Outline

Spatial and time series methods for national security applications

Soumendra N. Lahiri
Department of Statistics
Texas A&M University

1. Relevance of space/time statistical methods
2. Examples of applications involving spatial/time series data
3. Some relevant statistical methodology
4. Conclusions

Why consider spatial/time series methods?

- Almost every event occurring on earth (and beyond) has a space-time component to it. More often than not, we can gain valuable insights into solving a problem by incorporating the space-time information.
- For many problems, spatial and time series models and methods can help us
  - better understand the dynamics of the real life problem. e.g., through modeling complex interactions among 'neighboring' events,
  - to draw intelligent conclusions (i.e., inferences),
  - make better policy decisions,

  all in presence of uncertainty.
- This need is very commonplace in most security applications, where our understanding and knowledge of all the factors that contribute to a particular security challenge is rather *incomplete*.

Why consider spatial/time series?

- In such applications,
  - the cost of a wrongful decision may have huge ramifications for the public and businesses
  - mitigating policy decisions must be taken quickly.
- This calls for employing 'quick' and 'reliable' (statistical) methods that would allow us to minimize the time, cost and negative impacts under uncertainty.
- Here
  - 'quick' refers to being able to identify a potential threat quickly (high sensitivity or low probability of 'type I error'?) and provide an answer quickly (computationally efficient)
  - 'reliable' means the methodology gives a valid or reasonable solution for a wide range of situations, i.e., under minimal structural assumptions on the type of underlying uncertainty factors.
- Spatial and time series methods can be effective on all these counts.

Examples of potential Applications

- Detection of Finger-print features - applications of multivariate statistical curve estimation techniques
- Hot-spot detection/ anomaly detection
  A 'hot-spot' means a region or site of some special interests,e.g., due to its anomaly, aberration, outbreak, elevated clustering, critical resource area, etc.
  A common statistical device that has been proposed in this context is the Upper level set scan statistics (Patil, G. P., and Taillie, C. (2004), *Environmental and Ecological Statistics*, **11** 183–197.)
- Prediction of future hot spots for early warning
  - based on spatio-temporal modeling of the underlying process!

Potential Applications

- Time profile reconstruction - combining marginal information from multiple time series;
  Theory of best linear unbiased prediction for multiple time series can be adapted in a computationally efficient way for the prediction of multivariate time series at future time points!
- Monitoring and surveillance - based on spatio-temporal modeling!
  For example, these are relevant for

  (i) public health and disease surveillance,
  (ii) water resources and water services,
  (iii) transportation networks,
  (iv) contamination in food production and supply chain, etc.

- Border security
  - statistical resource optimization based on space-time models.

Surveillance and Monitoring applications

We will discuss a special statistic, called the spatial cumulative distribution function (SCDF) for these applications!

Background:

- Use of the SCDF was proposed by Overton (1989) *Technical Report No. 129, Department of Statistics, Oregon State University* in the context of analysis of survey data from the National Surface Water Surveys;
- SCDF has been used in environmental monitoring problems by Messer, J.J., Linthurst, R.A. and Overton, W.S. (1991), *Environmental Monitoring Assessment,* **17**, 67-78.
- Statistical inferential aspects are considered in Lahiri,S.N., Kaiser,M., Cressie, N., and N-J. Hsu. (1999), *Journal of the American Statistical Association*, **94** pp. 86-97.
- Bayesian treatment of the SCDF by Banerjee, Carlin, and Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Boca Raton, FL: Chapman & Hall/CRC.

Surveillance and Monitoring applications

Suppose that an attribute of interest over a region R is modeled as realizations of a spatial process $\{Z(\mathbf{s}) : \mathbf{s} \in R\}$. For example, $Z(\mathbf{s})$ may represent

1. the daily average chlorine concentration in water at location $\mathbf{s}$ in a reservoir;
2. the number of people affected by an infectious disease in a neighborhood centered at $\mathbf{s}$ in an urban area;
3. the level of Arsenic in the ground water at $\mathbf{s}$ in a rural area;
4. a crown defoliation index of trees in a plot centered at $\mathbf{s}$ in a forest, etc.

To get an overall sense of the attribute-values in the region R, one may summarize the behavior of the Z-process. The most common summary measure is the regional mean,

$$Z(R) \equiv \int_R Z(\mathbf{s}) d\mathbf{s}/|R|,$$

where $|R|$ denotes the area of R.

SCDF

Consider also the regional variance,

$$S^2(R) \equiv \int_R (Z(\mathbf{s}) - Z(R))^2 d\mathbf{s}/|R|,$$

which is a measure of how variable the Z-process is across R.

In comparison to these measures, the SCDF takes a more composite view of the attribute(s) of interest. Formally, the SCDF on R is defined as,

$$F_\infty(z; R) \equiv \int_R I(Z(\mathbf{s}) \leq z) d\mathbf{s}/|R|; z \in \boldsymbol{R}, \tag{1.1}$$

where $I(A)$ is the indicator function, equal to 1 if A is true and equal to 0, otherwise.

The SCDF shares all the properties of a cumulative distribution function (cdf) but with one major difference. It is a function of actual and potential observations and should be viewed as a random functional.

