PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

rsta.royalsocietypublishing.org

Research

Cite this article: Woodworth JT, Mohler GO, Bertozzi AL, Brantingham PJ. 2014 Non-local crime density estimation incorporating housing information. *Phil. Trans. R. Soc. A* **372**: 20130403.
http://dx.doi.org/10.1098/rsta.2013.0403

One contribution of 13 to a Theme Issue 'Partial differential equation models in the socio-economic sciences'.

Subject Areas:
applied mathematics, computational mathematics, differential equations, mathematical modelling, graph theory, statistics

Keywords:
crime hotspots, density estimation, graph Laplacian, maximum penalized likelihood estimation, non-local means, Nyström's extension

Author for correspondence:
J. T. Woodworth
e-mail: jwoodworth@math.ucla.edu

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsta.2013.0403 or via http://rsta.royalsocietypublishing.org.

Royal Society Publishing

Non-local crime density estimation incorporating housing information

J. T. Woodworth[1], G. O. Mohler[2], A. L. Bertozzi[1] and P. J. Brantingham[3]

[1]Department of Mathematics, and [3]Department of Anthropology, University of California, Los Angeles, CA 90095, USA
[2]Department of Mathematics and Computer Science, Santa Clara University, Santa Clara, CA 95053-0290, USA

ALB, 0000-0003-0396-7391

Given a discrete sample of event locations, we wish to produce a probability density that models the relative probability of events occurring in a spatial domain. Standard density estimation techniques do not incorporate priors informed by spatial data. Such methods can result in assigning significant positive probability to locations where events cannot realistically occur. In particular, when modelling residential burglaries, standard density estimation can predict residential burglaries occurring where there are no residences. Incorporating the spatial data can inform the valid region for the density. When modelling very few events, additional priors can help to correctly fill in the gaps. Learning and enforcing correlation between spatial data and event data can yield better estimates from fewer events. We propose a non-local version of maximum penalized likelihood estimation based on the H^1 Sobolev seminorm regularizer that computes non-local weights from spatial data to obtain more spatially accurate density estimates. We evaluate this method in application to a residential burglary dataset from San Fernando Valley with the non-local weights informed by housing data or a satellite image.

1. Introduction

In real-world applications, satellite images, housing data, census data and other types of geographical data become highly relevant for modelling the probability of a certain type of event. The methodology presented here provides us a general framework paired with fast algorithms for incorporating external information in density estimation computations.

In density estimation, one is given a discrete sample of event locations, drawn from some unknown density u on the spatial domain, and tries to approximately recover u [1]. Relating the events to the additional data allows one to search over a smaller space of densities, which can yield more accurate results with fewer events. We refer to the additional data source as the function $g(x)$ defined over the spatial domain Ω. We may typically assume two things about the relationship between g and u: (i) g informs the support of u via $g(x) = 0 \Rightarrow u(x) = 0$ and (ii) u varies smoothly with g in a non-local way (explained below). This method allows the additional information in g to significantly improve the recovery of u.

(a) Maximum penalized likelihood estimation

Although there are other classes of methods in the density estimation literature that are quite popular (such as average shifted histogram and kernel density estimation [2]), in this work we shall focus on maximum penalized likelihood estimation (MPLE). MPLE provides a general framework for finding an approximate density from sampled events. The likelihood of events occurring at the locations $\{x_i\}_{i=1}^n$ according to a proposed probability u is the product of the probability evaluated at each of those locations:

$$\mathcal{L}(u, \{x_i\}_{i=1}^n) = \prod_{i=1}^{n} u(x_i).$$

MPLE approximates u as the maximizer of a log-likelihood term combined with a penalty term, typically enforcing smoothness [3],

$$\hat{u} = \underset{u \geq 0,\ \int_\Omega u\,\mathrm{d}x = 1}{\operatorname{argmax}} \sum_{i=1}^{n} \log(u(x_i)) - P(u).$$

Without some kind of penalty term, the solution is just a weighted sum of Dirac deltas located at the training samples. Typical choices of $P(u)$ include the total variation (TV) norm, $P(u) = \lambda \int_\Omega |\nabla u|\,\mathrm{d}x$, and the H^1 Sobolev seminorm, $P(u) = (\lambda/2) \int_\Omega |\nabla u|^2\,\mathrm{d}x$. Here λ is the parameter that controls the amount of regularization. This is typically chosen via cross-validation, when it is computationally feasible.

(b) Maximum penalized likelihood estimation applied to crime

The H^1 seminorm is a common, well-understood regularizer in image processing related to Poisson's equation, the heat equation and the Weiner filter, producing visually smooth surfaces. For this reason, it is often a default choice when little is known about the data being modelled. H^1 MPLE has further justification in crime density estimation from the 'broken window' effect [4–6]. This observation states that, after a burglary has occurred at a given house, burglaries are more likely to occur at the same house or nearby houses for some period of time afterwards. Initial burglaries give criminals information about what valuables remain and the schedule of inhabitants in the area. Additionally, a successful burglary leaves environmental clues, such as broken windows, that indicate that an area is more crime-tolerant than others. This effect leads to repeat and near-repeat burglaries. More generally, criminals tend to move in a bounded region around a few key nodes and have limited awareness of potential for criminal activity outside of familiar areas [7–9]. Within neighbourhoods, risk factors are typically homogeneous [10–12]. All of this explains why observed incidence rates of burglaries are locally smooth.

However, local smoothness is not always appropriate, and in practice there is much room for improvement. In recent years, several studies on the application of MPLE to crime data [13–15] emphasize the fact that crime density should have boundaries corresponding to the local geography. Mohler *et al.* [13] and Kostić *et al.* [15] model this by choosing penalty functions that are edge-preserving, TV and Ginzburg–Landau, respectively. Smith *et al.* [14] more closely follow the idea presented here. That work introduces a modified H^1 MPLE, which based the penalty term on an additional component of the data. The method assumes that the valid region of the probability density estimate is known *a priori*. In their application to residential burglary, the valid region was the approximate support of the housing density in the region. If we denote the valid region by D, then the modified penalty term is just a standard H^1 MPLE with a factor z_ϵ^2 in the integral, where z_ϵ is a smooth Ambrosio–Tortorelli approximation of $(1-\delta(\partial D))$:

$$\hat{u} = \underset{u\geq 0,\, \int_\Omega u = 1}{\operatorname{argmin}} \frac{1}{2}\int_\Omega z_\epsilon^2 |\nabla u|^2 \,\mathrm{d}x - \mu \sum_{i=1}^{n} \log(u(x_i)),$$

$$z_\epsilon(x) = \begin{cases} 1 & \text{if } d(x,\partial D) > \epsilon, \\ 0 & \text{if } x \in \partial D. \end{cases}$$

(c) Graph-based methods

In spectral graph theory, data are represented as nodes of a weighted graph, where the weight on each edge indicates the similarity between the two nodes. Such data structures have been very successfully applied to data clustering problems and image segmentation [16–18]. The standard theory behind this is described in [19,20] and a tutorial on spectral clustering is given in [21]. A theory of non-local calculus was developed first by Zhou & Schölkopf in 2004 [22] and put in a continuous setting by Gilboa & Osher in 2008 [23]. Such methods were originally used for image denoising [23,24], but the general framework led to methods for inpainting, reconstruction and deblurring [25–29]. Compared with local methods, non-local methods are generally better able to handle images with patterns and texture. Further, by choosing an appropriate affinity function, the methods can be made suitable for a wide variety of different datasets, not just images.

