Panel Discussion of Quantitative and Statistical Methods for Defense and National Security

Myron J. Katzoff
Office of Research and Methodology
National Center for Health Statistics (NCHS)

February, 2006

Roles, Themes and Topic Areas

CDC/ATSDR Role

- Simply stated, the agency's role is to protect people's health

- Responsible for guiding public health in addressing dangers posed by chemical, biological, radiological/nuclear and mass trauma terrorist emergencies

Key Themes

- BioIntelligence

- Containment and response

Roles, Themes and Topic Areas *continued*

- Mitigation and Recovery

<u>Quantitative and Statistical Methods Needed for These Services</u>

- Information management and exchange

- Post-event remediation and health monitoring

A Couple of Personal Observations

- Caveats

  exposure to only a subset of dialogs ongoing for some time

  public health perspective

- Major emphases

  detection of biological threats involving use of microbial agents (*e.g.*, anthrax, small pox)

  measures/strategies for countering disease spread following intentional release

Personal Observations *continued*

- CDC has a large IT and data management effort of a scope that includes

  provision for storage and retrieval of massive quantities of data, the collection and distribution of which are controlled by state and local governments

  development of IT structures that would make it possible to share information across governmental boundaries whenever necessary or desirable

  generation of data in appropriate formats to enable the application of data-analysis tools

A “Short” List of Statistical Methods

- matching techniques to unduplicate records across files
- imputation methods for reporting delays, incomplete records
- adaptive sampling procedures for application to existing national surveys for emergency situations
- spatio-temporal methods for prediction, tracking and monitoring of BT events
- methods of inference for sparse data
- methods for social networks inference and analysis
- inference techniques for nonlinear dynamical systems

Some Reasons for Using Adaptive Sampling Procedures

- Ensure adequate numbers of sample units for elusive or hard-to-locate population members

- Sample enrichment for later secondary analyses or for detailed problem-related study

- Control of “case” content, to address perceptions as to the quality and relevance of a sample for general population estimates

Information for Dealing with Public Health Challenges

- Track the effectiveness of interventions by periodically surveying affected population groups

- Identify how a contagion is being passed from one community or population to another

- Monitor and forecast (*e.g.*, with state-space models) the numbers of individuals in the various stages of a disease

- Update estimates of the parameters of mathematical models used in analyzing the population dynamics of epidemics

The Distinctive and Important Characteristic of Adaptive Designs

- Different from conventional designs because the inclusion of units in the sample may depend on the values of variables observed during the survey

The term covers:

- link-tracing designs such as network sampling, snowball sampling, chain-referral sampling, random walk designs and adaptive cluster sampling

- active set adaptive designs, a somewhat new class of designs which includes those listed above

The NHIS Design for Adaptive Sampling

- National stratified multistage survey: households are selected at the final stage of selection. State is a stratifier.

- PSUs: counties, groups of counties, county equivalents (parishes and independent cities), towns, townships, minor civil divisions, or metropolitan statistical areas

- Some large PSUs are selected with certainty and, therefore, should be regarded as strata

- If necessary, samples for recent time periods can be grouped to ensure large enough initial samples needed for subsequent adaptive inclusion of sample units in order to cover 80+ areas designated as CMSAs, PMSAs and MSAs

Random Walk Designs

Problem: Estimate the proportion of a local population (CMSA, PMSA or MSA) exposed to an easily transmitted contagious biological agent

Nodes: members of the population

Links: defined by mechanisms which would enable contacts among members of the population

Random Walk Designs *continued*

- initial sample of persons drawn, for example, from households may be somewhat uncontrolled

- link-tracing procedure is monitored and controlled to asymptotically yield desired sample selection probabilities

- idea: work with the natural tendencies of populations and link-tracing procedures to provide just enough direction during the sampling so that simple estimates can be calculated which are representative of population as a whole

Targeting the Random Walk

- For the random walk design we want to have a Markov chain structure that is aperiodic and irreducible

