Article

Trivariate Kernel Density Estimation of Spatiotemporal Crime Events with Case Study for Lithuania

Michael Govorov [1],*, Giedrė Beconytė [2],* and Gennady Gienko [3]

1 Department of Geography, Vancouver Island University, Nanaimo, BC V9R 5S5, Canada
2 Institute of Geosciences, Vilnius University, LT-10223 Vilnius, Lithuania
3 Department of Geomatics, University of Alaska Anchorage, Anchorage, AK 99508, USA; ggienko@alaska.edu
* Correspondence: michael.govorov@viu.ca (M.G.); giedre.beconyte@gf.vu.lt (G.B.)

Abstract: The paper presents the results of the investigation of the applicability of spatiotemporal kernel density estimation (KDE) methods for density mapping of violent crime in Lithuania. Spatiotemporal crime research helps to understand and control specific types of crime, thereby contributing to Sustainable Development Goals. The target dataset contained 135,989 records of the events registered by the police of Lithuania from 2015–2018 that were classified as violent. The research focused on choosing appropriate KDE functions and their parameters for modeling the spatiotemporal point pattern of this particular type of crime. The aim was to estimate density, mass, and intensity function(s) so that they can be used in further confirmatory spatial modeling. The application-driven objective was to obtain reliable and practically interpretable KDE surfaces of crime events. Several options for improving and extending the investigated KDE methods are demonstrated.

Keywords: crime events; spatial point pattern; probability mass and density functions; bandwidth selectors; relative risk estimator

1. Introduction

Kernel density estimation (KDE) is a multipurpose non-parametric technique. It is used to estimate the probability density function (PDF), 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 the exploratory and confirmatory analysis of spatial and temporal data and cartographic visualization. KDE-based maps are used in spatial criminology to understand the spatial structures of crime, their distribution, differences, and temporal trends. Being an instrument for targeted crime prevention, crime mapping is important for the achievement of the Sustainable Development Goals: "Peace, justice and strong institutions" and "Sustainable cities and communities". A detailed understanding of the density and intensity of crime in a territory helps to plan prevention measures properly and to use anti-crime resources more efficiently.

Kernel smoothing with continuous PDF is a well-established approach for the estimation of density and intensity of spatial surfaces based on a sampled dataset [1]. Discrete kernel estimations of PMF have been far less investigated, especially for multivariate spatiotemporal estimations [2].

In this paper, we demonstrate the application of multivariate KDE that combines two types of kernel functions: spherical multivariate kernel function and product kernel function. We implemented the KDE procedures to estimate PDF and PMF for a spatiotemporal dataset of violent crime events in Lithuania. Different KDE techniques have been tested on the spatial point pattern process data. Our approach allows increasing the quality of KDE techniques by considering the data-driven characteristics of a spatiotemporal dataset. The originality of the approach lies in:

- application of different methods for bandwidth selection for estimation of PDF/PMF kernels;

- application of a bivariate radial kernel function for estimation in the spatial dimensions;
- using a product kernel function to combine temporal dimension with spatial dimension;
- applying an unconstrained bandwidth matrix for spatial anisotropic smoothing in directions different to those of the coordinate axes.

2. Background and Related Works

Univariate KDE is known as Rosenblatt's (1956) and Parzen's (1962) window method and is used to estimate the underlying PDF of a sample dataset with no assumptions on the underlying parametric distribution of the dataset [3,4]. KDE determines the contribution of each data point to the density function and can be applied to data drawn from a complicated distribution. It has been demonstrated that univariate KDE works well for observation data with inhomogeneous dispersions and thus is suitable for spatial and spatiotemporal point pattern datasets with high heterogeneity and anisotropy [5].

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. A multivariate kernel can be obtained by two common techniques: by a derivation of univariate kernels or by using a spherical or radial-symmetric kernels with l^2 Euclidean vector norm (length of the vector) multivariate kernel function [3,4,6]. In the first case, kernels over multidimensional inputs can be constructed by multiplying or averaging different types of univariate kernels [7]. The product of kernels is more appropriate for bounded or partially bounded distributions without correlation between the components [8]. In the second case, radial-symmetric kernels are estimated from data within a sphere around a location or grid point [3]. 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 l^2 norm.

The variables can be partially bounded (e.g., non-negative), completely bounded (e.g., in the unit interval), or of discrete (count, categorical) data types. Count data can be unordered or ordered. The classical symmetric PDF estimators (e.g., Epanechnikov and Gaussian kernels) presume that the underlying data are naturally continuous, which is frequently not the case. Thus, a symmetric kernel may be not suitable for discrete bounded datasets and, instead, other types of kernels should be used [2,7,9].

The estimation of a discrete kernel of a PMF has been far less often applied than the estimation of a continuous symmetric kernel of a PDF. For example, Aitchison and Aitken's kernel [10] can be used for unordered discrete or categorical variables; Wang and van Ryzin's kernel [11] can be used for ordered discrete variables [7]. Discrete asymmetric kernel methods have been extensively investigated by Kiessé, Kokonendji, and Somé [2,8]; they have constructed discrete kernels from known discrete PMFs such as Poisson, binomial, and negative binomial [2,12].

In a multivariate setting, a joint density function can be defined for a combination of discrete (unordered and ordered) and continuous variables, i.e., for both quantitative and qualitative data. The method of estimating a joint PDF/PMF has been extended using generalized product kernels [7] presuming the absence of correlation in its multivariate components. Kokonendji and Somé [8] introduced spherical kernel estimators of unknown densities on partially or fully bounded supports with correlation structures.

There are three main parameters of KDE. Most researchers agree that the most important component of the KDE function is the bandwidth h. The KDE function itself is arguably acknowledged to be of secondary importance to the bandwidth [3,4]. The third aspect of KDE is a method of edge correction that minimizes boundary bias. The consequences of the boundary problem in multivariate KDE can be much more severe than in univariate KDE because the boundary region increases with the number of dimensions [13].

Bandwidth selection is crucial in both univariate and multivariate estimations of a kernel of PDF, PMF, relative risk, and regression functions. There can be strong contextual justifications for choosing a particular bandwidth size h. For example, temporal bandwidth of one month may be appropriate for the analysis of seasonal changes.

