kernels (Kokonendji and Somé 2015). In the second case, radially symmetric kernels are constructed from data within a sphere around an event or a grid point (Silverman 1986). In the general case, a multivariate kernel of the second type can be constructed with other types of vector norms ($l^1, l^2, l^3, \ldots, l^\infty$), and not only with the sphere Euclidean l^2 norm.

The source data used to construct KDE can be partially bounded (e.g., all data are positive), completely bounded (e.g., data in the

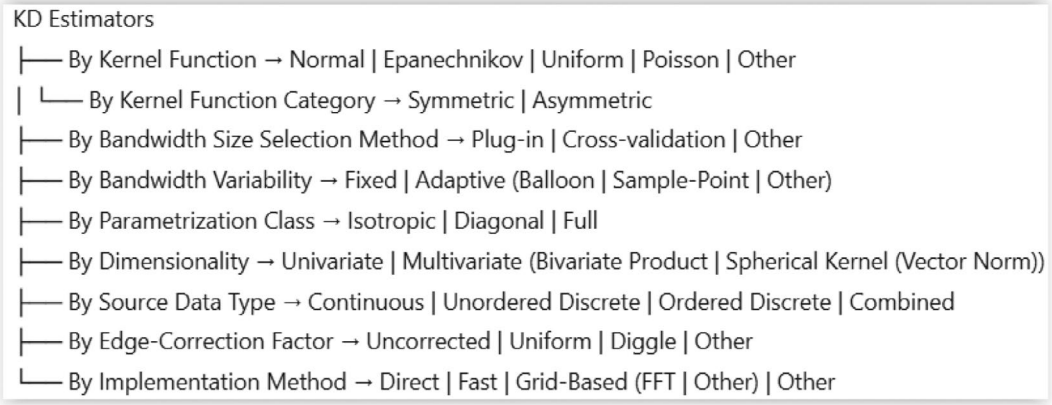

FIGURE 1 | Classification of KD estimators.

unit interval), or discrete (e.g., counts). Count data can be ordered or unordered. The classical symmetric PDF estimators (such as Epanechnikov or Gaussian kernel functions) presume that the underlying data is naturally continuous, which is often not the case. Thus, symmetric kernels may not be suitable for discrete bounded datasets; instead, other types of kernel functions should be used (Li and Racine 2007; Shimazaki and Shinomoto 2010; Kiessé 2017).

The Aitchison and Aitken's (1976) kernel can be used for unordered discrete or categorical variables, while the kernel proposed by Wang and van Ryzin (1981) can be used for ordered discrete variables (Li and Racine 2007). Several studies (Kokonendji and Kiessé 2011; Kiessé 2017) explore the use of asymmetric kernel functions where discrete kernels have been constructed from known discrete PMFs such as Poisson, binomial, and negative binomial.

In a multivariate setting, the joint density function can be defined as a combination of discrete (unordered and ordered) and continuous variables for both quantitative and qualitative data. The joint PDF/PMF estimation method has been extended using generalized product kernels (Li and Racine 2007), assuming no correlation in its multivariate components. Spherical joint kernel estimators are proposed by Kokonendji and Somé (2015) to estimate PDF/PMF on partially or fully bounded data with a correlation structure.

The *second feature* of KD estimators is the bandwidth. Most researchers agree that the most important component of kernel density estimation is the size of bandwidth h, and not the type of the kernel function itself (Silverman 1986; Wand and Jones 1995). There is a considerable amount of literature on selecting the optimal bandwidth, with many proposed selection rules; however, no single rule consistently outperforms the others. The choice of the optimal bandwidth largely depends on the true shape of the density being estimated and the criteria used to evaluate estimation quality.

In the case of spatial kernel, the *third feature* of KD estimators is the parametrization class of bivariate bandwidth matrix H, which controls the extent, shape, and orientation of smoothing. A constrained fixed matrix H is a diagonal or identity matrix scaled by a fixed scalar h, resulting in circular, isotropic kernel shapes. A constrained diagonal matrix allows for arbitrary ellipsoidal shapes but without rotation. An unconstrained full matrix H permits kernel functions with arbitrary orientation and ellipsoidal shapes, providing greater flexibility in smoothing anisotropic patterns.

The *fourth feature* of KD estimators is the method of edge correction, which minimizes the boundary bias due to the asymmetry of the weights. Compared to univariate KDE, the boundary problem in multivariate KDE can be much more problematic because the dimensionality increases the boundary region (Bouezmarni and Rombouts 2010). For more information on this aspect of KDE, refer to Section 4 of this paper (and also see Jones 1993; Diggle 1985; Davies et al. 2018).

Another useful approach to designing KD estimators is the use of an *adaptive* or *variable* bandwidth determined by the local density. There are two categories of adaptive KD estimators: balloon estimators (Breiman et al. 1977; Scott 2015), where bandwidths are determined at each evaluation location, and sample-point estimators (Abramson 1982; Davies and Baddeley 2018), where bandwidths are determined at each observation. Sample-point adaptive estimators result in densities that conserve mass and integrate to 1 over the study domain.

Figure 1 illustrates the variety of KD estimator forms, emphasizing the challenges in selecting the most suitable estimator for a specific application.

2.1 | Choosing the Kernel Function

For a two-dimensional homogeneous spatial point process, bivariate kernels are defined as the product of two univariate kernels. It is assumed here that the observed event points in the x and y coordinates do not exhibit autocorrelation. In the case where a radially symmetric kernel with the same bandwidth in the x and y directions is used, the amount of smoothing is the same in each coordinate direction. The underlying bivariate PDF is then estimated using the most

common kernel estimator $\hat{f}_h(s)$ without edge correction at a location $s(x, y)$ (Silverman 1986):

$$\hat{f}_h(s) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{s - S_i}{h}\right) \tag{1}$$

where s is the target (or test point) with 2-dimensional (2D) spatial coordinates within a bounded region; S_i are bivariate random independent identically distributed (i.i.d.) point samples taken from a common and usually unknown kernel function $\hat{f}$; $K()$ is a 2-dimensional, second-order, zero-centered, radially symmetric continuous bimodal fixed kernel function; n is the number of point samples; and h is the kernel bandwidth, scaled equally in all directions (or radius of the circle for a circular kernel). Outside the bounded region, the estimated probability density is zero.

Several studies examine kernel functions for bivariate *discrete* PMF kernels (Li and Racine 2007; Chu et al. 2017; Kiessé 2017; Belaid et al. 2018). The first class of two-dimensional discrete PMF kernels includes Dirac-type symmetric kernels such as discrete triangular (Belaid et al. 2018; Aitchison and Aitken 1976; Wang and van Ryzin 1981) and discrete Epanechnikov kernel functions (Chu et al. 2017). Another class of kernels built from Poisson, binomial, and negative binomial PMFs is based on non-Dirac-type discrete asymmetric kernels (Kiessé 2017).

In the general form, the kernel estimator for *count* data is expressed as.

$$\hat{f}_h(s) = \frac{1}{n} \sum_{i=1}^{n} L(S_i, s, h) \tag{2}$$

where $L(\cdot)$ is a discrete symmetric or asymmetric Dirac or non-Dirac-type kernel function suitable for smoothing discrete data; S_i is the location of the univariate event, s is the location of the estimate, and h is the kernel bandwidth. For example, a univariate discrete kernel function of ordered variable s (Wang and van Ryzin 1981) can be defined as.

$$L_h(S_i, s) = \begin{cases} 1-h, & \text{if } S_i = s \\ \frac{(1-h)}{2} h^{|S_i - s|}, & \text{if } S_i \neq s \end{cases} \tag{3}$$

For discrete ordered data, the Mean Integrated Squared Error (MISE) optimization method of univariate bandwidth selection was proposed by Shimazaki and Shinomoto (2010). It is assumed that the pattern of events is described by an inhomogeneous Poisson point process. This optimization technique can be applied to any Dirac-type kernel function.

The choice of kernel function can affect the quality of the KDE estimate (Kiessé 2017). Appropriate PMF kernels such as negative binomial kernels can be used for discrete data. For example, in the case of criminal events, it is important to consider that the data consist of discrete counts that are ordered in the temporal dimension but unordered in the spatial dimension.

2.2 | Choosing Kernel Density Estimator Based on Bandwidth Variability and Parameterization

Adaptive or variable KD estimator in the balloon category (Scott 2015; Abramson 1982) at a test location s is expressed as

$$\hat{f}_h(s) = \frac{1}{n} \sum_{i=1}^{n} \frac{K\left(\frac{(s - S_i)}{h_s}\right)}{h_s^2} \tag{4}$$

where h_s is the kernel bandwidth at the test point. In adaptive KD balloon estimators, each test point has its own bandwidth. Adaptive KD estimators enforce smoothing in areas where events are relatively sparse and reduce smoothing in areas with high event density. This method is expected to reduce bias, especially in the asymptotic context (van Lieshout 2022).

Choosing bivariate bandwidths for a spatially inhomogeneous and anisotropic point process is not a trivial task, especially when deciding on the appropriate amount of smoothing. There are several classes of *parameterization* of the bivariate bandwidth matrix (Wand and Jones 1995; Kokonendji and Somé 2015) that can be considered if the spatial point process is inhomogeneous and anisotropic. Although the use of diagonal bandwidth matrices (independent bandwidths in x and y directions) may be appropriate for heterogeneous processes, a full or unconstrained bandwidth matrix for smoothing in directions other than the directions of the coordinate axes (Duong and Hazelton 2003) may perform better in a particular anisotropic process. The spatial bivariate kernel can be used to form a multivariate KD estimator as follows:

$$\hat{f}_H(s) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|H|^2} K\left(H^{-1/2}\left(s - S_i\right)\right) \tag{5}$$

where $|H|$ is the determinant of the symmetric positive-definite 2×2 bandwidth matrix H.

For count data, the spatial kernel can be defined as a radially symmetric PDF kernel or obtained as a product of two unordered discrete univariate PMF kernels. In the first case, a radially symmetric kernel with a 2×2 full bandwidth matrix can be used to estimate the density surface. Such matrices include both kernel orientation and axis-specific bandwidths (Silverman 1986; Wand and Jones 1995), and assume autocorrelation structures in the data (Kokonendji and Somé 2015). In the second case, the product kernel assumes uncorrelated x and y coordinates, and the product kernel is defined by two bandwidths in x and y directions. However, to account for anisotropy in diagonal directions, the data can be pre-scaled or whitened, and then a diagonal bandwidth matrix is used for the product kernel, and the result is transformed back at the last step (Silverman 1986; Wand and Jones 1995). The bivariate product of two univariate kernel estimators is implemented as

$$\hat{f}_h(s) = \frac{1}{n} \sum_{i=1}^{n} \prod_{p=1}^{2} \frac{1}{h_p} K\left(\frac{(s_p - S_{ip})}{h_p}\right) \tag{6}$$

where h_p is the bandwidth in dimension p; s_p is a target univariate point in dimension p; S_{ip} are univariate point samples in dimension p.

The implementation of diagonal and unconstrained bandwidth matrices in adaptive settings presents some challenges (Davies et al. 2018). In practical applications, choosing the optimal kernel estimator with the optimal kernel function, parameterization, edge correction method, and bandwidth size (discussed below) is not a trivial task.

3 | KDE Validation

3.1 | Statistical Learning Methods for Model Assessment, Selection, and Validation

In machine learning, any quantitative confirmatory analysis has two important components: *assessment* of performance and *selection* of an acceptable method. *Assessment* of performance is an estimate of the method's prediction error for a data set in absolute terms. *Selection* of a method is the identification of the best method based on an assessment of its performance (prediction error) in relative terms (compared to other methods).

To solve the above two tasks, validation of a particular model is necessary. One of the common and convenient measures of assessment and selection of a particular model is the *test error* $\mathrm{Err}_{\mathrm{test}}$ or the *expected test error*. The test error is also known as the Mean Squared Error (MSE), extra-sample, or generalization error at the new independent test points $s\left(x_0, y_0\right)$. The *expected test error* between the new test point and the point fitted to the training sample points $S\left(x_i, y_i\right), i = 1; \ldots, n$ is defined as.

$$E\left(\mathrm{Err}_{\mathrm{test}}\right) = \mathrm{MSE}\left(\widehat{f}\,(s)\right) = E\left(f(s) - \widehat{f}\,(s)\right)^2 \tag{7}$$

where the expectation E is estimated over all random training set S and test point s; $f(s)$ is the true value at the new test point; $\widehat{f}\,(s)$ is the estimated value at the new test point. In this case, the error loss function used in (7) is the squared error l^2, but other performance metrics can be used. Training and testing points are assumed from the same independent identically distributed (i.i.d.) point samples.

Validation of method results can be "formal" or "informal" (Baddeley et al. 2016). *Formal* validation techniques are based on classical statistical inference with probabilistic assumptions about the data set and allow probabilistic statements to be made about the results. Such formal techniques include hypothesis testing, confidence intervals, and Bayesian model selection. *Informal* techniques are based on the philosophy of statistical learning. The outcomes of particular machine learning techniques with flexible modeling strategies must be validated by the ability of the generated model to predict new/hold-out datasets (Vapnik 2013; Hastie et al. 2017).

The validation includes informal diagnostics and model-specific validation procedures, such as residual analysis and the estimation of validation and test errors, among others. In this context, Vapnik (2013) emphasizes the importance of model validation through techniques like the structural risk minimization principle, which balances model complexity and training error to minimize generalization error and improve robustness. Meanwhile, Hastie et al. (2017) offer a comprehensive discussion on model validation, focusing on methods such as k-fold cross-validation and the use of metrics like mean squared error (MSE) to evaluate model performance and prevent overfitting.

Replication-based techniques, such as cross-validation (CV) and bootstrap, provide a stochastic estimate of model performance, blurring the line between "formal" and "informal" techniques—minimizing the AIC is similar to minimizing leave-one-out CV (Stone 1977), and minimizing the BIC corresponds to performing leave-K-out cross-validation (Shao 1997).

In the field of machine learning, three distinct methodologies are used to evaluate experimental validity for model assessment and selection. These methodologies aim to determine the degree to which measurements align with their intended representations. These methodologies include:

1. *Internal* numerical metrics (measures, indexes, scores, or criteria) are employed to assess the effectiveness of a model's structure, denoted as $f(\bullet)$, relying on inherent dataset features and quantities. There are various approaches for estimating the expected test error solely using the training dataset when the value at the test point $\left(x_0, y_0\right)$ remains unknown.
   a. Assessment and selection of a model can involve employing *analytical* techniques that rely on statistical assumptions regarding the dataset. These assumptions enable the generation of probabilistic optimizes about the results. The assessment of model results can encompass the analytical estimation of variance and its corresponding standard error relative to expected values, along with the creation of confidence pointwise intervals for $f(S)$. For model selection, hypothesis testing can be used, along with information criteria such as AIC and BIC, which account for both model complexity and performance. These criteria help identify the model that achieves the optimal balance between fit and complexity.
   b. A readily available estimate of the test error is the *training error* $\mathrm{Err}_{\mathrm{training}}$, often referred to as the *residual*. These residuals quantify the differences between predicted values and the actual values at each data sample point, denoted as $S\left(x_i, y_i\right)$. The training error represents the average loss across all training data points. The expected squared error within the set of sample points is defined as:

$$E\left(\mathrm{Err}_{\mathrm{training}}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}\left(f(S) - \widehat{f}(S)\right)^2\right) \tag{8}$$

where the expectation E is taken over all random training sets, $f(S)$ represents the observed value at a training point $S\left(x_i, y_i\right)$, $\widehat{f}(S)$ denotes the estimated value at a training point, and n represents the size of the training set.

In many cases, the training error may not serve as an adequate estimate of test error for two related reasons. The first reason is that the training error $\mathrm{Err}_{\mathrm{training}}$ will be consistently lower than the actual test error $\mathrm{Err}_{\mathrm{test}}$ because the same dataset S is used both for fitting a model and estimation of error. Consequently, $\mathrm{Err}_{\mathrm{training}}$ cannot be employed for the task of performance assessment.

The second reason is overfitting. In many cases, the training error $\mathrm{Err}_{\mathrm{training}}$ tends to decrease as the training dataset is fitted more rigorously. For some methods, $\mathrm{Err}_{\mathrm{training}}$ can even drop to zero if the method's complexity is increased significantly, but this often results in poor generalization. At the same time, even if the expected training error $E\left(\mathrm{Err}_{\mathrm{training}}\right)$ consistently deviates from the expected test error $E\left(\mathrm{Err}_{\mathrm{test}}\right)$, it can still be useful for model selection (Hastie et al. 2017).

c. In practical applications, the most widely employed data-driven techniques for directly estimating the expected prediction test error $E\left(\mathrm{Err}_{\mathrm{test}}\right)$ include *cross-validation* (CV), bootstrap, and other replication-based techniques. These techniques are also effective for comparing the results of various fitting methods and selecting the best one based on the minimum cross-validation error $\mathrm{Err}_{\mathrm{CV}}$. Notably, these techniques provide an expectation $E\left(\mathrm{Err}_{\mathrm{CV}}\right)$ that closely approximates $E\left(\mathrm{Err}_{\mathrm{test}}\right)$, and control overfitting. The K-fold CV test error estimate $\mathrm{Err}_{\mathrm{CV}}$ is defined as

$$\mathrm{Err}_{\mathrm{CV}} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{CV}_{\hat{f}^{-(k)}} \tag{9}$$

$$CV_{\hat{f}^{-(k)}} = \frac{1}{n_k}\sum_{i\in F_k}\left(f(S)-\hat{f}^{-(k)}(S)\right)^2$$

where K represents the number of randomly split folds F_i (so $F_1 \cup \ldots \cup F_k = \{1, \ldots n\}$)}), each containing approximately equal number of training points; $CV_{\hat{f}^{-(k)}}$ signifies the average validation error computed on the data points within the kth fold; the function $\hat{f}^{-(k)}(S)$ represents a predication function fitted on all training points except those in the kth fold; n_k represents the number of data points in the kth fold. In cross-validation (CV), the training involves $S\left(x_i, y_i\right)$ for $i \notin F_k$, while validation is performed on $S\left(x_i, y_i\right)$ for $i \in F_k$. Common choices for the value of K include 5, 10, and n, which corresponds to leave-one-out cross-validation (LOOCV). It is important to note that cross-validation effectively estimates only the average $\mathrm{Err}_{\mathrm{test}}$ (Hastie et al. 2017).