- We begin by defining a fundamental set of probabilities

$$q_{ij} = \begin{cases} (1-d)/N + d\ a_{ij}/a_{i.}, & \text{if } a_{i.} > 0 \\ 1/N, & \text{if } a_{i.} = 0 \end{cases}$$

  where $0 < d < 1$; $a_{ii} = 0$; and for $i \neq j$, a_{ij} is the number of links between units i and j; and $a_{i.} \overset{def}{=} \sum_j a_{ij}$, the out-degree of node i

Targeting the Random Walk *continued*

We invoke the method of Hastings (Biometrika, 1970) to guide the random walk so that it has a given stationary distribution $(\pi_1, \pi_2, \ldots)$ when, for $i \neq j$, its transition matrix has probabilities

$$P_{ij} = q_{ij}\alpha_{ij}$$

where

$$\alpha_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\};$$

and $P_{ii} = 1 - \sum_{j \neq i} P_{ij}$.

Targeting the Random Walk *continued*

- Using this procedure, we produce data to quantify the generalized ratio estimator of the mean
$$\widehat{\mu} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \Big/ \sum_{i=1}^{n} \frac{1}{\pi_i}$$
where n is the sample size with or without repeats.

- So far, simulation studies have indicated that the sampling distribution for this estimator has a mean that is very nearly that of the known mean for moderate values of n if there are no truly extreme values for the π_i.

Practical Methods for Spatio-Temporal Analysis

- build on well-established ideas and approaches

- easily taught because they are well-structured (*i.e*, can be applied in an orderly and systematic way)

- conform with common-sense notions about how circumstances in one location might be affected by what has been happening elsewhere

- results can be easily and clearly communicated

Outline of the STARMAX Method

• Suppose that we have a record of the frequency of occurrence of morbidity at each of several locations $s = 1, 2, \ldots, L$ and that $y_t(s)$ denotes a simple, variance-stabilizing transformation of morbidity frequency at location s at time t. (For example, if m_t denotes morbidity frequency at time t, then, dropping the s, $y \overset{\text{def}}{=} \sqrt{m_t + 1} - \sqrt{m_{t-1} + 1}$.)

• Suppose that we apply our favorite Box-Jenkins software to the $y_t(s)$ for each location with the result that an ARMA(1,1) models appears to be satisfactory for any of them. Then suppressing the index s, for each location we have

$$y_t = \phi y_{t-1} + z_t - \theta z_{t-1}$$

or, in state-space form

$$\begin{aligned} x_{t-1} &= \phi x_t + (\phi - \theta) z_t \quad \text{state equation} \\ y_t &= x_t + z_t \quad \text{observation equation} \end{aligned}$$

The Full Model Specification

For the L locations together, for the time series portion of the model one has

$$\begin{aligned} \overrightharpoon{X}_{t+1} &= \Phi \overrightharpoon{X}_t + \overrightharpoon{v}_t \\ \overrightharpoon{Y}_t &= \overrightharpoon{X}_t + \overrightharpoon{z}_t \end{aligned}$$

where $\overrightharpoon{v}_t = (\Phi - \Theta)\overrightharpoon{z}_t$ and Φ and Θ are appropriate diagonal matrices.

The Full Model Specification *continued*

To inject spatial dependence into the model framework, we introduce the $L \times L$ matrix D of spatial constraints with $1's$ along the diagonal such that the model specification is finally

$$\begin{aligned}\overrightarrow{X}_{t+1} &= D\,\Phi\overrightarrow{X}_t + D\overrightarrow{v}_t \\ \overrightarrow{Y}_t &= \overrightarrow{X}_t + \overrightarrow{z}_t\end{aligned}$$

where, again, $\overrightarrow{v}_t = (\Phi - \Theta)\overrightarrow{z}_t$ and Φ and Θ are, as before, appropriate diagonal matrices.

- Specification of the entries of $D = \{d_{ij}\}$ is left to the investigator

- Since spatial dependence is often expressed in terms of the spatial separation of locations, d_{ij} is often interpreted as an expression of the inverse "distance" between locations i and j

- For regularly spaced systems, the k-th order neighbors of a given location i are treated as equi-distant from i and the nearest neighbors have the most effect on each other.

- The inverse of some increasing nonnegative function of the Euclidean distance between locations or the inverse of the variogram as a function of distance have been used for irregularly spaced systems