There are numerous studies on optimal histogram bandwidth selection for particular data-driven choices. A measure of distance between the true density f and its kernel density estimator $\hat{f}_h(\cdot)$ is used to evaluate the performance of the estimator $\hat{f}_h(\cdot)$. Common automated methods of bandwidth h selection are based on the optimization of mean integrated squared error (MISE), integrated squared error (ISE), or other similar distance/error measures. Different optimization measures lead to different speculations on which bandwidth h is optimal. Based on ISE and MISE measures, two classes of methods are distinguished: the cross-validation (CV) methods and the plug-in (PL) methods ([14] and references therein). Many versions of CV and PL (including the rule-of-thumb methods) and their hybrids are used for density estimation [3–5,15–17]. New techniques for estimating optimal bandwidths for multivariate kernel functions emerge, such as likelihood cross-validation, Bayesian approaches, bootstrapping, extrapolation-based, mixing bandwidth selectors, neural networks, and others ([5,18] and references therein) that can be classified as CV, PL, or hybrid.

3. Method: Crime Events as Spatial Point Process

Criminal events are registered with spatiotemporal coordinates and can be represented as a set of 3D event points. Such events can be modeled statistically as a space–time process of a random point pattern in a time sequence dimension and a finite two-dimensional coordinate space. If the temporal extent (duration) of each crime event is negligible as each random event is observed within a short time interval, then such a temporal model is called an event process. The spatial point process collected over an observation period is referred to as the summary process. Mathematic expressions of time-dimensional and space-dimensional point processes differ. The natural order is characteristic of the temporal dimension while it is absent in spatial dimensions. The methods of analysis of the spatial point process are well established [16,19,20]. There are different approaches to how to extend these methods for the analysis of spatiotemporal point process data.

One approach for modeling spatiotemporal point process data is adding marks (labels) to the spatial point events. The quantitative marks describe the time of occurrences [20]. Otherwise, the spatial location may be viewed as a component of a multidimensional mark for a temporal point process [19]. In both ways, the spatiotemporal point process can be formalized as an inhomogeneous process. Inhomogeneous point patterns are designed specifically for applications in which non-uniformity of space and/or time is important.

A spatiotemporal point process with time marks can be modeled mathematically in different ways [15,16]. One of the possible methods is to use an inhomogeneous Poisson point process for the disjoint interval counts that are stochastically independent. The first-moment measure of the Poisson spatial point process is the estimated intensity function, proportional to the density or mass functions (PDFs or/and PMFs) [21].

The inhomogeneous Poisson point process model is a constant risk model that does not presume dependencies between crime events. In the inhomogeneous Poisson point process model, each person faces the same risk of being affected by crime during the observation period, regardless of location. Thus, the occurrences of crime are not uniformly distributed over space; on the contrary, they are concentrated in populated places. More crime events can be expected in the areas with higher population densities, and clusters appear. It is assumed that such events are Poisson spatial point processes with intensity proportional to the population density. It is the model of constant crime risk.

In KDE, there are a variety of methods to adjust (by ratio, re-weight, and transform) kernel density for underlying covariates such as human population. The common method of adjusting an inhomogeneous background due to spatial variation in the human population density can be implemented as the spatial relative risk or density ratio function first proposed by Bithell [22]. The function is the crime risk function at locations (x, y) derived as the ratio of two KDE bivariate functions—crime events and the population at risk over the same region. Then, the ratio of the PDF/PMF of the two-point processes is an indication of the spatially varying risk of crime.

In this study, we concentrated on the choice of bandwidth and kernel function for a trivariate KDE applied to crime events with two spatial and one temporal coordinate. The problem of edge correction has been left for future consideration. In this study, we have applied boundary corrections implemented in KDE algorithms.

Bandwidth selection. The main issues are: (1) the *width* of spatial (h) and temporal (λ) bandwidth and (2) whether the bandwidths should be of the same size or variable and, if variable, then whether they should be adaptive. Estimating the bandwidth matrix for three-dimensional time–space is not a trivial task, because time and space have different natural ordering and, in the case of crime events, are heterogeneous. The same bandwidth cannot be applied to time and space coordinate directions.

Time resolution (time interval width λ) and spatial resolution (spatial bandwidth h) can be chosen independently from each other using spatial or temporal PDF or/and PMF, correspondingly. If there are no preferences for the spatial or time intervals, a spatiotemporal PDF or/and PMF [7] can be used to jointly estimate optimal space–time bandwidths as a product of multivariate kernels [4]. In criminology, it is often assumed that crime events obey the Poisson distribution or over-dispersed variant of the Poisson probability function such as a negative binomial [23]. In our experiments, the bandwidth was selected based on the assumption of a Poisson point process.

Choice of the kernel function. A kernel function can affect the quality of the bandwidth estimation [2]. Appropriate PMF kernels, e.g., negative binomial kernels, should be used for discrete data types. In the case of crime events, it must be considered that the data are discrete counts, ordered in the time dimension, and unordered in the spatial dimensions.

Several studies investigated bandwidth selection for the estimation of *bivariate* discrete PMF kernels ([2,7,24,25] and references therein). The first class includes Dirac-type symmetric kernels such as discrete triangular [24], Aitchison–Aitken [10], Wang–van Ryzin [11], and discrete Epanechnikov [25] kernel functions. The other class of kernels contains discrete asymmetric non-Dirac type kernels constructed from PDFs such as Poisson, binomial, and negative binomial [2]. A kernel estimator for count data is generally estimated as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} L(X_i, x, \lambda)$$

where $L(\cdot)$ is a discrete symmetric or asymmetric, Dirac or non-Dirac type kernel function appropriate for smoothing discrete data; X_i is the location of the univariate event and x is the location of the estimate; λ is the kernel bandwidth parameter. For example, the univariate discrete kernel function of ordered time variable t [11] can be defined as

$$L(T_i, t, \lambda) = \begin{cases} 1 - \lambda, & \text{if } T_i = t \\ \frac{(1-\lambda)}{2}\lambda^{|T_i - t|}, & \text{if } T_i \neq t \end{cases}$$

For discrete-ordered data, the MISE optimization technique of *univariate* time bandwidth selection was proposed by Shimazaki and Shinomoto [26]. It is based on the assumption that the event pattern is described by an inhomogeneous Poisson point process [15]. This optimization technique can be applied to any kernel function of the Dirac type. The results of applying this and other ordered discrete kernels techniques for bandwidth selections are presented in the Results section.

For a two-dimensional homogenous spatial point process, *bivariate* product kernels based on the product of two univariate kernels can be used under the assumption that there are no interactions between x and y coordinates of the observed crime event locations. In the case when a radial-symmetric kernel with the same bandwidth in x and y directions is used, the amount of smoothing is the same in each coordinate direction. Then, the

underlying bivariate PDF is approximated by the most common KDE function $\hat{f}(x, y)$ without edge correction at a location x, y [3]:

$$\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{d_{(x,y),i}}{h}\right)$$

where x and y are 2-dimensional (2D) spatial coordinates within a bounded study region; K is a 2D, second-order, zero-centered, radially symmetric continuous bimodal fixed kernel function; $d_{(x,y),i}$ is the Euclidean distance between event point i and location (x, y); h is the kernel bandwidth that is scaled equally in all directions or the radius of the circle for a circular kernel.