The cross-validation techniques are applicable to any loss function, such as the one used for the training error $\mathrm{Err}_{\mathrm{training}}$. These techniques are also used for adaptive fitting methods such as nonparametric estimators. When the method includes a hyperparameter, denoted as h, $\mathrm{Err}_{\mathrm{CV}}(h)$ serves as a tool for estimating the error curve. It helps to determine the optimal tuning parameter by minimizing $\mathrm{Err}_{\mathrm{CV}}(h)$, thus facilitating the selection of the best-fit method.

The standard error of cross-validation $\mathrm{SE}_{\mathrm{CV}}$ is estimated at training sample points and serves as a valuable complement to $\mathrm{Err}_{\mathrm{CV}}$, offering an additional quantitative measure for assessing the variability of $\mathrm{Err}_{\mathrm{CV}}$. This concept of uncertainty intervals aids in the selection of the best-fit method. The standard error $\mathrm{SE}_{\mathrm{CV}}$ for the mean of $\mathrm{CV}_{\hat{f}^{-(k)}}$ can be estimated within the framework of the central limit theorem as.

$$\mathrm{SE}_{\mathrm{CV}} = \frac{sd\left\{\mathrm{CV}_{\hat{f}^{-(i)}}\right\}}{\sqrt{K}} \tag{10}$$

where $sd\left\{CV_{\hat{f}^{-(i)}}\right\}$ is the sample standard deviation of $\mathrm{CV}_{\hat{f}^{-(1)}}, \ldots, \mathrm{CV}_{\hat{f}^{-(k)}}$. Equation (10) is valid for small $K \ll n$ (e.g., $K = 5$ or 10) as folds' samples can be treated as originating from a single independent identical distribution (i.i.d.). According to the one standard error rule, the best model selection involves choosing the most regularized model with $\mathrm{Err}_{\mathrm{CV}}$ within one standard error $\mathrm{SE}_{\mathrm{CV}}$ of the minimal $\mathrm{argmin}\left(\mathrm{Err}_{\mathrm{CV}}(h)\right)$. And $\mathrm{Err}_{\mathrm{CV}} = \mathrm{argmin}\left(\mathrm{Err}_{\mathrm{CV}}(h)\right) \mp \mathrm{SE}_{\mathrm{CV}^{(h)}}$ (Hastie et al. 2017).

2. *External* numerical metrics are used to measure the degree to which the resulting modeling structure matches either the external test dataset or a predefined structure imposed on the dataset.

In a data-rich situation, a predictive strategy involves the random splitting of the original dataset into two distinct parts: a training set and a test (hold-out) set. The test set is strictly reserved for error estimation and is not utilized in the modelling process. This technique offers an impartial evaluation of the modelling error, and when the test sample size is substantial, it enhances precision. Utilizing extra-sample test errors allows for confident selection of the best method from a range of options, while ensuring that the chosen method does not exhibit overfitting. The extra-sample test error estimate is defined as follows:

$$E\left(\mathrm{Err}_{\mathrm{test}}\right) = E\left(\left(f(s)-\hat{f}\,(s)\right)^2\right) \tag{11}$$

where $f(s)$ represents the value at an independent test point, and $\hat{f}\,(s)$ signifies the estimated value at the same independent test point.

As a standard practice, a single random split, referred to as a hold-out, is employed to thin the original dataset into test and training sets. This is done to preserve the similarity in distributions between the training and test sets while ensuring the independence of the experiment results conducted on these sets and assessing how well the model generalizes to the hold-out dataset.

3. *Relative* numerical metrics are used to compare two different modeling structures that can be created using different algorithms or the same algorithm with different parameters.

When comparing KDE surfaces, special cases of f-divergence (such as KL-divergence, Hellinger distance, and total variation distance) can be used to measure the similarity of distributions, as shown in Moon and Hero (2014). Another set of relative or goodness-of-fit measures suitable for nonparametric testing using Integrated Squared Error (ISE) has been proposed in Li and Racine (2007), Martinez-Camblor and de Una-Alvarez (2009), and Chacón and Duong (2018). These goodness-of-fit metrics are valuable for comparing multiple estimated densities $\hat{f}(\bullet)$, against each other or against a known KDE function.

However, it is important to note that these measures are not applicable to evaluate how well KDE function $\hat{f}(\bullet)$ describes the (unknown) true density function

$f(\bullet)$. Typically, the density function $f(\bullet)$ is simulated from known distributions and used to test the fit of the method. The residual field can be used to evaluate the similarity of KDE surfaces.

3.2 | Assessment and Selection of the Kernel Density Estimators

3.2.1 | Internal Analytical Metrics of Kernel Density Estimation: ISE, MISE, and AMISE

The two most common error metrics for kernel density estimation are Integrated Squared Error (ISE) and Mean Integrated Squared Error (MISE). ISE is a stochastic variable that summarizes the performance of kernel density estimator $\widehat{f}_h(s)$ as a function of observed data. MISE, which is the expected value of ISE and a deterministic function of the bandwidth parameter h, is estimated given the distribution of the unknown density function $f(s)$ (Heidenreich et al. 2013; Gramacki 2017). ISE and MISE measure the overall estimation error instead of giving excessive importance to a small part of the support. These two different *global* measures lead to different specifications regarding the optimal bandwidth. ISE and MISE measure the proximity of $\widehat{f}_h(s)$ to the target density $f(s)$ and are defined as

$$\mathrm{ISE}\left(\widehat{f}_H(\cdot)\right) = \int_{\mathbb{R}^{2+}} \left(f(s) - \widehat{f}_H(s)\right)^2 du \tag{12}$$

$$\begin{aligned}\mathrm{MISE}\left(\widehat{f}_H(\cdot)\right) &= \int_{\mathbb{R}^{2+}} E\left(f(s) - \widehat{f}_H(s)\right)^2 ds = \int_{\mathbb{R}^{2+}} \mathrm{MSE}\left(\widehat{f}_H(s)\right) ds \\ &= \int_{\mathbb{R}^{2+}} \mathrm{Variance}\left(\widehat{f}_H(s)\right) ds + \int_{\mathbb{R}^{2+}} \mathrm{Bias}^2\left(\widehat{f}_H(s)\right) ds\end{aligned} \tag{13}$$

where s represents two vectors x_{0i} and y_{0i} of the point coordinates in 2D space, and $\widehat{f}_H(\bullet)$ indicates the global error which is the distance measure between $\widehat{f}_h(s)$ and $f(s)$ over the entire region instead of MSE at a specific point $s(x_0, y_0)$. MISE can be thought of as the mean of the global ISE measure relative to the sample density. MISE is a measure of estimation risk associated with the Mean Square Error $\mathrm{Err}_{\mathrm{test}}$. As a nonparametric technique, MISE incorporates the concept of "optimal" balancing between Bias and Variance, controlling under- and over-smoothing complexity of the density estimation. This approach aims to minimize an overall fitting measure such as $\mathrm{MSE}\left(\widehat{f}_H(s)\right)$ or $E\left(\mathrm{Err}_{\mathrm{test}}\right)$.

Nevertheless, using ISE and MISE directly for validation is impractical because they have no closed-form expression due to the unknown $f(s)$, except when $f(s)$ follows a normal mixture density and K is the normal kernel (Wand and Jones 1995). An alternative is to look for an approximation of MISE known as asymptotic MISE (AMISE). However, it is worth noting that AMISE also depends on the second derivative of the unknown density $f(s)$. Closed-form expressions for AMISE are only available when working with normal mixture density and can be calculated exactly. In scenarios involving other $f(s)$, commonly employed kernels would allow for obtaining only approximate AMISE values.

There are different AMISE implementations for different KD estimators, which makes AMISE a method-specific measure not directly suitable for comparing different KD estimators. The primary objectives of MISE/AMISE and ISE are:

1. Most modern methods for automatic bandwidth h selection are based on optimizing MISE/AMISE, ISE, or a combination of both. They aim to find $h_{\mathrm{AMISE}} = \mathrm{argmin}_{H\in\mathcal{F}}\ \mathrm{AMISE}\left(\widehat{f}_H(\bullet)\right)$ or $H_{\mathrm{LSCV}} = \mathrm{argmin}_{H\in\mathcal{F}}\ \mathrm{LSCV}(H)$, where $\mathcal{F}$ represents the space of all symmetric, positive definite $d \times d$ matrices; $\mathrm{LSCV}(H)$ denotes the least squares cross-validation objective function. These optimization criteria have led to the development of multiple bandwidth selectors to estimate the unknown $f(s)$ in Equations (12) and (13).
2. ISE is commonly employed for evaluating the performance of KDE techniques when applied to test datasets simulated from known density functions. In such scenarios, ISE is computed by numerically integrating the differences between the estimated density $\widehat{f}_H(s)$ and the true target density $f(s)$. Additionally, various performance metrics derived from ISE can also be used (Heidenreich et al. 2013).

KDE confidence intervals can be calculated analytically by using the KDE pointwise error, defined as the difference between $\widehat{f}_H(s)$ and $f(s)$. However, such confidence intervals remain impractical in real-world applications due to the unknown nature of $f(s)$. A straightforward approach is to substitute $f(s)$ with its estimate $\widehat{f}_H(s)$ in the asymptotic variance (Chen 2017). A more robust and alternative technique for estimating the asymptotic variance and creating the KDE confidence intervals involves employing the bootstrap. Several issues, such as bias under-coverage, and strategies to address these problems, are discussed in Chen (2017).

In numerous real-world applications, density functions tend to be complex and often remain unknown. Consequently, since it is not possible to compute ISE and MISE directly, assessing the accuracy of kernel density estimations directly through these metrics is not feasible.

ISE and MISE, which are error metrics for kernel density estimations, are typically employed as data-driven criteria for choosing the optimal bandwidths. Nevertheless, when considering a point pattern as an instance of a point process, the measure of error for estimating the KDE function can alternatively be based on the concept of the mass conservation property (Loader 1999; Baddeley et al. 2016; Cronie and van Lieshout 2018).

In the case of an inhomogeneous Poisson point process, the mass-preservation property of KDE is formally defined as $n(A \cap W) = \int_W \widehat{\lambda}(s)\, ds$, with the inclusion of a boundary correction to estimate the point pattern intensity $\widehat{\lambda}(s)$. Here, A represents a set of points $S(x_i, y_i)$ in the two-dimensional space $\mathbb{R}^{2+}$ of the observed point pattern; the term $n(A \cap W)$ denotes the count of points from the set $S(x_i, y_i)$ within the region W, while $\widehat{\lambda}(s)$ denotes the intensity of the fitted point process

estimated at any spatial location $s\left(x_0, y_0\right)$ (Cronie and van Lieshout 2018).

3.2.1.1 | Bandwidth Selection. The choice of bandwidth is essential when estimating kernel density, whether in univariate or multivariate scenarios. While there may be strong contextual justifications to select a specific bandwidth size h, in most applications, determining this value proves challenging and often unfeasible, especially in the context of bivariate KDE. The key questions that arise are:

- What is the optimal *size* of the spatial bandwidth?
- Should the bandwidth remain *fixed*, or would a *variable* bandwidth be more appropriate? If variable bandwidth, should it be adjustable and adaptive?
- Which class of *parametrization* matrix H should be employed for bivariate bandwidth?

Numerous fixed and variable bandwidth selection techniques are available. These techniques, often referred to as data-driven bandwidth selectors, aim to minimize various errors such as Mean Squared Error (MSE), Integrated Squared Error (ISE), and Mean Integrated Squared Error (MISE) (Wand and Jones 1995; Heidenreich et al. 2013; Gramacki 2017). When considering ISE and MISE distance measures, data-driven or fully automatic bandwidth selectors can be divided into two categories:

1. *Plug-in* selectors typically determine the optimal bandwidth h to minimize MISE. The plug-in selectors are based on the AMISE asymptotic equation, which provides an approximate MISE estimation for large samples. The AMISE equation includes only one unknown Ψ_4 quantity, which is estimated using various methods and assumptions (Gramacki 2017). Plug-in selectors employ internal data-derived measures to optimize the value of h.
2. Selectors based on *cross-validation* (CV) and bootstrapping typically aim to minimize ISE. A classic example is the least-squares (or unbiased) cross-validation (LSCV) selector. In cross-validation, a subset of the data is employed to evaluate another subset, effectively minimizing the ISE. The cross-validation technique often employs the classical leave-one-out approach when estimating $\widehat{f}_h(s)$ (Scott 2015; Davies and Lawson 2019).

A wide range of cross-validation and plug-in selectors, including rule-of-thumb techniques and their hybrids, employ various methods to estimate the unknown density function $f(s)$ (Silverman 1986; Wand and Jones 1995; Illian et al. 2008; Scott 2015; Davies et al. 2018). There are methods for estimating optimal bandwidths for multidimensional kernel functions based on Bayesian approaches, boot-strapping, extrapolation methods, the theory of spatial point processes (Berman and Diggle 1989; Loader 1999; Baddeley et al. 2016; Cronie and van Lieshout 2018), mixing bandwidth selectors, and neural networks (Heidenreich et al. 2013; Fernando and Hazelton 2014; Davies et al. 2018). These methods can be assigned into one of the cross-validation and plug-in categories.

3.2.1.2 | Issues Related to Selecting an Appropriate Bandwidth Based on ISE, MISE, and AMISE. When working with real-world datasets, it is not uncommon to observe that different selectors yield substantially different "optimal bandwidths". Typically, plug-in bandwidth selectors tend to over-smooth finite sample datasets, while significant sample variation can create challenges for nearly all cross-validation bandwidth selectors, often resulting in overfitting.

Certain challenges arise because ISE and MISE-based selectors rely on assumptions about the underlying data distribution, yet $f(s)$, in most cases, remains unknown and can be particularly complex, especially when dealing with spatial bivariate data. These selectors often use the normal or other known base density as a reference distribution function to estimate the unknown $f(s)$.

For example, rule-of-thumb selectors often use the normal distribution as a reference to replace the unknown density function $f(s)$. When dealing with a distribution that substantially deviates from the bivariate normal distribution (such as Poisson distributions), particularly in datasets containing outliers, the results can be highly inaccurate.

Plug-in based selectors require an additional pilot bandwidth parameter g_4 to estimate the unknown variable conventionally referred to as Ψ_4 in the AMISE formula (Wand and Jones 1995, Sect. 3.5; Gramacki 2017, 66). The estimation of the unknown Ψ_4 involves a multi-stage adjustment process, usually consisting of two stages. It is assumed that in the final stage, $\Psi_r\left(g_r\right)$ is computed using the normal scale formula, with r representing the derivative order and an even number. A notable challenge is the lack of analytical methods to determine the optimal number of stages (Gramacki 2017, 66).

In practice, cross-validation criteria can be derived if the integral is replaced by summation where the kernel function is convoluted with itself, which is appropriate for normal kernels (Gramacki 2017), and that is why most cross-validation implementations use the normal kernels. A well-known weakness of cross-validation selectors is that the objective function can have more than one local minimum. Another issue is that classical least-squares cross-validation selectors are unstable on large datasets and typically give substantially dissimilar outputs for different datasets having the same distribution (Gramacki 2017). Cross-validation does not work well on discrete data, and this is problematic as the real data is nearly always finite and discretized.

3.2.2 | Internal Analytical Error Metrics: Standard Error of the Spatial Point Pattern Intensity Function

Events that are recorded at specific locations, denoted as x and y, constitute a spatial point pattern. This pattern can be conceptualized as the result of a spatial point process, a framework used to understand the underlying generation mechanism of these points. A point process represents a random mechanism that can be mathematically formulated in various ways (Baddeley et al. 2016). For instance, one approach is to utilize

an inhomogeneous Poisson point process to model the counts of disjoint 2D intervals which are considered stochastically independent. The primary assumption in this context is that the individual points are statistically independent of each other.

Numerous research studies have confirmed the effectiveness of employing the inhomogeneous Poisson point process to represent various real-world, spatially variable independent random events. Examples of such events include crime events, traffic accidents, and noninfectious diseases, among others (Baddeley et al. 2016).

The measure of the first moment of an inhomogeneous Poisson spatial point process is its estimated intensity function $\lambda(s)$, representing the average number of points per unit area at point s. The entire Poisson point process can be fully specified by its intensity function, $\lambda(s)$. This intensity function can be integrated up to the expected number of points that fall within the region of interest W.

