A Comparison of Visualization Data Mining Methods for Kernel Smoothing Techniques for Cox Processes with Application To Spatial Decision Support Systems

David Rohde*, Ruth Huang**, Jonathan Corcoran**, Gentry White**

* Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

** School of Geography, Planning and Environmental Management, University of Queensland, Brisbane Australia

Abstract. Real world planning of complex logistical organisations such as the fire service is a complex task requiring synthesis of many different computational techniques, from artificial intelligence and statistical or machine learning to geographical information systems and visualization. A particularly promising approach is to apply established data mining techniques in order to produce a model and make forecasts. The nature of the forecast can then be rendered using visualization techniques in order to assess operational decisions, simultaneously benefiting from generic and powerful data mining techniques, and using visualization to understand these results in the context of the actual problem of interest which may be very specific. Previous approaches to visualization in similar contexts use iso surfaces to visualize densities, these methods ignore recent improvements in interactive 3D visualization such as volume rendering and cut-planes, these methods also ignore what is often a key problem of interest comparing two different stochastic processes, finally previous methods have not paid sufficient attention to differences between estimation of densities and point processes (or Cox processes). This paper seeks address all of these shortcomings and make recommendations for the trade-offs between visualization techniques for operational decision making. Finally we also demonstrate the ability to include interactive 3D plots within a paper by rendering an iso surface using 3D portable document format (PDF).

1 Introduction

Real world planning of logisticly complex organisations such as the fire service is a complex task that requires a synthesis of skills including intelligent data analysis techniques, geographical information systems and visualization. A promising way to harness the growing literature on sophisticated, but general data mining techniques with highly specific operational concerns is to use visualization data mining techniques in order to communicate the important trends the data mining algorithms have identified [13] these methods have proved powerful for the analysis of spatial [9] and spatial-temporal data [2].

An important class of problems for managing operational concerns is the Cox process or point process which is a model of events. Event modelling is of importance for many applications, the one pertinent to this study is modelling the occurrence in space and time of malicious hoax call fire events. An important contribution in combining data mining with visualization in this area was [4] which proposed a kernel smoothing method for estimating spatial diurnal (i.e. hour of day) aspects of a dataset and visualizing the model using 3D rendering of iso surfaces. While this work proposed a new kernel and demonstrated the utility of 3D visualization it had some shortcomings in that it left a number of important aspects of the model implicit including the distinction between Cox process estimation and density estimation and properties of smoothing in space time. In addition to this technology in 3D rendering has improved significantly in recent times. It is the goal of this paper to investigate the properties of kernel smoothing algorithms for Cox processes, and the relative merits of different visualization approaches, using the python [7] bindings for the Mayavi [11] visualization library and also demonstrating the capabilities of 3D PDF in latest PDF readers [1]. We also expand on previous work by considering Visual methods for comparing estimated stochastic processes.

In Section 2 we discuss kernel smoothing techniques for spatial-diurnal estimation of Cox processes. The method described is essentially that developed in [4], but we do make make explicit some aspects of the model including noting some of the important differences between estimating a Cox process and a density and illustrating lines of equal distance in space-time. In Section 3 implementation details are given for the statistical kernel smoothing and visualization methods using Mayavi and 3D portable document format. In Section 4 the spatial-diurnal Cox process is estimated for a dataset of hoax call fire events occurring in Australia. Visualization is carried out using three methods for rendering a scalar field, iso surfaces, volume rendering and cut planes. The ability of these algorithms to compare differences between stochastic processes is also evaluated. In Section 5 a discussion of the main findings are presented. Section 6 concludes.

2 Non-parametric estimation of Spatial-Diurnal Cox Processes

2.1 Kernel Smoothing for Density Estimation and Estimating Cox Processes

Kernel smoothing is a popular method originating and most popular when applied to density estimation [12] but which has also been demonstrated for the estimation of Cox processes [6].