In this article, we present non-local (NL) H^1 MPLE, which modifies the standard H^1 MPLE energy to account for spatial inhomogeneities, but unlike Smith *et al.* [14], we do so in a non-local way, which has the benefit of leveraging recent fast algorithms and the potential to generalize to other applications.

The organization of this article is as follows. In §2, we introduce the NL H^1 MPLE method and review the non-local calculus and numerical methods on which it is based. In §3, we demonstrate the advantages of NL H^1 MPLE by comparing it with standard H^1 MPLE when applied to modelling residential burglary. In §4, we summarize our conclusions and discuss directions for future research.

2. Non-local crime density estimation

We propose replacing the H^1 seminorm regularizer of H^1 MPLE with a linear combination of an H^1 regularizer and a non-local smoothing term $\iint_{\Omega\times\Omega} (\nabla_{w,s} u(x,y))^2 \,\mathrm{d}x\,\mathrm{d}y$, where $\nabla_{w,s}$ denotes the non-local symmetric-normalized gradient depending on an affinity function w derived from the spatial data, g. More details are found in §2*b*. The energy we optimize is thus

$$\hat{u} = \underset{u\geq 0,\, \int_\Omega u = 1}{\operatorname{argmax}} \sum_{i=1}^{n} \log(u(x_i)) - \alpha \iint_{\Omega\times\Omega} (\nabla_{w,s} u(x,y))^2 \,\mathrm{d}x\,\mathrm{d}y - \frac{\beta}{2}\int_\Omega |\nabla u(x)|^2 \,\mathrm{d}x. \tag{2.1}$$

The non-local term in equation (2.1) is tolerant of sharp changes in the probability density estimate, as long as they coincide with sharp non-local changes in the spatial data. The mathematical formulation of this statement follows from the definitions presented in the following sections and is presented in appendix A. Before reviewing the non-local calculus behind this energy, we motivate why a non-local regularizer is good for crime density estimation. Many cities grow in a dispersal colony-like fashion, i.e. colony patches start growing at dispersed locations at the same time with the same architectural or cultural model as a starting point, generating non-local similarities [30]. Dissimilar colony patches grow and meet to form diffuse interface-like boundaries [31]. Thus, housing data typically contain similar features spread across the domain, along with interfaces between different types of areas. Whereas opposite sides of these interfaces are spatially close, they are non-locally well separated.

The clearest advantage of non-local regularization is that it allows for sharp changes in crime density across interfaces of distinct housing regions. In particular, because the residential areas are non-locally well separated from the non-residential areas, the non-local regularized estimate correctly captures the support of the residential burglary density. This feature has been studied for its own sake in prior work, and non-local regularization addresses it in an automatic, hands-off way.

Another, more subtle advantage of non-local regularization is that it encourages distant, but non-locally similar regions (e.g. colony patches based on the same model) to have similar crime density values. The assumption is that the layout of a neighbourhood and its crime density are both tied to underlying socio-economic factors. When one has these relevant factors, one can perform risk terrain modelling [12], combining the factors in the way that is most consistent with the observed data. Non-local regularization implicitly measures correlation between housing features and levels of crime, presumably explained by these unknown factors. The regularization encourages those relationships to remain consistent across the entire domain and all data. In this work, we base the non-local similarity of two locations on the similarity of surrounding housing density patches. For the sake of simplicity, one could consider basing it on only the housing density in the immediate vicinity. This would encourage the crime density to be a smooth function of the immediate housing density. Probably, one would estimate residential burglaries as roughly proportional to the housing density. This would be a simple, but reasonable null model, assuming that burglary depends heavily on opportunity. One would balance the spatial smoothness and smoothness as a function of housing density with cross-validation, allowing for varying results depending on what the data show. Our non-local weights are based on housing density patches, which makes them more noise-robust and representative of more complex housing features. This approach is general, relates to previous work in image processing and produces favourable results.

(a) Non-local means

The technique of non-local means was originally developed for the application of image denoising, but can also be interpreted as an affinity function. The formula for the non-local means affinity, $w_{\mathbf{Im}}$, is given by [24]

$$w_{\mathbf{Im}}(x,y) = \exp\left(-\frac{(K_r * |\mathbf{Im}(x+\cdot) - \mathbf{Im}(y+\cdot)|^2)(0)}{\sigma^2}\right). \tag{2.2}$$

Here, **Im** is the image on which the non-local means weights are based, K_r is a non-negative weight kernel of size $(2r+1)\times(2r+1)$ and σ is a scaling parameter. This function measures similarity between two pixels based on a weighted ℓ^2 difference between patches surrounding them in the image. In our experiments, the image **Im** is either a housing image or a satellite image. In practical settings, computing and storing all function values of w is a very computationally intensive task, so we use the fast approximation: Nyström's extension (see §2*d*).

(b) Non-local calculus and graphs

Non-local calculus was introduced in its discrete form by Zhou & Schölkopf [22] and put in a continuous framework by Gilboa & Osher [23]. In these definitions, $w(x,y)$ is a general non-negative symmetric affinity function, which generally measures similarity between the points x and y.

Let $\Omega \subset \mathbb{R}^n$, and $u(x)$ be a function $u : \Omega \to \mathbb{R}$. Then, the non-local gradient of u at the point $x \in \Omega$ in the direction of $y \in \Omega$ is given by

$$(\nabla_w u)(x,y) = (u(y) - u(x))\sqrt{w(x,y)}.$$

This suggests an analogous generalization of divergence, which in turn leads to the following definition of the non-local Laplacian:

$$\Delta_w u(x) = \int_\Omega (u(y) - u(x))w(x,y)\,\mathrm{d}y. \tag{2.3}$$

Now, let $\{p_i\}_{i=1}^n$ be a discrete subset of Ω and let $w_{ij} = w(p_i, p_j)$ if $i \neq j$ and $w_{ii} = 0$. We then let $\{p_i\}_{i=1}^n$ be vertices and w_{ij} the edge weights on a weighted graph. Let $d_i = \sum_{j=1}^n w_{ij}$ be the weighted degree of the ith node. Then, the graph Laplacian applied to the function on the graph, u, is given by Lu where

$$L_{ij} = \begin{cases} d_i & \text{if } i = j, \\ -w_{ij} & \text{otherwise,} \end{cases} \quad \text{and so} \quad (Lu)_i = \sum_{j=1}^n (u_i - u_j)w_{ij}.$$

To keep the spectrum of the graph Laplacian in a fixed range as the number of samples in increased, and thus to guarantee consistency, we must normalize the graph Laplacian. See Bertozzi & Flenner [32] for a more in-depth discussion of this. We use the symmetric normalization.