The point process intensity function $\lambda(s)$ can be estimated nonparametrically using KDE. Estimating the intensity of a spatial pattern is similar to a bivariate estimate of the probability density. The intensity function is proportional to its probability density function as $f(s) = \lambda(s) / \int_W \lambda(s)ds$ (Baddeley et al. 2016). Here, $\int_W \lambda(s)ds$ generates the number n of independent identically distributed points for the Poisson process with the rate parameter $\lambda(s)$.

Thus, the KDE probability density surface can be transformed into an intensity surface by multiplying the probability density at each cell by the total number of incidents, expressed as $\lambda(s) = nf(s)$, and vice versa. Multiplying $f(s)$ by n scales the normalized density estimate to reflect the expected number of occurrences per unit area. The KDE probability density function $f(s)$ is normalized such that its total integral over the entire space equals 1.

Specific validation metrics can be used to analytically evaluate the performance of KDE when fitting a spatial point pattern. One such metric is the standard error (SE), which estimates the standard deviation of the error term. If the intensity function $\lambda(s)$ of the Poisson point process is estimated using an isotropic Gaussian kernel, then the standard error $SE_{\lambda(s,h)}$ of the estimate $\lambda(s)$ is (Baddeley et al. 2016):

$$SE_{\lambda(s,h)} = \frac{1}{h^2}\sqrt{\frac{1}{2\pi}\sum_{i=1}^{n} K\left(\frac{(s-S_i)}{h/\sqrt{2}}\right)} \tag{14}$$

The value of $SE_{\lambda(s,h)}$ is obtained from the estimate of the variance of $\lambda(s)$ at a target new point. Although the standard error of the intensity estimate is a measure of accuracy, it should be noted that this estimate is based on the assumptions of a particular point process.

3.2.3 | Internal Validation Metrics: Spatial Point Process Residuals From Training Samples

A Poisson point process, which describes the occurrence of random events, can be formally defined using a Poisson distribution. This allows for estimating the probability of a certain number of events occurring within a specified spatial region or time interval. In the context of validation of fitted Poisson regression models, various types of Poisson residuals are used for diagnostic purposes. These residuals include both raw residuals and adjusted/normalized residuals such as Pearson, Standardized, Studentized, Deviance, and Anscombe residuals (Cameron and Trivedi 2013; Hilbe 2014).

The basic or raw residual, denoted as r_i, is the difference between the observed response z_i and the expected response $E(z_i)$, which is similar to the use of residuals in classical linear regression models: $r_i = z_i - E(z_i)$. In the case of parametric Poisson maximum likelihood regression models, one common residual diagnostic involves comparing the fitted PMF with observed frequencies. The fitted frequency distribution is computed as the average over observations of the predicted probabilities, $\mathrm{E}(z_i) = \mu_i$, which are fitted for each count, where μ_i represents the fitted local conditional mean.

Even when dealing with count data in large samples, it is observed that the distribution of r_i exhibits heteroskedasticity and asymmetry (Cameron and Trivedi 2013). Consequently, analysts often rely on adjusted residuals to correct the inherent heteroskedasticity in raw residuals. These adjusted residuals are expected to be centered around zero and may also exhibit other desirable properties such as homoscedasticity and symmetry.

A commonly used adjustment for addressing heteroskedasticity is the *Pearson residual*, which is calculated as follows: $r_i^P = (z_i - \mu_i)/\sqrt{\widehat{\varphi}_i \mu_i}$, where $\widehat{\varphi}_i$ represents a parameter that helps control for overdispersion. It is worth noting that while Pearson residuals do result in adjusted residuals with zero mean and constant variance in large samples, the distribution of these residuals still tends to be asymmetric.

Deviance residuals serve as an estimate of the goodness of fit, and they are derived from the maximum likelihood estimation process. These residuals are computed as follows: $r_i^D = sgn(z_i - \mu_i)\sqrt{2(z_i \log(z_i/\mu_i) - (z_i - \mu_i))}$. The use of deviance residuals in the adjustment process achieves a distribution of residuals with desirable properties, including a zero mean, constant variance, and symmetry.

The *Anscombe residual* is a specific transformation of y_i, designed to make it closest to a normal distribution. Subsequently, it is standardized to have a zero mean and a variance of 1, represented as $r_i^A = 1.5\left(z_i^{2/3} - \mu_i^{2/3}\right)/\mu_i^{1/6}$. It's worth noting that when comparing the values of deviance and Anscombe residuals computed for the same model, they tend to be highly similar.

The concept of residuals derived from classical Poisson models can also be extended to Poisson point process models. Techniques for "informal" validation of parametric models of point processes, used to analyze spatial point pattern data, have been explored and discussed in Baddeley et al. (2005). Furthermore, the properties of such residuals were examined in Baddeley et al. (2008). In these "informal" techniques, no strict assumptions are imposed on the data (Baddeley et al. 2016). Typically, these techniques employ residuals to

validate a point process model, drawing an analogy to how residuals are used in classical parametric Poisson regression models.

Computing the raw residual of a point process model involves subtracting the integrated conditional intensity within the specified region B from the observed number of points, denoted as n, which represents the actual empirical count. The integrated conditional intensity corresponds to the fitted value or conditional mean of the intensity function. The raw residual is defined by the following equation (Stoyan and Grabarnik 1991; Baddeley et al. 2016):

$$R(B) = n(A \cap B) - \int_B \hat{\lambda}(s)\, \mathrm{ds} \tag{15}$$

where A is a set of training points $S(x_i, y_i)$ in two-dimensional space $\mathbb{R}^{2+}$ of the observed point pattern; $n(\mathrm{A} \cap B)$ is the number of points $S(x_i, y_i)$ in the sub-region B; and $\hat{\lambda}(s)$ is the fitted intensity of the fitted point process at any spatial position $s\,(x_0, y_0)$. Residuals in this context serve as metrics to quantify the discrepancy between the observed point pattern and the expected pattern, which is recalculated based on the estimated intensity $\hat{\lambda}(s)$ within the region B. Moreover, the concept of residual analysis is not limited to Poisson point processes; it can also be extended to non-Poisson point processes that involve interactions, such as Gibbs point processes (Baddeley et al. 2016).

The validation measure $R(B)$ in (15) considers not only the values at data points S, but also extends to locations s that do not correspond to observation points (Baddeley et al. 2005). This type of residual is referred to as *location-related* residuals (Illian et al. 2008), raw residuals (Baddeley et al. 2016), pixel-based (Gordon et al. 2015), or binned residuals (Lawson 1993). However, alternative *point-related* residuals have also been proposed, which involve calculation of local residuals specifically at the data observation points s (Stoyan and Grabarnik 1991; Lawson 1993; Illian et al. 2008), which will be explored in more detail below.

Accordingly, the total raw point residual $R(W)$ of the heterogeneous Poisson point process for the entire study area W is formally defined as.

$$R(W) = n(A \cap \mathrm{W}) - \int_W \hat{\lambda}(s)\, \mathrm{ds} \tag{16}$$

The Pearson point process residual is defined as:

$$R_P(B) = \sum_{S_i \in B} \frac{1}{\sqrt{\hat{\lambda}(S_i)}} - \int_B \sqrt{\hat{\lambda}(s)}\, \mathrm{ds} \tag{17}$$

for all instances where $\hat{\lambda}(S_i) > 0$. If the estimation is accurate, Pearson residuals are standardized with a mean of 0 and a variance of $|B|$.

The deviance point process residual for heterogeneous Poisson process (Lawson 1993) is defined as

$$R_D(T_i) = sgn\left(\frac{1}{|T_i|} - \hat{\lambda}_i(s)\right)\sqrt{d_i} \tag{18}$$

where $|T_i|$ is the area of the Thiessen polygon or Voronoi-Dirichlet tile, T_i, associated with the ith observation and d_i is the deviance contribution of the ith observation (Lawson 1993, 890–891). In this context, the log-likelihood serves as the loss function instead of the classical sum of squared errors. This choice is made because the log-likelihood is better suited for non-normally distributed response variables over a range of response density functions such as Poisson, gamma, exponential, log-normal, and others (Hastie et al. 2017).

Additionally, for comparing point process models, pixel-based deviances were proposed by Wong and Schoenberg (Gordon et al. 2015), which are defined as:

$$R_{DS}(B) = \sum_{S_i \in B} \log\left(\hat{\lambda}(S_i)\right) - \int_B \hat{\lambda}(s)\, \mathrm{ds} \tag{19}$$

According to Baddeley et al. (2005) and others, residuals are used in the analysis of fitted parametric models of point processes. Kernel density estimation can be considered as a fitted nonparametric point processing method. When dealing with an inhomogeneous point process that follows a Poisson distribution, the residuals of the point process can be used to test both the local (Equation 15) and global performance (Equation 16) of the spatial KDE.

The idea of optimizing residuals for performance *assessment* of a particular kernel density estimation method, similar to optimizing the fit of a linear regression model by minimizing the residual sum of squares, seems invalid. Trying to minimize Equation (15) by choosing an optimal bandwidth h for $\hat{\lambda}(u, v)$ will cause h to become extremely small, approaching zero. This is because aiming for the minimum residual sum of squares on the training dataset invariably results in zero residuals, making it unsuitable for determining the KDE bandwidth parameter h. Therefore, as h decreases, the level of overfitting in kernel density estimation increases.

Nevertheless, the idea of comparing different kernel density estimations by examining the residuals as the bandwidth approaches zero can be a valuable approach for *selecting* the appropriate KD estimator for a specific dataset. Additionally, creating a map and plotting the residuals as a function of different bandwidth sizes can be an effective tool for diagnosing the accuracy of kernel density estimations and even providing an informal rationale for choosing the optimal bandwidth. If the analysis is not primarily concerned with parameter estimates such as p-values, then a stepwise comparison of free parameters h with deterministic search may be a reasonable solution.

The proposed framework for comparing KD estimators is based on the mass conservation property defined as $n(A \cap \mathrm{W}) = \int_W \lambda(s)\mathrm{ds}$, where a boundary correction is considered to estimate $\hat{\lambda}(s)$. An adjusted mass conservation property has been introduced for bandwidth selection (Stoyan and Grabarnik 1991; Cronie and van Lieshout 2018). This property can also be expressed as

$\sum_{S_i \in B} \frac{1}{\lambda(s)} = |W|$, where $|W|$ represents the area of the observation window. The property is used to express inverse lambda residuals as follows:

$$R_I(W) = \sum_{S_i \in B} \frac{1}{\widehat{\lambda}(S_i)} - |W| \tag{20}$$

for all instances where $\widehat{\lambda}(S_i) > 0$. Calculating this residual metric is necessary only for the training data points S_i, and this metric can exhibit significant variance (Baddeley et al. 2016).

The classical global metrics ISE and MISE were developed to guide the selection of the optimal bandwidth based on data-driven properties for a particular KD estimator. KD estimators, developed specifically for spatial point processes, take into account spatial heterogeneity, point-to-point interactions (Diggle 1985), and covariance effects in fitting point patterns. However, the deviation from mass conservation serves as an absolute benchmark for assessing the performance of KD estimators, regardless of their data-driven properties and spatial effects. Using global or total residuals for diagnostics directly shows discrepancies between the fitted KDE surface and the spatial point process pattern.

Local residuals for all of the above residual metrics can be calculated for count data within sub-regions defined as regular (e.g., in quadrats of equal size and shape), natural, or administrative spatial units. The main drawback of the local approach is that the expected counts are influenced by the size and shape of the sub-regions. Creating a map of these local residuals can help interpret the variation of KDE residuals in different regions of the study area. When the residuals equal zero, the KDE surface has a perfect fit. Significant deviations from zero indicate a poor fit, highlighting the sub-regions where the KD estimator fails to accurately represent the data.

3.2.4 | Internal Validation Metrics: Estimating Cross-Validation Test Error or "Predicted Residual"

The raw residuals of the training point samples discussed above are estimates of the negative bias in intensity fitting (Baddeley et al. 2005). These residuals tend to overfitting when the bandwidth parameter h is small, leading to an artificially low expected training error $\mathrm{Err}_{\mathrm{training}}$. This overfitting is evident, given that the true test error $\mathrm{Err}_{\mathrm{test}}$ is not zero. Therefore, relying solely on the expected training error $\mathrm{Err}_{\mathrm{training}}$ to *assess* a KDE method and potentially make a *selection* may not be advisable.

Cross-validation (CV) estimators have an expected CV error $E(\mathrm{Err}_{\mathrm{CV}})$ that is closer to the expected test error, $E(\mathrm{Err}_{\mathrm{test}})$, than the expected training error $E(\mathrm{Err}_{\mathrm{training}})$ to $E(\mathrm{Err}_{\mathrm{test}})$. This behavior is due to the fact that $\widehat{f}^{-(i)}(S)$ is not a function of $S(x_i, y_i)$, $i \notin F_k$, and $\widehat{f}^{-(i)}(S)$ does not tend to overfit when the bandwidth parameter h is small. Consequently, it is recommended to use $\mathrm{Err}_{\mathrm{CV}}$, also known as the "predicted residual", rather than the $\mathrm{Err}_{\mathrm{training}}$ ordinary residuals for assessing and selecting a non-parametric KDE fit. The goal of cross-validation is to prevent overfitting and balance between bias and variance.

Classical cross-validation methods come in various forms, with two common forms being leave-one-out cross-validation (LOOCV) when K equals the total sample size ($K = n$), and k-fold cross-validation ($K > 1$) for other values of K, where K refers to the number of groups or folds.

LOOCV is used in many KDE methods for bandwidth selection. However, LOOCV tends to produce under-smoothed KDE surfaces, resulting in lower bias for $E(\mathrm{Err}_{\mathrm{test}})$, but a larger variance. This is primarily due to the fact that the n training sets generated in LOOCV are extremely similar to each other (Hastie et al. 2017).

The expected error of k-fold cross-validation $E(\mathrm{Err}_{k\text{-foldCV}})$ may deviate slightly more from the expected test error $E(\mathrm{Err}_{\mathrm{test}})$ compared to leave-one-out cross-validation $E(\mathrm{Err}_{\mathrm{LOOCV}})$. However, for sufficiently large sample sizes n, this discrepancy should not pose a serious problem. On the positive side, k-fold cross-validation provides the advantage of reducing the variance in the $\mathrm{Err}_{k\text{-foldCV}}$ estimate compared to the $\mathrm{Err}_{\mathrm{LOOCV}}$ estimate.

$\mathrm{Err}_{k\text{-foldCV}}$ is computed as the average of the validation errors obtained from k-fold cross-validation sets $CV_{\widehat{f}^{-(k)}}$, which tend to be less correlated than the $\mathrm{Err}_{\mathrm{LOOCV}}$ errors. This is due to the fact that the functions $\widehat{f}^{-(k)}(S)$, for $k = 1; \ldots, K$, do not depend on large overlapping k-folds, in contrast to LOOCV, where the functions $\widehat{f}^{-(k)}(S)$, $k = 1; \ldots, n$, use more overlapping data. However, one of the disadvantages of k-fold cross-validation is its sensitivity to the initial random splitting of samples (Hansen 2022).

Usually, the first step of k-fold cross-validation involves splitting the training dataset into k-folds of approximately equal size through a process of random shuffling of the training samples among these folds. Different random shuffling techniques may vary in how they generate the training and validation folds. By applying cross-validation to point process patterns, spatial cross-validation approaches consider spatial aspects of the data, including event locations and spatial effects such as spatial heterogeneity and spatial dependence.

There are some arguments that cross-validation partitioning for spatial point processes should be based on independent random thinning (Cronie et al. 2021). Independent random thinning is a process in which each point in a pattern of points is randomly removed or retained based on a specified probability function. The thinning process is independent, which means that the decision to remove or retain a particular point is made without considering the status of other points in the pattern. This process creates a reduced pattern of points that preserves certain statistical properties or meets certain requirements. For instance, when randomly thinning an inhomogeneous Poisson point process, the resulting point pattern remains a Poisson process with predefined probability density functions (Baddeley et al. 2016).

Using the classical k-fold cross-validation Equation (9) and the equation for raw total residuals (16), the total error $\mathrm{Err}_{k\text{-foldCV}}$ for the point process in the study area W can be determined as follows:

$$\mathrm{Err}_{k\text{-foldCV}} = \frac{1}{K} \sum_{k=1}^{K} R_{-(k)}(W) \tag{21}$$

$$R_{-(k)}(W) = n_{-(k)}(A \cap W) - \int_W \widehat{\lambda}\left(s_{-(k)}\right) \mathrm{ds}$$

Spatial point pattern thinning can be accomplished using various techniques, one of which is conventional k-fold cross-validation. In this technique, partitioning is achieved by randomly selecting cases from the learning set without replacement. Where the points are randomly marked from a multinomial distribution with independent identically distributed (i.i.d.) marks $m(S) \in \{1, \ldots, k\}$ and probabilities $p_1 = \ldots = p_k = 1/k$. Each fold corresponds to a specific mark, and these folds are mutually independent, with no overlapping elements.

Several spatial leave-one-group-out and leave-one-cluster-out cross-validation techniques have been proposed to address the spatial autocorrelation structure in the data, thereby eliminating spatial dependence.

1. One such technique is the *spatial l-block* cross-validation, denoted as SKBCV (Roberts et al. 2017). In the first step of SKBCV, the study area is divided into l spatially contiguous polygons or blocks, each of which can contain zero or more points. Blocks can have different sizes and shapes. Different techniques can be employed to create these blocks, such as applying unsupervised clustering methods to identify contiguity-constrained point clusters or employing regular or irregular grids to subdivide the spatial domain, etc.

