International Journal of Geo-Information

MDPI

Article

Considerations for Developing Predictive Spatial Models of Crime and New Methods for Measuring Their Accuracy

Chaitanya Joshi [1,2,*], Sophie Curtis-Ham [3], Clayton D'Ath [1] and Deane Searle [4]

[1] School of Computing and Mathematical Sciences, University of Waikato, Hamilton 3216, New Zealand; clayton.dath@gmail.com
[2] NZ Institute of Security and Crime Science, University of Waikato, Hamilton 3216, New Zealand
[3] Evidence Based Policing Centre, New Zealand Police, Wellington 3017, New Zealand; Sophie.CURTIS-HAM@police.govt.nz
[4] Waikato District, New Zealand Police, Hamilton 3216, New Zealand; Deane.Searle@Police.Govt.NZ
* Correspondence: cjoshi@waikato.ac.nz

Abstract: A literature review of the important trends in predictive crime modeling and the existing measures of accuracy was undertaken. It highlighted the need for a robust, comprehensive and independent evaluation and the need to include complementary measures for a more complete assessment. We develop a new measure called the penalized predictive accuracy index (PPAI), propose the use of the expected utility function to combine multiple measures and the use of the average logarithmic score, which measures accuracy differently than existing measures. The measures are illustrated using hypothetical examples. We illustrate how PPAI could identify the best model for a given problem, as well as how the expected utility measure can be used to combine different measures in a way that is the most appropriate for the problem at hand. It is important to develop measures that empower the practitioner with the ability to input the choices and preferences that are most appropriate for the problem at hand and to combine multiple measures. The measures proposed here go some way towards providing this ability. Further development along these lines is needed.

Keywords: predictive crime models; measures of accuracy; model selection; predictive accuracy index (PAI); expected utility

Citation: Joshi, C.; Curtis-Ham, S.; D'Ath, C.; Searle, D. Considerations for Developing Predictive Spatial Models of Crime and New Methods for Measuring Their Accuracy. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 597. https://doi.org/10.3390/ijgi10090597

Academic Editor: Wolfgang Kainz

Received: 25 June 2021
Accepted: 8 September 2021
Published: 10 September 2021

1. Introduction

Crime prevention is key to reducing crime. One crime prevention strategy is to preempt crime and use targeted policing and other measures to prevent it from occurring. The success of this strategy partially rests on how accurately one can predict the time and location of a particular crime before it happens. Such prediction should be possible, in principle, given relevant information (e.g., the location and date/time of past crime, and the socio-demographic or environmental factors that are likely to stimulate crime). However, crime is a dynamic and evolving process complexly related to a wide array of factors, many of which are also dynamic and evolving. While numerous statistical models have been built in the last two decades to predict crime with varying degrees of success, evidence of a rigorous model building process has not necessarily been demonstrated in each instance. Any predictions need to satisfy operational constraints so that they can be acted on. Otherwise, a model that may look good on paper could turn out to be ineffective in practice. As a result, building crime models that predict crime with a great degree of accuracy and are of practical use remains an ongoing research problem [1–3].

Another important issue is how to appropriately measure the accuracy of a crime model. Statistical theory contains many standard measures to assess model fit and the predictive accuracy. While some of these measures have been employed in the criminology literature, new predictive accuracy measures have also been developed specifically for

crime models. However, some of these measures may have drawbacks and there is, as yet, no consensus on which accuracy measures are the most appropriate for a crime model.

The aims of this paper are first, to propose new, complementary measures that address some limitations of existing measures and second, to highlight some additional challenges and key considerations involved in developing and assessing predictive crime models. The structure of the paper is as follows. Section 2 reviews existing predictive crime model approaches and measures of accuracy proposed so far in the criminology literature. Section 3 proposes two new measures of accuracy for crime models, namely, the penalized predictive accuracy index (PPAI) and the average logarithmic score (ALS). In Section 4, we propose the use of the expected utility function to combine multiple measures. Section 5 discusses some additional considerations involved in building and testing predictive crime models. Finally, in Section 6, we conclude by highlighting our key points and implications for future research and practice.

2. Predictive Crime Models and Measures—A Review

Common to all the models discussed in this paper is the concept of the spatial prediction of crime hotspots. Hotspots are areas with more crime than average; in other words, areas where crime concentrates [4–6]. Much empirical evidence suggests that crime is highly concentrated, such that hotspots account for disproportionately large amounts of crime [7] and that the prioritization of police resources to these hotspots can lead to significant crime prevention benefits [8]. Many different types of spatial and spatio-temporal models have been developed to predict hotspots. We briefly describe these next, before reviewing existing methods to evaluate their accuracy in predicting hotspots. Given that the police need to prioritize resources to places where crime is indeed most likely to occur, having a clear understanding of the accuracy of a model's predictions is important.

2.1. A Brief Review of Crime Prediction Models

A large number of different crime models have been developed over time. It is not possible to include all of them in a brief review, and more comprehensive reviews exist, e.g., [9]. Here, we aim to highlight models that exemplify some of the important types of models that have been proposed.

Early approaches were primarily based on time series analysis and attempted to study how crime rates, as well as factors that could influence crime (e.g., unemployment rates, drug use, deterrence and legislative changes) evolved over time, in order to explain the level of crime, e.g., [10–13]. Such models have limited utility as predictive models because the causal structure is often weak or partially incorrect [14] and because they make community-level predictions, which may not be as actionable as models that make space- and time-specific predictions.

The terms *retrospective* and *prospective* have been used to classify the crime models [15,16]. While such classification has no formal statistical meaning, the terminology is useful to distinguish the models based on the predictive rationale used.

Retrospective models use past crime data to predict future crime. These include hotspot-based approaches which assume that yesterday's hotspots are also the hotspots for tomorrow. This assumption has empirical justification: research has shown that while hotspots may flare up and cool down over relatively short periods of times, they tend to occur in the same places over time [17]. Hotspot models have typically been spatial models only, not explicitly accounting for temporal variations, e.g., [10], so seasonal or cyclical patterns could be missed. Retrospective time series models have also been proposed, e.g., [18,19], and while the more complex of these methods are able to capture various patterns in crime over time, they also increasingly become less user friendly and have to be aggregated to a community level [20], limiting their use for informing patrol patterns.

Prospective models use not just past data, but attempt to understand the root causes of crime and build a mathematical relationship between the causes and the levels of crime. Prospective models are based on criminological theories and model the likely prospect

of crime based on the underlying causes. It is therefore expected that these models may be more meaningful and provide predictions that are more 'enduring' [15]. Prospective models developed so far are based on either socio-economic factors (e.g., RTM [15]) or the *near-repeat* phenomenon (e.g., *Promap* [16]; *PredPol* [21]). The term *near-repeat* refers to the widely observed phenomenon (especially in relation to crimes such as burglary) where a property or the neighboring properties or places are targeted again shortly after the first crime incident [16].

Employing a near-repeat approach, Johnson et al. [22] modeled the near-repeat phenomenon (i.e., for how far and for how long is there an increased risk of crime) and produced a predictive model named *Promap*. Mohler et al. [21] modeled the near-repeat phenomenon using self-exciting point processes which had earlier been used to predict earthquake aftershocks. This model is available in a software package *PredPol*. While these two models consider the near repeat phenomenon, they do not consider longer-term historical data, taking into account the overarching spatial and temporal patterns. They also do not take into account the socio-demographic factors that can result in crime, and long-term changing dynamics in suburbs/communities.