$$L_{\mathrm{sym}} := D^{-1/2}LD^{-1/2}, \quad D_{ij} = \begin{cases} d_i & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Because we express our energy as applied to functions over continuous domains, we also introduce the following notation for the symmetric-normalized non-local gradient:

$$\nabla_{w,s} u(x,y) := \frac{\nabla_w u(x,y)}{(\int_\Omega w(x,z)\,\mathrm{d}z \int_\Omega w(y,z)\,\mathrm{d}z)^{1/4}}.$$

(c) Numerical optimization

We must numerically find an approximate solution. The unconstrained energy has gradient flow

$$u_t = \alpha \Delta_{w,s} u + \beta \Delta u + \frac{1}{u}\sum_{i=1}^n \delta(x - x_i).$$

We evolve this equation, projecting onto the space of probability densities after each step. We discretize the equation as

$$\frac{u^{k+1} - u^k}{\delta t} = -\alpha L_{\mathrm{sym}} u^{k+1} + \beta \Delta_h u^{k+1} + \frac{1}{u^k}\sum_{i=1}^n \delta(x - x_i).$$

Here, Δ_h denotes the discrete Laplacian from the five-point finite difference stencil with mesh size $h = 1$. Solving for u^{k+1} yields

$$u^{k+1} = (I + \alpha\delta t L_{\mathrm{sym}} - \beta\delta t \Delta_h)^{-1}\left(\frac{\delta t}{u^k}\sum_{i=1}^n \delta(x - x_i) + u^k\right).$$

Table 1. Non-local H^1 MPLE algorithm.

Nyström ($\mathbf{Im}_g$) $\rightarrow \Phi, \Lambda$: $L_{\text{sym}} \approx \Phi \Lambda \Phi^{\mathrm{T}}$.
Initialize $u^0 \equiv 1/|\Omega|$, succDiff $= \infty, k = 0$.
while succDiff $> 10^{-7}$ and $k <$ maxSteps $= 800$

- $k = k + 1$
- $\boldsymbol{b} = \Phi^{\mathrm{T}} \left[u^{k-1} + \dfrac{\delta t}{u^{k-1}} \displaystyle\sum_{i=1}^{n} \delta(x - x_i) \right]$
- $a_i = \dfrac{b_i}{1 + \alpha(\delta t/2)\lambda_i}$
- $\boldsymbol{u}^{k-1/2} = \Phi \boldsymbol{a}$
- $\boldsymbol{b} = \text{fft2} \left[u^{k-1/2} + \dfrac{\delta t}{u^{k-1/2}} \displaystyle\sum_{i=1}^{n} \delta(x - x_i) \right]$
- $a_i = \dfrac{b_i}{1 + 2\beta\delta t\pi^2(m^2 + n^2)}$, $i \sim (m, n)^{\text{th}}$ Fourier mode,
  $a_1 = 1$ (guarantees integral 1 constraint)
- $\boldsymbol{u}^k = \text{ifft2}(\boldsymbol{a})$
- $\boldsymbol{u}^k = \max(\boldsymbol{u}^k, 0)$
- succDiff $= \|u^k - u^{k-1}\|_2^2 / \|u^k\|_2^2$

To approximate this, we use a split-time method:

$$u^{k+1/2} = \left(I + \alpha \frac{\delta t}{2} L_{\text{sym}}\right)^{-1} \left(\frac{\delta t}{u^k} \sum_{i=1}^{n} \delta(x - x_i) + u^k\right)$$

and

$$u^{k+1} = \left(I - \beta \frac{\delta t}{2} \Delta_h\right)^{-1} \left(\frac{\delta t}{u^{k+1/2}} \sum_{i=1}^{n} \delta(x - x_i) + u^{k+1/2}\right).$$

To apply these operators, we use a spectral method. This has two advantages over forming and multiplying the matrices. First, we can approximate the projection onto the constraint by using the spectral decomposition of the discrete Laplacian (shown in table 1). Second, the computation required to form and apply the entire symmetric graph Laplacian is also intensive. Fortunately, we can apply Nyström's extension (discussed in §2*d*), which is a popular method for approximating a portion of the eigenvectors and eigenvalues that approximate the operator well. To project onto the eigenvectors of Δ_h, we apply the two-dimensional fast Fourier transform.

In both the case of applying $(I + \alpha(\delta t/2)L_{\text{sym}})^{-1}$ and $(I - \beta(\delta t/2)\Delta_h)^{-1}$, we are applying operators of the form $(I + \delta t P)^{-1}$, where P is symmetric and positive semidefinite. In general, if P has spectral decomposition $P = \Phi \Lambda \Phi^{\mathrm{T}}$, then we apply $(I + \delta t P)^{-1}$ to $\boldsymbol{w}$ by first projecting onto the eigenvectors, $\boldsymbol{a} = \Phi^{\mathrm{T}} \boldsymbol{w}$, updating the coefficients $\tilde{a}_m = a_m/(1 + \delta t \lambda_m)$, and finally transforming back to the standard basis, $(I + \delta t P)^{-1} \boldsymbol{w} = \Phi \tilde{\boldsymbol{a}}$. We summarize the steps of our algorithm in table 1.

(d) Nyström's extension

To apply the spectral method described in the previous section, we need to approximate the eigenvectors and eigenvalues of the symmetric graph Laplacian. Here, we present Nyström's extension method and refer the reader to [25,32,33] for further discussion and analysis. Nyström's extension is a technique for performing matrix completion, well known within the spectral graph theory community. In this setting, Nyström's extension is applied to the normalized affinity matrix $W_{\text{sym}} = D^{-1/2} W D^{-1/2}$, where the (i, j)th entry of W is the affinity between nodes i and

j. Note that the matrices W_{sym} and L_{sym} have the same eigenvectors, and λ is an eigenvalue of W_{sym} if and only if $1-\lambda$ is an eigenvalue of L_{sym}.

We let N denote the set of nodes in our complete weighted graph, then take X to be a small random sample from N, and Y its complement. Up to a permutation of the nodes, we can write the affinity matrix as

$$W = \begin{pmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YY} \end{pmatrix},$$

where the matrix $W_{XY} = W_{YX}^{\mathrm{T}}$ consists of weights between nodes in X and nodes in Y, W_{XX} consists of weights between pairs of nodes in X, and W_{YY} consists of weights between pairs of nodes in Y. Nyström's extension approximates the eigenvalues and eigenvectors of the affinity matrix by manipulating the approximation:

$$W \approx \hat{W} = \begin{pmatrix} W_{XX} \\ W_{YX} \end{pmatrix} W_{XX}^{-1} (W_{XX} \quad W_{XY}).$$

This approximates $W_{YY} \approx W_{YX} W_{XX}^{-1} W_{XY}$. The error due to this approximation is determined by how well the rows of W_{XY} span the rows of W_{YY}. If the affinity matrix W is positive semidefinite, then we can write it as a matrix transpose times itself: $W = V^{\mathrm{T}} V$. In [34], the authors show that the Nyström extension thus approximates the unknown part of V (corresponding to W_{YY}) by orthogonally projecting it onto the range of the known part (corresponding to W_{XY}). In this setting, it is clear that, as the size of X grows, the approximation improves. Further, a random choice of X is likely to yield W_{XY} full rank if the rank of W is sufficiently large.