SCDF

In particular, the SCDF is *not* the theoretical cdf,

$$G(z;\mathbf{s}) \equiv P\{Z(\mathbf{s}) \leq z\}; z \in \boldsymbol{R},$$

associated with the Z-process at the spatial location $\mathbf{s}$. (Indeed, $EF_\infty(z;R) = G(z;\mathbf{0})$ for all $z \in \boldsymbol{R}$, if the Z-process is stationary).
It is easy to check that

$$\begin{aligned} Z(R) &= \int z dF_\infty(z;R); \\ S^2(R) &= \int (z - Z(R))^2 dF_\infty(z;R). \end{aligned}$$

Thus, all spatial moments, areal proportions, and spatial quantiles can be recovered from the SCDF.

Prediction of the SCDF

Since the SCDF is unobservable, we need to predict the SCDF based on a finite sample, $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_N)\}$, from $\{Z(\mathbf{s}) : \mathbf{s} \in R\}$. A basic predictor of the SCDF is the *weighted empirical cdf*

$$F_N(z; R) \equiv \sum_{i=1}^{N} u(\mathbf{s}_i) I(Z(\mathbf{s}_i) \leq z) / \sum_{i=1}^{N} u(\mathbf{s}_i), \tag{1.2}$$

where $u(\mathbf{s}_i)$'s is a set of known weights. Examples of weights include

- $u(\mathbf{s}_i) = 1$ for all i.
- $u(\mathbf{s}_i) =$ the inverse inclusion probabilities from a sampling design for location $\mathbf{s}_i$, etc.

Since the SCDF gives a region-specific picture of the attribute values at all levels, it is

- more sensitive than a regional average, where large/small values of z receive equal weights, making it hard to detect the changes in the extreme regions;

Advantages and uses of the SCDF

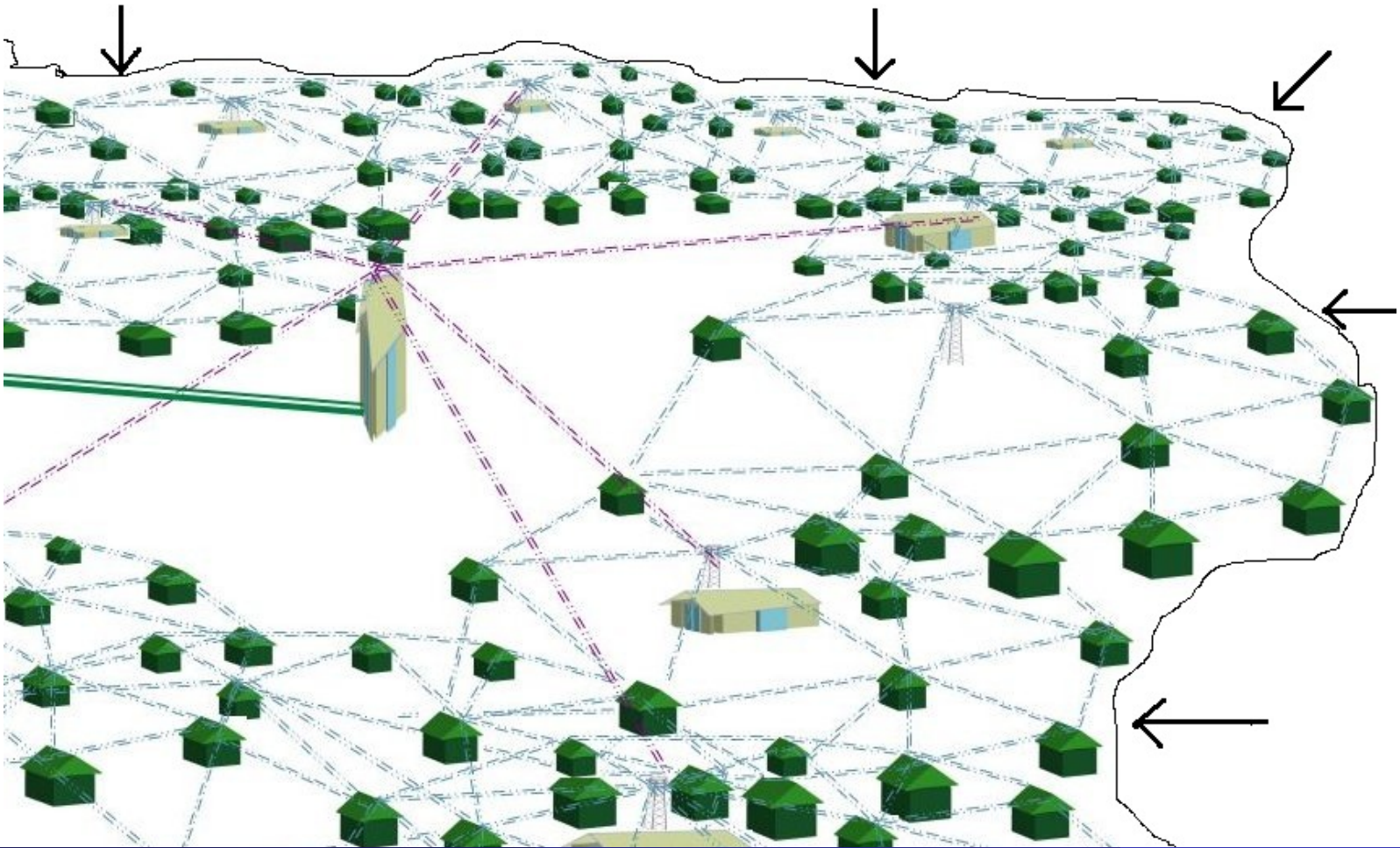
- more adaptive than using high level quantiles, where the true action may be taking place either below or above the pre-specified levels; by focusing on suitable intervals of the z-values, one can identify the levels where most changes have taken/are taking place, adaptively;