   There are various strategies for assigning block points to the corresponding cross-validation fold. For instance, the number of blocks can be equal to the number of folds, that is, $k = l$ (where each block serves as its own fold). Alternatively, several blocks with points can be systematically or randomly assigned to the fold, resulting in $k \neq l$ (Valavi et al. 2019).

   However, the SKBCV approach presents several challenges. One of the main challenges relates to how folds should be defined. If grid subdivision is used, new hyperparameters such as block size and shape need to be estimated. There is also the possibility of edge effects, which should be corrected for, and extrapolation beyond blocks/folds may be necessary as well.

   Additionally, some blocks may remain empty, not containing any points. One notable limitation of this approach is its inability to account for point pattern heterogeneity. Blocks may vary significantly in the number of points they contain, and changes in point density are not taken into consideration.

2. To overcome some of the limitations associated with the formation of contiguous blocks, *space filling curves* (SFCs) can be used to group observation points within local neighborhoods. SFC starts with a base curve consisting of a set of segments positioned on an n-dimensional regular or irregular grid. This curve traverses each grid vertex exactly once, ensuring that it does not intersect itself. It has two free ends that can be joined to other paths. Formally, the SFC is a continuous function with endpoints whose domain is the unit interval (0, 1) (Sagan 1994). A SFC completely fills the region of interest. Notable examples of SFCs include the Peano curve and the Hilbert curve.

   The base curve is initially set to level/order 1. To construct a level i curve, each vertex of the base curve is replaced by a level $i-1$ curve, which can be appropriately rotated and aligned to fit the higher-level curve. Level $i-1$ vertices located in close proximity in space are assigned to the corresponding spatial cross-validation block. Each block corresponds to a level of sub-squares within the curve.

   The SFC approach has several benefits: there is no need to define block shapes; preserves the heterogeneity of point patterns; it eliminates the concern of extrapolation beyond blocks or folds; and there are no empty blocks.

3. The next spatial cross-validation technique is known as *buffering leave-one-out cross-validation*, $l = n$, denoted as BLOOCV (Le Rest et al. 2014; Pohjankukka et al. 2017). BLOOCV involves an additional step compared to classic LOOCV: in addition to excluding the point intended for validation, it also excludes other points that exhibit high autocorrelation with the validation point. The remaining points then form the training set and are used for estimations and fit validation.

   However, the BLOOCV approach comes with several challenges. This requires the estimation of a new hyperparameter, related to the autocorrelation range or buffer size. Additionally, it may require large computational resources, among other considerations.

4. Another approach that combines elements from both SKBCV and BLOOCV techniques, while overcoming some of their limitations, is the use of *Delaunay triangulation*. First, Delaunay triangulation is applied to all training points. Spatial folds are then created by excluding the validation point and its nearest neighbors, which are connected to the validation point in the triangle network, while the remaining points are used for training. To expand the fold size, additional nearest neighbors of the initial set of nearest neighbors can be excluded from the spatial fold.

It should be noted that there are no exact rules for determining the number of folds k and, for example, it has been discussed that it should be chosen based on the sample size n. A five-fold or ten-fold cross-validation is commonly recommended as a good compromise (Hastie et al. 2017), and is often used in practice. An even more challenging task is the choice of the number of spatial blocks l for the spatial cross-validation, as well as the size and shape of the block. A common approach is to treat the number of folds k and blocks l as hyperparameters and tune them, for example, using grid search tools such as GridSearchCV from the scikit-learn Python package (Pedregosa et al. 2011). However, the number of folds/blocks in CV can potentially lead to overfitting if not chosen carefully. The appropriate number of folds should be selected based on the dataset size and complexity (Hastie et al. 2017).

3.2.5 | Internal Validation Metrics: Point-Related Residuals

The spatial point process residuals from the training samples, as discussed in Sections 3.2.3 and 3.2.4, are computed at each location (x_0, y_0) on a fine grid within the bounded region

W. These residuals are known as *location-related residuals* (Illian et al. 2008). However, residuals can also be computed at the observation points $S(x_i, y_i)$ of the process, and these are called *point-related residuals* (Stoyan and Grabarnik 1991; Lawson 1993; Illian et al. 2008, 283). Similar to residuals in classical regression, these point-related residuals measure the difference between the observed points and their predicted values.

Few approaches have been proposed for calculating point-related residuals. One such approach, suggested by Illian et al. (2008, 283), involves using the "residual radius", which is the distance from an observation point (x_i, y_i) to the nearest point $s(x_0, y_0)$ with the maximum intensity value. A good fit of the point process intensity surface is indicated when the length of the residual radius is small. However, this method requires determining a control parameter, such as the residual radius, and identifying local maxima within it, if they exist, which makes the implementation challenging.

Another approach is the Lawson deviance residual, which provides a maximum likelihood estimate for each data point, as defined in Equation (18) (Lawson 1993). This approach is implemented based on the assumption that a point process is defined within convex polygons generated by tessellation. Thus, the observation window area $|W|$ is divided into Thiessen polygons, with the area of each polygon $|T_i|$ used to estimate the expected intensity as $1/|T_i|$ within the polygon. One limitation of implementing Thiessen polygons is that duplicate points are not allowed.

Similarly, a deviance residual can be calculated for KDE. The intensity at the data points corresponding to each Thiessen polygon in Equation (18) can be estimated using the leave-one-out cross-validation estimator, which introduces a slight negative bias, as shown below:

$$\widehat{\lambda_h}\left(s_i^{-(i)}\right) = \frac{|W|}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{K\left(\frac{s_i - s_j^{-(i)}}{h_{s_i^{-(i)}}}\right)}{h^2_{s_i^{-(i)}}} \tag{22}$$

$$|W| = \sum_{i=1}^{n} |T_i|$$

where $\widehat{\lambda_h}\left(s_i^{-(i)}\right)$ represents the kernel estimator of the intensity function fitted on all training points except the ith point. Bias corrections for edge effects can also be applied. This technique relies on the adjusted mass conservation property. However, its main limitations include the crude assumption that intensity is constant within each Thiessen polygon, as well as the difficulties in correcting for edge effects, particularly for truncated Thiessen tiles at the boundary of the observation window.

Additionally, Barr and Schoenberg (Gordon et al. 2015) proposed Voronoi residuals to diagnose the performance of spatial point pattern models. Similar to the Lawson deviance residual, Voronoi residuals are estimated within Thiessen polygons surrounding observed event points. The raw Voronoi residual for a point in the point process and its corresponding Thiessen polygon T_i is given by:

$$R_V(T_i) = 1 - \int_{T_i} \widehat{\lambda}(T_i) d\mu \tag{23}$$

where μ is the Lebesgue measure used to assign an area measurement within the Thiessen cell (Gordon et al. 2015).

3.2.6 | External Validation Metrics

The ability of a learning method to make accurate predictions on an independent test dataset is evidence of its generalization performance. In cross-validation resampling techniques, each test fold subset is actually used in the fitting process. To estimate the cross-validation error $E(\mathrm{Err}_{\mathrm{CV}})$, the expectation is performed across all random training and test point sets. To support the central assumption that the training and testing datasets are completely independent, a holdout approach can be used to split the entire learning data set into training and testing subsets. Therefore, the hold-out test data set is used solely for testing and is not incorporated in the training process. In addition, the hold-out test dataset does not depend on the distribution of the training set. Otherwise, error estimates will tend to be overly optimistic, leading to method selection favoring excessively complex models.

In a scenario with a "sufficiently large" dataset, one hold-out point dataset can be extracted from the entire learning point dataset using random thinning or spatial cross-validation splitting techniques, as described in Section 3.2.4.

In a more complex scenario, the entire learning dataset can be split into three subsets: the training subset, the validation subset (for cross-validation testing), and a test subset that includes hold-out points that remain entirely separate from the training process. To perform such splitting, various techniques can be used, such as independent random multinomial p-thinning with varying retention probabilities, subsampling/extrapolation technique or creating spatial blocks and then allocating them into subsets using various scenarios (Valavi et al. 2019; Wang 2019; Cronie et al. 2021).

In the previously described scenarios, $\mathrm{Err}_{\mathrm{test}}$ can be viewed as an *extra-sample error* since the test set does not have to overlap with the training set. Cross-validation and bootstrap techniques also provide direct estimates of the extra-sample error (Hastie et al. 2017).

It is worth emphasizing again that there are no precise guidelines for determining the optimal percentage allocation between training, validation, and testing subsets. Suggestions exist, for example, to include 10% to 30% of learning cases in the test set, and the remaining 90% to 70% of cases in the training set. Alternatively, a typical split might include allocating 50% to training and 25% each to validation and testing (Hastie et al. 2017).

Since the testing points are excluded from the KDE training process, point-related residuals must be calculated for these hold-out points. Three methods for calculating the point-related residuals are outlined in Section 3.2.5.

Furthermore, the inverse lambda point-based residuals for each Thiessen polygon T_i can be estimated using the technique described below:

$$R_I(T_i) = \frac{K}{M} \sum_{T_i} \frac{1}{\hat{\lambda}(T_i)} - |T_i| \tag{24}$$

where K is the number of points in the removed pattern (points that are discarded) and M is the number of points in the thinned pattern (points that are retained). The number of points removed (not retained) is $K = n - M$. The ratio $\frac{K}{M}$ represents the retention probability ratio between the *uniformly* thinned training and hold-out point patterns.

The technique involves the following steps. The intensity is estimated from the M-point training set for the entire window W, which is then divided using tessellation for the K hold-out points. In Equation (24), $\frac{K}{M} \sum_{T_i} \frac{1}{\hat{\lambda}(T_i)}$ represents the sum of intensities within the Thiessen polygon T_i of the hold-out testing point pattern, considering only those polygons where $\sum_{T_i} \frac{1}{\hat{\lambda}(T_i)} > 0$.

This technique assumes that if an inhomogeneous Poisson pattern with intensity function $\hat{\lambda}(S_i)$ is subjected to uniform random thinning, where the probability of retaining a point at location S_i is p, and the outcome for each point is independent of others, the resulting process of retained points will also be Poisson.

After uniform random thinning, the expected number of points remaining in region W is $M = p * \int_W \hat{\lambda}(S_i) dS$. The intensity function of the thinned process will be $p * \hat{\lambda}(S_m)$. The expected number of points remaining after uniform random thinning is also given by $M = n * p$, where n is the total number of points in the original pattern before thinning.

4 | Case Study: The Crime of Violence in Lithuania

From a criminological perspective, the study of the spatial dispersion of crime draws its theoretical framework from environmental criminology, which examines the relationship between crime and the physical environment, such as population density, land use, and urban design. Environmental criminology highlights the importance of geographic patterns in understanding and preventing crime. It informs national policies on urban planning, policing strategies, and crime prevention programs by emphasizing place-based interventions rather than just offender-focused approaches (Ceccato 2024).

The violent crime in Lithuania was selected for this case study because the crime statistics in Lithuania are notably higher than the average within the European Union, particularly in certain cities. In 2022, Lithuania's violent crime rate was reported at 2.21 homicides per 100,000 inhabitants, while the European Union's average rate was 0.86 per 100,000 inhabitants (Eurostat 2024). The value of this indicator increased to 2.63 in 2023.

Crime maps provide insights into environmental factors influencing criminal behavior. Utilizing KDE and estimated crime risk surfaces at a national level allows law enforcement and policymakers to develop proactive, data-driven strategies aimed at reducing violent crime in high-risk areas. The risk surface can be combined with other geographical data layers, enabling users to evaluate the crime rate in their area of interest. KDE can be utilized to aggregate and smooth various types of crime data, both spatially and temporally.

The comprehensive geocoded dataset documenting criminal incidents reported to Lithuanian police for 2022 includes millions of records, including attributes for various types of crime. Previous studies have found significant dependencies between these records and spatial factors, including socio-demographic patterns of the population, infrastructure patterns, and urban and landscape structures. This correlation is evident across various types of crimes, including violent crime (Vasiliauskas and Beconyte 2016; Beconytė et al. 2020).

For the KDE experiments, the records available for 2022 were selected, which include acts of violence such as assault, physical abuse, threatening behavior, home invasion, murder, and manslaughter, resulting in a total of 108,670 records. Figure 1 illustrates the spatial distribution of those events. The violent crime rate varies significantly between densely populated areas (more than 100 people per square kilometer, accounting for 77.4% of all events) and sparsely populated areas (< 100 people per square kilometer, which make up 22.6% of all events). The average violent crime rate per 1000 population is 36.59 for densely populated areas and 48.30 for sparsely populated areas.

Visual inspection of the point distribution in Figure 2 suggests that KDE will need to estimate a mixture of distributions, which may correspond to a mixed multivariate normal distribution.

4.1 | Crime Events as a Spatial Point Process

Criminal incidents are registered with geographic coordinates and represented as a collection of two-dimensional event points (Figure 2). Statistical analysis of these events can be done by considering them as a random point pattern derived from the realization of a spatial point process within a finite two-dimensional coordinate space (Daley and Vere-Jones 2003; Diggle 2006; Illian et al. 2008, etc.).

Figure 2 clearly indicates that the spatial pattern of crime events is not completely random. The spatial distribution demonstrates a tendency toward clustering, and this tendency can be attributed to first and/or second-order spatial effects (O'Sullivan and Unwin 2010). In this scenario, the first-order spatial effect can be readily explained by the spatially varying density of the human population, which differs significantly between urban and rural areas. Clearly, occurrences of crime are linked to populated places, as depicted in Figure 2. However, with respect to the second-order spatial effect, it is conceivable to assume that the location of one violent crime event is independent of the location of another violent crime event.

Given the above assumptions, the crime point pattern can be conceptualized and modeled as an inhomogeneous Poisson

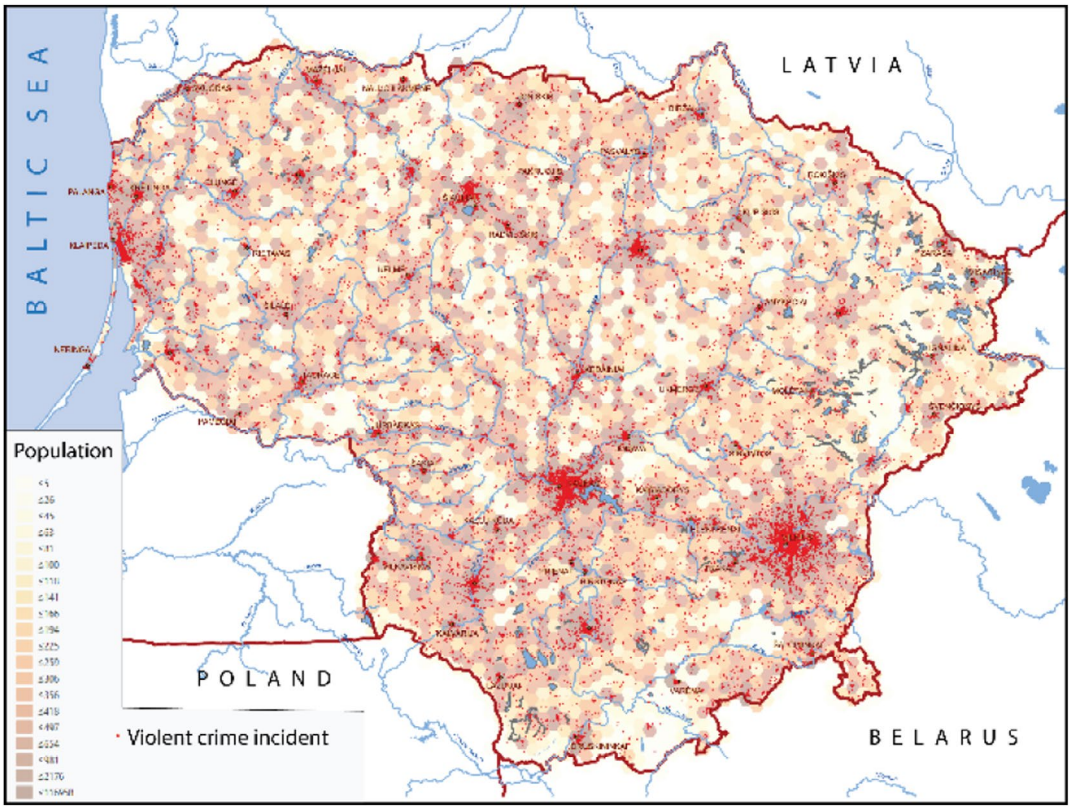

FIGURE 2 | Distribution of violent crime (108,670 incidents) in relation to Lithuania's population (2022).

process. The *Poisson* probability distribution is the classical law that governs the frequency of rare events, such as crime incidents. In criminology, there is a common assumption that crime events follow the Poisson distribution or an over-dispersed variant of the Poisson probability function, such as the negative binomial (Berk and MacDonald 2008).

Inhomogeneous point patterns are designed for situations where spatial heterogeneity is a significant factor. In the inhomogeneous Poisson point process model, under the assumption of constant risk and no interactions between events, each individual has an equal probability of being affected by crime during the observation period, regardless of location. In regions with a higher population at risk, a higher number of crime cases can be anticipated. A population covariate can be used to model the intensity of the Poisson crime process using various forms of the Poisson model, such as the baseline or constant risk model, where the crime intensity is proportional to the covariate.

The next section examines the relationship between population and crime events and evaluates the consistency of the crime point pattern as a realization of an inhomogeneous Poisson process. Following that, validation experiments were carried out under the same assumptions.