Adaptive KDE in location (x, y), which is referred to as the balloon estimator [17], is

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{K\left(\frac{d_{(x,y),i}}{h_i}\right)}{h_i^2}$$

This adaptation reduces the smoothing in areas of high event density (densely populated places) and increases the smoothing in areas where the events are relatively sparse (in rural areas).

The selection of bivariate bandwidths for a spatially inhomogeneous and anisotropic point process presents many challenges when it comes to selecting the correct amount of smoothing. There are several parametrization classes of bivariate bandwidth matrix [4,8] that can be considered if the spatial point process is inhomogeneous and anisotropic. Choosing a diagonal bandwidth matrix (independent bandwidths in x and y directions) will sometimes be adequate for an inhomogeneous process, however, for an isotropic process, a full bandwidth matrix to smooth in directions different from those of the coordinate axes [27] can result in better performance. A spatial bivariate kernel can be used to form a multivariate KDE as

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|H|} K\left(H^{-1} d_{(x,y),i}\right)$$

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

In the case of the count data of crime events, the spatial kernel can be obtained as a radial-symmetric PDF kernel or as a product of two unordered discrete bivariate PMF kernels. In the first case, a radial-symmetric kernel can be used to estimate a 2×2 full bandwidth matrix that considers both the axis-specific smoothing and the relative orientation of the kernel [3,4]. The full bandwidth matrices consider different correlation structures in the data sample [8]. In the second case, a product kernel assumes independence between the x and y coordinates, and the product kernel is defined by two bandwidths in the x and y directions. However, to consider anisotropy in diagonal directions, the data can be pre-rotated, and then a diagonal bandwidth matrix can be used for a product kernel [4]. A bivariate product of two univariate kernel estimators is implemented as

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \prod_{p=1}^{2} \frac{1}{h_p} K\left(\frac{d_{(x,y),i,p}}{h_p}\right)$$

where h_p is the bandwidth in dimension p and $d_{i,(x,y),p}$ is the Euclidean distance between event point i and location (x, y) in the dimension p.

The estimation of space–time bandwidths for crime point events can be based on a product of PDFs/PMFs. Multiplying two kernels, each of which depends only on a single input dimension, results in a prior probability distribution over functions that vary across both dimensions. Thus, a spatiotemporal trivariate function can be a product of the discrete-ordered (time) kernel and the second kernel that can be (*a*) a product of two spatial discrete unordered PMF kernels or (*b*) a radial-symmetric PDF kernel. Optimized

data-driven bandwidths, in both space and time dimensions, can be estimated from such a *trivariate* function [5,7,24].

The generalized trivariate product KDE for the spatiotemporal domain at x, y, t is defined as:

$$\hat{f}(x,y,t) = \frac{1}{nh^2\lambda}\sum_{i=1}^{n} K\left(\frac{d_{(x,y),i}}{h}\right)L\left(\frac{d_{t,i}}{\lambda}\right)$$

where K and L are kernel probability densities or/and mass functions; t is the time of occurrence of the event at location x, y; h and λ are spatial and temporal bandwidths. Jointly optimal bandwidth selection methods for spatiotemporal density functions can be naturally extended from the spatial-only setting [5].

The spatial relative risk function can be employed for handling the inhomogeneity in the distribution of the data. If both the kernel bivariate densities of crime events $K_e(\cdot)$ and the population at risk $K_p(\cdot)$ are estimated through their KDE processes, then the joint spatial relative risk function $\hat{r}(x,y)$ can be expressed as the ratio of densities describing, respectively, the spatial distribution of crime events and the controls (population at risk background) as:

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

where h_e and h_p are the bandwidths for both kernel functions. Application of this function to spatiotemporal crime events and population data is based on the assumption that the population density remains unchanged over the observational period that is within the 10-year census interval. Then, the trivariate relative risk KD estimator with a constant denominator becomes

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

where T is the length of the period under consideration [18]. Estimation of individual jointly optimal bandwidths h_e, h_p, and λ for a spatiotemporal relative risk function is complicated, especially with time-static control densities [5]. According to [28], there are benefits of using a common jointly optimal spatial bandwidth $h_e = h_p$ for both events and background controls. For locally adaptive spatial kernel density, the use of a single bandwidth function for both events and controls is recommended [18,28]. An example of such a function is

$$h_i(x,y) = h_0\alpha_i(x,y)$$

where h_0 is the global bandwidth and $\alpha_i(x,y)$ is the i th local bandwidth factor (1, 1982). Bandwidth selection methods for standalone spatiotemporal density functions can be extended from the spatial-only setting, but data-driven, to jointly optimize bandwidth selection for a spatiotemporal relative risk estimate which is not straightforwardly adaptable [5].

4. Case Study: Crime of Violence in Lithuania from 2015–2018

The above-described KDE techniques have been tested with the spatiotemporal dataset of the crime events in Lithuania. The entire geocoded dataset of criminal events registered by Lithuanian police from 2015–2018 contains 2.78 million records, out of which 1.36 million belong to groups that have been in previous studies recognized as strongly dependent on spatial factors (socio-demographic patterns of population, patterns of infrastructure, urban and landscape structures): crime of violence, theft, destruction of or damage to property, drug-related crime, and public nuisance [9].

The reason for studying these data is the crime statistics in Lithuania, which are relatively extensive compared to the European Union, especially in particular cities. It is often argued that whereas international institutions use generalized official statistics and present Lithuania as a dangerous country, the real picture is different: the majority of crime in open space is thefts and public nuisance; homicides are mostly committed without aforethought malice, due to excessive consumption of alcohol or, more rarely, drugs. Thus, we selected the records that represent the most serious crimes only—crimes of violence that encompass assault, physical abuse (including sexual abuse), threatening behavior, home invasion, murder, and manslaughter; a total of 135,989 records, with the yearly average of 34,000. The violent crime rate significantly differs in densely (more than 100 people per square kilometer, 70.5% of all events) and sparsely (less than 100 people per square kilometer, 29.5%) populated areas: relative averages per 1000 population are 10.7 and 13.1, correspondingly [29].

Common kernel density maps show the peculiarities of the spatial distribution over the territory; however, they are often compiled without paying due attention to the actual spatial and temporal distribution of data, kernel function, and bandwidth. This study targets these aspects and is aimed at optimal density estimation for this part of crime data that is most sensitive.