In contrast, *Risk Terrain Modeling* (RTM) [15] combines a number of socio-demographic and environmental factors using a regression-based model to predict the likelihood of crime in each grid cell. However, this model does not consider historical crime data and thus may not accurately capture the overarching spatial and temporal patterns in crime. It also does not take near repeats into account and thus does not consider short-term risks at specific locations. A meta-analysis [23] found that RTM is an effective forecasting method for a number of different crime types. However, research has also demonstrated that RTM can be less accurate than machine learning methods that better model the complexity of interactions between input variables, such as Random Forest [24].

Ratcliffe et al. [25] argue that a model that includes both short-term (near-repeat) as well as long-term (socio-demographic factors and past crime data) components has a superior 'parsimony and accuracy' compared to models that only include one of those. While this argument is logical, their assertion is based on comparing models using their BIC (Bayesian Information Criterion) values. While BIC is a standard statistical measure to compare models, it measures how well a given model 'fits' or 'explains' the data (e.g., [26]) and does not directly measure the predictive accuracy (e.g., [27]). We discuss this point further in Section 2.2. Therefore, the assertion made by Ratcliffe et al. [25] still remains to be verified.

In recent years, several attempts have been made to build predictive crime models using artificial neural networks-based machine learning algorithms [6,28–30]. These studies report encouraging results, indicating that neural-network-based models could play an important role in predicting crime in the future. Neural networks are often considered as 'black box' models and a common criticism of such models is that they cannot explain causal relationships. Thus, while a neural network model may be able to predict crime with good accuracy, it may not be able to highlight the underlying causal factors and could lack transparency in how it works.

Lee et al. [1] argued that transparency in exactly how an algorithm works is just as important a criterion as predictive accuracy and operational efficiency. They point out that many of the available crime models are complex, proprietary and lack transparency. They propose a new Excel-based algorithm that is fully transparent and editable. It combines the principal of population heterogeneity in the space–crime context with the principal of state dependency (near-repeat victimization). The authors claim their algorithm outperforms existing crime models on operational efficiency, but not on accuracy. However, they do point to further improvements that could potentially lead to better accuracy.

While individual authors have argued the strengths of their respective methods, there have been few independent comparative evaluations. Perry [31] and Uchida [32] concluded that statistical techniques used in predictive crime analytics are largely untested and are yet to be evaluated rigorously and independently. Bennett Moses and Chan [33] reviewed

the assumptions made while using predictive crime models and issues regarding their evaluations and accountability. They concluded by emphasizing the need to develop better understanding, testing and governance of predictive crime models. Similarly, Meijer and Wessels [34] concluded that the current thrust of predictive policing initiatives is based on convincing arguments and anecdotal evidence rather than on systematic empirical research, and call for independent tests to assess the benefits and the drawbacks of predictive policing models. Most recently, a systematic review of spatial crime models [9] concluded that studies often lack a clear report of study experiments, feature engineering procedures, and use inconsistent terminology to address similar problems. The findings of a recent randomized experiment [35] suggested that the use of predictive policing software can reduce certain types of crime but also highlighted the challenges of estimating and preventing crime in small areas. Collectively, these studies support the need for a robust, comprehensive and independent evaluation of predictive crime models.

2.2. Measures for Comparing Crime Models

Focusing crime prevention efforts effectively relies on identifying models that accurately forecast where crime is likely to occur. Note, however, that the reported *accuracy* of a model depends on the data to which it was applied, as it could be less accurate or more accurate on a different dataset. More importantly though, the accuracy also depends on how it is measured. Not all measures are created equal.

Criminology research has employed both the standard statistical measures as well as measures developed specifically for predictive crime models. There is, however, little consensus as to how to best measure and compare the performance of predictive models [36]. Here, we provide a brief review of some of these measures. It is not an exhaustive review, but highlights measures that exemplify the different types of approaches that have been proposed.

Some of these measures assess the ability of predictive models to accurately predict crime (i.e., whether the predictions come true), while others assess their ability to yield operationally efficient patrolling patterns: minimizing patrol distance whilst maximizing potential prevention gain.

In predictive modeling, one typically uses two sets of data. A *training* dataset—the data used to fit the model, and a *testing* dataset—the data that will be used to test the model predictions. Model assessment is typically based on comparing predictions derived from the training dataset to observed crimes in the test dataset. In applications such as crime modeling, where a process evolves over time, the testing and the training datasets typically correspond to data from two distinct time periods. To account for variability in predictive accuracy over time, the reported value is often the average value of the measure over several test time periods. Using separate testing data ensures that the predictive accuracy is correctly measured.

One of the reasons why standard statistical model fitting measures such as AIC, BIC and R^2 cannot measure the predictive accuracy of a model very accurately [27] is that they only use the training data. This is because their main objective is to measure how well a model explains a given set of data and not how well it can predict unseen or future data. Because we specifically focus on the measures of predictive accuracy or operational efficiency, we do not include measures of model fit, such as the BIC, in this review.

The average logarithmic score (ALS) was first proposed by Good [37] and later advocated by Gneiting et al. [38] for its technical mathematical properties. ALS computes the average joint probability of observing the testing data under a given model. In simple terms, if the ALS for model A is higher than the ALS for model B, then model A is more likely to produce the testing data than model B. Thus, ALS directly measures the predictive accuracy of a model.

$$\mathrm{ALS} = \frac{1}{N}\sum_{i=1}^{N} \log f(\tilde{x}_i, \theta), \tag{1}$$

where, $\tilde{x}$ denotes the testing data, θ denotes the model parameters and $f(\tilde{x}_i, \theta)$ denotes the probability of observing the testing observation x_i under the model. ALS was used [39] to measure the predictive accuracy of a spatio-temporal point process model to predict ambulance demand in Toronto, Canada. Similar models have been used for crime (e.g., PredPol, [21]). Thus, ALS is a natural candidate to measure accuracy for such crime models.

Another approach to assess the predictive accuracy of a model is by looking at the distribution of True Positives (TPs), False Positives (FPs), True Negatives (TNs) and False Negatives (FNs). Several of the accuracy measures proposed in the crime literature are indeed based on one or more of these quantities. This includes the 'hit rate' [4,6,21,22,31,36,40–43], which is the proportion of crimes that were correctly predicted by the model out of the total number of crimes committed in a given time period (TP/(TP + FN)), and is typically applied to hotspots identified by the model. Similarly, a measure termed 'precision'—defined as the proportion of crimes that were correctly predicted by the model out of the total number of crimes predicted by the model (TP/(TP + FP))—has also been proposed [6,44]. Finally, a measure termed 'predictive accuracy' (PA)—which measures the proportion of crimes correctly classified out of the total number of crimes ((TP + TN)/(TP + FP + TN + FN))—has also been used [45–47].

Note that while the terms used, namely, hit rate, precision and predictive accuracy, may appear to be novel, the measures themselves are well established in the statistical literature (e.g., [48]). Thus, hit rate refers to what is also commonly known as the 'sensitivity' of the model, whereas precision refers to what is also commonly referred to as the 'positive predictive value', or the PPV. Finally, predictive accuracy refers to what is often referred to as 'accuracy' [48].

A contingency table approach, which takes into account all four—TP, TN, FP and FN—quantities, has also been used [19], and statistical tests of association on such contingency tables have also been applied [15,25,42]. Rummens et al. [6] used receiver operating characteristic (ROC) analysis, which is fairly common in other domains such as science or medicine (e.g., [44,48]). ROC analysis plots the hit rate (sensitivity) against the false positive rate 1-specificity (FP/(TN + FP)) at different thresholds.