4.2 | Exploratory Data Analysis

As part of exploratory data analysis, the inhomogeneity, independence, and clustering of crime events were examined. Additionally, the baseline effect of the human population covariate on the distribution of crime was investigated.

As anticipated, the *p*-values obtained from Pearson chi-squared, likelihood ratio G^2, Freeman-Tukey $T2$, and Monte Carlo tests of homogeneity (Baddeley et al. 2016), using quadrat counts with rectangles of equal area, indicate the rejection of the null hypothesis of complete spatial randomness or homogeneous intensity of crime events. This rejection is made under the assumption that the point process follows a Poisson distribution, or that the points are independent of each other.

The *p*-value of the Average Nearest Neighbor (ANN) obtained from a Monte Carlo test indicates that the distribution of ANN values, simulated from the population density background, does not entirely support the notion that the clustering of crime events can be explained by an only random process when population density is the single controlling factor. Nevertheless, when the distribution of ANN values is adjusted for human population density, it exhibits a closer proximity to the observed ANN value. This suggests the possibility that other variables, such as household income, may contribute to explaining the clustering of crime events further.

Figure 3 illustrates variations in crime event density corresponding to higher human population density, generated through quantile tessellation of population counts using quadrat counts of crime events. Visual inspection of the KDE maps reveals overall similarities in the distribution of crime density and population density patterns (Figure 4), which are further confirmed by the explanatory data analysis presented in the following section.

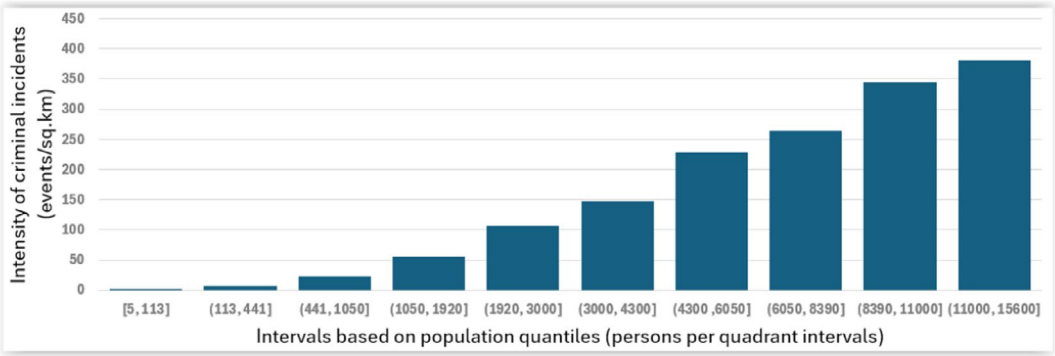

FIGURE 3 | The graph presents the distribution of crime event intensity against population intensity, grouped by 10-quantile intervals.

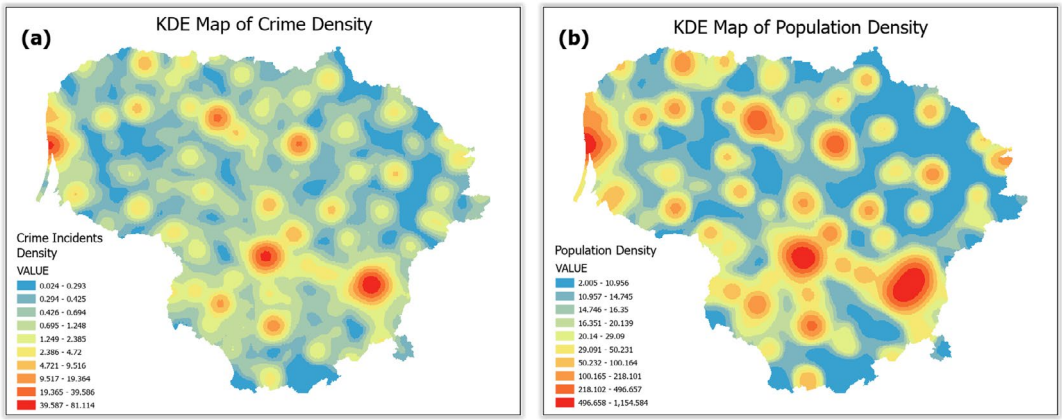

FIGURE 4 | Isotropic kernel density surfaces for (a) crime and (b) population, using a 5 km bandwidth. The densities were grouped using the geometric interval classification method, with 10 classes for each surface.

Two formal tests, the chi-squared test and the Monte Carlo quadrat counting test, were conducted to assess the (non-)dependence on a covariate (Baddeley et al. 2016). Both tests suggested that the density of crime events depends on population density, with a two-sided $p \leq 0.01$, supporting the alternative hypothesis.

Additional more robust tests were performed to explore the influence of a covariate on which crime intensity may depend, rather than assuming homogeneity. The estimation of Ripley's second moment function $K(r)$ derived from the distribution of crime points reveals a departure from a complete spatial randomness or a homogeneous Poisson point process. The estimation of the inhomogeneous K-function (Baddeley et al. 2016) for a nonstationary point pattern indicates that the crime points pattern aligns with an inhomogeneous Poisson process characterized by the chosen density. This suggests the possibility of correlation-stationarity within the point process, or fluctuations due to changes in underlying covariates such as population density. Consequently, the similarity in the results of the inhomogeneous K-function between the two halves of the dataset implies the validity of assumption of correlation-stationarity.

Cumulative distribution function (CDF) tests, including Kolmogorov–Smirnov, Cramér-Von Mises, and Anderson-Darling (Baddeley et al. 2016), evaluate CSR based on covariate values at data points. Those tests reject the homogeneity of crime points and support the alternative hypothesis of dependence on the human population covariate. Furthermore, Berman $Z1$ and $Z2$ tests (Baddeley et al. 2016) demonstrate the dependency of the crime point process on the human population.

The CDF test also measures the strength of the effect of a covariate in terms of area under the receiver operating characteristic curve (AUC). The AUC values close to 1 or 0 signify robust discrimination, while a value of 0.5 indicates no discriminatory power. In this CDF analysis, the obtained AUC ≈ 0.945 suggests that high densities of crime events are anticipated at elevated values of population density.

To further test and measure the relationships between the spatial distribution of crime events and the human population, several parametric point process models are run and evaluated against the null hypotheses using ANOVA likelihood ratio tests and compared using AIC and AUC measures (Baddeley et al. 2016):

- The homogeneous Poisson model, also known as the constant intensity model with $\widehat{\lambda}(u,v) \equiv \lambda = 1.674$, is used as a reference in certain tests; the AIC value is 105,366.
- The offset model, expressed as $\widehat{\lambda}(u,v) = 0.05160782\widehat{z}(u,v)$, defined by an intensity that is scaled proportionally with the baseline, namely the human population density. The ANOVA test rejects the null hypothesis, supporting the notion that crime varies with population density; the AIC value is −344,222 and the AUC value is 0.8674.

- The log-linear Poisson regression model, given by $\log\left(\widehat{\lambda}(u,v)\right) = 0.256 + 0.00279\widehat{z}(u,v)$, incorporates the human population as a covariate. The ANOVA test rejects the null hypothesis, supporting the notion that crime varies with population density, with an AIC value of −296,543 and an AUC value of 0.8674.

In the above equations, $\widehat{\lambda}(u,v)$ denotes the crime intensity per square kilometer, and $\widehat{z}(u,v)$ denotes the human population value at a given spatial location u,v. The last two models demonstrate a significant ($p < 0.01$) influence of population density on crime distribution. The area under curve (AUC) measures show a high separation in the spatial domain, delineating areas with high and low densities of crime points corresponding to human population. The offset model with the lowest AIC value outperformed the log-linear Poisson regression model.

The spatial relative risk function was used to illustrate the spatial interactions between crime events and the underlying at-risk population. If the kernel bivariate densities of crime events, denoted as $K_e(\bullet)$, and the human population at risk, denoted as $K_p(\bullet)$, are both estimated through their own KDE processes, then the joint spatial relative risk function, denoted as $\hat{r}(x,y)$, can be expressed as the ratio of densities describing the spatial distribution of crime events and the population at risk background controls (Bithell 1991; Davies et al. 2018) as follows:

$$\hat{r}(x,y) = \frac{\frac{1}{n_e h_e^2}\sum_{i=1}^{n_e} K_e\left(\frac{d_{(x,y),i}}{h_e}\right)}{\frac{1}{n_p h_p^2}\sum_{j=1}^{n_p} K_p\left(\frac{d_{(x,y),j}}{h_p}\right)} \tag{25}$$

where bandwidths for both kernel functions are denoted as h_e and h_p, respectively. As per Davies et al. (2016), employing a common jointly optimal spatial bandwidth $h_e = h_p$ for both events and background controls offers several advantages.

The density surfaces of relative risk with bandwidths $h_e = h_p = 1$ km are illustrated in Figure 5. Areas with an average risk density $\hat{r}(x,y) \cong 0$ and $\widehat{f}_c(x,y) \cong \widehat{f}_p(x,y)$ are shown in yellow. Figure 5b highlight areas where $\hat{r}(x,y) > 0$ with a higher localized concentration of crime relative to the population density, and areas with a relatively low crime rate where $\hat{r}(x,y) < 0$ are outlined in Figure 5c.

Maps in Figure 5 illustrate asymptotic p-value surfaces delineating tolerance areas from upper-tailed and lower-tailed tests, respectively, at significant 5% thresholds of elevated (Figure 5b) and reduced (Figure 5c) risk (Davies and Lawson 2019). The highlighted areas outline areas where anomalous crime activity may occur, indicating significantly increased or decreased crime risk compared to the background population density.

Nonetheless, the relationship between crime event densities and human population densities is not directly proportional, as evident from the relative risk maps (Figure 5). Therefore, using an inhomogeneous Poisson offset model with an intensity proportional to the baseline population can serve as a simplification of the relationship between crime events and population.

Based on exploratory data analysis, it appears that the crime event pattern closely follows an inhomogeneous Poisson process, with clustering being explained by the baseline effect of the human population.

4.3 | KDE of the Crime of Violence in Lithuania

A series of experiments was carried out to investigate kernel estimations for crime count data. These experiments advanced through three stages: (1) estimating bandwidths; (2) computing kernel functions using the estimated bandwidths and visualizing kernel surfaces as density maps; and (3) calculating validation metrics for the evaluation of results. The key reference in this section is Baddeley et al. (2016), which includes extensive discussion on model validation for point patterns. They introduce methods for assessing the fit of a theoretical model to observed data, such as analyzing residuals and using formal statistical tests.

4.3.1 | Bandwidth Estimation Experiments

The most popular bandwidth selectors from plug-in (PL), cross-validation (CV), and hybrid methods underwent testing. Some of these selectors had foundations in spatial point process theory. Different techniques were used to estimate univariate and bivariate bandwidths in the spatial domain, involving isotropic, diagonal, and full bandwidth matrices, as well as fixed, adaptive, and mixed bandwidths. Boundary corrections implemented within the KDE algorithms were applied. The most applicable outcomes of bandwidth estimations for the specified dataset are presented in Table 1.

The bandwidth values presented in Table 1 show considerable variations, ranging from approximately 100 m (based on

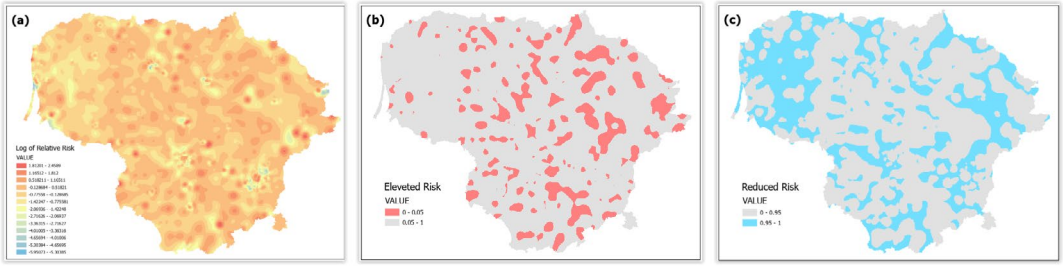

FIGURE 5 | (a) Estimated log-transformed relative risk surface for relative crime/population density; (b) areas with high relative risk ($p \leq 0.05$); and (c) areas with low relative risk ($p \leq 0.05$).

TABLE 1 | Bandwidth selection results.

Bandwidth selectors	Univariate/ isotropic Spatial fixed or adaptive (interval), m	Anisotropic/ diagonal, m	Full matrix, m
Plug-in (PI) including "rule-of-thumb"			
Oversmoothing (OS) bandwidth selector, Terrell (1990) rule-of-thumb, as implemented in the "sparr" package	11,656		
Normal scale bandwidth selector by Silverman (1986) rule-of-thumb, "sparr" and "sm" packages	10,747	$\begin{bmatrix} 12{,}299 & 0 \\ 0 & 9195 \end{bmatrix}$	
Bivariate bandwidth selector by Scott (1992) rule-of-thumb, "spatstat" package	10,634	$\begin{bmatrix} 12{,}299 & 0 \\ 0 & 9195 \end{bmatrix}$	
Rule-of-thumb for bandwidth selector for the pair correlation function (Stoyan and Stoyan 1995), "spatstat" package[a]	52		
Direct Sheathe and Jones' rule-of-thumb bandwidth selector at level 2, 2D data, (Wand and Jones 1995), "ks" package		$\begin{bmatrix} 1341 & 0 \\ 0 & 1003 \end{bmatrix}$	$\begin{bmatrix} 1563 & -929 \\ -929 & 1168 \end{bmatrix}$
Normal scale bandwidth selector, (Chacón et al. 2011), "ks" package		$\begin{bmatrix} 12{,}144 & 0 \\ 0 & 8131 \end{bmatrix}$	$\begin{bmatrix} 12{,}299 & -7225 \\ -7225 & 9195 \end{bmatrix}$
Normal mixture bandwidth selector with four mixture components, (Cwik and Koronacki 1997), "ks" package		$\begin{bmatrix} 12{,}008 & 0 \\ 0 & 8872 \end{bmatrix}$	$\begin{bmatrix} 12{,}270 & -7306 \\ -7306 & 9268 \end{bmatrix}$
Normal scale bandwidth selector over product kernel with the Silverman rule-of-thumb, (Li and Racine 2007), "np" package		$\begin{bmatrix} 13{,}028 & 0 \\ 0 & 9740 \end{bmatrix}$	
Cross-validation (CV)			
Unbiased least squares CV (LSCV) selector for bivariate, edge-corrected bandwidth (Davies and Baddeley 2018), "sparr" package	222		
Likelihood CV (LIK) selector for bivariate, edge-corrected bandwidth (Davies and Baddeley 2018), "sparr" package	333		
MSE CV bandwidth selector of point process density (Berman and Diggle 1989), assumes a Cox process, "spatstat" package[a]	13		
The likelihood leave-one-out CV (LCV) selector (Loader 1999), assumes an inhomogeneous Poisson process, "spatstat" package[a]	530		
The likelihood CV bandwidth selector based on preservation mass criterion (Cronie and van Lieshout 2018), "spatstat" package[a]	26,447		
The global adaptive likelihood CV bandwidth selector based on preservation mass criterion (van Lieshout 2022), "spatstat" package[a]	530		
Abramson-Hall-Marron's (1982, Hall and Marron 1988) adaptive bandwidth selector uses a global bandwidth derived from the LCV selector, "spatstat" package	120–2650		
Least squares CV (LSCV) bandwidth selector derived from a single value (Bowman and Azzalini 1997), "sm" package	6182	$\begin{bmatrix} 7405 & 0 \\ 0 & 5536 \end{bmatrix}$	

(Continues)

TABLE 1 | (Continued)

Bandwidth selectors	Univariate/ isotropic Spatial fixed or adaptive (interval), m	Anisotropic/ diagonal, m	Full matrix, m
Biased CV (BSV) bandwidth selector for bivariate data (Sain et al. 1994), "ks" package		$\begin{bmatrix} 11{,}742 & 0 \\ 0 & 8620 \end{bmatrix}$	$\begin{bmatrix} 12{,}655 & 7523 \\ 7523 & 9461 \end{bmatrix}$
Unbiased CV (UCV) bandwidth selector for bivariate data (Bowman 1984), "ks" package			$\begin{bmatrix} 9471 & 4573 \\ 4573 & 2207 \end{bmatrix}$
Smoothed CV (SCV) bandwidth selector (Chacón and Duong 2018), "ks" package		$\begin{bmatrix} 1640 & 0 \\ 0 & 1274 \end{bmatrix}$	$\begin{bmatrix} 1852 & -1021 \\ -1021 & 1433 \end{bmatrix}$
Bootstrap			
Bootstrap-estimated MISE, edge-corrected fixed and global bandwidth selectors (Davies and Baddeley 2018), "sparr" package	2907 5330		
Mixed (Mammen et al. 2011)			
$\alpha = \beta = 1$	3162		
$\alpha = 2, \beta = 1$	2154		
$\alpha = 1, \beta = 2$	4642		

[a]Bandwidth selectors based on point process theory rely on assumptions regarding point dependencies, such as assuming an inhomogeneous Poisson process or assuming a Cox process, etc.

cross-validation) to nearly 12,000 m (using the over-smoothing Terrell plug-in selector) for fixed bandwidths. Notably, the selector based on the preservation mass criterion (Cronie and van Lieshout 2018) produces unreliable results. Choosing the suitable bandwidth for a particular application is a practical concern, as no single selection rule dominates others. Therefore, the results presented in Table 1 highlight the impracticality of using ISE and MISE optimization criteria to unquestionably select the optimal bandwidth.