As was expected, spatial exploratory analysis (p-values of Pearson chi-squared, likelihood ratio, and Freeman–Tukey tests of homogeneity using the quadrat counts method) led to the rejection of the null hypothesis of complete spatial randomness (CSR) for the crime events. The alternative hypothesis is that the process is an inhomogeneous Poisson process. The likelihood ratio test for a Poisson point process model with a covariate effect [15] showed that population density significantly ($p < 0.01$) contributes to the distribution of crime.

A series of experiments on kernel estimations for the crime count data were performed. The experiments were conducted in four steps: (1) bandwidths were calculated; (2) kernel functions were derived based on the estimated bandwidths; (3) the kernel surfaces were calculated and plotted as density maps; (4) point process residual R(B) was calculated to compare the results.

Experiments with bandwidth. The most popular automatic CV, PL, and hybrid bandwidth selection methods were tested. The set of tested methods included the methods for estimations of univariate, bivariate, and multivariate bandwidths in the spatiotemporal domain; isotropic bandwidth, diagonal bandwidth, and full/unconstrained bandwidth matrix; fixed, adaptive/variable, and mixed bandwidths. In Table 1, the most representative results of bandwidth estimations for the target dataset are presented.

Selecting an optimal amount of smoothing is cast into a formal mathematical framework as an example of a bias–variance trade-off. The choice of a particular optimal bandwidth is related to the data sample size and the complexity of the data distribution. Several bandwidth selection methods, mainly CV-based (SCV and LCV unconstrained selector) methods, yielded completely unsatisfactory results and are not included in Table 1. The failure can be due to (a) discretization effects and data rounding of 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, multimodalities, and asymmetries.

Table 1. Experimental results of bandwidth selection.

Bandwidth Selectors	Univariate/ Isotropic		Anisotropic/Diagonal, Meters	Full Matrix, Meters
	Spatial Fixed or Adaptive (Interval), Meters	Time, Days		
Cross-validation (CV)				
Least squares CV (LSCV) for bivariate, edge-corrected KDE [30], "sparr" package	195			
Likelihood CV (LCV) for bivariate, edge-corrected KDE [30], "sparr" package	315			
Likelihood CV (LCV) [31], "spatstat" package	630			
Biased CV (BSV) for bivariate data [32], "ks" package			11,769 8187	12,706 7599 7599 9056
Smoothed CV (SCV) [33,34], "ks" package		58.9	4207 3169	4651 −2785 −2785 3541
Abramson–Hall–Marron [35] rule's adaptive selector over global bandwidth from LCV, "spatstat" package	205–3150			
CV for spatial smoothing of marks [30], "spatstat" package		19.5		
Least squares CV (LSCV) derived from a single value [36], "sm" package	5877		7312 5212	
Plug-in (PL)				
Normal scale selector by Silverman rule-of-thumb [3], 'sparr' and "sm" packages	10,574	35.5	12,348 8801	
Normal scale selector over product kernel with the Silverman rule-of-thumb [7], "np" package			13,080 9322	
Direct Sheathe and Jones' rule-of-thumb at level 2, 2-dimensional data [4], "ks" package		15.0	1182 843	1408 −842 −842 1004
Direct Sheathe and Jones' rule-of-thumb at level 2, 3-dimensional data [27], "ks" package		49.8	4007 2934	4125 −2463 28 −2463 3132 −14 28 −14 53
Normal mixture with four mixture components [37] "ks" package			12,166 8404	12,368 −7338 −7338 8765
Normal scale [33], "ks" package			11,397 8123	12,348 −7385 −7385 8801
Bivariate by Scott rule-of-thumb [17], "spatstat" package	10,425		12,347 8802	
Abramson–Hall–Marron [35] (1988) rule's adaptive selector over global bandwidth from Scott's rule, "spatstat" package	3248–50,000			
Oversmoothing Terrell "rule-of-thumb" [5], "sparr" package	11,469	35.5		
Cronie and van Lieshout's criterion [38], "spatstat" package	10,845			
Shimazaki and Shinomoto [26] PL method		29.2		
Mixed [39]				
$\alpha = \beta = 1$	2500			
$\alpha = 2, \beta = 1$	1500			
$\alpha = 1, \beta = 2$	3900			

The results confirm the expectation that fixed CV-based methods (especially ordinary LSCV) would produce very small undersmooth isotropic bandwidths. Most of the tested selectors estimate bandwidths to be less than 1000 m. Thus, CV methods lead to a small bias but a large variance—they are not appropriate for data with high variability. The LSCV selectors do not perform well for large samples either [14]. Modified CV methods such as SCV and BCV account for much less variation without increasing too much in bias, in our case, they produced more realistic bandwidths.

The PL selectors always have a smaller (asymptotic) variance compared to CV methods but often a larger bias. The PL estimates are more stable and generate oversmooth bandwidths. There is an assumption that practically no other bandwidth selector has to date outperformed the asymptotic properties of the sophisticated PL methods [14]. From our tests, most PL selectors suggest spatial bandwidths of around 10,000 m. Only Sheathe and Jones' rule-of-thumb generates relatively small bandwidths, however, there are no objective methods to choose the arbitrary level parameter.

These findings suggest investigating mixing methods that combine different bandwidths and/or KDE estimators. Bandwidth mixtures might produce stable results. The mixture can be carried out by different methods, one of the simplest being using the different CV and PL bandwidth proportions on a logarithmic multiplicative scale described in [14]. Even in practical settings, a simple average of CV and PL bandwidths may perform much better than its two alternatives [39]. The combination of CV and PL is a compromise giving biases lying between the ISE and the MISE minimizing methods. The following formula can be used to calculate three types of CV and PL mixture bandwidths:

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

where the three possible combinations are $\alpha = \beta = 1$, $\alpha = 1, \beta = 2$, and $\alpha = 1, \beta = 2$. In Table 1, combinations of $\hat{h}_{PL} = 10,000$ and $\hat{h}_{CV} = 600$ are used.

PDF/PMF estimation and mapping: A set of rounded fixed (10,000, 5000, 2000, and 600 m) and adaptive bandwidths and three kernel functions (Gaussian, Epanechnikov, and quartic) were tested to generate kernel estimators. The study resulted in a series of maps outlining kernel density regions of violent crime events. The outcomes of violent crime KDEs with classic bivariate radial-symmetric kernels with isotropic and anisotropic fixed and isotropic adaptive bandwidths are shown in Table 2.