A related measure is the *Predictive Accuracy Index* (PAI) [4], which is defined as

$$\text{PAI} = \frac{n/N}{a/A}, \tag{2}$$

where a denotes the area of the hotspot/s, A, the total area under consideration, n, the number of crimes observed in the hotspot/s and N, the total number of crimes observed in the area under consideration. Thus, a/A denotes the *relative area,* namely the proportion of area covered by the hotspot/s and n/N denotes the *hit rate,* namely the proportion of crimes in that hotspot/s. Hit rate does not take into account the operational efficiency associated with patrolling the hotspot identified. For example, a large hotspot may have a high hit rate simply because it accounts for more crime, yet such a hotspot will have very little practical value in terms of preventing the crime because it may not be effectively patrolled. PAI overcomes this drawback by scaling the hit rate using the coverage area. If two hotspots have a similar hit rate, the one that has a smaller coverage area will have a higher PAI. Thus, PAI factors in both the predictive accuracy as well as the operational efficiency of the model. PAI has been widely used in the crime literature [6,36,41,49–53].

Several other attempts have been made to incorporate operational efficiency into a measure. Bowers et al. [40] proposed measures such as the *Search Efficiency Rate* (SER), which measures the number of crimes successfully predicted per km2, and *Area-to-Perimeter-Ratio* (APR), which measures how compact the hotspot is and gives higher scores for more compact hotspots. Hotspots may be compact, but if they are evenly dispersed over a wide area, then they would still be operationally difficult to patrol compared to if the hotspots were clumped together, for example. The *Clumpiness Index* (CI; [1,36,43,54]) attempts to solve this problem by measuring the dispersion of the hotspots. A model that renders hotspots that are clustered together will achieve a higher CI score compared with a model

that predicts hotspots that are more dispersed. The *Nearest Neighbour Index* (NNI; [22,55]) provides an alternative approach to measure dispersion based on the nearest neighbour clustering algorithm.

A model that predicts hotspots that change little over consecutive time periods may be operationally preferred over a model where the predicted hotspots vary more. Measures have been proposed to measure the variation of the hotspots over time. These include the *Dynamic Variability Index* (DVI; [36]) and the *Recapture Rate Index* (RRI; [41,49–53,56]). One advantage of the DVI is that it is straightforward to calculate and does not require specialized software. However, if the actual crime exhibits spatial variation over time, then one would expect a good predictive model to capture it, and hence the DVI would be higher for that model compared to (say) another model that did not capture this variation, and thus had a lower predictive accuracy. Thus, measures such as the DVI need to be considered in conjunction with the predictive accuracy of the model, and not independently. Finally, the *Complementarity* [36,57] is a visual method to investigate how a number of different crime models complement each other by predicting different crime hotspots. Here, a Venn diagram is used to display the hotspots that are jointly predicted by all the models, as well as the hotspots that are uniquely predicted by each individual model.

Some measures may be arguably superior to others because they account for more aspects of accuracy. For example, PAI could be considered as superior to the hit rate because it also accounts for the corresponding hotspot area. However, other measures capture fewer aspects of accuracy: as mentioned above, hit rate equates to sensitivity; precision equates to PPV. In such cases, which measure is more appropriate depends on the particular application and the subjective opinions of the analysts. Rather than aiming to find just *one* measure, we recommend using multiple measures to ensure that all aspects of accuracy are assessed. Kounadi et al. [9] also argued in favor of including complimentary measures. In fact, as we illustrate later in this paper, some measures can be combined in the desired way using the expected utility function. The model with the highest *expected utility* can be considered to be the best model.

The predictive accuracy of a given model will vary over time because the data considered in building the model and the actual number of crimes that happened during the prediction period will vary with time. Therefore, in practice, accuracy obtained over time will have to be somehow summarized. Considering the mean value is important but will not measure the variation in the accuracy observed over time. Therefore, it is also important to consider the standard deviation of the accuracy as well.

A final issue is to test whether the differences in accuracy (however measured) between models are statistically significant. Adepeju et al. [36] employed the Wilcoxon signed-rank test (WSR) to compare the predictive performance of two different models over a series of time periods. WSR is a non-parametric hypothesis test that can be applied to crime models under the assumption that the difference in the predictive accuracy of the two methods is independent of the underlying crime rate. When comparing multiple models, a correction method such as Bonferroni's has to be applied to ensure that the probability of false positives (in relation to whether the difference is significant or not) is maintained at the desired (usually 5%) level.

3. A New Measure for Crime Models—Penalized Predictive Accuracy Index (PPAI)

A limitation of PAI is that as long as the model is correctly identifying a hotspot, it will prefer the model whose hotspot area is smaller simply because of the way PAI is formulated. This drawback is best illustrated by a simple hypothetical example. Suppose an urban area contains 15 hotspots, which together account for 42% of the area, but 83% of the crime based on historical data. They are listed in Table 1 in decreasing order of their individual PAI (the top hotspots have the highest proportion of crime and the smallest area). As defined earlier, we use a/A to denote the *relative area*, namely the proportion of area covered by hotspot/s and n/N to denote the *hit rate*, namely the proportion of crimes

in that hotspot/s, where n denotes the number of crimes in that hotspot and N the total number of crimes in the urban area.

Table 1. Hotspots, their relative area (n/N) and hit rate (a/A) for the hypothetical example.

Hotspot	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
a/A	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03
n/N	0.1	0.09	0.08	0.07	0.06	0.06	0.05	0.05
Hotspot	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Total**
a/A	0.03	0.03	0.04	0.04	0.05	0.05	0.05	**0.42**
n/N	0.05	0.04	0.04	0.04	0.04	0.03	0.03	**0.83**

Suppose that we are comparing four models: models M-I, M-II, M-III and M-IV. The hotspots identified, their collective relative area, hit rate and PAI are listed in Table 2.

Table 2. Hotspots identified, their combined relative area (n/N) and hit rate (a/A) for the four models in the hypothetical example.

Models	**Hotspots Identified**	a/A	n/N	**PAI**
M-I	1, 2, 3	0.03	0.27	9.00
M-II	1,	0.01	0.10	10.00
M-III	1, 5, 10, 15	0.11	0.23	2.09
M-IV	13, 14, 15	0.15	0.10	0.67

Here, model M-I correctly identifies the top three hotspots (each covering only 1% of the area, and together, they account for 27% of the crime). On the other hand, model M-II is only able to identify the top hotspot (accounting only for 10% of the crime) and yet, the PAI for model M-II is higher than the PAI for model M-I.

$$\text{PAI (M-I)} = \frac{0.27}{0.03} = 9, \qquad \text{PAI (M-II)} = \frac{0.1}{0.01} = 10.$$

There could be situations where identifying a single hotspot can be considered to be desirable, so long as the hotspot is not too large to cover given the resources available. However, it could also be argued that model M-I is superior to model M-II. This is because identifying smaller or fewer hotspots, by definition, could also mean capturing a relatively smaller proportion of crime. As illustrated by the above example, the PAI could fail to penalize a model that only identifies a subset of the hotspots compared to another model that may be able to correctly capture more hotspots.

To overcome this limitation, we propose a Penalized PAI (PPAI) that penalizes the identification of a total hotspot area that is too small.

$$\text{PPAI} = \frac{n/N}{(a/A)^{\alpha}}, \qquad 0 \leq \alpha \leq 1. \tag{3}$$

The penalization is performed by using an extra parameter, α, whose value is fixed by the user. The value lies between 0 and 1.

Mathematical properties of PPAI:

1. As α goes to 0, PPAI will converge to the hit rate,

$$\lim_{\alpha \to 0} \text{PPAI} = \frac{n}{N} = \text{hit rate.}$$

2. As α goes to 1, PPAI will converge to PAI,

$$\lim_{\alpha \to 1} \text{PPAI} = \frac{n}{N} / \frac{a}{A} = \text{PAI}.$$

3. Hit rate < PPAI < PAI, when $0 < \alpha < 1$.

The value of α is related to the importance of the collective size of the hotspots. In the extreme case when $\alpha = 0$, the size of the hotspots is not important at all. Mathematically, for $\alpha = 0$, the denominator becomes 1 and PPAI reduces to the hit rate, i.e., attaching 0 weight to the hotspot area ensures that it is not considered. At the other extreme, $\alpha = 1$ indicates that the hotspot size is extremely important. Mathematically, a unit weight represents no penalty and PPAI reduces to the PAI. Thus, both the hit rate and the PAI can be considered as special cases of PPAI. For $0 < \alpha < 1$, PPAI would prefer a model that strikes the desired balance between capturing enough hotspots and yet ensuring that the collective hotspot size is not too large for practical considerations. Next, we provide two ways of determining the value of α and two different usages of PPAI.