When applied to density estimation a kernel smoothing algorithm applied to N data points is

[1] e.g. Adobe Acrobat reader 8 or later.

$$P(X_{N+1}|X_1, ..., X_N) \approx \frac{1}{N} \sum_{n=1}^{N} \mathcal{K}(X_{N+1}, X_n, \sigma^2)$$

The kernel function $\mathcal{K}()$ measures the distance between X_{N+1} and all previous points i.e. is high for low values and decays to zero. Smoothing is achieved such that points close to observed points are more probable. In general the kernel can be viewed as computing distances between pairs of points or alternatively as placing a probability distribution on top of every previously observed point.

A Cox process or a Point process [5] has some important similarities and differences with a standard density. A Cox process is a stochastic process, the practical significance is that under a stochastic process you must specify the space of interest, in our application this is metropolitan Australia in a given region of time (and different regions will be of interest when estimating and forecasting the Cox process). Once the region of interest is specified an integral must be computed under the region of interest under an intensity function, once this integral is computed Poisson distributions on quantities of interest can be specified. The fact that two integrals need to be computed to determine probabilities one under the intensity function and one under the density produced means that this model is referred to as 'doubly stochastic'.

A Cox process models the count of incidents q within a region of time and space R as a Poisson distribution with the expectation of the Poisson model specified as an integral under the intensity function $\lambda(s)$ where s specifies a point in 2D space and diurnal time.

$$q \sim \text{Poisson}\left(\int_{\text{R}} \lambda(\text{r})\text{dr}\right)$$

If two or more disjoint regions are specified then the Poisson distributions for the respective regions must be independent.

Like a density the intensity function of a Cox process is non-negative, although it need not integrate to one. An important contribution of [6] is that standard density estimation algorithms can be applied in order to estimate Cox process intensity functions. Although the density approach has been shown to be effective there are a number of alternatives such as Bayesian mixture models [10], Bayesian Gaussian Cox Processes [1] which employ other models and sophisticated computational techniques such as Markov chain Monte Carlo.

2.2 Kernel Smoothing in Euclidean Space

First we consider kernel smoothing using standard Euclidean space and ignoring the diurnal component. Distance in space follows from simple geometry where the distance between $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x}_2 = (x_2, y_2)$ is given by:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

A line of constant distance is given by a circle. A kernel smoothing algorithm is obtained by specifying a bandwidth σ and computing

$$
\begin{aligned}
&f(x_1, y_1; x_2, y_2) \\
&\quad = e^{-\frac{1}{2\sigma}^2 d(x_1,y_1;x_2,y_2)^2} \\
&\quad = \mathcal{N}(x_1 - x_2; 0, \sigma^2)\mathcal{N}(y_1 - y_2; 0, \sigma^2)
\end{aligned}
$$

where $\mathcal{N}()$ is the normal distribution, this results in the standard kernel smoothing algorithm using spherical Gaussian kernels.

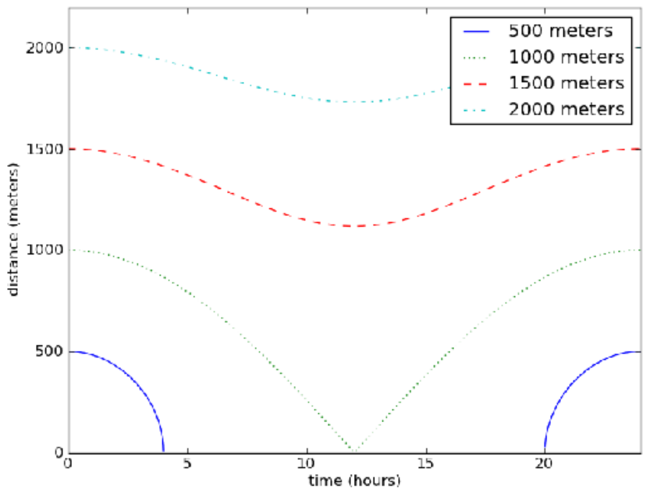

Fig. 1. Plots of equal distance using the proposed space time distance function. Note that 500 meters away is equivalent to 4 hours away. The maximum possible distance in time (12 hours) is 1000m.