Next, we must incorporate the normalization factors into the above approximation. The degrees are approximated by applying their definition to the approximation. Note that $d_i = \sum_{j=1}^{n} w_{ij}$ can also be written as $d = W\mathbf{1}_n$, where $\mathbf{1}_n$ is the length-n vector of ones. This yields

$$\begin{aligned} \hat{d}_X &= W_{XX}\mathbf{1}_{|X|} + W_{XY}\mathbf{1}_{|Y|}, \\ \hat{d}_Y &= W_{YX}\mathbf{1}_{|X|} + W_{YX}W_{XX}^{-1}W_{XY}\mathbf{1}_{|Y|}. \end{aligned}$$

In this way, we approximate the degrees without forming any matrices of size larger than $|X| \times |Y|$. Define also the vectors $s_X = d_X^{-1/2}$ and $s_Y = d_Y^{-1/2}$. Normalizing our approximation of W gives

$$W_{\text{sym}} \approx \hat{W}_{\text{sym}} = \begin{pmatrix} W_{XX} \odot (s_X s_X^{\mathrm{T}}) & W_{XY} \odot (s_X s_Y^{\mathrm{T}}) \\ W_{YX} \odot (s_Y s_X^{\mathrm{T}}) & (W_{YX} W_{XX}^{-1} W_{XY}) \odot (s_Y s_Y^{\mathrm{T}}) \end{pmatrix},$$

where $\odot$ denotes component-wise product. For notational convenience going forward, let us define $W_{XX}^{\text{sym}} = W_{XX} \odot (s_X s_X^{\mathrm{T}})$ and $W_{XY}^{\text{sym}} = W_{XY} \odot (s_X s_Y^{\mathrm{T}})$.

In practice, one uses a diagonal decomposition of such a formula to avoid forming and applying the full matrix. It follows from the analysis discussed in [33] that, if W_{XX}^{sym} is positive definite, the diagonal decomposition of the approximation is given by $\hat{W}_{\text{sym}} = V \Lambda_S V^{\mathrm{T}}$, where

$$S = W_{XX}^{\text{sym}} + (W_{XX}^{\text{sym}})^{-1/2} W_{XY}^{\text{sym}} W_{YX}^{\text{sym}} (W_{XX}^{\text{sym}})^{-1/2},$$

S has diagonal decomposition $S = U_S \Lambda_S U_S^{\mathrm{T}}$, and

$$V = \begin{bmatrix} W_{XX}^{\text{sym}} \\ W_{YX}^{\text{sym}} \end{bmatrix} (W_{XX}^{\text{sym}})^{-1/2} U_S \Lambda_S^{-1/2}.$$

Note that S is size $|X| \times |X|$ and V is size $|N| \times |X|$. Their computation never requires computing or storing matrices larger than size $|N| \times |X|$. Thus, V is a matrix of $|X|$ approximate eigenvectors of W_{sym} with corresponding eigenvalues Λ_S. For more detailed discussion on Nyström's extension, see [25,32,33].

(e) Cross-validation

Cross-validation is a methodology for choosing the smoothing parameter λ which yields probability densities that are predictive of the missing data [35]. Because our method consists primarily of simple coefficient updates after mapping to different eigenspaces, it is fast relative to methods with similar goals [14]. This speed increase allows us to perform 10-fold cross-validation, which requires many evaluations of the density estimation method. In V-fold cross-validation, we randomly partition the data points into V disjoint subsets $X = \bigcup_{v=1}^{V} X_v$ with complements $X_{-v} = X \backslash X_v$. We let $u_{\lambda,-v}$ denote the density estimate using parameter λ trained on the data X_{-v}. The objective we minimize is an application of the Kullback–Leibler divergence, an asymmetric distance measure for probabilities given by

$$D_{\text{KL}}(p, q) = \int_{\Omega} \log\left(\frac{p(x)}{q(x)}\right) p(x)\, \mathrm{d}x.$$

We select the parameter λ that minimizes the average KL divergence between the density estimates, $u_{\lambda,-v}$, and the discrete distributions on the withheld data points,

$$p_v(x) = \frac{1}{|X_v|} \sum_{x_i \in X_v} \delta(x - x_i).$$

This yields the following optimization:

$$\begin{aligned} \hat{\lambda} &= \operatorname*{argmin}_{\lambda} \frac{1}{V} \sum_{v=1}^{V} D_{\text{KL}}(p_v, u_{\lambda,-v}) \\ &= \operatorname*{argmax}_{\lambda} \frac{1}{V} \sum_{v=1}^{V} \sum_{x_i \in X_v} \log(u_{\lambda,-v}(x_i)). \end{aligned}$$

The result can also be interpreted as maximizing the average log-likelihood that the missing events are drawn from the corresponding estimated densities. We approximate this optimization via a grid search (note that $\lambda = (\alpha, \beta)$ is two-dimensional). The search requires the computation of all the density estimates $u_{\lambda,-v}$. In particular, for 10-fold cross-validation, we must compute $10 \times |\alpha \text{ values}| \times |\beta \text{ values}|$ densities.

When evaluating the energy, it is important to ensure that non-negativity and sum-to-unity constraints hold strictly for the input densities. If a density is slightly negative somewhere, it could add complex terms to the objective, and if a density has sum slightly larger than unity, it could unfairly achieve a slightly higher objective. Further, in the strictest interpretation, if a density has a value zero at the location of a missing event, the objective will take value $-\infty$. We relax this penalty by replacing $u_{\lambda,-v}(x_i)$ with $\max\{u_{\lambda,-v}(x_i), 10^{-16}\}$.

3. Numerical experiments

Here, we demonstrate the advantage of the NL H^1 MPLE method over standard H^1 MPLE by evaluating its performance on residential burglary data from San Fernando Valley, Los Angeles, CA, using corresponding housing data and a satellite image to inform the non-local weights.