The first option is to choose $\alpha = n/N$ (hit rate). Choosing $\alpha = n/N$ means that the hotspot area identified by a model is weighed by the relative proportion of crimes (n/N) that take place in that area. Thus, if a model identifies hotspots where fewer crimes take place, then it will be penalized more (α smaller) than a model that identifies hotspots where more crimes take place. We illustrate this by calculating PPAI with $\alpha = n/N$ for the hypothetical example considered above.

$$\text{PPAI (M-I)} = \frac{0.27}{(0.03)^{0.27}} = 0.6958,$$
$$\text{PPAI (M-II)} = \frac{0.1}{(0.01)^{0.1}} = 0.1584.$$

Because model M-II identifies hotspots that account for a much smaller proportion of crimes, it is penalized more and as a result has a much smaller PPAI value than model M-I. Suppose we have two additional models to test: models M-III and M-IV. As listed in Table 2, Model M-III identifies the top hotspot and three other smaller hotspots, in total accounting for 23% of the crime (much higher than model M-II); however, these hotspots are also much larger, accounting for 11% of the area together. Model M-IV is a clearly an inferior model that identifies the bottom three hotspots. Model M-IV achieves the lowest score on both PAI as well as PPAI, as one would expect. Model M-III, however, achieves a higher PPAI but lower PAI compared to model M-II, because PPAI was calculated using $\alpha =$ hit rate and M-III has a much higher hit rate than M-II.

$$\text{PPAI (M-III)} = \frac{0.23}{(0.11)^{0.23}} = 0.3821,$$
$$\text{PPAI (M-IV)} = \frac{0.1}{(0.15)^{0.1}} = 0.1209.$$

The second option is to find the optimal value of α so that PPAI will peak at the desired hotspot area. For example, operationally, it may only be possible to effectively patrol, say, 2% of the area at a time. The objective then is to find a model that will identify the best hotspots totaling to 2% of the area. To do this, one first finds the optimal value of α using a grid search algorithm and then uses this value of α to evaluate PPAI for all models under consideration. The model with the highest PPAI is the most suitable for the hotspots totaling 2% of the area in terms of the predictive accuracy. The value of α that is optimal according to this criterion will differ from dataset to dataset.

We illustrate how the grid search method can be used to find the optimal α value using our hypothetical example. First, we list the hotspots in an increasing order of their size (smallest first) and within that, in a decreasing order of their hit rate (smallest area

with highest hit rate first), as shown in Table 3. We then identify the top hotspots that account for 2% of the area. In our case, the top two hotspots together account for the 2% of the area, as desired, and together, they account for 19% of all crime. The optimal α value is then found by calculating the PPAI for all the cumulative hotspot levels for each value of α from 0.01 to 0.99 and finding the α value that yields the highest PPAI at the desired cumulative level (in this case, 2%). Often, there may be a range of values that satisfy this criterion. For instance, for our example, any value of α between 0.87 and 0.92 will ensure that the PPAI for the 2% cumulative level is higher than at any other level (say, 1% or 3% or higher). Within this range of values, we find the value for which PPAI for the 2% level is the most different from the neighboring levels (1% and 3%). That value of α is the optimal value, and in our case, it turns out to be 0.9. The PPAI value of 6.42 that we obtain for this optimal value of α is the highest possible PPAI value for this specific data. Over time, as crime evolves, the historical data will change and so will the optimal α value. The grid search described here can be performed using a basic spreadsheet package, such as Microsoft Excel, by entering the hotspots, as in Table 3, and then finding the value of PPAI using different values of α. A spreadsheet that includes all the calculations performed for this example and illustrates the grid search algorithm has been included in the Supplementary Materials.

Table 3. The relative area and hit rate, the cumulative area and hit rates as well as the PAI and PPAI (for α = 0.9) for the cumulative hotspot covers.

Hotspot	a/A	n/N	Cumulative a/A	Cumulative n/N	PAI	PPAI ($\alpha = 0.9$)
1	0.01	0.1	0.01	0.1	10.00	6.31
2	0.01	0.09	0.02	0.19	9.50	6.42
3	0.01	0.08	0.03	0.27	9.00	6.34
4	0.01	0.07	0.04	0.34	8.50	6.16
5	0.02	0.06	0.06	0.4	6.67	5.03
6	0.02	0.06	0.08	0.46	5.75	4.47
7	0.02	0.05	0.1	0.51	5.10	4.05
8	0.03	0.05	0.13	0.56	4.31	3.51
9	0.03	0.05	0.16	0.61	3.81	3.17
10	0.03	0.04	0.19	0.65	3.42	2.90
11	0.04	0.04	0.23	0.69	3.00	2.59
12	0.04	0.04	0.27	0.73	2.70	2.37
13	0.05	0.04	0.32	0.77	2.41	2.15
14	0.05	0.03	0.37	0.8	2.16	1.96
15	0.05	0.03	0.42	0.83	1.98	1.81

Using $\alpha = 0.9$ gives us the following PPAI scores for the four models:

$$\begin{aligned} \text{PPAI (M-I)} &= 6.3380, & \text{PPAI (M-II)} &= 6.3096, \\ \text{PPAI (M-III)} &= 1.6767, & \text{PPAI (M-IV)} &= 0.5515. \end{aligned} \tag{4}$$

Because the PPAI was asked to find optimal models for the top 2% hotspot area, the scores are now very different. Model M-I is still the best model, but it is now closely followed by M-II, whereas model M-III, which identified hotspots of much larger area, has a much smaller PPAI score. Model M-IV is clearly still the worst-rated model, as expected.

By determining the highest possible value of PPAI for these data and this α, this approach provides a scale on which all models can be compared not just in relative terms, but in absolute terms. For instance, in this case, we know that model M-I is not just the best among the four models considered, but a very good model for these data and for a 2% hotspot area among all possible models, because its PPAI is very close to the maximum possible value of 6.42 that was found for these data using the grid search method and as listed in Table 3. Thus, PPAI parameter α provides flexibility around the importance of the

hotspot area and even allows the choice of α that is optimized for a certain desired hotspot level, constituting an important improvement over PAI. Computing PPAI is straightforward and does not require any additional resources than those required for PAI.

4. Using Expected Utility to Combine Multiple Measures

As discussed, different measures consider different aspects of predictive accuracy and operational efficiency. While there could be reasons why one of those measures could be considered to be the most appropriate measure for a given analysis, in general, using multiple measures will provide a more complete picture of the performance of a model. It is also possible that multiple measures will rate the given models differently and may not all agree on which model is better. Consider the four measures discussed earlier—the true positives (TPs), false positives (FPs), true negatives (TNs) and false negatives (FNs). Each of these measure the predictive performance of a model in their unique way. Taking all four of them in to account will provide a more complete picture of the overall performance of a model. However, this now requires either a multi-dimensional evaluation (because a method with a very high true positives rate could also have a very high false negative rate, for example) or summarizing the four measures into a single measure in an appropriate way.