2.3 Kernel Smoothing in Euclidean Diurnal Space

We now imagine that an additional diurnal variable is available and as such we have $\mathbf{x}_1 = (x_1, y_1, t_1)$ and $\mathbf{x}_2 = (x_2, y_2, t_2)$. Where time is a time given in hours, it is a real value although we will only model the periodic hourly component. In order to extend the model to include periodic time we define a periodic function which maps a diurnal hour of day variable to be between zero and one:

$$
\Delta(t_1, t_2) = \frac{1}{2}\cos(2\pi/24(t_1 - t_2)) - 1/2.
$$

This function returns 1 if $t_1 - t_2$ is an even multiple of 12 hours (the maximum possible. This represents a distance in time in order to be able to smooth in time space in order to achieve this we introduce the constant c which gives an equivalent distance in meters equivalent to a distance in time of 12 hours. We will see shortly that if we adopt the following distance kernel that we will obtain the kernel smoothing algorithm applying the Von-Mises kernel in the time dimension

$$d(x_1, y_1, t_1; x_2, y_2, t_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 - c^2 \Delta(t_1, t_2)}.$$

We can explain the presence of squares for the Euclidean distance, but the absence of squares for the periodic component by the fact that we need a non-negative distance and the proposed periodic distance function is already non-negative.

Intuition about this function can be gained by plotting lines of constant distance in space-time i.e. taking one spatial dimension and one dimension in time. This is demonstrated in Figure 1 with $c = 1000$m. From this plot we can see that a zero distance in space and a distance of 4 hours in time is equivalent to a distance of 500 meters in space and 0 hours in time. Similarly as a direct consequence of setting $c = 1000m$ a distance of 0 meters in space and 12 hours in time is equivalent to a distance of 1000m in space and 0 hours in time. It is also illustrative to consider distances larger than 1000m for which it is impossible to achieve such a large distance without some contribution in space as the largest distance time can contribute is 1000m. Another point to note about this kernel is that it is periodic and therefore points that are near 0 hours are correctly smoothed to points near 23.9 hours. In real smoothing situations assuming data is plentiful a bandwidth σ may be set to a relatively small value, this means that the tail behavior i.e. the behavior at large distances might correspond to many multiples of σ and therefore have negligible effect on the smoothing. Therefore the behavior at large distances while interesting is of relatively little importance and the most important contribution of this distance function is its periodic nature.

Kernel smoothing can be developed by analogous to Euclidean space considering the exponential of the negative square of the distance function.

$$\begin{aligned}
&f(x_1, y_1, t_1; x_2, y_2, t_2) \\
&\quad \propto e^{-\frac{1}{2\sigma^2} d(x_1, y_1, t_1; x_2, y_2, t_2)^2} \\
&\quad \propto N(x_1 - x_2; 0, \sigma^2) N(y_1 - y_2; 0, \sigma^2) \mathcal{V}(2\pi(t_1 - t_2)/24; 0, c/(2\sigma^2))
\end{aligned}$$

where $\mathcal{V}()$ is the Von Mises distribution (see [8] for details) which has a period of 2π and is defined

$$\mathcal{V}(\theta; \vartheta, \gamma) \propto e^{\gamma \cos(\theta - \vartheta)}.$$

The normalization constant is available in analytical form if you employ a special function (the Bessel function) although this is not important for our purposes.

This kernel was applied to a Cox process in [4] with 3D iso surfaces, but the discussion neglected to identify the role of the c parameter, the distinction between density estimation and Cox processes and did not consider more modern visualization techniques nor differences between different intensity functions.

3 Implementation

Both the statistical model and the visualization were implemented in a Python environment [7] using [11] for visualization. Kernel density estimation was achieved by adding the Von Mises kernel to existing scientific python (scipy) libraries to produce a three dimensional kernel surface. We set a distance of 12 hours to be equivalent to a distance of 100km, the bandwidth was set to 10km. An intensity function was estimated for the data combined and for weekends and weekdays separately.