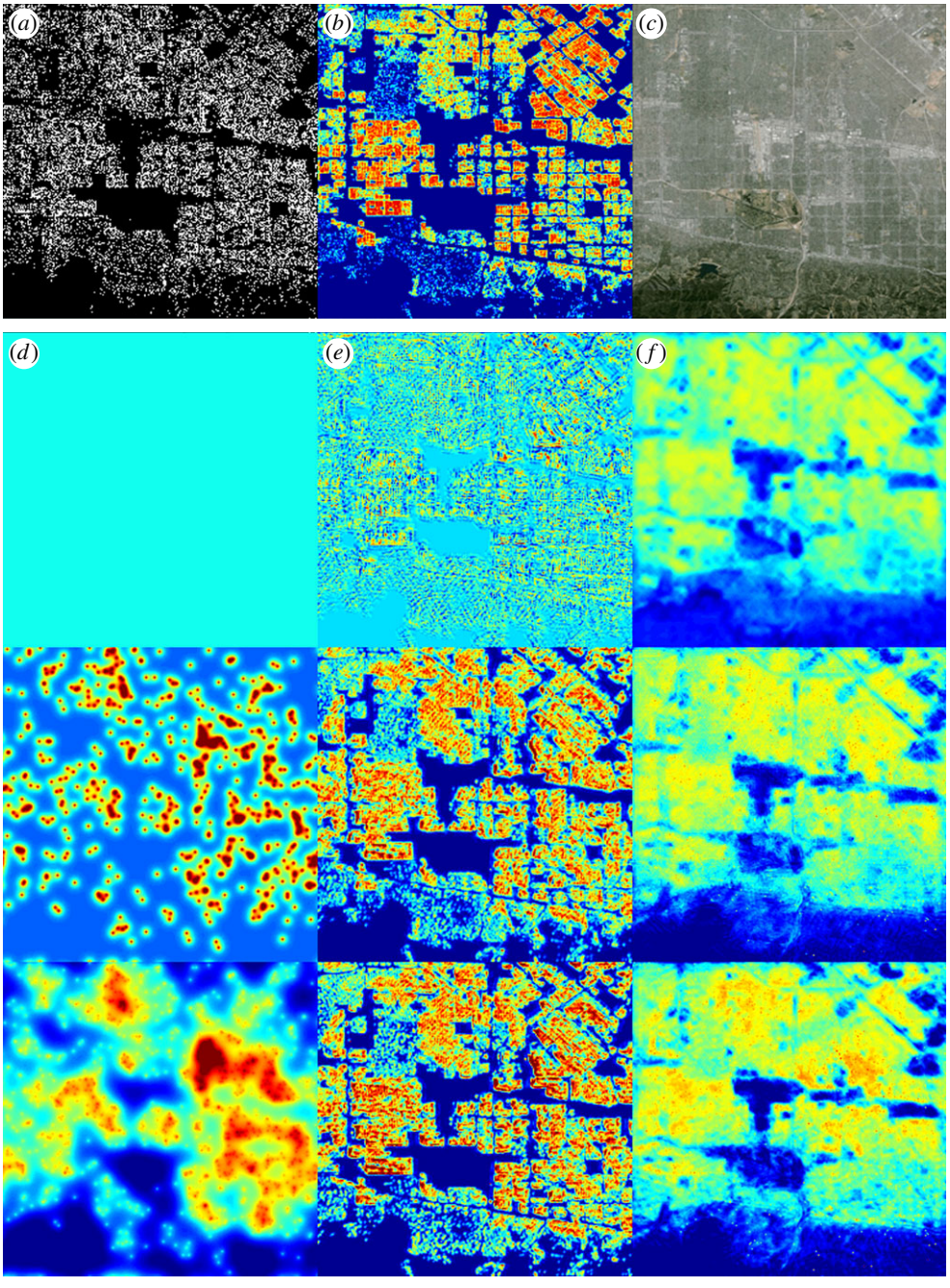

Figure 1. Top row: data. (*a*) 2005–2013 residential burglaries in San Fernando Valley (from LAPD). (*b*) San Fernando Valley log(min(# housing units, 7) + 1) (from LA County Tax Assessor). (*c*) Satellite image of San Fernando Valley (from Google Maps). Bottom three rows: MPLE of 50, 500 and 1000 random samples from 2008 residential burglaries. (*d*) Column 1: H^1 MPLE. (*e*) Column 2: housing NL H^1 MPLE. (*f*) Column 3: satellite NL H^1 MPLE. (Online version in colour.)

(a) Residential burglary

We perform experiments on residential burglary data from San Fernando Valley in 2005–2013, getting substantially different results from those shown in [13–15]. In figure 1, we show the data used (locations of residential burglaries in figure 1*a*, housing in figure 1*b* and satellite image in

Table 2. Log-likelihood of densities on residential burglaries from 2009 to 2013 (corrected and raw). The maximum value in each row is shown in *italics*.

training dataset (corrected)	scaled histogram	H^1	housing NL H^1
50 random from 2008	-3.6039×10^5	*-1.3386×10^5*	-1.3396×10^5
100 random from 2008	-3.5991×10^5	*-1.3369×10^5*	-1.3369×10^5
500 random from 2008	-3.5197×10^5	-1.3282×10^5	*-1.3004×10^5*
1000 random from 2008	-3.4350×10^5	-1.3246×10^5	*-1.2953×10^5*
2008	-3.1905×10^5	-1.3189×10^5	*-1.2888×10^5*
2007–2008	-2.9846×10^5	-1.3174×10^5	*-1.2850×10^5*
2006–2008	-2.8152×10^5	-1.3136×10^5	*-1.2815×10^5*
2005–2008	-2.6847×10^5	-1.3121×10^5	*-1.2774×10^5*
training dataset (raw)	**scaled histogram**	**H^1**	**satellite NL H^1**
50 random from 2008	-3.6959×10^5	-1.3733×10^5	*-1.3553×10^5*
100 random from 2008	-3.6822×10^5	-1.3732×10^5	*-1.3553×10^5*
500 random from 2008	-3.6342×10^5	-1.3583×10^5	*-1.3524×10^5*
1000 random from 2008	-3.5733×10^5	-1.3598×10^5	*-1.3525×10^5*
2008	-3.3313×10^5	-1.3535×10^5	*-1.3494×10^5*
2007–2008	-3.1326×10^5	-1.3525×10^5	*-1.3482×10^5*
2006–2008	-2.9630×10^5	-1.3496×10^5	*-1.3449×10^5*
2005–2008	-2.8334×10^5	-1.3488×10^5	*-1.3431×10^5*

figure 1*c*), and H^1 MPLE (figure 1*d*), housing-based NL H^1 MPLE (figure 1*e*) and satellite-based NL H^1 MPLE (figure 1*f*) density estimates on increasing subsets of data from 2005 to 2008. To evaluate performance, we compute the log-likelihood of each density on the residential burglaries from 2009 to 2013 (shown in table 2).

As one would predict, the locations of residential burglaries in figure 1*a* are primarily restricted to the support of the housing density image in figure 1*b*. There are some locations in the burglary dataset that correspond to locations with no residences (4173 events out of 23 725 total), which we attribute to imprecision in the burglary data. Most such misplaced events occur on streets, suggesting that the actual event took place at a residence facing that street. Because of this inconsistency between the datasets, for the experiments that use the housing data, we adjust the residential burglary data for training and testing (for both H^1 and NL H^1), moving each event to the nearest house if it is within two pixels, and dropping the event otherwise. This results in 603 dropped events. For the experiments that do not use housing data, we work with the raw burglary data for training and testing.

We implement H^1 MPLE by applying our algorithm, described in table 1, with $\alpha = 0$ and $\Phi =$ Id. We choose the value of the regularization parameter β for each training dataset by performing 10-fold log-likelihood cross-validation, searching over $\beta =$ `[0,10.^(-2:8)]`. We apply H^1 MPLE to both the raw and corrected burglary data.

For housing-based NL H^1 MPLE, we perform Nyström's extension with non-local means applied to g, the housing density image shown in figure 1*b*. We use 400 random samples for Nyström's extension. We use the first 300 eigenvectors and eigenvalues in our computations. The non-local means weights are based on differences between patches of size 11×11 and $\sigma = 1 \cdot \mathrm{std}(g)$, the standard deviation of the housing image. The weight kernel K_r, $r = 5$, is given as follows:

$$K_r(1+r+i, 1+r+j) = \frac{1}{r} \sum_{d=\max(|i|,|j|,1)}^{r} \frac{1}{(2d+1)^2}, \quad i,j = -r, \ldots, r.$$

To choose the regularization parameters α and β, we perform 10-fold log-likelihood cross-validation, searching over $\alpha = $ `[0,10.^(-2:12)]` and $\beta = $ `[0,10.^(-2:8)]`. We apply housing NL H^1 MPLE to the corrected burglary data.