Potential uses in the security context:

- Set a region-specific **baseline benchmark** for the attribute of interest!
- Compare (predicted) SCDFs to **detect changes** at the regional level during attacks/outbreaks;
- Group regions with similar benchmarks for **formulating common emergency response policies**;
- **Set priorities for remedial actions** based on the level of changes in the critical range of the attribute values.

Application:Border Security

Envision a part of the land border that is subjected to infiltration, where violations must be intercepted/stopped. A cost-effective mechanism to monitor the border may be obtained using a network of wireless sensors and manned-stations, in a formation like : –



Network structure

Suppose that the network consists of sensors and level-1 and level-2 manned stations with the following structure:

- **Wireless Sensors:** Sense intrusions and communicate the information to the level-1 stations;
- **Level-1 stations:** Verify the information for false alarms (by combining inputs from more than one WS); If true, communicate with the relevant level-2 stations and send force to intercept within its assigned range;
- **Level-2 stations:** Co-ordinate the efforts of level-1 stations and intercept in case level-1 stations are busy/unavailable.

The problem is to optimize the cost of the 24/7 operation by designing an optimal network.

Statistical Formulation

Let $X(A, T)$ denote the number of infiltrations through any given stretch A of the border over a given time interval T. Note that $X(A, T)$ is random variable. We suppose that $X(\cdot, \cdot)$ is an (inhomogeneous) spatio-temporal Poisson random field satisfying

$$E\{X(A, T)\} = \int_T \int_A \lambda(\mathbf{s}, t) d\mathbf{s} dt.$$

The rate function $\lambda(\mathbf{s}, t)$ typically has a strong periodic behavior, with hourly, daily and yearly components . Thus, we may write

$$\lambda(\mathbf{s}, t) = g(\mathbf{s}) \exp\left(h_{24}(\mathbf{s}, t) + d_7(\mathbf{s}, t) + a_{365}(\mathbf{s}, t)\right)$$

where $g(\cdot)$ is a non-negative function, and for each $\mathbf{s}$, $h_{24}(\mathbf{s}, \cdot)$, $d_7(\mathbf{s}, \cdot)$, and $a_{365}(\mathbf{s}, \cdot)$ are periodic functions with periods 24, 7 and 365, respectively.

Statistical optimization

Ignoring the cost of operating the WSs, the main problem here is to determine the resource levels at the level-1 and -2 manned-stations such that with high probability, 100% interception is possible at all hours. Let $Z(T)$ and $Y_1(T), \ldots, Y_k(T)$ respectively denote the resources to be maintained at a level-2 station and at its k-many reporting level-1 stations, over a time interval T. If A_i is the assigned border-stretch under the control of the ith level-1 station, then one possible formulation requires

$$P(Y_i(T) \geq \chi_\alpha(A_i, T)) = 1 \quad \text{for all} \quad i$$

and

$$P(Z(T) \geq e_\alpha(T)) = 1,$$

where $\chi_\alpha(A_i, T)$ is the α-quantile of POISSON$(\lambda(A_i, T))$ distribution and $e_\alpha(T)$ is a multiple of the expected number of un-intercepted cases reported by the k level-1 stations to the level-2 station in the time interval T.

Statistical issues

The Optimization problem: Given c_1 and c_2 - (unit costs for stations 1 and 2), choose α and k to optimize the *Minimal cost of operation*:

$$c_1 \sum_{i=1}^{k} \chi_\alpha(A_i, T) + c_2 e_\alpha(T).$$

This formulation reduces the cost by exploiting the variation in the rate function, as a function of space and time.
The optimization problem here eventually depends on the knowledge of the rate function $\lambda(\cdot, \cdot)$. But typically, this function is unknown. Statistical methodology can be employed here further for:

1. Estimation of the rate function $\lambda(\mathbf{s}, t)$.
2. Boundary hot-spot detection.
3. Estimation of the direction/tracking individuals using the WS-information, etc.

Conclusions

In this talk, we have demonstrated, through some examples, that statistical methodology is highly relevant for national security applications. In particular, spatial/temporal methods can be used in a variety of security problems for

- better surveillance of the nation's infrastructure and resources
- detection of an impending emergency situation
- policy formulation for region-specific response in handling in an emergency situation
- early warning of a potential emergency situation through future hot-spot prediction
- optimizing the cost of surveillance operations in a random environment, etc.

Conclusions

Thank you!