Chi-squared statistics can be used to summarize these four measures into one using a contingency table [18]. While such an application of the Chi-squared test will determine if the predictive accuracy of a model is statistically significantly different to a random assignment of hotspots, the Chi-squared test does not test for the direction of the difference, nor is it able to be used to identify which of the two models is better. Further, a Chi-squared test is not able to weigh the four outcomes differently based on their relative importance. These drawbacks are also applicable to other alternatives for the Chi-squared test, such as a Fisher's exact test. Alternatively, one can use the receiver operating characteristics (ROC) analysis [6]; however, this uses just two of the four measures: true positives and false positives.

One way to combine multiple measures while being able to account for the importance of each measure is by using the concept of expected utility. The notion of utility was first introduced in the context of game theory [58]. It is now widely used in Bayesian statistics, game and decision theory, data mining and economics (e.g., [59–62]). More recently, expected utility has been used in Adversarial Risk Analysis models, which model the actions of the strategic adversary and find the optimal actions for the defender (e.g., [63–66]).

Finding the expected utility of a model involves considering the utilities (costs, gains or losses) associated with each outcome and then taking a weighted sum (expected value) of the utilities, where the weights are proportional to the probabilities of the outcomes to arrive at the net average gain/loss of using the model. The utility can be either positive (indicating that gains outweigh losses) or negative (losses outweigh gains) or 0 (losses = gains). One then chooses the model that has the maximum expected utility. The concept of expected utility can be applied to crime models in three different ways. We describe them below.

4.1. Cost Matrix Approach

Consider the problem of combining the four measures considered above, namely the TP, FP, TN and FN. The cost matrix approach, popular in data mining [62], can be used for models that classify a cell/area as belonging to one class or another. For instance, a predictive crime model will label a cell/area as '+' (crime likely to happen) or '−' (crime unlikely to happen). Applying this approach, each of the four measures can be weighted and combined. We first find the expected utility of a positive/negative label:

$$\begin{aligned} \text{Expected utility } (+) &= \%\text{TP} \times \text{utility (TP)} + \%\text{FP} \times \text{utility (FP)}, \\ \text{Expected utility } (-) &= \%\text{TN} \times \text{utility (TN)} + \%\text{FN} \times \text{utility (FN)}. \end{aligned} \tag{5}$$

and then find the expected utility of the model:

$$\text{Expected utility (model)} = [\%\text{'+'}] \times [\text{Expected utility } (+)] + [\%\text{'}-\text{'}] \times [\text{Expected utility } (-)]. \tag{6}$$

In policing, the utilities will be subjective and there are likely no fixed norms about what utility should be attached to which outcome. This is a decision that requires consideration of not only the associated costs and gains of crime and policing responses but also the kind of police response that is considered appropriate for a given community [67]. These considerations are likely complex and utilities could be different for different crime types, locations and time periods. As an example, the utility associated with an FP prediction could be perceived differently by different police services. The expected utility measure allows each police service to input the utilities that are realistic and relevant to them. Thus, the subjectivity associated with determining utilities will make model assessment more realistic and relevant for each analysis. This feature sets the expected utility measure apart from most other measures discussed.

In illustrating this measure, for the sake of simplicity, we elected to attach relative utility values between the range +1 and −1, where +1 indicates the most desirable outcome and −1, the least desirable Table 4. For example, a correct prediction, whether a TP or a TN, may be considered as highly valuable, and therefore, utility (TP)= utility (TN) = +1. However, an FP, where the model predicted the cell to be a hotspot but no crime happened during the prediction window, could be considered as a more acceptable and a smaller loss (the cost of resources spent) compared to an FN, where the model predicted the cell to be not a hotspot but a crime did happen, resulting in the cost of that crime to the victim and society (in investigating and dealing with the offense). Therefore, we may assign utility (FP) = −0.5, and utility (FN) = −1.

Suppose we have two models. Model A has a very high TP rate but also a very high FN rate. Model B has a slightly lower TP rate but a substantially lower FN rate. Expected utility will enable us to identify which model is better according to the utility criteria. We assume that both the models predict 5% of cells to be hotspots (i.e., % '+' = 0.05).

Table 4. The TP, FP, TN and FN measures for the two hypothetical models.

Model	TP%	FP%	TN%	FN%
Model A	85	15	30	70
Model B	75	25	45	55

Model A

$$\begin{aligned} \text{Expected utility } (+) &= 0.85 \times 1 + 0.15 \times (-0.5) = 0.775 \\ \text{Expected utility } (-) &= 0.3 \times 1 + 0.7 \times (-1) = -0.4 \end{aligned} \tag{7}$$

Expected utility (A) = 0.05 × 0.775 + 0.95 × (−0.4) = **−0.34125**

Model B

$$\begin{aligned} \text{Expected utility } (+) &= 0.75 \times 1 + 0.25 \times (-0.5) = 0.625 \\ \text{Expected utility } (-) &= 0.45 \times 1 + 0.55 \times (-1) = -0.1 \end{aligned} \tag{8}$$

Expected utility (B) = 0.05 × 0.625 + 0.95 × (−0.1) = **−0.06375**

In the above example, both models have an expected utility that is negative (as a result that 95% of the cells are labeled '-' and the FN rate is high). However, it also reveals that model B has much smaller negative utility despite having a smaller TP rate. Thus, we now know that model B provides better predictions overall after taking account of all the possible prediction outcomes and the utilities associated with each possible outcome.

In addition to TP, FP, TN and FN, some other related measures can also be combined using expected utility. For example, sensitivity (hit rate) can be combined with specificity (TN/(TN + FP)) and PPV (precision) can be combined with negative predictive value (NPV), which is TN/(FN + TN) and so on.

4.2. Replacing Probabilities with Weights

A limitation of the cost matrix approach is that it cannot be applied to all accuracy measures. By definition, expected utility can only be applied to measures that correspond to mutually exclusive events. It is not possible to combine, say, the PPAI and the ALS using expected utility.

An alternative option is to define a weight measure inspired by the expected utility concept. The probabilities can be replaced by weights that are positive and sum to 1 (just like probabilities) and represent the relative importance of each measure. We illustrate this method with the same hypothetical example, where we have the two models Table 5, A and B, but we now want to use two different measures, namely hit rate and precision. We can combine them by taking the weighted sum of their scores. Using the hypothetical TP, FP, TN and FN values already considered, we can calculate the hit rate and precision for these two models as shown below. Here, model A is the better model according to precision but model B is the better model according to the hit rate.

Table 5. The hit rate and the precision for the two hypothetical models.

Model	Hit Rate	Precision
Model A	0.06	0.85
Model B	0.067	0.75

Note that because we assume %'+' =0.05 for both the models, hit rate (A) = $85 \times 0.05/(85 \times 0.05 + 70 \times 0.95) = 0.06$ and similarly, hit rate (B) $= 75 \times 0.05/(75 \times 0.05 + 55 \times 0.95) = 0.067$. Furthermore, note that in this case, because each model yields hotspots of the same area (5%), the PAI ranking will be exactly same as the hit rate ranking. An analyst might want to give a considerably high weight to the hit rate (say, $w = 0.7$) and the remaining $(1 - w = 0.3)$ to the precision. The weighted aggregates for the two models then become:

$$\begin{aligned} \text{Model A}: &\quad 0.7 \times 0.06 + 0.3 \times 0.85 = 0.297 \\ \text{Model B}: &\quad 0.7 \times 0.067 + 0.3 \times 0.75 = 0.291 \end{aligned} \tag{9}$$

Taking into consideration the relative importance of the two measures, we can now see that model A is the better model overall. The fact that this ranking does not match the ranking obtained using expected utility on TP, FP, TN and FN is not surprising because different measures are used here and combined differently.