Mayavi approximates a 3D scalar field by the use of a three dimensional array with samples taken at intervals, this structure was created by evaluating the kernel estimate over the grid. Modifications of existing Mayavi examples means it is relatively easy to render the density with iso surfaces, volume rendering and cut planes. Both the densities and their differences were rendered. A coastline of Australia was added in order to improve the visualization.

The rendering of the iso surfaces for all the data was also converted to 3D PDF format. This involved saving the data in wavefront (obj) format from Mayavi and then converting to universal 3D (u3d) format finally for importing into latex with the movie15 package. Judging from comments by other researchers the software for producing these files is improving, although it remains non-trivial. The result is shown in Fig 2 and requires a recent PDF viewer.

4 Visualization of estimates

An iso surface estimate of the intensity function identifying the highest 90% of the intensity function is demonstrated in an interactive figure in Figure 2. There are some obvious advantages in presenting the visualization to the reader in such a way with potential draw backs being that Mayavi remains a better visualization environment and there remain technological issues in both the creation and viewing of such figures. File size requirements being one such limitation which restricts the inclusion of any more 3D figures in this document.

The fire incidence datasets were implemented to visualize the incident regions in time and space with 67% and 50 % of probability contour. The kernel smoothing technique is applied to estimate the weekend and weekday fire incidence datasets from malicious hoax calls and accompanied by the 3D visualizing iso surface (Figure 4) using the Mayavi library. This method allows visualizing and identifying the intensity of the fire risk at certain locations and times. For example in Figure 4 (top) in Brisbane, there are more fire incidents happening approximately during 12:00 to 18:00 of the day which shows the higher intensity on the weekday than the weekend (see Figure 3 for location of cities). It is also confirmed by Figure 5 that between 12:00 to 18:00 time period in Brisbane, it shows yellow colour which refers to the higher fire incidents during the weekday than the weekend. By using two different colours to refer to the higher fire incidents during the weekend or weekday, it can be easily visualized and given the information including the time of the day, the location and intensities of the incident. It also can represent the comparison of these two datasets (weekends and weekdays) by subtracting the kernel (Figure 5) as well as the volume rendering (Figure 6). The volume rending is generated by using the kernel smoothing identifies the difference of spatial-time between two datasets with the same colour.

