Statistical Science
2018, Vol. 33, No. 3, 327–329
https://doi.org/10.1214/18-STS653
Main article: https://doi.org/10.1214/17-STS629

Self-Exciting Point Processes: Infections and Implementations

Sebastian Meyer

Abstract. This is a contribution to the discussion of Reinhart's "Review of Self-Exciting Spatio-Temporal Point Processes and Their Applications" [*Statist. Sci.* **33** (2018)], which synthesizes developments from various research fields. Here, I discuss some experiences from modeling the spread of infectious diseases. Furthermore, I try to complement the review with regard to the availability of software for the described models, which I think is essential in "paving the way for new uses."

Key words and phrases: Spatio-temporal modeling, infectious disease epidemiology, statistical software.

POINT PROCESS MODELS FOR INFECTIOUS DISEASE SPREAD

For notifiable diseases, public health surveillance data is routinely available in aggregated form as time series of infection counts. Such data are typically approached with autoregressive models using a negative binomial distribution, or assuming the counts as approximately Gaussian after a suitable transformation to adopt classical ARIMA models or even Facebook's Prophet procedure (see Held and Meyer, 2018 for an assessment). For *multivariate* time series stratified by region, spatial epidemic models can account for varying demographic and environmental factors, and enable spatially explicit predictions. Höhle (2016) provides a recent overview of spatio-temporal infectious disease models.

Taylor et al. (2013) propose to tackle even such aggregate-level surveillance data with point process methods (specifically, a log-Gaussian Cox process model with Bayesian data augmentation). However, for "mechanistic," self-exciting point process models to unfold in infectious disease epidemiology, individual-level data are indispensable. A distinction is between a point process indexed in a continuous spatial domain, such as in the ETAS model, versus a multivariate temporal point process operating on a discrete set of interacting locations/individuals, that is, on a network. Reinhart mentions recent applications of such multivariate processes in social networks. It is important to note that similar models have also been developed in the infectious disease context, where they are not that much "in its infancy." For instance, the models described in Diggle (2006), Scheel et al. (2007) and Höhle (2009) all describe the spread of livestock diseases among farms using distance-based transmission kernels. Such spatial distances could just as well be replaced by geodesic distances to quantify the coupling between the individual infection processes, for example, using movement networks as in Schrödle, Held and Rue (2012) or contact networks as mentioned by Reinhart. Aldrin, Huseby and Jansen (2015) use a combination of spatial distances and local contact networks.

In what follows, I focus on spatially *continuous* self-exciting point process models for the spread of infectious diseases in human populations. Such models come with several caveats, on three of which I would like to elaborate.

Limited Spatio-Temporal Data Resolution

The available spatial resolution of case reports is often limited by data protection. This constrains the detail with which spatial interaction can be estimated. "Areal censoring" (e.g., to the postcode level) may yield events that apparently occurred at the same location, which is impossible in simple point processes.

Sebastian Meyer is a Research Fellow at the Institute of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany (e-mail: seb.meyer@fau.de).

Equivalently, interval censoring of the infection times results in concurrently observed events, making it impossible to ascertain which infection predates the other. Furthermore, the situation is complicated by the fact that event times only correspond to the date of specimen sampling or notification to public health authorities. As latent periods and reporting delays differ between cases, the observed ordering of the events may not always properly reflect the infection chain.

One way of dealing with tied event times and locations is to add random jitter with an amount corresponding to the level of censoring in the data, and ideally conduct a sensitivity analysis or use model averaging over several random seeds. Breaking ties will affect estimates of the triggering function as well as it will remove spikes in the distribution of rescaled temporal residuals (see Meyer, Elias and Höhle, 2012, Figure 4). These are described in Ogata (1988), Section 3.3, and supplement the spatial diagnostics discussed by Reinhart.

The Meaning of Location

Even if the data provided the georeferenced place of residence of each patient, would that be a suitable proxy for the "epicentre"? It may neither be the place where the individual initially became exposed nor the location receiving the highest triggering rate during the infectious period. Nevertheless, it is probably the best available proxy. A more realistic triggering function would obviously need to employ social contacts rather than spatial displacement. This is possible in the multivariate models for $\lambda_i(t)$ above but not for $\lambda(s,t)$, as there is no mapping of locations $s \in X$ to contact rates. Using a spatio-temporal point process model for human infections thus entails the assumption that geographic distance reflects interaction good enough, which is (at least) supported by the findings of Brockmann, Hufnagel and Geisel (2006) and Read et al. (2014).