4.3. Using Ranks

Another challenge is that when combining different measures, one needs to ensure that the results are not distorted due to scaling. This issue did not arise when combining hit rate and precision, because both will always take values between 0 and 1. However, other measures could yield large positive values or even large negative values (for instance, in case of ALS, which is an aggregation of log probabilities). To avoid distortion due to scaling, weighted aggregation may only be performed after standardizing the scores (0 mean and unit standard deviation), where possible.

Alternatively, models could be ranked according to each measure and the weighted aggregation may be performed on the ranks. Consider the four models, M-I, M-II, M-III and M-IV, discussed in Section 3. In addition to PPAI (using $\alpha = 0.9$), suppose the analyst also used the ALS to measure the predictive accuracy of these models. Because the two measures take values on a different scale, it would not be appropriate to simply combine them using weights, as in Section 4.2. Instead, the analyst could rank the models according to each of the measures. Let us assume that the ranks are as listed in Table 6. Because the models are ranked differently according to each measure, the analyst can combine the two

ranks using weights that add up to 1. Let us say that they use weights 0.6 and 0.4 for PPAI and ALS, respectively, thus favoring the PPAI more than ALS.

Table 6. Hypothetical model ranks according to PPAI and ALS.

Model	PPAI Rank	ALS Rank	Weighted Aggregate Rank
Model M-I	1	2	1.4
Model M-II	2	1	1.6
Model M-III	3	4	3.4
Model M-IV	4	3	3.6

Then, the weighted aggregate rank can be obtained to determine which model is the best overall according to the measures and the weights used. For example, the weighted aggregate rank for model M-I = $1 \times 0.6 + 2 \times 0.4 = 1.4$, and so on. Thus, the expected utility approach can be generalized to combine multiple measures and can be used to combine any and all measures as long as the models can be numerically ranked according to those measures.

5. Additional Considerations When Choosing and Comparing Models

The development and comparison of predictive models is not straight-forward; it requires the careful consideration of various issues of both a technical and practical nature. Further to our proposed measures, to support a more appropriate and comprehensive evaluation of crime prediction models, here, we elaborate on some additional issues to consider.

5.1. Technical Considerations

The performance of a predictive model depends on several key technical factors. These include the type of model used, the variables included in the model, the tuning or calibration of any free parameters, the sparsity of the data, the predictive window used and finally, the way in which its accuracy is measured.

From a purely technical point of view, a model should only be considered if it is appropriate given the aims of the analysis and if any founding mathematical assumptions made by the model have been satisfied by the data. Additionally, one must consider the practicalities of the data pre-processing or data-coding needed, the ease or complexity of implementing the models and the computational and financial costs needed to run these models.

During the model building process, one or more free parameters may need to be assigned a value. Common examples include the bandwidth in a kernel density estimation model or a smoothing parameter in a time-series model. A common approach with crime data is to aggregate it over rectangular grid cells and then use this aggregated data for analysis. Here, the cell size is also a free parameter. Predictions will likely be sensitive to the values assigned to such free parameters [5,36]. Their values must thus be assigned based on some theoretical or empirical justification and not arbitrarily. It is also advisable to perform some robustness analysis to understand the sensitivity of the predictions to the assigned values.

The quality and usability of the predictions are usually a product of the type of the model used and the factors included. For example, a model that considers socio-demographic factors but not spatial or temporal factors may yield predictions that are independent of space and time. Therefore, it may be of lesser operational value than a model that also considers space and time and offers predictions specific to a given location and time. However, such a model is also likely to require a larger and denser dataset in order to provide accurate predictions. Further, the predictions of a model are usually only valid for the range of the predictor variables considered in the analysis. These are not likely to be valid or accurate for values of the predictor variables outside that range as well as

time periods too far in the future. The factors used may change with space and over time; the model assumptions may not be valid in the future or to a different area.

The size of the dataset and the density or sparsity of the data are also critical factors. Size relates not only to the number of observations, but also to the amount of information available for each observation and its spread across space and time. Typically, larger data contain more information and allow for more advanced models incorporating more variables. Its density or sparsity depends on how many observations are available with each level of a given factor. For example, in a small town with low crime rate, it may only be reasonable to include either the spatial factor or the temporal factor, but not both. Implementing a model on data that are too small or sparse may lead to over-fitting and hence, poor predictive accuracy. Typically, data become sparser as more factors are considered, affecting the optimal cell size and other free parameters.

Multiple measures are available to measure the performance of a model and compare multiple models. Some measures consider the goodness of fit, some others look at the predictive accuracy and indeed some others may take into account the operational utility. The same model may fare differently when assessed using different measures. Therefore, one needs to identify the most appropriate measure for the study and then use that measure to identify the best performing model. For example, as discussed earlier, traditional goodness of fit measures such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) may be more appropriate when comparing models to see which one of them 'explains' the data better. However, as noted above, these do not measure predictive accuracy and are therefore not the most appropriate measures for finding the model that has the highest predictive accuracy.

Another, more complex issue is how to identify the best model when different measures point to different models. We have proposed one solution, considering the expected utility measure or the weighted aggregate measure discussed in Section 4. Either of these options involves making a subjective choice, meaning that someone with a different set of weights or utilities could choose a different model as the best model for the same data and using the same set of measures. Further, different analysts may prefer to use other accuracy measures altogether, and hence are likely to arrive at different conclusions about the relative model performance.

Analysts should recognize this subjectivity and provide a clear rationale for the choice of models, accuracy measures as well as the utilities and weights used to combine these measures. This will go towards providing more transparency and reproducibility in studies of crime prediction models.

5.2. Other Considerations

Although the focus of this manuscript is on technical considerations, there are several practical issues to consider in deciding on the most suitable predictive crime model, including ethical and legal aspects. As discussed earlier, [1] argued that transparency in exactly how an algorithm works is just as important a criterion such as predictive accuracy and operational efficiency. However, it is also important that the data used are obtained using best practices, are accurate and are not a product of racially biased or unlawful practices. Recent research [68] has highlighted the ramifications of using predictive policing tools informed by 'dirty data' (data obtained during documented periods of flawed, racially biased and sometimes unlawful practices and policies). Predictive policing models using such data could not only lead to flawed predictions but also increase the risk of perpetuating additional harm via feedback loops. Such problems do not negate the potential utility of hotspot prediction in principle, but highlight problems that can occur in the development and implementation of predictive models [69]. Those developing spatial crime prediction models need to be mindful of these issues as well as the technical considerations discussed in this paper.

Finally, the choice of model is also a choice of policing theory. As Ferguson [67] argues, when purchasing a particular predictive technology, police are not simply choosing the

most sophisticated predictive model; the choice reflects a decision about the type of policing response that makes sense in their community. Foundational questions about whether we want police officers to be agents of social control, civic problem solvers, or community partners lie at the heart of any choice of which predictive technology might work best for any given jurisdiction.

6. Discussion and Summary

Significant research efforts have focused on developing predictive crime models that can inform police decisions about where to prioritize often limited resources, to achieve the biggest possible reductions in crime. Many different models have been proposed and claims have been made about the superiority of a given model for informing such decisions. As pointed out by several literature reviews [9,31–35], these claims have often not been based on rigorous, independent and impartial assessment. Finding or developing appropriate measures to assess and compare the performance of these models is therefore important, yet this has not received nearly equal attention, with only a few measures having been developed specifically for crime application. Further, these measures have limitations, necessitating the development of additional, complementary measures. The ALS measure described above is one possible measure. It has been used in other domains, measures aspects not measured by existing measures and could be used for a wide variety of crime models.

It is worth emphasizing that there can be no single model, or indeed, no single measure that is superior over all others at all times. Future studies should explicitly explain why certain measures were considered the most appropriate for the problem at hand and demonstrate how a given model performs according to those measures. As discussed, the accuracy achieved not only depends on the model but on the quality of the data and their density or sparsity. The quality of the data refers not only to omissions or inaccuracies in the data but also on whether the data reflect flawed or biased practices. The choice of model not only concerns the predictive accuracy or the operational efficiency possible, but also the choice of policing theory and what is appropriate for the community in question.