Table 2. Bivariate radial-symmetric kernels with isotropic and anisotropic fixed and isotropic adaptive bandwidths.
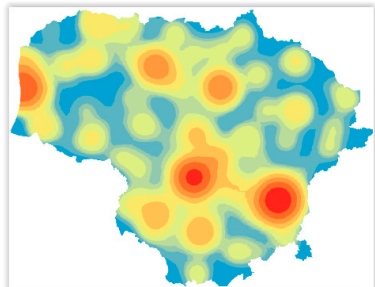
KDE Estimation Method	Bandwidth, Meters	Point Process Residual $R(B)$ of 135,984 Events	Density Map
Spatial Bivariate Gaussian Isotropic	10,000 (fixed)	8284 (6.09%)	

Table 2. *Cont.*

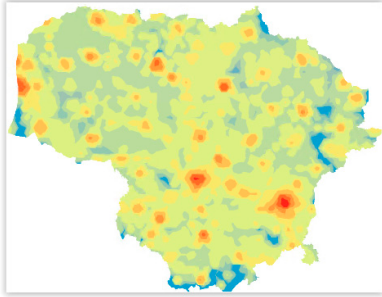| KDE Estimation Method	Bandwidth, Meters	Point Process Residual $R(B)$ of 135,984 Events	Density Map
Spatial Bivariate Gaussian Isotropic	5000 (fixed)	2806 (2.06%)	
Spatial Bivariate Epanechnikov Isotropic	5000 (fixed)	5771 (4.24%)	
Spatial Bivariate Gaussian Isotropic	2000 (fixed)	2362 (1.74%)	
Spatial Bivariate Gaussian Diagonal Anisotropic	12,000 8500 (diagonal matrix)	8590 (6.32%)	
Spatial Bivariate Gaussian Diagonal Anisotropic	8000 5700 (diagonal matrix)	6918 (5.09%)	

Table 2. *Cont.*
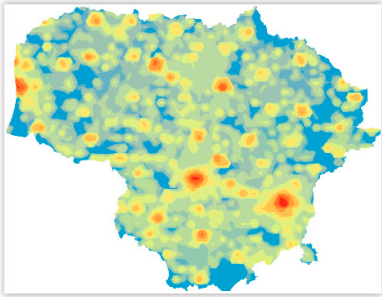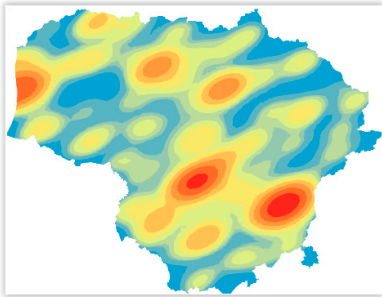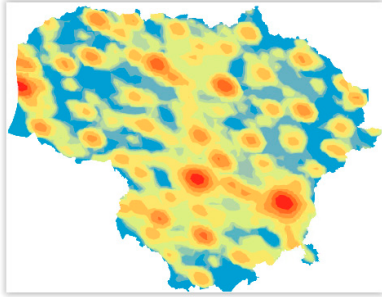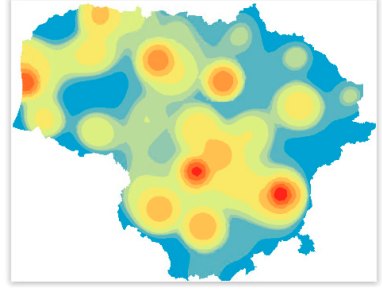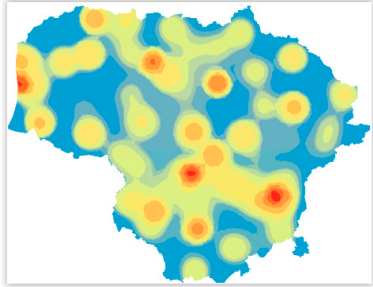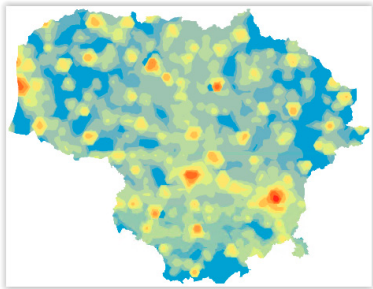
KDE Estimation Method	Bandwidth, Meters	Point Process Residual $R(B)$ of 135,984 Events	Density Map
Spatial Bivariate Gaussian Diagonal Anisotropic	2000 1400 (diagonal matrix)	2254 (1.66%)	
Spatial Bivariate Gaussian Unconstrained Anisotropic BSV	12,706 7599 7599 9056 (full matrix)	9074 (6.67%)	
Spatial Bivariate Gaussian Unconstrained Anisotropic SCV	4651 −2785 −2785 3541 (full matrix)	4958 (3.65%)	
Spatial Bivariate Gaussian Unconstrained Anisotropic SJ	1408 −842 −842 1004 (full matrix)	1219 (0.90%)	
Isotropic Spatial Gaussian Adaptive Abramson Scott's Rule	10,000 (global h_0) 5500 (pilot h_{pl})	11,033 (8.11%)	

Table 2. *Cont.*

KDE Estimation Method	Bandwidth, Meters	Point Process Residual $R(B)$ of 135,984 Events	Density Map
Isotropic Spatial Gaussian Adaptive Abramson LSCV	5500 (global h_0) 5500 (pilot h_{pl})	6833 (5.02%)	
Isotropic Spatial Gaussian Adaptive Abramson 3D FFT Multiscale Estimation [30]	600 (global h_0) 600 (pilot h_{pl})	104 (0.08%)	

Three sets of bandwidths, estimated with PL, CV, and mixture selectors, were tested to create KDE surfaces. As expected, surfaces with smaller bandwidths produced smaller biases. At the same bandwidth scale, the best estimates are produced with the full matrix/unconstrained bandwidths and adaptive bandwidths.

The KDE surfaces with unconstrained bandwidth selectors work better than their isotropic and diagonal counterparts as the crime point patterns have mass oriented obliquely to the coordinate axes. Isotropic KDE surfaces with adaptive bandwidth also yield improvements compared to fixed bandwidth isotropic surfaces as, in practice, the smoothing regimen rule of Abramson [35] is well-suited to spatial data, which often exhibit marked (in this case population density) heterogeneity [30]. At present, using unconstrained bandwidth matrices in adaptive settings remains a challenge.

Validation of estimates: Several types of errors can be considered for bandwidth selection and measurement of KDE performance [4,40]: pointwise error (e.g., the mean squared error (MSE)), uniform error (e.g., integrated squared error (ISE), mean integrated square error (MISE) and its approximation asymptotic mean integrated square error (AMISE)), the integrated standard error for the intensity function, etc.