Selecting the optimal bandwidth for a specific case relies on the actual shape of the density being estimated and the criteria employed to evaluate the estimate's quality. Furthermore, the choice of a particular optimal bandwidth is related to the data sample size and the complexity of the data distribution. Several bandwidth selectors, primarily cross-validation-based methods (SCV and LCV unconstrained selectors) yielded completely unsatisfactory results and are not included in Table 1. The failure can be due to (a) discretization effects and data rounding in a very large dataset, and/or (b) intrinsic assumptions about the dependence between points that are not true for the test dataset with density regions of different shapes and sizes, multimodality, and asymmetry.

However, the data in Table 1 reveal several consistent patterns, with some exceptions. The results confirm the anticipation that fixed cross-validation-based methods, notably classic LSCV, would generate isotropic bandwidths with very small undersmoothing. The majority of the assessed selectors estimate bandwidths to be under 1000 m. Cross-validation methods might not be suitable for highly variable data due to their tendency to yield estimates with minimal bias but substantial variance. Moreover, LSCV selectors do not perform effectively for large samples (Heidenreich et al. 2013). Modified cross-validation methods, such as SCV, account for much less variation without significantly increasing bias; in our study, they generated more realistic bandwidths.

Compared to cross-validation methods, plug-in selectors always have a smaller, asymptotic variance, but relatively large bias. The plug-in estimates tend to be more stable, resulting in over-smoothed surfaces. There is an assumption that the asymptotic properties of the sophisticated plug-in methods make those methods hard to compete, and other bandwidth selectors are usually underperformed (Heidenreich et al. 2013). From our tests, most plug-in selectors suggest spatial bandwidths with sizes around 10,000 m. Sheathe and Jones' rule-of-thumb generates relatively modest bandwidths; nevertheless, there are no objective criteria available for selecting its arbitrary level parameter.

Furthermore, experiments were conducted by mixing methods that combine various bandwidths and/or KD estimators. Such bandwidth mixtures have some potential to yield stable results. The mixture can be done by different methods; one of the simplest ones is using cross-validation and plug-in bandwidth in different proportions on a logarithmic multiplicative scale, as described in Heidenreich et al. (2013). Even in practical scenarios, a straightforward average of cross-validation and plug-in bandwidths may outperform their individual alternatives, as noted in Mammen et al. (2011). Given the range of biases inherent in the ISE and MISE minimizing methods, a combination of cross-validation and plug-in can represent a viable compromise. The following formula can be used

TABLE 2 | Bivariate radial-symmetric kernels with isotropic fixed bandwidths.

Map no.	Kernel density estimation method	Bandwidth, m (fixed)	Total raw point residuals (number of points)	Total point raw point residuals (% out of 108,670 events)
1.	Spatial bivariate Gaussian isotropic	10,000	6566	6.04
2.	Spatial bivariate Gaussian isotropic	5000	3625	3.34
3.	Spatial bivariate Epanechnikov isotropic	5000	3790	3.49
4.	Spatial bivariate Quartic isotropic	5000	3756	3.46
5.	Spatial bivariate Gaussian isotropic	2000	1453	1.34

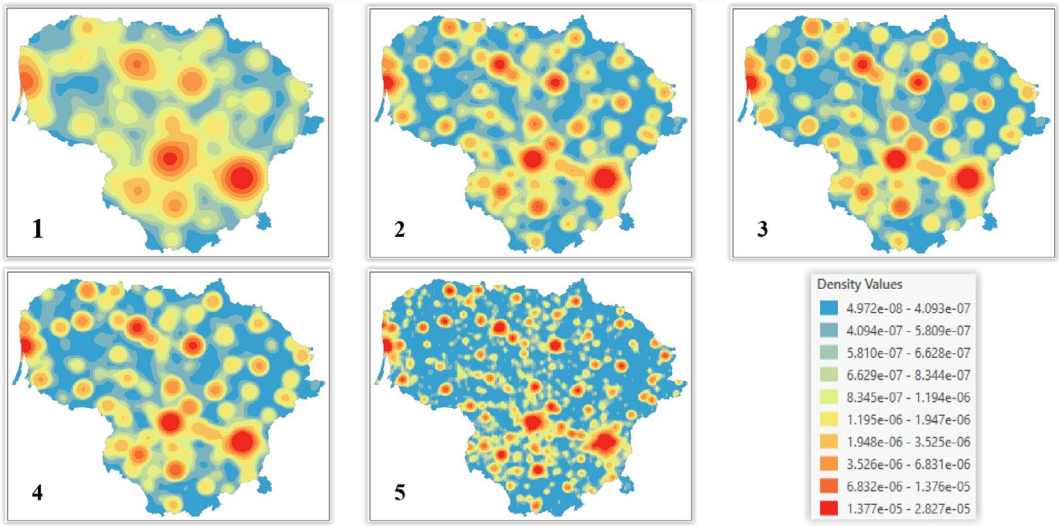

FIGURE 6 | Kernel density maps generated using spatial bivariate isotropic KDEs with different bandwidths and kernel functions, as specified in the settings provided in Table 2.

to calculate three types of mixture of cross-validation and plug-in bandwidths:

$$h_{\text{mix}} = \left(\hat{h}_{\text{CV}}^{\alpha}\hat{h}_{\text{PL}}^{\beta}\right)^{\frac{1}{\alpha+\beta}} \tag{26}$$

where the three possible combinations are $\alpha = \beta = 1, \alpha = 1,\ \beta = 2$ and $\alpha = 1,\ \beta = 2$. In Table 1, combinations of $\hat{h}_{\text{PL}} = 10{,}000$ m and $\hat{h}_{\text{CV}} = 1000$ m were used.

4.3.2 | Experiments With Variants of Kernel Density Estimators

The experiments involve a variety of kernel functions with different parameterizations and bandwidth sets to estimate kernel densities for violent crime events, followed by the calculation of internal validation metrics and visualization of kernel density estimation surfaces. The experiments were carried out using fixed isotropic, diagonal, and full matrix bandwidths, as well as adaptive bandwidths as specified below.

Based on the range of isotropic bandwidth values, estimated using plug-in, cross-validation, and mixture selectors (see Table 1), the fixed bandwidths of 2000, 5000, and 10,000 m were chosen for testing. For each bandwidth value, Gaussian, Epanechnikov, and quartic kernel functions were used to generate kernel density surfaces. These density estimates were not adjusted for edge effect bias.

Table 2 and Figure 5 display selected outcomes of violent crime density surfaces using classic bivariate radial-symmetric kernels with isotropic fixed bandwidths. In all maps, the classification and visualization of density values remain consistent, as indicated in the legend in Figure 5. The geometric interval classification method with 10 classes is used throughout. Additionally, Table 2 presents the total raw point residuals $R(W)$ for each estimate.

Maps in Figure 6 show that density surfaces constructed with these bandwidths exhibit significant visual differences. Reducing the bandwidth for estimates (as shown in maps #1, 2, and 5 in Figure 6) enhances the surface details. As a result,

TABLE 3 | Bivariate radial-symmetric kernels with anisotropic fixed diagonal and full bandwidths.

Map no.	KDE estimation method	Bandwidth matrices, m	Total raw point residuals (number of points)	Total point raw point residuals, (% out of 108,670 events)
6.	Spatial bivariate Gaussian diagonal anisotropic	$\begin{bmatrix} 10{,}000 & 0 \\ 0 & 7375 \end{bmatrix}$	5909	5.44
7.	Spatial bivariate Gaussian diagonal anisotropic	$\begin{bmatrix} 5000 & 0 \\ 0 & 3687 \end{bmatrix}$	3383	3.11
8.	Spatial bivariate Gaussian diagonal anisotropic	$\begin{bmatrix} 2000 & 0 \\ 0 & 1475 \end{bmatrix}$	1385	1.27
9.	Spatial bivariate Gaussian unconstrained anisotropic	$\begin{bmatrix} 10{,}000 & -5643 \\ -5643 & 7375 \end{bmatrix}$	5730	5.27
10.	Spatial bivariate Gaussian unconstrained anisotropic	$\begin{bmatrix} 5000 & -2822 \\ -2822 & 3687 \end{bmatrix}$	3258	3.00
11.	Spatial bivariate Gaussian unconstrained anisotropic	$\begin{bmatrix} 2000 & -1129 \\ -1129 & 1475 \end{bmatrix}$	1317	1.21

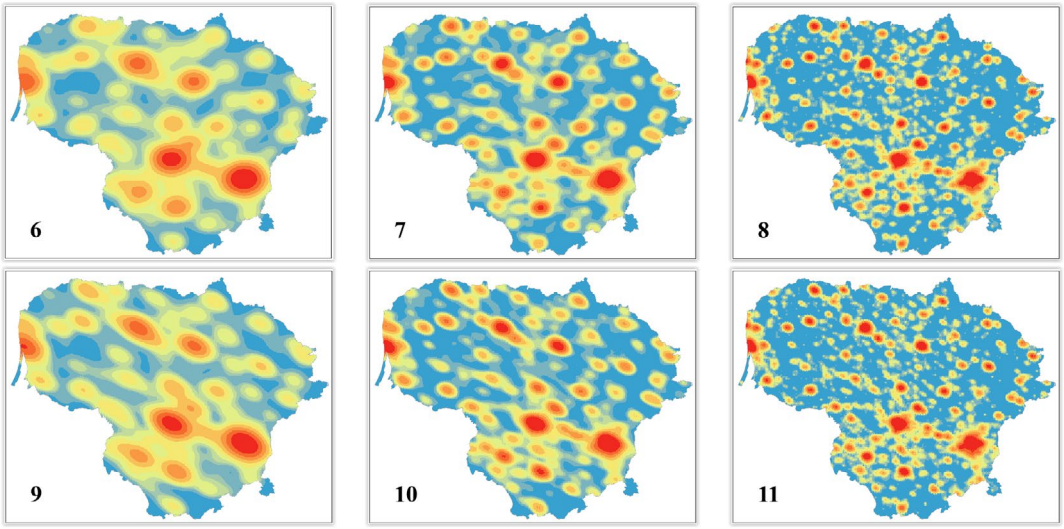

FIGURE 7 | Kernel density maps generated using spatial bivariate anisotropic Gaussian KDEs with different bandwidths and parameterization classes, as specified in the settings defined in Table 3.

the density spots around major population areas become more condensed and less rounded.

As expected, the choice of kernel functions (maps #2–4 in Figure 6) does not have a significant impact, as evidenced by the differences in the total point raw residual $R(W)$ values as well. However, with the same other parameters, the Gaussian kernel yields slightly better results in terms of the total point raw residual $R(W)$. The Gaussian kernel is one of the smoothest KD estimators, often requiring a larger optimal bandwidth h. When applied to non-normal data, it tends to cause over-smoothing.

Density surfaces, corrected for edge effects using both "uniform" correction and Diggle's correction techniques (Davies et al. 2018; Diggle 1985), appear visually similar to the corresponding uncorrected ones, with differences observed only along the boundary of the study area.

Table 3 and maps in Figure 6 display selected outcomes of violent crime density surfaces using classic bivariate radial-symmetric kernels with anisotropic fixed bandwidths, without boundary corrections. The anisotropy ratio between the horizontal h_{11} and vertical h_{22} directions in the bandwidth matrix is consistent across all estimated bandwidth sizes in Table 1, the average ratio value is approximately 1.356. The average ratio between matrix elements h_{11} and h_{12} is approximately 1.772. These ratios were used to compute the bandwidth matrix elements in Table 3 and to plot the residuals in the following subsection.

Surfaces of the diagonal and full parametrization classes (maps #6–11 in Figure 7) exhibit anisotropic stretching from the northwest to the southeast. The major range direction of anisotropy is confirmed through trend analysis using global polynomial interpolation, estimated to be approximately 120°.

TABLE 4 | Bivariate radial-symmetric kernels with isotropic adaptive bandwidths.

Map no.	KDE estimation method	Global bandwidths (h_0), m	Total raw point residuals (number of points)	Total point raw point residuals, (% out of 108,670 events)
12.	Isotropic spatial Gaussian adaptive Abramson	10,000	8943	8.23
13.	Isotropic spatial Gaussian adaptive Abramson	5000	4887	4.50
14.	Isotropic spatial Gaussian adaptive Abramson	2000	1955	1.80

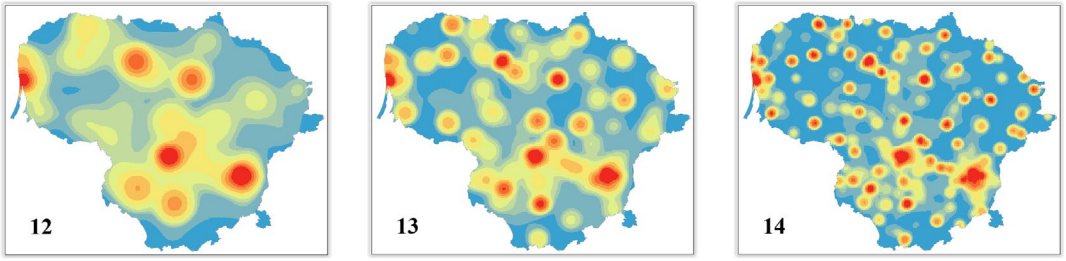

FIGURE 8 | Kernel density maps generated using spatial bivariate isotropic Abramson Gaussian KDE with different bandwidths, as outlined in the settings provided in Table 4.

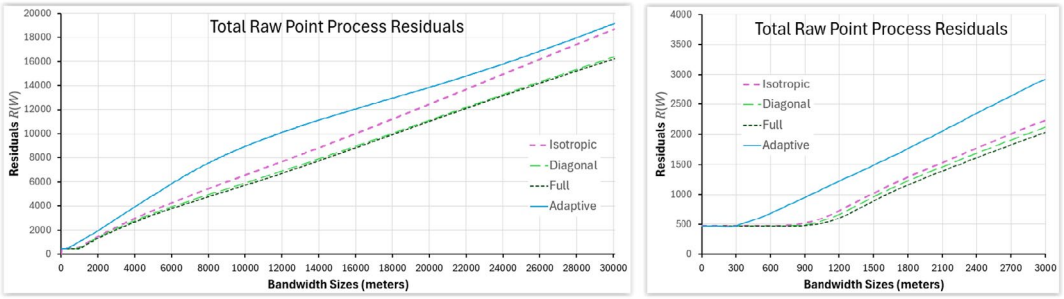

FIGURE 9 | Variation of the total raw point residuals $R(W)$ (left), and its zoomed-in view at lower values (right), displayed against bandwidth size in meters, without adjustments for edge effects.

Considering the anisotropic tendency of the phenomena, the total point process residual values decrease in the diagonal KD estimations compared to the isotropic KD estimations and decrease further in the full parametrization KDEs (Tables 2 and 3), indicating an improvement in estimations.

The adaptive bandwidth surfaces (maps #12–14 in Table 4 and Figure 8), generated with Abramson's adaptive sample-point technique (Abramson 1982) and without boundary corrections, are structurally similar to the surfaces with fixed bandwidth sizes. However, their densities tend to be more concentrated in areas with higher populations. Interestingly, the estimates with variable bandwidths perform worse than both fixed bandwidth isotropic and anisotropic KDEs.

4.3.3 | KDE Global Errors for Selecting the Best Estimation

The global errors of KDE were plotted against varying bandwidth sizes to investigate their changes. Figure 9 shows the total raw point residuals against various bandwidth sizes for different kernel density estimators. These density estimates were made without adjusting for edge effect bias.

As expected, the total raw residuals $R(W)$ decrease with smaller bandwidths but do not converge to zero, likely due to the numerical precision of density computations. Spatial bivariate radial-symmetric kernels with anisotropic fixed bandwidths outperform both isotropic fixed and adaptive KDEs.

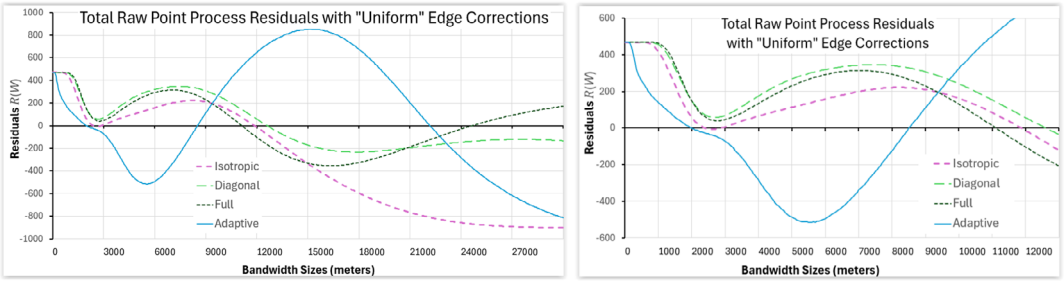

FIGURE 10 | Variation of the total raw point residuals $R(W)$ (left), and its zoomed-in view at lower values (right), displayed against bandwidths with "uniform" adjustments for edge effects.

As shown in Figure 8, the total raw point residuals $R(W)$ for the spatial bivariate Gaussian KD estimators with fixed bandwidths stabilize around 1000 m. This suggests that a fixed smoothing bandwidth of <1000 m may be sufficient. Meanwhile, the residuals $R(W)$ for the isotropic spatial Gaussian adaptive KD estimators, where global and pilot bandwidths vary under the condition of $h_0 = h_{pl}$, stabilize at bandwidth $h \approx 300$ meters. Similar to bandwidths derived from cross-validation selectors, the stabilization of $R(W)$ at small bandwidths results in minimal bias.