For satellite-based NL H^1 MPLE, we perform Nyström's extension with non-local means applied to g, the Google Maps image shown in figure 1*c*. In applying non-local means to a colour image, we interpret the image as a vector-valued function with three components (one for each colour channel) and so, in equation (2.2), the expression $|\mathbf{Im}(x+\cdot) - \mathbf{Im}(y+\cdot)|^2$ has size $(2r+1)\times(2r+1)\times 3$. We use 800 random samples for Nyström's extension. We use the first 600 eigenvectors and eigenvalues in our computations. The non-local means weights are based on differences between patches of size 11×11 and $\sigma = 1\cdot \mathrm{std}(g)$, the standard deviation of the Google Maps image. The weight kernel is as in the previous case, but repeated on each colour channel. To choose the regularization parameters α and β for each training set, we perform 10-fold log-likelihood cross-validation, searching over $\alpha = $ `[0,10.^(-2:12)]` and $\beta = $ `[0,10.^(-2:8)]`. We apply satellite NL H^1 MPLE to the raw burglary data.

The H^1 MPLE results transition from a completely smooth uniform density to a probability density with more apparent structure as the amount of training data increases. The NL H^1 MPLE housing and satellite results exhibit a similar trend, but are able to better approximate the correct support of the density with many fewer data points. The measurable benefit of non-local smoothing is shown by the log-likelihood values in table 2. NL H^1 generally gets higher log-likelihood than H^1. This means that the densities estimated by housing NL H^1 on corrected 2005–2008 data are more congruous with the corrected 2009–2013 data than the H^1 densities, and the densities estimated by satellite NL H^1 on raw 2005–2008 data are more congruous with the raw 2009–2013 data than the H^1 densities.

The added complexity of our algorithm results in an increase in run time from the standard H^1 MPLE, but the difference is not too substantial. We compare run times on a laptop with one Intel core i7 processor that has two cores with processor speed 2.67 GHz and 4 GB of memory. The run time for Nyström applied to the housing image is typically about 17 s. The run time for Nyström applied to the satellite image is typically about 36 s. For cross-validation purposes, Nyström can be run once outside of the loop and the results used for all combinations of datasets and parameters. The run time for H^1 MPLE with parameters as chosen by cross-validation on the residential burglaries from 2005 to 2008 is typically about half a second. The run time for housing NL H^1 MPLE with parameters as chosen by cross-validation on the residential burglaries from 2005 to 2008 is typically about 2.3 s. The run time for satellite NL H^1 MPLE with parameters as chosen by cross-validation on the residential burglaries from 2005 to 2008 is typically about 1.5 s. The cross-validation run times depend on what range of parameters are being tested, but can easily be run in parallel across several computing nodes.

(b) Synthetic density

To further verify that NL H^1 MPLE is correctly performing density estimation, we test the method's ability to recover a given density. We start with a known density, draw events from it and attempt to recover it. Because the method assumes a relationship between the spatial data g and the density u, we generate a synthetic density that is closely related to the housing data, shown in the bottom left of figure 2. This density is given by taking a random linear combination of the first five approximated eigenvectors of the graph Laplacian (with weights based on the housing image) and then shifting and normalizing the result to yield a probability density. The coefficients are chosen uniformly at random in $[0,1]$ and the non-local weights are based on the housing data as they were in the previous section. This randomly generated density was chosen over others because it looks like a potential probability density for residential burglary. It should be noted that this choice of synthetic density is quite ideal for the proposed method. The hope is that very good density recovery of ideal probability densities extends to good density recovery of less ideal probability densities.

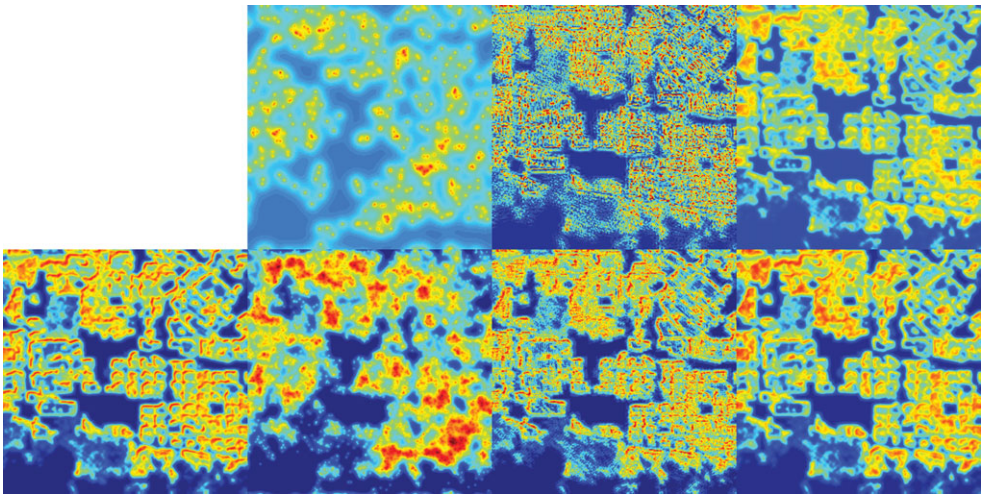

Figure 2. Synthetic density recovery (see §3*b*). Top row: density estimates based on 400 samples from synthetic density. $\overline{|\text{error}|}$: H^1 7.12473×10^{-6}, NL H^1 5.26617×10^{-6}, NL H^1 restricted 2.55042×10^{-6}. Bottom row: synthetic density and density estimates on 4000 samples. $\overline{|\text{error}|}$: H^1 5.05662×10^{-6}, NL H^1 2.52831×10^{-6}, NL H^1 restricted 1.36416×10^{-6}. (Online version in colour.)

We sample the events according to this density by generating numbers uniformly at random in $[0, 1]$ and inverting the cumulative distribution function associated with the density. In the top row of figure 2, we show the H^1 MPLE result on the 400 events ($\beta = 5 \times 10^4$), the housing NL H^1 MPLE result on the 400 events ($\alpha = 100, \beta = 0$) and the NL H^1 MPLE result on 400 events restricted to the first five eigenvectors. In the bottom row of figure 2, we show the synthetic density, the H^1 MPLE result on the 4000 events ($\beta = 10^5$), the housing NL H^1 MPLE result on the 4000 events ($\alpha = 10^8, \beta = 0$) and the NL H^1 MPLE result on 4000 events restricted to the first five eigenvectors. In all the cases, smoothing parameters were chosen to minimize the mean absolute error of the probability density. The NL H^1 results and the restricted NL H^1 results do a substantially better job at recovering the probability density than H^1 MPLE. This is expected of course, from the construction of the probability. The comparison merely suggests that, if the correct density is well approximated by a combination of eigenvectors of the graph Laplacian, enforcing non-local smoothness can substantially improve recovery of the density. It is, in general, difficult to determine when a density is well approximated by a graph Laplacian's eigenbasis. The assumption is that the primary and non-local data have some meaningful, consistent connection. We refer the reader to §2 for heuristics on this connection and the appendix for some more precise formulations. It is also worth noting that, if unrelated non-local data are used, cross-validation will probably yield $\alpha = 0$, reverting the model back to standard H^1 MPLE.