In this case, the accuracy of the KDE approximations cannot be directly measured by ISE, as the true density functions are complex and unknown. The MISE/AMISE is generally used in bandwidth selection to measure the overall performance of a bandwidth selector for a particular kernel function and PDF. There are different realizations of AMISE, which are difficult to compare. If the model is an inhomogeneous point Poisson process, to compare performances of the above KDE surfaces, the point process residuals for the study area B are used. The point process residual $R(B)$ for the study region B is defined as the observed minus the expected number of points falling in the study area as per:
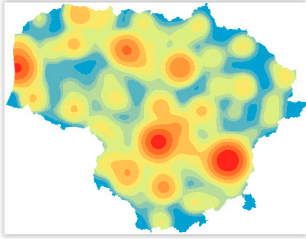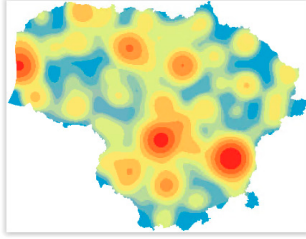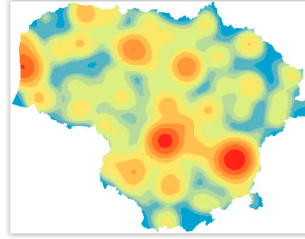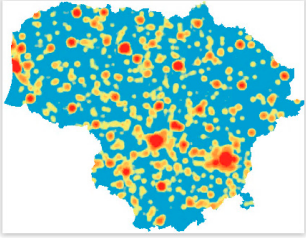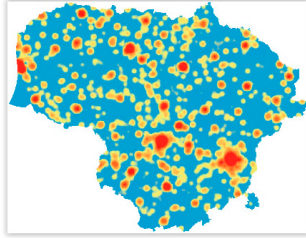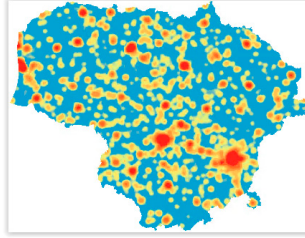
$$R(B) = n(A \cap B) - \iint_B \hat{\lambda}(u,v)\mathrm{d}u\mathrm{d}v$$

where $A = \{x, y\}$ is the set of points (x, y) in two-dimensional space $\mathbb{R}^2$—the observed point pattern; $n((x, y) \cap B)$ is the number of points of (x, y) in the region B; and $\hat{\lambda}(u, v)$

is the intensity of the fitted KDE in any spatial location. The probabilities can be converted into intensities by multiplying the probabilities in any one cell by the total number of incidents.

The values of the point process residual $R(B)$ are shown in the second column of Table 3. Residuals can be considered as measures of bias between the observed and expected (recalculated from the estimated intensity $\hat{\lambda}(u,v)$ within B) point patterns. Clear winners from all surfaces by the $R(B)$ values are adaptive and full matrix anisotropic undersmooth estimates.

Table 3. Spatiotemporal slices of the joint density at a given space–time location $(x; t)$ with isotropic spatial bandwidths and temporal bandwidths.

Dates	1 January 2018	1 March 2018	17 June 2018
Number of crime events	324	37	172
Spatiotemporal fixed bandwidth isotropic slice, $h = 10,000, \lambda = 30$			
Spatiotemporal isotropic slice, $h = 2000, \lambda = 15$			

The choice of the kernel function type has not much impact on the visual appearance of a KDE surface. However, with the same other parameters, Gaussian density produces better results in terms of point process residual $R(B)$.

The Gaussian kernel is one of the smoothest possible KDEs and so the optimal value of h is usually large. If this is then applied to non-normal data, it tends to induce oversmoothing.

Spatiotemporal estimations: Several methods were used to estimate the fixed temporal bandwidth (see Table 1). The most prominent outcomes are: (1) using Sheathe and Jones' rule-of-thumb resulted in temporal bandwidth of 15 days, (2) the bandwidth for smoothing of the temporal margin computed using the Shimazaki and Shinomoto [26] PL method with the Poisson assumption is 30 days, and (3) Silverman [3] and oversmoothing "rule-of-thumb" methods produced temporal bandwidth of 35 days. Temporal slices of trivariate density surfaces show seasonal changes in the distribution of violent crime. As an example, trivariate joint density surfaces for the days with the lowest, medium, and the highest total number of crime events in 2018 are presented in Table 3. These surfaces were calculated using Fernando and Hazelton's [18] function which is the product of spatial and temporal KDEs. Isotropic smoothing with fixed bandwidth in the spatial margin was used. The graph in Figure 1 illustrates the periodical character of biweekly crime event counts.
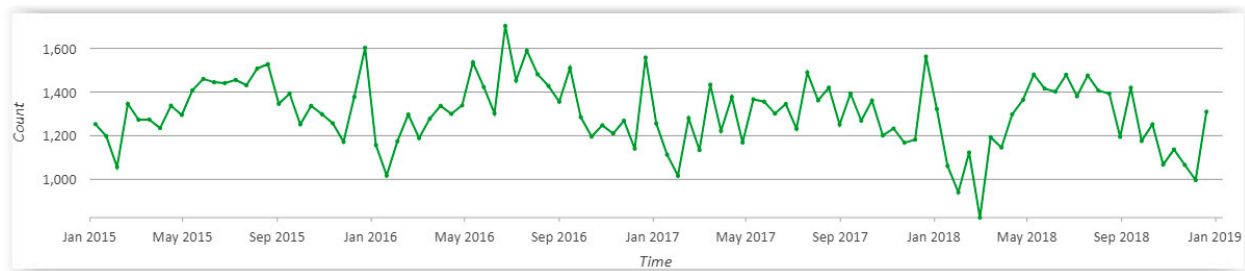
Figure 1. Variation of total violent crime count over time.

Relative risk estimation: A population point pattern was generated inside a 1 sq. km polygon grid that covers the entire territory of the country. Points were placed randomly within the grid cells. The number of points was generated proportionally to the population within each grid cell.

Jointly optimal fixed and adaptive bandwidths for the spatial relative risk function were calculated by using the LSCV selector. The common jointly optimal fixed spatial bandwidths for the Gaussian KDE relative risk function result approximately at 2375, 850, and 1400 m by using selectors based on MISE, a weighted MISE, and AMISE, respectively.

The relative risk density surfaces are presented in Table 4. Areas of average risk density where $\hat{r}(x,y) \cong 0$ and $\hat{f}_c(x,y) \cong \hat{f}_p(x,y))$ are represented in yellow. Peaks (red) in the surface $\hat{r}(x,y) > 0$ show a higher localized concentration of crime relative to the population density in the suburbs, while depressions (blue) $\hat{r}(x,y) < 0$ indicate territories with relatively low crime rates.