Underreporting

Public health surveillance data suffer from considerable underreporting (Gibbons et al., 2014). The consequence is that the self-exciting model component will be underestimated while the background process might partially capture cases caused by unobserved sources. This is similar to the boundary effects discussed in the review. Indeed, there *is* a background process "producing new cases from nowhere," meaning immigration of infectives from outside the observation region (e.g., sick tourists or contaminated food) or via antigenic drifts. To identify such events, stochastic declustering is also of interest in infectious disease epidemiology, but is less useful in practice because of the biases from underreporting.

A similar limitation holds for a key epidemiological parameter, the basic reproduction number R_0, estimated as the space–time integral of the triggering function. Underreporting and implemented control measures imply that this estimate is only a lower bound for the *effective* reproduction number. So yes, self-exciting models of infectious disease spread do require careful interpretation, especially since pathogens in humans are not nearly as well observable as earthquakes.

SOFTWARE

In synthesizing estimation and inference techniques, the review covers relevant topics for the analysis of spatio-temporal point patterns from epidemic phenomena. I found one crucial aspect to be missing though: software. Providing implementations of statistical methods or at least the code for the specific analysis at hand is essential for scientific progress today, as it enables others to reproduce the findings and use the described approaches in their own data-analysis pipelines.

Unsurprisingly, most publicly available implementations of self-exciting point process models are related to the ETAS model. Several implementations exist for estimating purely temporal versions, for example, the Fortran code `etas_solve` by Kasahara, Yagi and Enescu (2016), and the R packages `SAPP` (The Institute of Statistical Mathematics, 2016), `PtProcess` (Harte, 2010), and `bayesianETAS` (Ross, 2017, see Section 3.5 of the review). A general-purpose implementation to estimate and simulate purely spatial cluster process models is provided in the R package `spatstat` (Baddeley and Turner, 2005). The `ETAS` package (Jalilian, 2018) provides access to a C/C++ port of Zhuang's Fortran routines for stochastic declustering in spatio-temporal ETAS models. There are two sophisticated software packages, which support both temporal and spatio-temporal ETAS models: `SEDA` (Lombardi, 2017) is a Matlab-based GUI (currently documented to require Mac OS) for Fortran routines employing simulated annealing for maximum likelihood estimation, and `etasFLP` (Adelfio and Chiodi, 2015) is an R package using the estimation approach described in Section 3.2.2 of the review.

In principle, these ETAS packages could also be used for nonseismological applications. However, they

often do not allow for different parametric forms of the triggering function, and the modified Omori formula is not necessarily applicable in other contexts. For instance, different formulations have been used in crime (Section 4.2) and epidemic (Section 4.3) forecasting. At least for epidemiological models, the R package `surveillance` (Meyer, Held and Höhle, 2017) fills the gap. Apart from the multivariate model of Höhle (2009), it can also estimate and simulate the spatio-temporal model of Meyer, Elias and Höhle (2012) mentioned in the review. Various spatial triggering functions are supported, including Gaussian, power law, student and (piecewise) constant kernels (custom forms are possible as well, but will usually be much slower to estimate). A Newton-type optimizer with analytical derivatives is used to maximize the log-likelihood. Efforts have been made to avoid vague approximations of the contained integrals $\int_X f(s - s_i)\,\mathrm{d}s$ over the polygonal observation region X. Assuming all these integrals to equal 1 is inappropriate for events close to the boundary and for heavy-tailed kernels in general. So we compute these integrals, but use an efficient cubature method for isotropic spatial interaction functions f, which only requires one-dimensional numerical integration (see Meyer and Held, 2014, Supplement B, and the C implementation available via the R package `polyCub`).

CLOSING COMMENT

I hope that Reinhart's review will be as infectious as its content and trigger further applications of such models to epidemic phenomena. Readily available, well-documented, open-source software facilitates this process.