4. Conclusion and future work

In this paper, we have looked at the problem of obtaining spatially accurate probability density estimates. The need for new approaches is demonstrated by the inadequate performance of standard techniques such as H^1 MPLE.

Our proposed solution accomplishes this by incorporating a non-local regularity term based on the H^1 regularizer and non-local means, which fuses geographical information into the density estimate. Our experiments with the San Fernando Valley residential burglary dataset demonstrate that this method does yield a probability density estimate with the correct support that also gives favourable log-likelihood results. Further, our results based on the Google Maps image suggest

that we can apply NL H^1 MPLE to a wide variety of geographical regions without obtaining specialized geographical data.

There are several other aspects of this and related problems to explore. In general, testing the method on other datasets would be interesting. This may present the added challenge of dealing with other types of geographical information, because high-resolution housing density data may not be readily available. In modelling the density of other types of events, the geographical data may not be related to housing at all. As the problem dictates, the non-local weights can be replaced with whatever weights seem appropriate for the data at hand. We have yet to incorporate time, leading indicators of crime, or census data into the model. Any of these could further improve results and allow one to use density estimation in place of risk terrain modelling.

Finally, our method need not stand alone. Several sophisticated spatio-temporal models for probabilistic events make use of density estimation, typically using the standard methods [36–38]. By replacing the standard density estimation techniques with a non-locally regularized MPLE such as ours, the density estimates in these models could improve, thus improving the overall result of the resulting simulation.

Acknowledgements. The authors thank the LAPD for the residential burglary dataset, and the NSF Human Social Dynamics Program (BCS-0527388) for purchasing the housing data from the LA County Tax Assessor. The authors obtained the satellite image from Google Maps.

Data accessibility. The crime data cannot be shared because it contains human subject data. The housing data, satellite image and synthetic density are all available in the electronic supplementary material.

Funding statement. This work was supported by NSF grants no. DMS-0968309 and DMS-1417674, the W. M. Keck Foundation, ONR grants no. N000141210040 and N000141210838, AFOSR MURI grant no. FA9550-10-1-0569, ARO grant no. W911NF1010472 and NSF grant no. DGE-1144087.

Appendix A

To examine the effect of the non-local regularization term, we compute an alternative formulation of the NL H^1 MPLE problem and derive an inequality that solutions must satisfy. Recall from equation (2.1) that NL H^1 MPLE applied to the event samples $X = \{x_i\}_{i=1}^n$ with parameters $\alpha, \beta \geq 0$ is given by the following optimization:

$$u_{\alpha,\beta,X} := \underset{u\geq 0,\, \int_\Omega u = 1}{\operatorname{argmax}} \sum_{i=1}^{n} \log(u(x_i)) - \alpha \iint_{\Omega\times\Omega} (\nabla_{w,s} u(x,y))^2 \,\mathrm{d}x\,\mathrm{d}y - \frac{\beta}{2}\int_\Omega |\nabla u(x)|^2 \,\mathrm{d}x.$$

For every such X, α, β, one can show that there exist non-negative constants C_1 and C_2 such that $u_{\alpha,\beta,X}$ is also the solution to a more constrained optimization:

$$u_{\alpha,\beta,X} = \operatorname{argmax} \sum_{i=1}^{n} \log(u(x_i)) \quad \text{subject to}$$
$$\left\{ u \geq 0,\ \int_\Omega u = 1,\ \iint_{\Omega\times\Omega} (\nabla_{w,s} u(x,y))^2 \,\mathrm{d}x\,\mathrm{d}y \leq C_1,\ \frac{1}{2}\int_\Omega |\nabla u(x)|^2 \,\mathrm{d}x \leq C_2 \right\}. \tag{A 1}$$

It can further be shown that, for X and $\beta \geq 0$ fixed, C_1 is a non-increasing function of $\alpha \geq 0$ and, for X and $\alpha \geq 0$ fixed, C_2 is a non-increasing function of $\beta \geq 0$.

Any solution of equation (A 1) satisfies $\iint_{\Omega\times\Omega} (\nabla_{w,s} u(x,y))^2 \,\mathrm{d}x\,\mathrm{d}y \leq C_1$, and likewise in the discrete setting we have the following:

$$\sum_{i,j\in\Omega} (u_i - u_j)^2 \frac{w_{ij}}{\sqrt{d_i d_j}} \leq C_1.$$

Thus, for some non-negative discrete function $f : \Omega \times \Omega \to \mathbb{R}^{\geq 0}$ with $\sum_{i,j\in\Omega} f_{ij} \leq C_1$, we have the following:

$$\forall\, i,j \in \Omega, \quad (u_i - u_j)^2 \leq f_{ij} \frac{\sqrt{d_i d_j}}{w_{ij}}. \tag{A 2}$$

Recalling that, in our application, we set the weights w_{ij} to be non-local means applied to a housing image, $g : \Omega \to \mathbb{R}$, we can interpret what this means. Up to some factors constrained by the parameter C_1, the squared difference between the density at pixels i and j is bounded by $\sqrt{d_i d_j}/w_{ij}$. Thus, the bound is made restrictive when: d_i and d_j are small, which means the patches of g around pixels i and j are very different from the rest of the image; and when w_{ij} is large, which means the neighbourhoods of g around pixels i and j are similar to each other.

It is also worth noting that, by constraint, the left-hand side of (A 2) is always smaller than or equal to 1. Thus, for the inequality to be non-trivial, we must have $f_{ij} < w_{ij}/\sqrt{d_i d_j}$ for some pair $i, j \in \Omega$. Thus, C_1 must be sufficiently small (or α sufficiently large) in order to guarantee that the non-local smoothing will have any effect on u.