As anticipated, at the same bandwidth scales, KD estimators that utilize diagonal and unconstrained bandwidth selectors outperform their isotropic counterparts, resulting in lower residuals $R(W)$ due to the oblique orientation of the crime point pattern relative to the coordinate axes. Additionally, the KD estimators with full matrices show slightly better performance than those with diagonal matrices in terms of $R(W)$.

An interesting finding was that isotropic KD estimators with adaptive bandwidths did not show any improvement over KD estimators with fixed bandwidths. The smoothing regimen rule (Abramson 1982) is typically considered well-suited for processing spatial data, which often exhibit significant heterogeneity due to underlying processes like population density (Davies and Baddeley 2018). However, with the settings used $h_0 = h_{pl}$, the $R(W)$ values for the adaptive KD estimators are higher than those for estimators with fixed bandwidths—as seen in Figure 8, all the graphs converge into a single line at a bandwidth of approximately 300 m.

The total raw point residuals $R(W)$ were calculated for density estimations without considering edge or boundary effects. The raw kernel density estimates may exhibit significant negative bias near the boundary, which can fluctuate depending on the chosen bandwidth values, especially when larger bandwidth values are used. This boundary bias may need to be corrected, depending on the asymptotic properties of the estimator.

Several solutions have been proposed to address this boundary bias issue, such as using special kernels, incorporating bias-correction terms into $\hat{f}_h(s)$, applying domain transformations, etc. (Karunamuni and Alberts 2005). In the experiments, both "uniform" edge correction at the test point s and Diggle's edge correction based on observation S_i (Baddeley et al. 2016) have been used. The corresponding KD estimators and the incorporated edge-correction terms are defined as follows:

$$\hat{f}_h^{\,U}(s) = \frac{1}{nh^2 g_h(s)} \sum_{i=1}^{n} K\left(\frac{s - S_i}{h}\right) \text{ where } g_h(s) = \frac{1}{h^2} \int_W K\left(\frac{u - s}{h}\right) du \tag{27}$$

$$\hat{f}_h^{\,D}(s) = \frac{1}{nh^2} \sum_{i=1}^{n} \frac{1}{g_h(S_i)} K\left(\frac{s - S_i}{h}\right) \text{ where } g_h(S_i) = \frac{1}{h^2} \int_W K\left(\frac{u - S_i}{h}\right) du \tag{28}$$

where $\hat{f}_h^{\,U}(s)$ and $\hat{f}_h^{\,D}(s)$ denote the uniformly global edge-corrected and Diggle's local edge-corrected KD estimators at a point s, along with their respective edge-correction terms. The edge correction term $g_h(.)$ is defined as the reciprocal of the kernel mass across the entire study area W. The edge correction term rescales the current estimate.

Figure 10 displays the total raw point residuals against different bandwidth sizes for various KDE variants with "uniform" edge correction. This estimator is biased in general and is only unbiased when the density is constant in a homogeneous point pattern process (Baddeley et al. 2016). However, the values of the total raw point residuals after "uniform" scaling are significantly lower than those without corrections.

The graphs of the uniformly adjusted residuals first reach local minima at around 2000 to 2500 m (see Figure 9). The second set of local minima for the residual graphs using kernel density estimators with fixed bandwidths occurs between 11,000 and 12,500 m. Meanwhile, the adaptive residual graph displays local minima at 2000 and 8500 m.

Comparing these local minima to values in Table 1, it can be seen that the fixed KDE bandwidths in the plug-in selectors closely correspond to the second set of local minima. Additionally, some cross-validation selectors also estimate bandwidths near 11,000 m.

The results for the cross-validation selectors in Table 1 are inconsistent. However, it can be noticed that some cross-validation selectors, as well as the Bootstrap and Mixed selectors, correspond to the first minima observed in the graphs.

In terms of performance, the trends of residuals with "uniform" scaling are different from the trends of residuals without scaling. Notably, Figure 10 shows that isotropic fixed kernel uniformly adjusted residuals exhibit smaller values compared to anisotropic fixed kernel uniformly adjusted

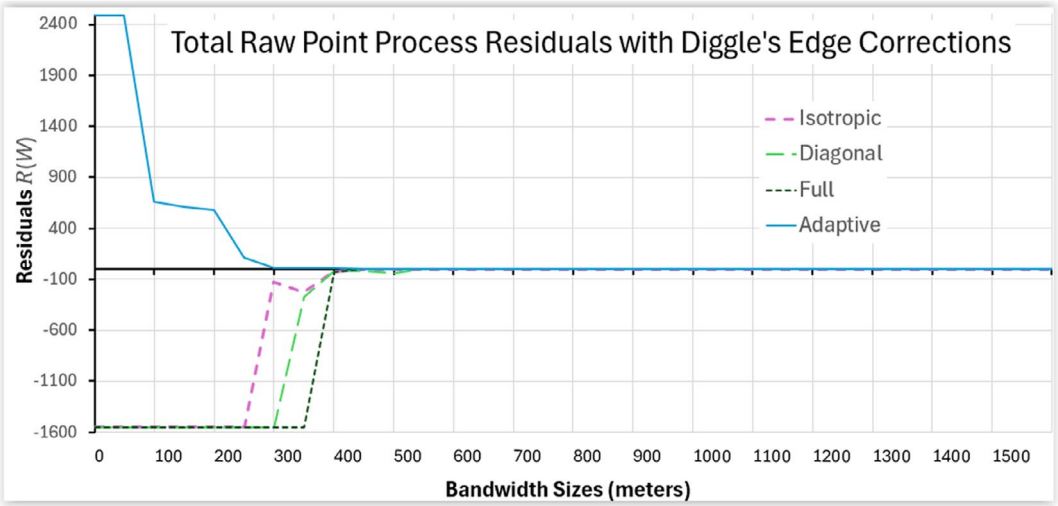

FIGURE 11 | The total raw point residuals $R(W)$, plotted against bandwidths with Diggle's adjustments for edge effects.

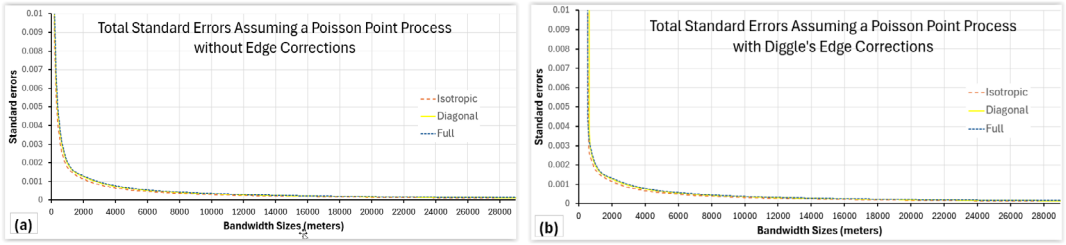

FIGURE 12 | The total standard errors for kernel estimates: (a) without edge corrections, and (b) with Diggle's edge-corrected adjustments. The graph for the Uniform edge-corrected adjustment closely resembles the uncorrected graph and is not shown in these plots.

residuals across the bandwidths up to 15,000 m. Additionally, the kernel for anisotropic diagonal bandwidths outperforms the kernel with full matrix bandwidths. However, once again, isotropic KD estimators with adaptive bandwidths show the weakest performance.

Figure 11 illustrates the total raw point residuals $R(W)$ against different bandwidths for various KD estimators with Diggle's edge correction. Diggle's scaling normalizes the integral of $\hat{f}_h^D(s)$ over the entire study area W, ensuring it is exactly equal to the observed number of points (Baddeley et al. 2016). The estimators seem unbiased within the bandwidth intervals > 400 m. However, this plot does not yield meaningful insights for comparing the performance of KD estimators.

Figure 12 shows the standard error $\mathrm{SE}_{\lambda(s,h)}$ (Equation 14) against different bandwidths for various kernel density estimators, including both uncorrected and edge-corrected estimates. These graphs provide limited information for selecting the best-performing kernel estimators.

Leave-one-out cross-validation is used to estimate the "predicted residual" with uncorrected, Uniform, and Diggle's edge-corrected adjustments. However, computation becomes very costly when processing large datasets. The total LOOCV errors for uncorrected estimations closely match the total raw point residuals, with the error curves being nearly identical.

Figure 13 shows the distribution of leave-one-out cross-validation $\mathrm{Err}_{\mathrm{LOOCV}}$ errors with uniform correction across various bandwidths, parameterization classes, and types of bandwidth variability.

There appear to be some variations between plots showing the LOOCV errors and the total raw point residuals; however, their relative positions remain consistent.

Experiments were conducted using k-fold cross-validation with independent random thinning, with K values of 2, 5, 10, and 100 folds. Figure 14 illustrates the distribution of total cross-validation error $\mathrm{Err}_{k-\mathrm{foldCV}}$ for $K = 10$, across various bandwidths, parameterization classes, and bandwidth variability types. In all cases, the total k-fold cross-validation errors are larger than the total raw point residuals $R(W)$, however, the shapes of the curves remain nearly identical (Figure 14).

Figure 15 illustrates an example of spatial leave-one-group-out validation. Space-filling curves were used to sort points into local neighborhoods, and the sorted list was subsequently divided into blocks, each containing 96 points at the next-to-lowest level. In this scheme, the algorithm first visits 96 locations within the next-to-lowest level neighborhood before moving on to the next neighborhood. This process generated 1132 blocks, which were then used to calculate the leave-one-block-out KDE errors.

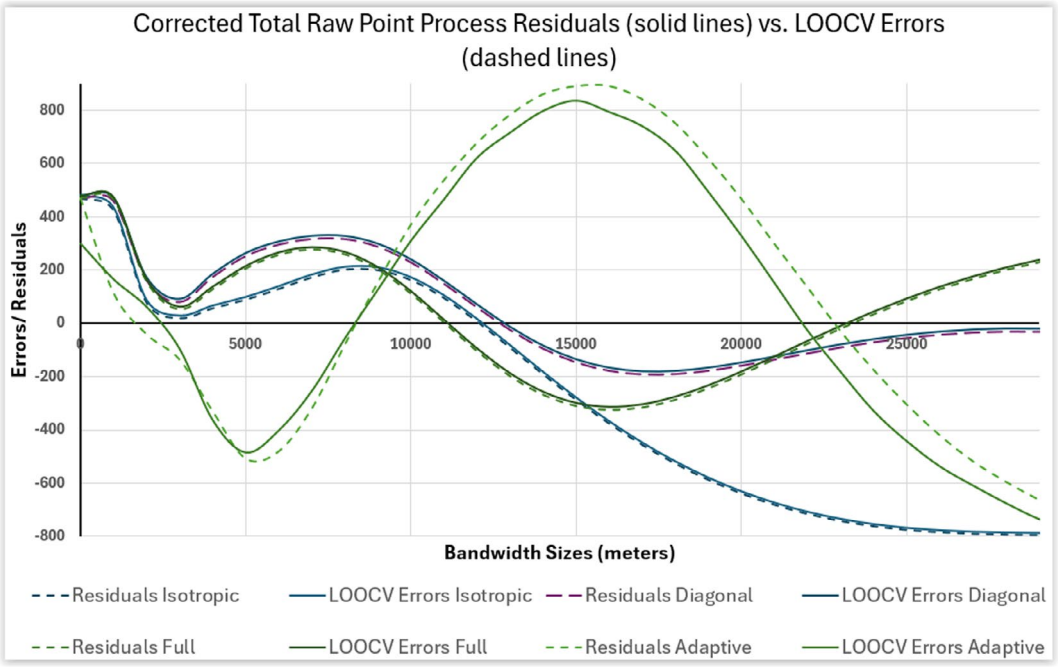

FIGURE 13 | Comparison of total point process residuals $R(W)$ (dushed lines) and total $\mathrm{Err}_{\mathrm{LOOCV}}$ errors (solid lines) for uniformly corrected estimations.

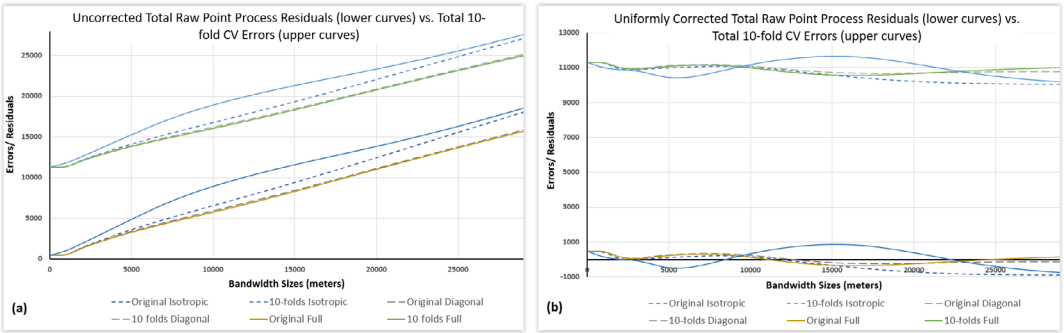

FIGURE 14 | Comparison between total point process residuals $R(W)$ (lower sets of darker color graphs) and total 10-fold CV $\mathrm{Err}_{k-\mathrm{foldCV}}$ errors (upper sets of lighter color graphs): (a) based on uncorrected estimations, (b) based on uniformly corrected estimations.

Once again, in most cases, the total $\mathrm{Err}_{\mathrm{SFC-BlockCV}}$ errors are larger than the total raw point residuals $R(W)$; however, while the shapes of the error curves remain similar, they exhibit more deviations compared to LOOCV and random k-fold cross-validation errors.

4.3.4 | Selecting the Best Estimation

Figure 8 shows that, based on the total raw point residuals $R(W)$, the best kernel density estimator is the one with an anisotropic full matrix parametrization, which effectively captures the anisotropy of smooth phenomena. Interestingly, fixed kernels outperformed the adaptive kernels in all tests. While adaptive kernels may provide the most accurate density estimates at the observation locations, fixed kernels deliver the best overall surface estimate.

Figure 9 shows that fixed bandwidth KDEs with uniform corrections exhibit the smallest errors with bandwidths between 2000 and 3000 m. This aligns well with the performance of kernels with bandwidths, estimated using bootstrap-based MISE and mixed calculations. Interestingly, the fixed isotropic kernel slightly outperforms the unconstrained kernel within this bandwidth range, while the unconstrained kernel still performs better than the diagonal kernel. The adaptive kernel performs the worst again.

For the fixed corrected kernels in the bandwidth interval of 10,000–13,000 m, which aligns with the optimal range according to most plug-in selectors (Table 1), the error curves show the second minimum in residuals. Within this bandwidth range, the fixed corrected unconstrained kernels achieve the shortest bandwidth with zero residuals. However, this clearly represents an over-smoothed solution.

Figures 12 and 13 show that the expected training error $E\left(\mathrm{Err}_{\mathrm{training}}\right)$ consistently deviates from the expected test error $E\left(\mathrm{Err}_{\mathrm{test}}\right)$ by a fixed margin. As a result, the expected training error $R(W)$ can serve as a reliable substitute for the expected test error $\mathrm{Err}_{\mathrm{LOOCV}}$, as discussed above.

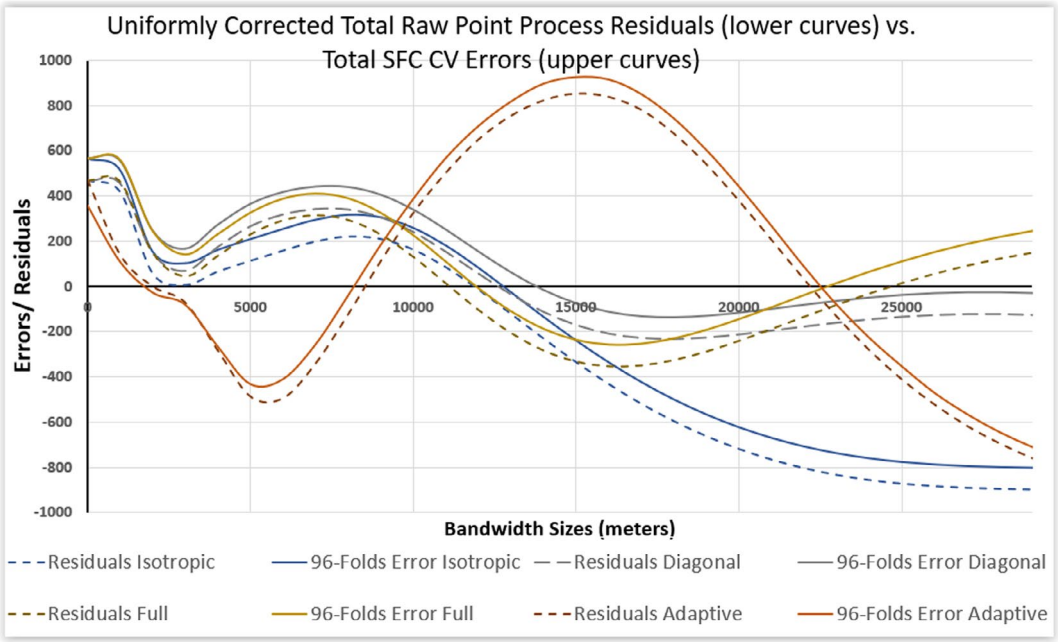

FIGURE 15 | Comparison of total point process residuals $R(W)$ (dashed lines) and total SFS-block CV $\mathrm{Err}_{\mathrm{SFC-BlockCV}}$ errors (solid lines), based on uniformly corrected estimations.

