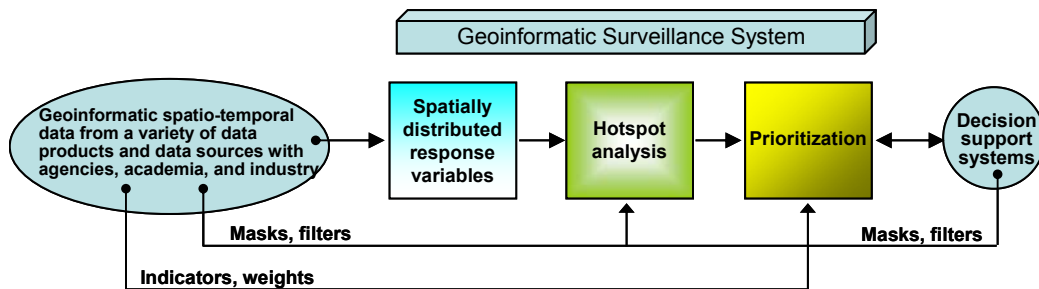


Geoinformatic Surveillance Hotspot Prioritization Using Linear Extensions of Partially Ordered Sets for Multi-criterion Ranking with Multiple Indicators

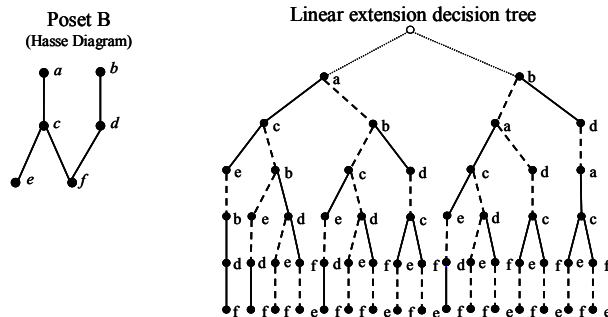
G. P. Patil and C. Taillie

Center for Statistical Ecology and Environmental Statistics
 Department of Statistics, Penn State University
 University Park, PA 16802 USA
 Email: gpp@stat.psu.edu

The prioritization system of the geoinformatic surveillance project is concerned with the question of ranking a finite collection of objects when a suite of indicator values is available for each member of the collection. The objects can be represented as a cloud of points in indicator space, but the different indicators (coordinate axes) typically convey different comparative messages and there is no unique way to rank the objects while taking all indicators into account. A conventional solution is to assign a composite numerical score to each object by combining the indicator information in some fashion. Consciously or otherwise, every such composite involves judgments (often arbitrary or controversial) about tradeoffs or substitutability among indicators.

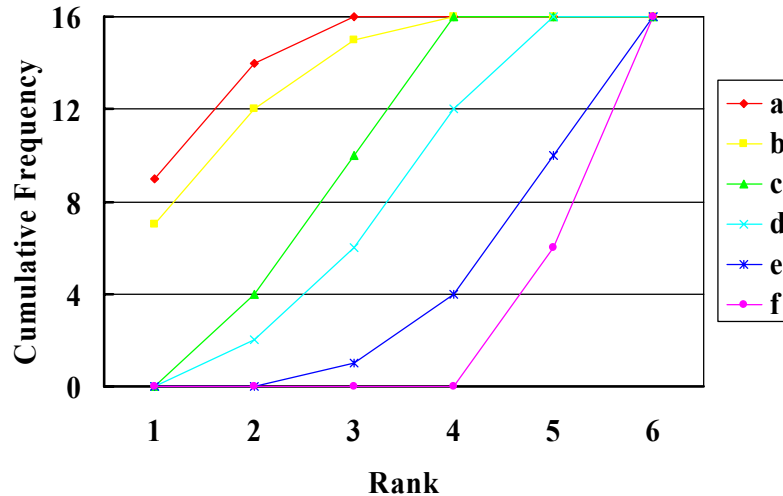


Rather than trying to combine indicators, we take the view that the relative positions in indicator space determine only a *partial ordering* and that a given pair of objects may not be inherently comparable. Working with *Hasse diagrams* of the partial order, we study the collection of all rankings that are compatible with the partial order (*linear extensions*). In this way, an interval of possible ranks is assigned to each object. The intervals can be very wide, however. Noting that ranks near the ends of each interval are usually infrequent under linear extensions, a probability distribution is obtained over the interval of possible ranks. This distribution, called the *rank-frequency* distribution, turns out to be unimodal (in fact, log-concave) and represents the degree of ambiguity involved in attempting to assign a rank to the corresponding object.



Stochastic ordering of probability distributions imposes a partial order on the collection of rank-frequency distributions. This collection of distributions is in one-to-one correspondence with the original collection of objects and the induced ordering on these objects is called the cumulative rank-frequency (CRF) ordering; it extends the original partial order. For example, the figure above shows the Hasse diagram for a small partially ordered set (poset) with six objects (labeled *a* through *f*). The decision tree on the right

enumerates all possible linear extensions of the poset (each path through the tree determines a linear extension). In this example, there are a total of 16 linear extensions. Object *a* is assigned rank 1 by nine of those extensions, rank 2 by five of the extensions, and rank 3 by the remaining two extensions. The cumulative rank frequencies for object *a* are thus 9, 9+5=14, and 9+5+2=16. These determine a cumulative rank profile for object *a* as shown in the figure below (and similarly for the other five objects).



For this example, the six profiles are stacked one-above-the-other, thus determining a linear ordering of the objects. Although the CRF ordering need not be linear in general, it can be iterated to yield a fixed point of the CRF operator. We hypothesize that the fixed points of the CRF operator are exactly the linear orderings. The CRF operator treats each linear extension as an equal “voter” in determining the CRF ranking. It is possible to generalize to a weighted CRF operator by giving linear extensions differential weights either on mathematical grounds (e.g., number of jumps) or empirical grounds (e.g., indicator concordance). Explicit enumeration of all possible linear extensions is computationally impractical unless the number of objects is quite small. In such cases, the rank-frequencies can be estimated using discrete Markov chain Monte Carlo (MCMC) methods.

The resulting prioritization system will have the following innovative features:

- Ability to rank and prioritize hotspots
- Utilizes multiple indicator and stakeholder criteria without integrating indicators into an index
- Employs Hasse diagrams, partially ordered sets, and Markov Chain Monte Carlo computations

leading to several key applications, including:

- Early warning systems
- Identification of critical areas for focused investigation.

In the area of Health Policy, Health Statistics, and Disease Etiology, the prioritization component will be combined with a hotspot detection component to yield a three-stage surveillance system:

- **First stage screening**
Identification of significant clusters (hotspots) by an upper level set version of the scan statistic
- **Second stage screening**
Rank and prioritize significant hotspots using likelihood values and other attributes such as raw intensity values, remediation-feasibility scores, socio-economic and demographic factors, etc.
- **Third stage screening**
Follow up hotspots for etiology and/or intervention

Keywords. Geoinformatic Surveillance, Prioritization, Multicriteria Ranking, Multiple Indicators