REFERENCES

ADELFIO, G. and CHIODI, M. (2015). FLP estimation of semi-parametric models for space–time point processes and diagnostic tools. *Spat. Stat.* **14** 119–132. MR3429716

ALDRIN, M., HUSEBY, R. B. and JANSEN, P. A. (2015). Space–time modelling of the spread of pancreas disease (PD) within and between Norwegian marine salmonid farms. *Preventive Veterinary Medicine* **121** 132–141.

BADDELEY, A. and TURNER, R. (2005). `spatstat`: An R package for analyzing spatial point patterns. *J. Stat. Softw.* **12** 1–42.

BROCKMANN, D., HUFNAGEL, L. and GEISEL, T. (2006). The scaling laws of human travel. *Nature* **439** 462–465.

DIGGLE, P. J. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Stat. Methods Med. Res.* **15** 325–336. MR2242245

GIBBONS, C. L., MANGEN, M.-J. J., PLASS, D., HAVELAAR, A. H., BROOKE, R. J., KRAMARZ, P., PETERSON, K. L., STUURMAN, A. L., CASSINI, A., FÈVRE, E. M. and KRETZSCHMAR, M. E. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: A comparison of methods. *BMC Public Health* **14** 1–17.

HARTE, D. (2010). `PtProcess`: An R package for modelling marked point processes indexed by time. *J. Stat. Softw.* **35** 1–32.

HELD, L. and MEYER, S. (2018). Forecasting based on surveillance data. In *Handbook of Infectious Disease Data Analysis* (L. Held, N. Hens, P. D. O'Neill and J. Wallinga, eds.) Chapman & Hall/CRC, Boca Raton, FL. To appear.

HÖHLE, M. (2009). Additive-multiplicative regression models for spatio-temporal epidemics. *Biom. J.* **51** 961–978. MR2744450

HÖHLE, M. (2016). Infectious disease modelling. In *Handbook of Spatial Epidemiology* (A. B. Lawson, S. Banerjee, R. P. Haining and M. D. Ugarte, eds.) 477–500. CRC Press, Boca Raton, FL. MR3586838

JALILIAN, A. (2018). `ETAS`: Modeling earthquake data using ETAS model. R package version 0.4.4, Comprehensive R Archive Network.

KASAHARA, A., YAGI, Y. and ENESCU, B. (2016). `etas_solve`: A robust program to estimate the ETAS model parameters. *Seismological Research Letters* **87** 1143.

LOMBARDI, A. M. (2017). SEDA: A software package for the statistical earthquake data analysis. *Sci. Rep.* **7** 44171.

MEYER, S., ELIAS, J. and HÖHLE, M. (2012). A space–time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics* **68** 607–616. MR2959628

MEYER, S. and HELD, L. (2014). Power-law models for infectious disease spread. *Ann. Appl. Stat.* **8** 1612–1639. MR3271346

MEYER, S., HELD, L. and HÖHLE, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package `surveillance`. *J. Stat. Softw.* **77** 1–55.

OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.

READ, J. M., LESSLER, J., RILEY, S., WANG, S., TAN, L. J., KWOK, K. O., GUAN, Y., JIANG, C. Q. and CUMMINGS, D. A. T. (2014). Social mixing patterns in rural and urban areas of southern China. *Proc. R. Soc. Lond., B Biol. Sci.* **281**.

ROSS, G. J. (2017). `bayesianETAS`: Bayesian estimation of the ETAS model for earthquake occurrences. R package version 1.0.3, Comprehensive R Archive Network.

SCHEEL, I., ALDRIN, M., FRIGESSI, A. and JANSEN, P. A. (2007). A stochastic model for infectious salmon anemia (ISA) in Atlantic salmon farming. *J. R. Soc. Interface* **4** 699–706.

SCHRÖDLE, B., HELD, L. and RUE, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics* **68** 736–744. MR3055178

TAYLOR, B., DAVIES, T., ROWLINGSON, B. and DIGGLE, P. (2013). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in R. *J. Stat. Softw.* **63** 1–48.

THE INSTITUTE OF STATISTICAL MATHEMATICS (2016). `SAPP`: Statistical analysis of point processes. R package version 1.0.7, Comprehensive R Archive Network.