It is therefore important to develop measures that empower the practitioner with a certain level of flexibility to tweak the measure so that it is the most appropriate for a given situation. The penalty parameter in the PPAI measure that we propose does precisely that. It empowers the practitioner to choose the right balance between capturing enough crime and operational efficiency for the problem at hand. Since different measures measure different aspects of accuracy and efficiency, it may be desirable to use multiple measures to assess models. However, this can be challenging because different measures could rank models differently and combining the measures may not be straightforward. A possible solution is a mathematical function that empowers the user with flexibility to combine multiple measures in a way that is desirable for the problem at hand. We suggest using the expected utility function and its extension, the weighted aggregate for this purpose.

The concept of utility or weights reflects subjective inputs and therefore could be perceived as undesirable. However, they in fact model the human decision-making process. Decision makers have different sets or preferences and value systems, which are often reflected in the decisions they make. The use of weights enables the practitioner to translate their thought process in an objective mathematical equation and ensures that the equation will identify the correct model once the weights have been elicited according to the user's preferences. Because every dataset, prediction problem and context is unique, a seemingly 'objective' or 'one size fits all' solution is unlikely to be the right approach. Instead, we advocate solutions that empower practitioners to tailor their assessment to their particular problem and to clearly document their decision making. In doing so, future studies are more likely achieve standards of transparency, reproducibility and independence that will further crime prediction as a field of study and practice.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijgi10090597/s1, A spreadsheet named *PPAI Illustrations* is provided along with this manuscript. This spreadsheet includes the PAI and PPAI calculations included in this manuscript and enables the reader to re-produce the corresponding tables. It also enables the reader to find an optimum value of α for a given criteria, as illustrated in the manuscript.

Author Contributions: Conceptualization, Chaitanya Joshi; Formal analysis, Chaitanya Joshi; Investigation, Sophie Curtis-Ham and Clayton D'Ath; Methodology, Chaitanya Joshi, Sophie Curtis-Ham and Clayton D'Ath; Resources, Sophie Curtis-Ham, Clayton D'Ath and Deane Searle; Writing—original draft, Chaitanya Joshi; Writing—review & editing, Chaitanya Joshi, Sophie Curtis-Ham and Deane Searle. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank Jordan Tomkins and Muhammad Ejaz for their help in formatting this manuscript and the anonymous referees for their valuable comments and suggestions.

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

References

1. Lee, Y.; SooHyun, O.; Eck, J.E. A Theory-driven algorithm for real-time crime hot spot forecasting. *Police Q.* **2020**, *23*, 174–201. [CrossRef]
2. Santos, R.B. Predictive Policing: Where is the Evidence? In *Police Innovation: Contrasting Perspectives*; Cambridge University Press: Cambridge, UK, 2019; p. 366.
3. Ratcliffe, J.H. Predictive Policing. In *Police Innovation*; Wesiburd, D., Braga, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2019.
4. Chainey, S.; Tompson, L.; Uhlig, S. The utility of hotspot mapping for predicting spatial patterns of crime. *Secur. J.* **2008**, *21*, 4–28. [CrossRef]
5. Chainey, S.P. Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime. *Bull. Geogr. Soc. Liege* **2013**, *60*, 7–19.
6. Rummens, A.; Hardyns, W.; Pauwels, L. The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. *Appl. Geogr.* **2017**, *86*, 255–261. [CrossRef]
7. Lee, Y.; Eck, J.E.; SooHyun, O.; Martinez, N.N. A Theory-driven algorithm for real-time crime hot spot forecasting. *Crime Sci.* **2017**, *6*, 23. [CrossRef]
8. Braga, A.A.; Turchan, B.; Papachristos, A.V.; Hureau, D.M. Hot spots policing of small geographic areas effects on crime. *Campbell Syst. Rev.* **2019**, *15*, e1046. [CrossRef]
9. Kounadi, O.; Ristea, A.; Araujo, A.; Leitner, M. A systematic review on spatial crime forecasting. *Crime Sci.* **2020**, *9*, 1–22. [CrossRef]
10. Adams-Fuller, T. Historical hOmicide Hot Spots: The Case of Three Cities. Ph.D. Thesis, Howard University, Washington, DC, USA, 2001.
11. Cantor, D.; Land, K.C. Unemployment and crime rates in the post-World War II United States: A theoretical and empirical analysis. *Am. Sociol. Rev.* **1985**, *50*, 317–332. [CrossRef]
12. Corman, H.; Mocan, H.N. A time-series analysis of crime, deterrence, and drug abuse in New York City. *Am. Econ. Rev.* **2000**, *90*, 584–604. [CrossRef]
13. Sridharan, S.; Vujic, S.; Koopman, S.J. *Intervention Time Series Analysis of Crime Rates*; Technical Report, Tinbergen Institute Discussion Paper; TI2003-040/4; Elsevier: Amsterdam, The Netherlands, 2003.
14. Greenberg, D. Time series analysis of crime rates. *J. Quant. Criminol.* **2001**, *17*, 291–328. [CrossRef]
15. Caplan, J.M.; Kennedy, L.W.; Miller, J. Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting. *Justice Q.* **2011**, *28*, 360–381. [CrossRef]
16. Johnson, S.D.; Birks, D.J.; McLaughlin, L.; Bowers, K.J.; Pease, K. Prospective crime mapping in operational context: Final report. In *Home Office*; Great Britain Home Office Research Development and Statistics: London, UK, 2007.
17. Spelman, W. The severity of intermediate sanctions. *J. Res. Crime Delinq.* **1995**, *32*, 107–135. [CrossRef]
18. Gorr, W.; Olligschlaeger, A.; Thompson, Y. Assessment of crime forecasting accuracy for deployment of police. *Int. J. Forecast.* **2000**, 743–754. [CrossRef]
19. Gorr, W.; Olligschlaeger, A. *Crime Hot Spot Forecasting: Modeling and Comparative Evaluation, Final Project Report*; US Department of Justice: Washington, DC, USA, 2001.
20. Groff, E.R.; La Vigne, N.G. Forecasting the future of predictive crime mapping. *Crime Prev. Stud.* **2002**, *13*, 29–58.
21. Mohler, G.O.; Short, M.B.; Brantingham, P.J.; Schoenberg, F.P.; Tita, G.E. Self-exciting point process modeling of crime. *J. Am. Stat. Assoc.* **2011**, *106*, 100–108. [CrossRef]
22. Johnson, S.D.; Bowers, K.J.; Birks, D.J.; Pease, K. Predictive mapping of crime by ProMap: Accuracy, units of analysis, and the environmental backcloth. In *Putting Crime in Its Place*; Springer: New York, NY, USA, 2009; pp. 171–198.