References

1. Silverman BW. 1986 *Density estimation for statistics and data analysis*, Monographs on Statistics and Applied Probability, vol. 26. London, UK: Chapman & Hall.
2. Scott DW. 1992 *Multivariate density estimation*. London, UK: Wiley.
3. Eggermont PPB, LaRiccia VN. 2001 *Maximum penalized likelihood estimation*, vol. II, *Regression*. Springer Series in Statistics. Berlin, Germany: Springer.
4. Wilson JQ, Kelling GL. 1982 Broken windows. *Atlantic Mon.* **249**, 29–38.
5. Short MB, D'Orsogna MR, Pasour VB, Tita GE, Brantingham PJ, Bertozzi AL, Chayes LB. 2008 A statistical model of criminal behavior. *Math. Models Methods Appl. Sci.* **18**, 1249–1267. (doi:10.1142/S0218202508003029)
6. Short MB, Brantingham PJ, Bertozzi AL, Tita GE. 2010 Dissipation and displacement of hotspots in reaction–diffusion models of crime. *Proc. Natl Acad. Sci. USA* **107**, 3961–3965. (doi:10.1073/pnas.0910921107)
7. Block R, Bernasco W. 2009 Finding a serial burglar's home using distance decay and conditional origin–destination patterns: a test of empirical Bayes journey-to-crime estimation in the Hague. *J. Invest. Psychol. Offender Profiling* **6**, 187–211. (doi:10.1002/jip.108)
8. Bernasco W, Nieuwbeerta P. 2005 How do residential burglars select target areas? A new approach to the analysis of criminal location choice. *Br. J. Criminol.* **45**, 296–315. (doi:10.1093/bjc/azh070)
9. Short M, D'Orsogna M, Brantingham P, Tita G. 2009 Measuring and modeling repeat and near-repeat burglary effects. *J. Quant. Criminol.* **25**, 325–339. (doi:10.1007/s10940-009-9068-8)
10. Townsley M, Homel R, Chaseling J. 2003 Infectious burglaries. A test of the near repeat hypothesis. *Br. J. Criminol.* **43**, 615–633. (doi:10.1093/bjc/43.3.615)
11. Liu H, Brown DE. 2003 Criminal incident prediction using a point-pattern-based density model. *Int. J. Forecast.* **19**, 603–622. (doi:10.1016/S0169-2070(03)00094-3)
12. Kennedy LW, Caplan JM, Piza E. 2011 Risk clusters, hotspots, and spatial intelligence: risk terrain modeling as an algorithm for police resource allocation strategies. *J. Quant. Criminol.* **27**, 339–362. (doi:10.1007/s10940-010-9126-2)
13. Mohler GO, Bertozzi AL, Goldstein TA, Osher SJ. 2011 Fast TV regularization for 2D maximum penalized likelihood estimation. *J. Comput. Graph. Stat.* **20**, 479–491. (doi:10.1198/jcgs.2010.09048)
14. Smith LM, Keegan MS, Wittman T, Mohler GO, Bertozzi AL. 2010 Improving density estimation by incorporating spatial information. *EURASIP J. Adv. Signal Process.* **2010**, 265631. (doi:10.1155/2010/265631)
15. Kostić T, Bertozzi AL. 2013 Statistical density estimation using threshold dynamics for geometric motion. *J. Sci. Comput.* **54**, 513–530. (doi:10.1007/s10915-012-9615-6)
16. Cour T, Benezit F, Shi J. 2005 Spectral segmentation with multiscale graph decomposition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR 2005*), *San Diego, CA, 20–25 June 2005*, vol. II, pp. 1124–1131. Los Alamitos, CA: IEEE. (doi:10.1109/CVPR.2005.332)
17. Grady L, Schwartz EL. 2006 Isoperimetric graph partitioning for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 469–475. (doi:10.1109/TPAMI.2006.57)
18. Shi J, Malik J. 2000 Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 888–905. (doi:10.1109/34.868688)

19. Chung FRK. 1997 *Spectral graph theory*, CBMS Regional Conference Series in Mathematics, no. 92. Providence, RI: American Mathematical Society.
20. Mohar B. 1991 The Laplacian spectrum of graphs. *Graph Theory, Combin. Appl.* **2**, 871–898.
21. Von Luxburg U. 2007 A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416. (doi:10.1007/s11222-007-9033-z)
22. Zhou D, Schölkopf B. 2004 A regularization framework for learning from graph data. In *ICML 2004 Workshop on Statistical Relational Learning and its Connections to Other Fields, Banff, Alberta, Canada, 8 July 2004*, pp. 132–137.
23. Gilboa G, Osher SJ. 2008 Nonlocal operators with applications to image processing. *Multiscale Model. Simul.* **7**, 1005–1028. (doi:10.1137/070698592)
24. Buades A, Coll B, Morel JM. 2005 A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.* **4**, 490–530. (doi:10.1137/040616024)
25. Merkurjev E, Kostić T, Bertozzi AL. 2013 An MBO scheme on graphs for classification and image processing. *SIAM J. Imaging Sci.* **6**, 1903–1930. (doi:10.1137/120886935)
26. Gilboa G, Osher SJ. 2007 Nonlocal linear image regularization and supervised segmentation. *Multiscale Model. Simul.* **6**, 595–630. (doi:10.1137/060669358)
27. Zhang X, Chan TF. 2010 Wavelet inpainting by nonlocal total variation. *Inverse Probl. Imaging* **4**, 191–210. (doi:10.3934/ipi.2010.4.191)
28. Peyré G, Bougleux S, Cohen L. 2008 Non-local regularization of inverse problems. In *Computer Vision–ECCV 2008, Part III, Proc. 10th European Conf. on Computer Vision, Marseille, France, 12–18 October* (eds D Forsyth, P Torr, A Zisserman). Lecture Notes in Computer Science, vol. 5304, pp. 57–68. Berlin, Germany: Springer.
29. Lou Y, Zhang X, Osher SJ, Bertozzi AL. 2010 Image recovery via nonlocal operators. *J. Sci. Comput.* **42**, 185–197. (doi:10.1007/s10915-009-9320-2)
30. Herold M, Goldstein NC, Clarke KC. 2003 The spatiotemporal form of urban growth: measurement, analysis and modeling. *Remote Sens. Environ.* **86**, 286–302. (doi:10.1016/S0034-4257(03)00075-0)
31. Batty M, Longley P, Fotheringham S. 1989 Urban growth and form: scaling, fractal geometry, and diffusion-limited aggregation. *Environ. Plann. A* **21**, 1447–1472. (doi:10.1068/a211447)
32. Bertozzi AL, Flenner A. 2012 Diffuse interface models on graphs for classification of high dimensional data. *Multiscale Model. Simul.* **10**, 1090–1118. (doi:10.1137/11083109X)
33. Fowlkes C, Belongie S, Chung F, Malik J. 2004 Spectral grouping using the Nystrom method. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 214–225. (doi:10.1109/TPAMI.2004.1262185)
34. Belongie S, Fowlkes C, Chung F, Malik J. 2002 Spectral partitioning with indefinite kernels using the Nyström extension. In *Computer Vision–ECCV 2002, Part III, Proc. 7th European Conf. on Computer Vision, Copenhagen, Denmark, 28–31 May* (eds A Heyden, G Sparr, M Nielsen, P Johansen). Lecture Notes in Computer Science, vol. 2352, pp. 531–542. Berlin, Germany: Springer.
35. Sardy S, Tseng P. 2010 Density estimation by total variation penalized likelihood driven by the sparsity ℓ_1 information criterion. *Scand. J. Stat.* **37**, 321–337. (doi:10.1111/j.1467-9469.2009.00672.x)
36. Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE. 2011 Self-exciting point process modeling of crime. *J. Am. Stat. Assoc.* **106**, 100–108. (doi:10.1198/jasa.2011.ap09546)
37. Lewis E, Mohler GO, Brantingham PJ, Bertozzi AL. 2011 Self-exciting point process models of civilian deaths in Iraq. *Secur. J.* **25**, 244–264. (doi:10.1057/sj.2011.21)
38. Wang X, Brown DE. 2012 The spatio-temporal modeling for criminal incidents. *Secur. Inf.* **1**, 1–17. (doi:10.1186/2190-8532-1-2)