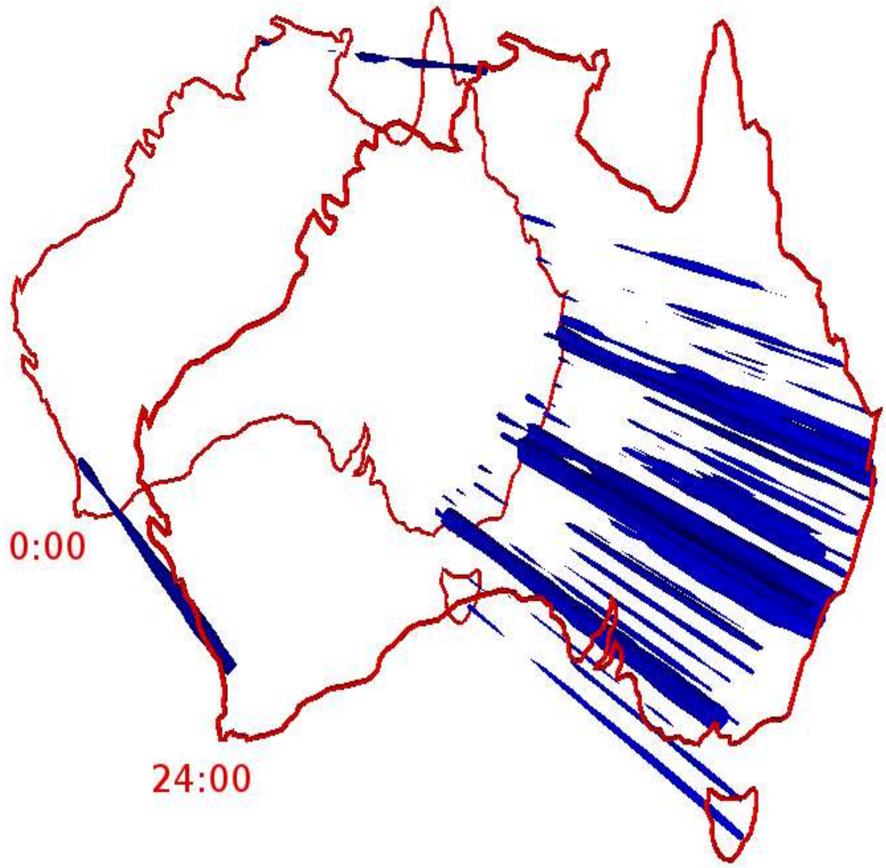

Fig. 2. Iso surface for all fire events, the contour represents the highest 90% . Please click in order to rotate (requires Acrobat Reader 8 or greater).

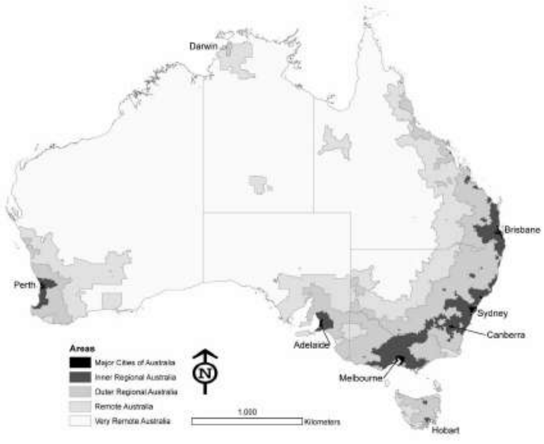

Fig. 3. Map of Australia

5 Discussion

Kernel smoothing is a simple, but very powerful data mining technique for non-parametric analysis of large datasets. Kernel smoothing refers to two similar techniques, the first and most common is density estimation, the second is estimation of a Cox process, which is a significantly more complex 'doubly stochastic' model. Methodologically these can be considered identical, but the treatment of time in particular in a Cox process is more subtle. In a density estimation scenario a kernel smoothed surface can be seen as a summary of history that can be used to forecast the future. A spatial temporal Cox process is a stochastic process indexed by both space and more importantly time, as the stochastic process is indexed by time, it is not in general meaningful to use a smoothed history to forecast future events. A way around this problem is to use only the hour of day or diurnal component of time as in [4], this simplifies the problem such that under reasonable assumptions a smoothed history in space and periodic time can be used for making forecasts using the insight in [6] that kernel density estimation techniques can also be applied to Cox processes. The visualization of the intensity surfaces using iso surfaces was also demonstrated in [4].

In this work we have made use of the space circular-time kernel in [4] and made explicit some of the issues of applying kernel smoothing to Cox processes estimation rather than to density estimation. We also considered visualization of the Cox process using not only iso surfaces, but also volume rendering and cut planes, these techniques were applied not only to estimates of the intensity of Cox process, but also to differences in Cox processes in order to consider graphical methods for comparing estimates of stochastic processes. The use of 3D PDF was demonstrated to be a useful way to distribute these results.

We found that the iso surface remains the tool of choice in a number of important situations. The iso surface is the most interpretable of the graphics to