23. Marchment, Z.; Gill, P. Systematic review and meta-analysis of risk terrain modelling (RTM) as a spatial forecasting method. *Crime Sci.* **2021**, *10*, 12. [CrossRef]
24. Wheeler, A.P.; Steenbeek, W. Mapping the risk terrain for crime using machine learning. *J. Quant. Criminol.* **2021**, *37*, 445–480. [CrossRef]
25. Ratcliffe, J.H.; Taylor, R.B.; Perenzin, A. *Predictive Modeling Combining Short and Long-Term Crime Risk Potential (Final Report)*; U.S. Department of Justice: Washington, DC, USA, 2016.
26. Dobson, A.J.; Barnett, A.G. *An Introduction to Generalized Linear Models*, 3rd ed.; Chapman and Hall/CRC: London, UK, 2008.
27. Shmueli, G. To explain or to predict? *Stat. Sci.* **2010**, *25*, 289–310. [CrossRef]
28. Chun, S.A.; Avinash Paturu, V.; Yuan, S.; Pathak, R.; Atluri, V.; Adam, N.R. Crime prediction model using deep neural networks. In Proceedings of the 20th Annual International Conference on Digital Government Research, Dubai, United Arab Emirates, 18–20 June 2019; pp. 512–514.
29. Tumulak, J.A.U.; Espinosa, K.J.P. Crime modelling and prediction using neural networks. In *Theory and Practice of Computation: Proceedings of Workshop on Computation: Theory and Practice WCTP2015*; World Scientific: Singapore, 2017; pp. 218–228.
30. Wang, B.; Zhang, D.; Zhang, D.; Brantingham, P.J.; Bertozzi, A.L. Deep learning for real time crime forecasting. In *International Symposium on Nonlinear Theory and Its Applications, NOLTA*; Cornell University: Ithaca, NY, USA, 2017.
31. Perry, W.L. *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*; Rand Corporation: Santa Monica, CA, USA, 2013.
32. Uchida, C.D. Predictive policing. In *Encyclopedia of Criminology and Criminal Justice*; Springer: Berlin, Germany, 2014; pp. 3871–3880.
33. Bennett Moses, L.; Chan, J. Algorithmic prediction in policing: Assumptions, evaluation, and accountability. *Polic. Soc.* **2018**, *28*, 806–822. [CrossRef]
34. Meijer, A.; Wessels, M. Predictive policing: Review of benefits and drawbacks. *Int. J. Public Adm.* **2019**, *42*, 1031–1039. [CrossRef]
35. Ratcliffe, J.H.; Taylor, R.B.; Askey, A.P.; Thomas, K.; Grasso, J.; Bethel, K.J.; Fisher, R.; Koehnlein, J. The Philadelphia predictive policing experiment. *J. Exp. Criminol.* **2020**, *17*, 1–27. [CrossRef]
36. Adepeju, M.; Rosser, G.; Cheng, T. Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions-a crime case study. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 2133–2154. [CrossRef]
37. Good, I. Rational Decisions. *J. R. Stat. Soc. Ser. B* **1952**, *14*, 107–114. [CrossRef]
38. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
39. Zhou, Z.; Matteson, D.S. Predicting ambulance demand: A spatio-temporal kernel approach. In Proceedings of the 21th ACM SIGKDD International Conference On Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 2297–2303.
40. Bowers, K.J.; Johnson, S.D.; Pease, K. Prospective hot-spotting: The future of crime mapping? *Br. J. Criminol.* **2004**, *44*, 641–658. [CrossRef]
41. Hart, T.C.; Zandbergen, P.A. *Effects of Data Quality on Predictive Hotspot Mapping*; National Criminal Justice Reference Service: Washington, DC, USA, 2012
42. Kennedy, L.W.; Caplan, J.M.; Piza, E. Risk clusters, hotspots, and spatial intelligence: Risk terrain modeling as an algorithm for police resource allocation strategies. *J. Quant. Criminol.* **2011**, *27*, 339–362. [CrossRef]
43. Lee, J.; Gong, J.; Li, S. Exploring spatiotemporal clusters based on extended kernel estimation methods. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1154–1177.
44. Brown, C.D.; Davis, H.T. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemom. Intell. Lab. Syst.* **2006**, *80*, 24–38. [CrossRef]
45. Araújo, A.; Cacho, N.; Bezerra, L.; Vieira, C.; Borges, J. Towards a crime hotspot detection framework for patrol planning. In Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Exeter, UK, 28–30 June 2018; pp. 1256–1263.
46. Malik, A.; Maciejewski, R.; Towers, S.; McCullough, S.; Ebert, D.S. Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1863–1872. [CrossRef] [PubMed]
47. Mu, Y.; Ding, W.; Morabito, M.; Tao, D. Empirical discriminative tensor analysis for crime forecasting. In *Proceedings of the International Conference on Knowledge Science, Engineering and Management*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 293–304.
48. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
49. Drawve, G. A metric comparison of predictive hot spot techniques and RTM. *Justice Q.* **2016**, *33*, 369–397. [CrossRef]
50. Harrell, K. The Predictive Accuracy of Hotspot Mapping of Robbery Over Time And Space. Ph.D. Thesis, University of Salford, Manchester, UK, 2014.
51. Hart, T.; Zandbergen, P. Kernel density estimation and hotspot mapping. *Polic. Int. J. Police Strateg. Manag.* **2014**, *37*, 305–323. [CrossRef]
52. Levine, N. The "Hottest" part of a hotspot: Comments on "The utility of hotspot mapping for predicting spatial patterns of crime". *Secur. J.* **2008**, *21*, 295–302. [CrossRef]

53. Van Patten, I.T.; McKeldin-Coner, J.; Cox, D. A microspatial analysis of robbery: Prospective hot spotting in a small city. *Crime Mapp. A J. Res. Pract.* **2009**, *1*, 7–32.
54. Turner, M.G. Landscape ecology: The effect of pattern on process. *Annu. Rev. Ecol. Syst.* **1989**, *20*, 171–197. [CrossRef]
55. Levine, N. Chapter 5: Distance Analysis I and II. In *CrimeStat IV: A Spatial Statistics Program for the Analysis of Crime Incident Locations, Version 4*; U.S. Department of Justice: Washington, DC, USA, 2013; Volume 4.
56. Drawve, G.; Moak, S.C.; Berthelot, E.R. Predictability of gun crimes: A comparison of hot spot and risk terrain modelling techniques. *Polic. Soc.* **2016**, *26*, 312–331. [CrossRef]
57. Caplan, J.M.; Kennedy, L.W.; Piza, E.L. Joint utility of event-dependent and environmental crime analysis techniques for violent crime forecasting. *Crime Delinq.* **2013**, *59*, 243–270. [CrossRef]
58. Von Neumann, J.; Morgenstern, O. *Theory of Games and Economic Behavior*; Princeton University Press: Princeton, NJ, USA, 1947.
59. Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.
60. Peterson, M. *An Introduction to Decision Theory*; Cambridge University Press: Cambridge, UK, 2013.
61. Robert, C.P. Bayesian Point Estimation. In *The Bayesian Choice*, 2nd ed.; Springer: New York, NY, USA, 2007; pp. 165–221.
62. Lomax, S.; Vadera, S. A Cost-Sensitive Decision Tree Learning Algorithm Based on a Multi-Armed Bandit Framework. *Comput. J.* **2016**, *60*, 941–956. [CrossRef]
63. Gil, C.; Rios Insua, D.; Rios, J. Adversarial risk analysis for urban security resource allocation. *Risk Anal.* **2016**, *36*, 727–741. [CrossRef] [PubMed]
64. Joshi, C.; Aliaga, J.R.; Insua, D.R. Insider threat modeling: An adversarial risk analysis approach. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 1131–1142. [CrossRef]
65. Rios Insua, D.; Ríos, J.; Banks, D. Adversarial risk analysis. *J. Am. Stat. Assoc.* **2009**, *104*, 841–854. [CrossRef]
66. Rios, J.; Insua, D.R. Adversarial risk analysis for counterterrorism modeling. *Risk Anal. Int. J.* **2012**, *32*, 894–915. [CrossRef] [PubMed]
67. Ferguson, A.G. *Predictive Policing Theory in the Cambridge Handbook of Policing in the United States*; Cambridge University Press: Cambridge, UK, 2019.
68. Richardson, R.; Schultz, J.M.; Crawford, K. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online* **2019**, *94*, 15.
69. Philips, P.J.; Pohl, G. Algorithms, human decision-making and predictive policing. *SN Soc. Sci.* **2021**, *1*, 1–21. [CrossRef]