Table 4. Estimated log-transformed relative risk surfaces for relative crime/population density.

Jointly Optimal Bandwidth for Gaussian KDE Relative Risk Function	Estimated Log-Transformed Relative Risk Surface	Regions with Significantly Increased Crime Risk
Fixed smoothing. Jointly optimal global bandwidth at 2000 m.		
Symmetric pooled adaptive smoothing. Jointly optimal global and pilot bandwidths at 2000 m. Adaptive bandwidths in the interval of 500–1000 m.		

The third column of Table 4 shows asymptotic *p*-value surfaces with tolerance areas (in blue) from an upper-tailed test at a significant 5% threshold of heightened risk [1]. Tolerance areas represent regions of potentially anomalous activity with significantly increased crime risk relative to the background population density.

The experiments have been conducted using *R Studio* with the contributed packages *sparr*, *spatstat*, *ks*, *np*, *kedd*, *KernSmooth*, *sm*, and *Ake*, and QGIS plugins. ESRI ArcGIS

software has been used for dataset preparation and visualization purposes. R scripts are available upon request.

5. Discussion and Conclusions

Detailed analysis of the distribution of crime in Lithuania [29,41] allowed asserting that crime patterns are relatively homogeneous within a 500 to 2000 m radius and very strongly correlated with population density, especially in densely populated areas. Variations of violent crime are inhomogeneously distributed and usually have seasonal character. Thus, the bandwidth estimation methods for KDE that have yielded a spatial bandwidth smaller than 5000 m and temporal bandwidth of 15 to 60 days appear adequate and applicable. The likelihood CV fixed bandwidth selector can be considered the best choice. Of the tested KDE functions with a 2000 m bandwidth, *adaptive* and *unconstrained* estimators yielded the smallest point process residual.

The KDE of violent crime events in Lithuania of 2015–2018 allowed for outlining the regions with different intensities of crime. Naturally, the clusters appear in densely populated areas, but different levels of smoothing allow for representations of a crime surface at different scales. For a better understanding of the situation, crime events are modeled as an inhomogeneous Poisson point process with population counts as covariates. In such a setup, the KDE relative risk surfaces are estimated, and it becomes possible to highlight the regions with heightened and reduced risk of crime on the population background.

The outcomes—PMF surfaces—can be used in different applications beyond the simple visualization of the estimated density. Multivariate analysis methods, including kernel regression, kernel discriminant analysis, trend dynamic estimation, density level-set estimation, unsupervised and supervised learning, and density ridge estimation for PCA [33,42], can be applied to refine the outcomes. PDF surfaces can be a building block for more complicated semiparametric and other models.

Crime analysis usually involves big data. In the presented study, 135,984 crime event points and 586,952 population points (one point represents 5 people) were analyzed. In such settings, KDE estimations are very computationally expensive, especially for selecting optimal bandwidths. Memory issues occurred for bandwidth estimations using bootstrapping and some cross-validation methods in adaptive, mixed data type, and spatiotemporal settings.

Sometimes it is possible to determine bandwidth subjectively (usually one-dimensional and fixed isotropic bandwidth settings). In 2–3D, adaptive, and anisotropic situations it is very beneficial to have the bandwidth automatically selected from the data even with some prior knowledge about the structure of the data.

It is known that KDE performs poorly for high-dimensional data ($d > 3$), but even in 3D, many bandwidth selectors are unstable and produce unrealistic outputs. Different optimization measures lead to different definitions of optimal bandwidth h or/and λ. If the dataset is very large and inhomogeneous, CV selectors might fail. There are no exact rules on how to choose some parameters of the selectors, e.g., pilot bandwidths for adaptive estimation often have arbitrary specifications and might have an impact on KDE estimations. Thus, there is no universally reliable procedure to select the optimal bandwidth and only an expert in the field might confirm the validity of the result.

A key finding of this study is that it definitely makes a difference which bandwidth selector is chosen; not only in numerical terms but also for the quality of density estimation for which this bandwidth is used. Different bandwidth selectors have generated very different results that are difficult to validate quantitatively. Especially challenging is the validation of the methods that use unconstrained bandwidth matrices, that cannot be assessed by an expert.

Enhanced bandwidth selectors can significantly improve the residual $R(B)$ in the point process. The general conclusion from these experiments is that adaptive bandwidths work better for a large dataset with complex spatial patterns. An unconstrained bandwidth

matrix is useful when a large probability mass is oriented away from the coordinate directions, as in the case of our dataset. The use of adaptive bandwidth selectors with an unconstrained bandwidth matrix would potentially be the best choice.

The developed methodology, selected bandwidths, and generated PDF surfaces will be used for further analysis to investigate correlations between different types of crime events, specific covariate effects on the process of crime pattern distribution, regionalization of multitype spatial point pattern by using neural networks [43], and evaluation of trend dynamics of crimes. These methods and the maps they produce enhance research and application of crime analysis and public safety.