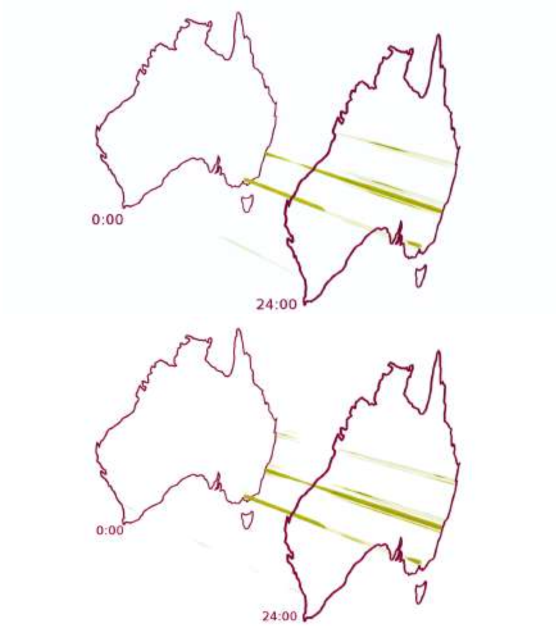

Fig. 4. Iso surface showing the contours covering 67% and 50% of the intensity function for weekends (top) and weekdays (bottom).

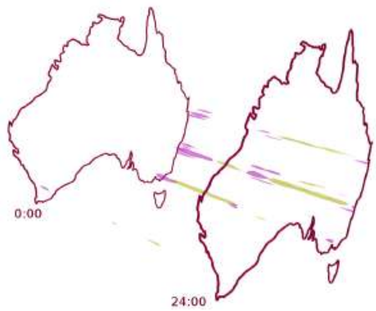

Fig. 5. The difference between weekends and weekdays i.e. Fig 4 (top) and (bottom). The pink colour area is represented for the probability density of the fire incidence during the weekend occurred more than during the weekday, yellow represents the incidence during the weekday occurred more than the weekend.

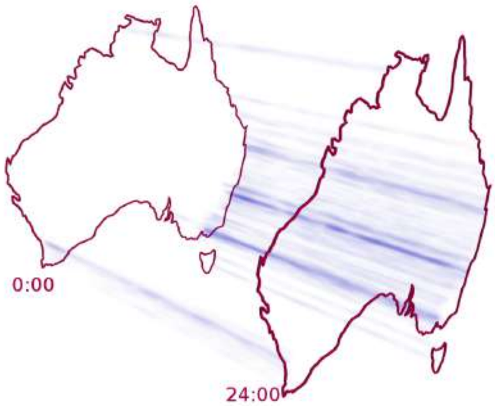

Fig. 6. Volume rendering showing the difference between (a) and (b) with the kernel smoothing using the Normal-Normal-von Mises kernel. Rendering used the Mayavi library.

display in a non-interactive journal paper, this provides a strong motivation for continuing to use the iso surface for many authors, however there are now many other forums for presenting information such as oral presentations alternatively it is possible to produce interactive electronic papers in the portable document format such as [3] . An additional advantage of modern implementations of the iso surface is that colour and opacity can be used to allow multiple surfaces to be displayed simultaneously, this can be used to give an indication of the intensity function at different levels. The problem of visually comparing two iso surfaces can be accomplished by displaying the two surfaces side by side at the same orientation or by rendering as an iso surface the difference between the two stochastic processes. Yet another advantage of the iso surface is apparent for this application as colour can effectively be used in order to give sign information when rendering a difference between cox processes. The main weakness of iso surfaces is that information is only available at a fixed number and perhaps a single threshold of the intensity surface, this means that some details of the intensity surface are necessarily absent from the display. Volume rendering and cut planes suffer less from this particular problem.

Volume rendering also showed promise for a range of scientific visualization. It is arguably more difficult to interpret a static display of a intensity surface, but some information is still apparent from displaying a static volume rendering image, particularly if colour is available. While there are no arbitrary thresholding in volume rendering, the colour scheme itself is arbitrary and the image can be difficult to interpret in a precise way. The appeal of volume rendering seems to be in an interactive display to get an overall intuition about the nature of the intensity function of the Cox process. Volume rendering offers the advantages over iso surfaces of being more intuitive and requiring no arbitrary thresholds, but should primarily be used in interactive settings and in preliminary parts of the study where precision on the intensity function is not critical. Volume rendering can also be used for comparison of stochastic processes again by side by side display or by considering differences between intensity functions. While colour could in principle be used to represent a signed difference between intensities, our experience was that this was not very intuitive and we recommend that volume rendering only be used for unsigned differences.