The comparison analysis above suggests that for the dataset used in this study, the unconstrained kernel density estimator with a bandwidth of approximately 2800 m will be appropriate.

4.4 | Software Used

The experiments were conducted using R Studio, along with packages spatstat, sparr, raster, sf, ks, np, kedd, KernSmooth, sm, cvTools, Ake, and cvTools, all of which are freely available from the CRAN website (cran.r-project.org). ESRI ArcGIS Pro was used for dataset preparation and pre-processing, point sorting with space-filling curves, and visualization. The R scripts used for this study is not included in the paper but is available upon request.

5 | Conclusion

GIS modeling extends beyond basic visualizations of point density and subjective assessments. When integrating Kernel Density Estimation (KDE) surfaces into the modeling process, it is essential to consider the errors associated with KDE in order to improve the overall accuracy of the model. Although many GIS software packages offer tools for kernel density estimation, these tools often do not provide sufficient capabilities to help users select appropriate settings for generating accurate kernel density surfaces and validating their accuracy. Typically, parameters such as bandwidth are estimated using rules of thumb, which do not allow users to effectively assess the accuracy of the resulting kernel density estimates.

The findings from the case study highlight that there is no single, definitive method for selecting the best kernel density estimator for a specific point dataset, especially when using current rules and "optimal" bandwidth selectors. In real-world scenarios involving large datasets with extreme value mixtures, it is common to observe multiple modes or centers of activity, each exhibiting varying heights, widths, and directions due to baseline effects. This observation is relevant not only to crime events but also extends to various phenomena, such as rare diseases, patterns of mobile phone calls, animal sightings, tree distributions in forests, and many other occurrences.

The choice of bandwidth selector is important, but it may not be the primary factor affecting evaluation errors. Instead, the selected parameterization class and the type of bandwidth variability can have a significant impact. Additionally, the choice of kernel function may also influence the validation results.

Selection of the optimal kernel density estimator and the corresponding parameters for a specific dataset and application can only be accomplished using the approaches described above, as there is no direct method to calculate MISE/AMISE/ISE for real-world datasets, particularly in anisotropic situations and adaptive KDE cases. The benchmark for this selection can be the internal and external error measures derived from the mass conservation property, which can be expressed in terms of the number of points (Equations 15–17, 21–22), the area of the observation window (Equations 18 and 20), or the area of the Thiessen polygons (Equations 23–24).

When large datasets are involved, such as the 108,670 crime event points used in this study, the total raw point residuals can be used for selection. Leave-one-out cross-validation errors are more suitable for small datasets. However, certain types of leave-k out cross-validation, where k increases with n, will remain consistent because as the dataset size grows, leaving out a proportional number of data points (with k increasing relative to n) helps ensure that the validation process becomes more stable and accurately reflects the data's underlying structure.

The primary objective of this study was to identify the optimal KDE surface using error measures. This selected surface will serve as an intermediate layer for spatial regression and classification in a neural network (Govorov et al. 2019). Validated Probability Density Function (PDF) surfaces can also serve as a component in more advanced semi-parametric and other models (Chacón and Duong 2018).

Kernel smoothing methods are effective tools for visualizing and understanding patterns in crime data. Crime risk maps can assist policymakers and law enforcement agencies in managing crime at both national and local levels. Additionally, these maps can help citizens evaluate the safety of their living environment and consider the crime context when choosing the location of their property.

Acknowledgments

Thanks are due to the Lithuanian Police for the provision of tabular data on reported crime.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study were provided by the Police of Lithuania. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors with the permission of the Police Department under the Ministry of the Interior of the Republic of Lithuania.

References

Abramson, I. S. 1982. "On Bandwidth Estimation in Kernel Estimates—A Square Root Law." *Annals of Statistics* 10, no. 4: 1217–1223.

Aitchison, J., and C. G. G. Aitken. 1976. "Multivariate Binary Discrimination by the Kernel Method." *Biometrika* 63, no. 3: 413–420.

Baddeley, A., J. Møller, and A. G. Pakes. 2008. "Properties of Residuals for Spatial Point Processes." *Annals of the Institute of Statistical Mathematics* 60: 627–649.

Baddeley, A., E. Rubak, and R. Turner. 2016. *Spatial Point Patterns: Methodology and Applications With R*. Chapman and Hall/CRC Press.

Baddeley, A., R. Turner, J. Møller, and M. Hazelton. 2005. "Residual Analysis for Spatial Point Processes." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 67: 617–666.

Beconytė, G., D. Vasiliauskas, and M. Govorov. 2020. "Lietuvos policijos 2015-2019 m. registruotų įvykių erdvinė sklaida ir dinamika [Spatial Distribution and Dynamics of Events Registered by Lithuanian Police in 2015–2019]." *Filosofija. Sociologija* 2: 175–185.

Belaid, N., S. Adjabi, C. C. Kokonendji, and N. Zougab. 2018. "Bayesian Adaptive Bandwidth Selector for Multivariate Discrete Kernel Estimator." *Communications in Statistics - Theory and Methods* 47, no. 12: 2988–3001. https://doi.org/10.1080/03610926.2017.1346807.

Berk, R., and J. M. MacDonald. 2008. "Overdispersion and Poisson Regression." *Journal of Quantitative Criminology* 24: 269–284.

Berman, M., and P. Diggle. 1989. "Estimating Weighted Integrals of the Second-Order Intensity of a Spatial Point Process." *Journal of the Royal Statistical Society, Series B* 51: 81–92.

Bithell, J. F. 1991. "Estimation of Relative Risk Functions." *Statistics in Medicine* 10: 1745–1751.

Bouezmarni, T., and J. V. K. Rombouts. 2010. "Nonparametric Density Estimation for Multivariate Bounded Data." *Journal of Statistical Planning and Inference* 140: 139–152.

Bowman, A. 1984. "An Alternative Method of Cross-Validation for the Smoothing of Kernel Density Estimates." *Biometrika* 71: 353–360.

Bowman, A. W., and A. Azzalini. 1997. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach With S-Plus Illustrations*. Oxford University Press.

Breiman, L., W. Meisel, and E. Purcell. 1977. "Variable Kernel Estimates of Multivariate Densities." *Technometrics* 19, no. 2: 135–144.

Cameron, A., and P. Trivedi. 2013. "Regression Analysis of Count Data." In *Econometric Society Monographs*, 2nd ed. Cambridge University Press. https://doi.org/10.1017/CBO9781139013567.

Ceccato, V. 2024. "Introduction to the Special Issue: Implementing Environmental Criminology for Crime Prevention." *Crime Prevention and Community Safety* 26: 133–138. https://doi.org/10.1057/s41300-024-00203-w.

Chacón, J. E., and T. Duong. 2018. "Multivariate Kernel Smoothing and Its Applications." In *Monographs on Statistics and Applied Probability*, vol. 160. CRC Press. https://doi.org/10.1201/9780429485572.

Chacón, J. E., T. Duong, and M. P. Wand. 2011. "Asymptotics for General Multivariate Kernel Density Derivative Estimators." *Statistica Sinica* 21: 807–840.

Chen, Y.-C. 2017. "A Tutorial on Kernel Density Estimation and Recent Advances." *Biostatistics & Epidemiology* 1, no. 1: 161–187. https://doi.org/10.1080/24709360.2017.1396742.

Chu, C.-Y., D. J. Henderson, and C. F. Parmeter. 2017. "On Discrete Epanechnikov Kernel Functions." *Computational Statistics and Data Analysis* 116: 79–105. https://doi.org/10.1016/j.csda.2017.07.003.

Cronie, O., M. Moradi, and C. A. Biscio. 2021. "Statistical Learning and Cross-Validation for Point Processes." *arXiv preprint arXiv:2103.01356*.

Cronie, O., and M. N. M. van Lieshout. 2018. "A Non-Model-Based Approach to Bandwidth Selection for Kernel Estimators of Spatial Intensity Functions." *Biometrika* 105, no. 2: 455–462.

Cwik, J., and J. Koronacki. 1997. "A Combined Adaptive-Mixtures/Plug-In Estimator of Multivariate Probability Densities." *Computational Statistics and Data Analysis* 26: 199–218.

Daley, D. J., and D. Vere-Jones. 2003. *Introduction to the Theory of Point Processes*. 2nd ed, Vol. I and II. Springer.

Davies, T. M., and A. Baddeley. 2018. "Fast Computation of Spatially Adaptive Kernel Estimates." *Statistics and Computing* 28, no. 4: 937–956.

Davies, T. M., K. Jones, and M. L. Hazelton. 2016. "Symmetric Adaptive Smoothing Regimens for Estimation of the Spatial Relative Risk Function." *Computational Statistics and Data Analysis* 101: 12–28.

Davies, T. M., and A. B. Lawson. 2019. "An Evaluation of Likelihood-Based Bandwidth Selectors for Spatial and Spatiotemporal Kernel Estimates." *Journal of Statistical Computation and Simulation* 89, no. 7: 1131–1152. https://doi.org/10.1080/00949655.2019.1575066.

Davies, T. M., J. C. Marshall, and M. L. Hazelton. 2018. "Tutorial on Kernel Estimation of Continuous Spatial and Spatiotemporal Relative Risk." *Statistics in Medicine* 37: 1191–1221. https://doi.org/10.1002/sim.7577.

Diggle, P. J. 1985. "A Kernel Method for Smoothing Point Process Data." *Applied Statistics* 34: 138–147.

Diggle, P. J. 2006. "Spatio-Temporal Point Processes: Methods and Applications." In *Statistical Methods for Spatio-Temporal Systems.*

Monographs on Statistics & Applied Probability, 1–45. Chapman and Hall/CRC.

Duong, T., and M. L. Hazelton. 2003. "Plug-In Bandwidth Matrices for Bivariate Kernel Density Estimation." *Journal of Nonparametric Statistics* 15: 17–30. https://doi.org/10.1080/10485250306039.

Eurostat. 2024. *Crime Statistics—Recorded Crime by Offence Category. Data Was Extracted in April 2024*. https://ec.europa.eu/eurostat/statistics-explained/index.php/Crime_statistics.

Fernando, W. T. P. S., and M. L. Hazelton. 2014. "Generalizing the Spatial Relative Risk Function." *Spatial and Spatio-temporal Epidemiology* 8: 1–10. https://doi.org/10.1016/j.sste.2013.12.002.

Gordon, J. S., R. A. Clements, F. P. Schoenberg, and D. Schorlemmer. 2015. "Voronoi Residuals and Other Residual Analyses Applied to CSEP Earthquake Forecasts." *Spatial Statistics* 14b: 133–150.

Govorov, M., G. Beconytė, G. Gienko, and V. Putrenko. 2019. "Spatially Constrained Regionalization With Multilayer Perceptron." *Transactions in GIS* 23: 1048–1077. https://doi.org/10.1111/tgis.12557.

Gramacki, A. 2017. *Nonparametric Kernel Density Estimation and Its Computational Aspects. Studies in Big Data*. Springer International Publishing.

Hall, P., and J. S. Marron. 1988. "Variable Window Width Kernel Density Estimates of Probability Densities." *Probability Theory and Related Fields* 80: 37–49.

Hansen, B. 2022. *Econometrics*, 1080. Princeton University Press.

Härdle, W., and M. Müller. 2000. "Multivariate and Semiparametric Kernel Regression." In *Smoothing and Regression*, 357–391. Wiley. https://doi.org/10.1002/9781118150658.ch12.

Hastie, T., R. Tibshirani, and J. Friedman. 2017. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. 2nd ed, 745. Springer.

Heidenreich, N., A. Schindler, and S. Sperlich. 2013. "Bandwidth Selection for Kernel Density Estimation: A Review of Fully Automatic Selectors." *AStA Advances in Statistical Analysis* 97: 403–433. https://doi.org/10.1007/s10182-013-0216-y.

Hilbe, J. 2014. *Modeling Count Data*. Cambridge University Press. https://doi.org/10.1017/CBO9781139236065.

Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan. 2008. *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470725160.

Jones, M. C. 1993. "Simple Boundary Correction for Kernel Density Estimation." *Statistics and Computing* 3: 135–146.

Karunamuni, R. J., and T. Alberts. 2005. "On Boundary Correction in Kernel Density Estimation." *Statistical Methodology* 2, no. 3: 191–212.

Kiessé, T. S. 2017. "On Finite Sample Properties of Nonparametric Discrete Asymmetric Kernel Estimators." *Statistics* 51, no. 5: 1046–1060. https://doi.org/10.1080/02331888.2017.1293060.

Kokonendji, C. C., and T. S. Kiessé. 2011. "Discrete Associated Kernels Method and Extensions." *Statistical Methodology* 8, no. 6: 497–516.

Kokonendji, C. C., and S. M. Somé. 2015. "On Multivariate Associated Kernels for Smoothing Some Density Function." *arXiv: 1502.01173*.

Lawson, A. 1993. "A Deviance Residual for Heterogeneous Spatial Poisson Processes." *Biometrics* 49: 889–897.

Le Rest, K., D. Pinaud, P. Monestiez, J. Chadoeuf, and V. Bretagnolle. 2014. "Spatial Leave-One-Out Cross-Validation for Variable Selection in the Presence of Spatial Autocorrelation." *Global Ecology and Biogeography* 23: 811–820. https://doi.org/10.1111/geb.12161.

Li, Q., and J. S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. Vol. 3, 1–88. Princeton University Press, Economics Books.

Loader, C. 1999. *Local Regression and Likelihood*. Springer.

Mammen, E., M. D. Martínez-Miranda, J. P. Nielsen, and S. Sperlich. 2011. "Do-Validation for Kernel Density Estimation." *Journal of the American Statistical Association* 106: 651–660.

Martinez-Camblor, P., and J. de Una-Alvarez. 2009. "Non-Parametric-Sample Tests: Density Functions vs Distribution Functions." *Computational Statistics and Data Analysis* 53, no. 9: 3344–3357.

Moon, K., and A. Hero. 2014. "Multivariate f-Divergence Estimation With Confidence." *Advances in Neural Information Processing Systems* 27: 2420–2428.

O'Sullivan, D., and D. Unwin. 2010. *Geographic Information Analysis*. Wiley.

Pedregosa, F., G. Varoquaux, A. Gramfort, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12, no. 85: 2825–2830.

Pohjankukka, J., T. Pahikkala, P. Nevalainen, and J. Heikkonen. 2017. "Estimating the Prediction Performance of Spatial Models via Spatial k-Fold Cross Validation." *International Journal of Geographical Information Science* 31: 2001–2019. https://doi.org/10.1080/13658816.2017.1346255.

Roberts, D. R., V. Bahn, S. Ciuti, et al. 2017. "Cross-Validation Strategies for Data With Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40: 913–929. https://doi.org/10.1111/ecog.02881.

Sagan, H. 1994. *Space-Filling Curves*. Springer-Verlag.

Sain, S. R., K. A. Baggerly, and D. W. Scott. 1994. "Cross-Validation of Multivariate Densities." *Journal of the American Statistical Association* 82: 1131–1146.

Scott, D. W. 1992. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley.

Scott, D. W. 2015. *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley and Sons.

Shao, J. 1997. "An Asymptotic Theory for Linear Model Selection." *Statistica Sinica* 7: 221–264.

Shimazaki, H., and S. Shinomoto. 2010. "Kernel Bandwidth Optimization in Spike Rate Estimation." *Journal of Computational Neuroscience* 29: 171–182. https://doi.org/10.1007/s10827-009-0180-4.

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

Stone, M. 1977. "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion." *Journal of the Royal Statistical Society: Series B: Methodological* 39, no. 1: 44–47.

Stoyan, D., and P. Grabarnik. 1991. "Second-Order Characteristics for Stochastic Structures Connected With Gibbs Point Processes." *Mathematische Nachrichten* 151: 95–100.

Stoyan, D., and H. Stoyan. 1995. *Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics*. John Wiley and Sons.

Terrell, G. R. 1990. "The Maximal Smoothing Principle in Density Estimation." *Journal of the American Statistical Association* 85: 470–477.

Valavi, R., J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. 2019. "blockCV: An R Package for Generating Spatially or Environmentally Separated Folds for k-Fold Cross-Validation of Species Distribution Models." *Methods in Ecology and Evolution* 10: 225–232.

van Lieshout, M. N. M. 2022. "Non-Parametric Adaptive Bandwidth Selection for Kernel Estimators of Spatial Intensity Functions." *Annals of the Institute of Statistical Mathematics* 76: 313–331.

Vapnik, V. 2013. *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Vasiliauskas, D., and G. Beconyte. 2016. "Cartography of Crime: Portrait of Metropolitan Vilnius." *Journal of Maps* 12, no. 5: 1236–1241. https://doi.org/10.1080/17445647.2015.1101404.

Wand, M. P., and M. C. Jones. 1995. *Kernel Smoothing*. Chapman & Hall/CRC.

Wang, M. C., and J. van Ryzin. 1981. "A Class of Smooth Estimators for Discrete Distribution." *Biometrika* 68, no. 1: 301–309. https://doi.org/10.1093/biomet/68.1.301.

Wang, Q. 2019. "Extrapolation-Based Bandwidth Selectors: A Review and Comparative Study With Discussion on Bivariate Applications." *International Statistical Review* 87, no. 1: 127–151. https://doi.org/10.1111/insr.12276.