Author Contributions: Conceptualization, G.B., M.G. and G.G.; Writing—original draft, M.G. and G.B.; Writing—review & editing, G.G. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Sample datasets are available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Davies, T.M.; Lawson, A.B. An evaluation of likelihood-based bandwidth selectors for spatial and spatiotemporal kernel estimates. *J. Stat. Comput. Simul.* **2019**, *89*, 1131–1152. [CrossRef]
2. Kiessé, T.S. On finite sample properties of nonparametric discrete asymmetric kernel estimators. *Statistics* **2017**, *51*, 1046–1060. [CrossRef]
3. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman & Hall: London, UK, 1986.
4. Wand, M.P.; Jones, M.C. *Kernel Smoothing*; Chapman & Hall/CRC: Boca Raton, FL, USA, 1995.
5. Davies, T.M.; Marshall, J.C.; Hazelton, M.L. Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk. *Stat. Med.* **2018**, *37*, 1191–1221. [CrossRef] [PubMed]
6. Härdle, W.; Müller, M. Multivariate and semiparametric Kernel regression. In *Smoothing and Regression*; Wiley: New York, NY, USA, 2000; pp. 357–391. [CrossRef]
7. Li, Q.; Racine, J.S. *Nonparametric Econometrics: Theory and Practice*; Princeton University Press: Princeton, NJ, USA, 2007.
8. Kokonendji, C.C.; Somé, S.M. On multivariate associated kernels for smoothing some density function. *arXiv* **2015**, arXiv:1502.01173.
9. Vasiliauskas, D.; Beconyte, G. Cartography of crime: Portrait of metropolitan Vilnius. *J. Maps* **2016**, *12*, 1236–1241. [CrossRef]
10. Aitchison, J.; Aitken, C.G.G. Multivariate binary discrimination by the kernel method. *Biometrika* **1976**, *63*, 413–420. [CrossRef]
11. Wang, M.C.; van Ryzin, J. A class of smooth estimators for discrete distribution. *Biometrika* **1981**, *68*, 301–309. [CrossRef]
12. Kokonendji, C.C.; Kiessé, T.S. Discrete associated kernels method and extensions. *Stat. Methodol.* **2011**, *8*, 497–516. [CrossRef]
13. Bouezmarni, T.; Rombouts, J.V.K. Nonparametric density estimation for multivariate bounded data. *J. Stat. Plan. Inference* **2010**, *140*, 139–152. [CrossRef]
14. Heidenreich, N.; Schindler, A.; Sperlich, S. Bandwidth selection for kernel density estimation: A review of fully automatic selectors. *AStA Adv. Stat. Anal.* **2013**, *97*, 403–433. [CrossRef]
15. Baddeley, A.; Rubak, E.; Turner, R. *Spatial Point Patterns: Methodology and Applications with R*; Chapman and Hall/CRC Press: London, UK, 2016.
16. Illian, J.; Penttinen, A.; Stoyan, H.; Stoyan, D. *Statistical Analysis and Modelling of Spatial Point Patterns*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2008. [CrossRef]
17. Scott, D.W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley and Sons: Hoboken, NJ, USA, 2015.
18. Fernando, W.T.P.S.; Hazelton, M.L. Generalizing the spatial relative risk function. *Spat. Spatio-Temporal Epidemiol.* **2014**, *8*, 1–10. [CrossRef] [PubMed]
19. Daley, D.J.; Vere-Jones, D. *Introduction to the Theory of Point Processes*, 2nd ed.; Springer: New York, NY, USA, 2003; Volumes I and II.
20. Diggle, P.J. Spatio-temporal point processes: Methods and applications. In *Statistical Methods for Spatio-Temporal Systems. Monographs on Statistics & Applied Probability*; Chapman and Hall/CRC: London, UK, 2006; pp. 1–45. ISBN 9781584885931.
21. Waller, L.A.; Gotway, C.A. *Applied Spatial Statistics for Public Health Data*; John Wiley & Sons: New York, NY, USA, 2014; 520p. [CrossRef]

22. Bithell, J.F. Estimation of relative risk functions. *Stat. Med.* **1991**, *10*, 1745–1751. [CrossRef] [PubMed]
23. Berk, R.; MacDonald, J.M. Overdispersion and Poisson regression. *J. Quant. Criminol.* **2008**, *24*, 269–284. [CrossRef]
24. Belaid, N.; Adjabi, S.; Kokonendji, C.C.; Zougab, N. Bayesian adaptive bandwidth selector for multivariate discrete kernel estimator. *Commun. Stat.-Theory Methods* **2018**, *47*, 2988–3001. [CrossRef]
25. Chu, C.-Y.; Henderson, D.J.; Parmeter, C.F. On discrete Epanechnikov kernel functions. *Comput. Stat. Data Anal.* **2017**, *116*, 79–105. [CrossRef]
26. Shimazaki, H.; Shinomoto, S. Kernel Bandwidth Optimization in Spike Rate Estimation. *J. Comput. Neurosci.* **2010**, *29*, 171–182. [CrossRef] [PubMed]
27. Duong, T.; Hazelton, M.L. Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparametric Stat.* **2003**, *15*, 17–30. [CrossRef]
28. Davies, T.M.; Jones, K.; Hazelton, M.L. Symmetric adaptive smoothing regimens for estimation of the spatial relative risk function. *Comput. Stat. Data Anal.* **2016**, *101*, 12–28. [CrossRef]
29. Beconytė, G.; Vasiliauskas, D.; Govorov, M. Lietuvos policijos 2015–2019 m. registruotų įvykių erdvinė sklaida ir dinamika (Spatial Distribution and Dynamics of Events Registered by Lithuanian Police in 2015–2019). *Filos. Sociol.* **2020**, *31*, 175–185.
30. Davies, T.M.; Baddeley, A. Fast computation of spatially adaptive kernel estimates. *Stat. Comput.* **2018**, *28*, 937–956. [CrossRef]
31. Loader, C. Bandwidth selection: Classical or plug-in? *Ann. Stat.* **1999**, *27*, 415–438. [CrossRef]
32. Sain, S.R.; Baggerly, K.A.; Scott, D.W. Cross-validation of multivariate densities. *J. Am. Stat. Assoc.* **1994**, *89*, 807–817. [CrossRef]
33. Chacón, J.E.; Duong, T. Multivariate Kernel Smoothing and Its Applications. In *Monographs on Statistics and Applied Probability*; CRC Press: Boca Raton, FL, USA, 2018; Volume 160. [CrossRef]
34. Jones, M.C.; Marron, J.; Park, B. A simple root n bandwidth selector. *Ann. Stat.* **1991**, *19*, 1919–1932. [CrossRef]
35. Abramson, I.S. On bandwidth estimation in kernel estimates—A square root law. *Ann. Stat.* **1982**, *10*, 1217–1223. [CrossRef]
36. Bowman, A.; Azzalini, A. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*; Oxford University Press: Oxford, UK, 1997.
37. Ćwik, J.; Koronacki, J. Multivariate density estimation: A comparative study. *Neural Comput. Appl.* **1997**, *6*, 173–185. [CrossRef]
38. Cronie, O.; van Lieshout, M.N.M. A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika* **2018**, *105*, 455–462. [CrossRef]
39. Mammen, E.; Martínez-Miranda, M.D.; Nielsen, J.P.; Sperlich, S. Do-validation for kernel density estimation. *J. Am. Stat. Assoc.* **2011**, *106*, 651–660. [CrossRef]
40. Chen, Y.-C. A tutorial on kernel density estimation and recent advances. *Biostat. Epidemiol.* **2017**, *1*, 161–187. [CrossRef]
41. Eismontaitė, A.; Beconytė, G. Šalies įvykiai ir nusikalstamumas—Viešosios informacijos pateikimas žemėlapiuose (Crime events and representation of public information on maps in 2010). *Filos. Sociol.* **2011**, *24*, 405–413.
42. Gilbert, R.O. *Statistical Methods for Environmental Pollution Monitoring*; Wiley: Hoboken, NJ, USA, 1987.
43. Govorov, M.; Beconytė, G.; Gienko, G.; Putrenko, V. Spatially constrained regionalization with multilayer perceptron. *Trans. GIS* **2019**, *23*, 1048–1077. [CrossRef]