Cut planes are primarily an interactive way to visually explore a Cox process, static realizations of a cut plane are neither interpretable or very attractive. The advantage of a cut plane is that like volume rendering and unlike the iso surface there are no arbitrary thresholds although again like volume rendering an arbitrary colour mapping is required, but this mapping is made much more explicitly as a color bar is available and transparency does not confuse the colour at any given location. The cut plane is the most difficult to use in an interactive environment and some orientations of the cut plane do not seem very interpretable in this application i.e. if the plane sweeps through time-space the meaning of the plane is not necessarily very clear.

6 Conclusion

Visualization of intensity of spatial-diurnal Cox processes estimated with kernel smoothing is a technique which combines powerful data mining approaches with sophisticated 3D visualization in order to assist real world decision makers. Previous work demonstrated the effectiveness of this technique with application to rendering spatial-diurnal Cox processes. Here we considered the same kernel, articulated some of its underlying assumptions in terms of identifying lines of constant distance in space-time and discussed some of the subtle differences between density estimation a Cox process estimation.

Despite requiring one or more arbitrary thresholds be chosen iso surfaces are likely to be 3D visualization tool of choice for visualizing Cox processes, however alternatives such as volume rendering and cut planes are likely to see an expanded role in scientific visualization particularly in interactive environments. In terms of rendering differences between the intensity functions of Cox processes, the iso surface again shows advantages in representing signed differences, but volume rendering in particular shows promise in visualizing absolute differences. It was argued that operational decisions could be evaluated using these visualization techniques.

Improvements in graphical software will make more widespread application of 3D visualization techniques possible. Our experience with Mayavi has been very positive, and its flexible licensing terms and easy integration into scientific python are both big advantages, as is the possibility to change between different methods for rendering a scalar field.

An obvious route to consider for future work is in applying visualization techniques to other statistical methods for Cox process. A potential advantage of some other modeling techniques is to allow a more sophisticated treatment of time within the model, we are currently investigating these models.

Bibliography

[1] R. P. Adams, I. Murray, and D. J. C. MacKay. Tractable nonparametric Bayesian inference in poisson processes with Gaussian process intensities. In *In In Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, Montreal,Quebec, 2009.

[2] G. Andrienko, N. Andrienko, and P. Gatalsky. Visual mining of spatial time series data. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD '04, pages 524–527, New York, NY, USA, 2004. Springer-Verlag New York, Inc.

[3] D. G. Barnes and C. J. Fluke. Incorporating interactive three-dimensional graphics in astronomy research papers. *New Astronomy*, 13(8):599–605, Nov. 2008.

[4] C. Brunsdon, J. Corcoran, and G. Higgs. Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems*, 31(1):52–75, 2007.

[5] D. R. Cox. Some Statistical Methods Connected with Series of Events. *Journal of the Royal Statistical Society. Series B*, 17(2):129–164, 1955.

[6] P. J. Diggle. A kernel method for smoothing point process data. *Applied Statistics*, 34:13847, 1985.

[7] P. F. Dubois. Guest editors introduction: Python: Batteries included. *Computing in Science and Engineering*, 9(3):79, 2007.

[8] N. I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1996.

[9] D. A. Keim, C. Panse, M. Sips, and S. C. North. Visual data mining in large geospatial point sets. *IEEE Comput. Graph. Appl.*, 24(5):36–44, Sept. 2004.

[10] A. Kottas and B. Sansó. Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, 137:31513163, 2009.

[11] P. Ramachandran and G. Varoquaux. Mayavi: 3d visualisation of scientific data. *Computing in Science & Engineering*, 13(2):40–51, 2011.

[12] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1992.

[13] K. Zhao, B. Liu, T. M. Tirpak, and W. Xiao. A visual data mining framework for convenient identification of useful knowledge. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, pages 530–537, Washington, DC, USA, 2005. IEEE Computer Society.