

PENNSSTATE



Center for Statistical Ecology and Environmental Statistics

Multiscale Detection of Localized Anomalous Structure In Aggregate Disease Incidence Data

By Mary M. Louie¹ and Eric D. Kolaczyk²

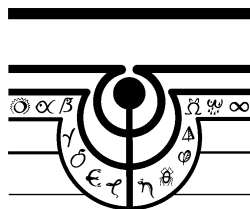
¹National Center for Health Statistics, Hyattsville, MD

²Department of Mathematics and Statistics, Boston University, Boston MA

[Presented at the 7th Annual International Conference on Digital Government Research]

This material is based upon work partially supported by the National Science Foundation under Grant No. 0307010. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

Technical Report Number 2006-0534
TECHNICAL REPORTS AND REPRINTS SERIES
May 2006



Department of Statistics
The Pennsylvania State University
University Park, PA 16802

G. P. Patil
Distinguished Professor and Director
Tel: (814)865-9442 Fax: (814)865-1278
Email: gpp@stat.psu.edu

<http://www.stat.psu.edu/~gpp>
<http://www.stat.psu.edu/hotspots>
[DGOnline News](#)

Multiscale Detection of Localized Anomalous Structure in Aggregate Disease Incidence Data*

Mary M. Louie
Office of Research and Methodology
National Center for Health Statistics
3311 Toledo Road, Rm 3215
Hyattsville, MD 20782
mlouie@math.bu.edu

Eric D. Kolaczyk
Department of Mathematics and Statistics
Boston University
111 Cummington Street
Boston, MA 02215
kolaczyk@math.bu.edu

ABSTRACT

We present a modeling framework for detection of potentially anomalous structure in aggregate spatial disease incidence data in a manner sensitive to localization at multiple scales and/or positions. The key technical contribution is the re-casting of the components of a multiscale disease mapping methodology into a form appropriate for hypothesis testing. In particular, we describe how hypotheses of spatially clustered variations in disease incidence may be linked in one-to-one correspondence with collections of hypotheses on the values of certain multiscale parameters. A Bayesian hypothesis testing methodology is developed in the context of a standard Poisson measurement model, over the collection of possible multiscale hypotheses. The methodology is illustrated on both simulated and real data.

Keywords

multiscale, Bayesian, hypothesis testing, aggregate spatial incidence data, clustered variation

1. INTRODUCTION

It has been noted (e.g., [2]) that in spatial epidemiology, as in most other geo-spatial fields of study, the concept of *scale* plays an important role in analysis and inference. Accordingly, we have recently introduced a framework for multiscale disease mapping that is, to the best of our knowledge, the first such framework [3]. Here, we re-cast our previous framework for the purpose of testing for localized anomalous spatial variations in disease incidence data in a similarly multiscale fashion.

Our proposed framework exploits the fact that, under the standard Poisson measurement model for aggregate count data, hypotheses of spatially localized variation in disease

*Full version of this paper is published in *Statistics in Medicine* 2006; 25(5):787-810.

incidence can be usefully linked in one-to-one correspondence with collections of hypotheses on the values of certain multiscale parameters associated with a user-defined hierarchy of nested partitions of an overall spatial region. Hence, the problem of finding localized anomalous spatial variations in the data is replaced by one of finding various patterns within a collection of local hypotheses indexed by position and scale. We develop a Bayesian hypothesis testing machinery to evaluate and compare models within this collection.

2. MULTISCALE FACTORIZATION AND REPARAMETERIZATION

Let D be a spatial region of interest. As in [3], our notion of a multiscale model is associated with a collection $\mathcal{B}^{(J)} = \{\{B_{j,k}\}_{k=1}^{N_j}\}_{j=0}^J$ of $J+1$ nested partitions of D . That is, the $B_{j,k}$ represent subregions in D at spatial scales $j = 0, 1, \dots, J$ and relative positions $k = 1, \dots, N_j$ within scale, such that

$$\bigcup_{k=1}^{N_j} B_{j,k} = D \quad \text{and} \quad \bigcup_{k' \in ch(k)} B_{j+1,k'} = B_{j,k}. \quad (1)$$

Here, for a given choice of j and k , $ch(k)$ denotes the set of indices k' for which $B_{j+1,k'} \subseteq B_{j,k}$.

Note that for each successive spatial aggregation of subregions $B_{j,k}$, from one scale to another, a similar aggregation of the measurement and mean variables may be defined:

$$Y_{j,k} = \sum_{k' \in ch(k)} Y_{j+1,k'} \quad \text{and} \quad \mu_{j,k} = \sum_{k' \in ch(k)} \mu_{j+1,k'}. \quad (2)$$

A simultaneous analysis of the $Y_{j,k}$ at all scales is desirable, but is complicated by the dependence of these variables. A multiscale factorization of the data likelihood can be used to induce a particularly useful decoupling. Let $\mathbf{Y}_j = (Y_{j,1}, \dots, Y_{j,N_j})^T$ and let $\mathbf{Y}_{j+1,ch(k)}$ denote those measurements $Y_{j+1,k'}$ for whom $B_{j+1,k'} \subseteq B_{j,k}$. Define $\boldsymbol{\mu}_j$ and $\boldsymbol{\mu}_{j+1,ch(k)}$ similarly. Then for $\mathbf{Y}_J | \boldsymbol{\mu}_J \sim \text{Poisson}(\boldsymbol{\mu}_J)$, one can write

$$\Pr(\mathbf{Y}_J | \boldsymbol{\mu}_J) = \Pr(Y_{0,1} | \mu_{0,1}) \times \prod_{j=0}^{J-1} \prod_{k=1}^{N_j} \Pr(\mathbf{Y}_{j+1,ch(k)} | Y_{j,k}, \boldsymbol{\omega}_{j,k}), \quad (3)$$

where $Y_{0,1} | \mu_{0,1} \sim \text{Poisson}(\mu_{0,1})$ and $\mathbf{Y}_{j+1,ch(k)} | Y_{j,k}, \boldsymbol{\omega}_{j,k} \sim \text{Multinomial}(Y_{j,k}; \boldsymbol{\omega}_{j,k})$, with

$$\boldsymbol{\omega}_{j,k} = \mu_{j,k}^{-1} \times \boldsymbol{\mu}_{j+1,ch(k)}. \quad (4)$$

3. MULTISCALE DETECTION OF LOCAL DEVIATIONS IN RELATIVE RISK: A BAYESIAN TESTING FRAMEWORK

Detecting localized deviations in the relative risk from a constant θ^* is equivalent to detecting deviations in the $\omega_{j,k}$ from the values $e_{j+1, ch(k)}/e_{j,k}$. Let $\gamma \equiv \{\{\gamma_{j,k}\}_{k=1}^{N_j}\}_{j=0}^{J-1}$ be a collection of Bernoulli random variables. Letting $\gamma_{j,k}$ indicate whether (1) or not (0) the parameter vector $\omega_{j,k}$ ‘deviates’ from $e_{j+1, ch(k)}/e_{j,k}$, we define the local null and alternative hypotheses $H_{j,k}^{(0)} : \gamma_{j,k} = 0$ and $H_{j,k}^{(1)} : \gamma_{j,k} = 1$. Now let \mathcal{H} be the set of all models γ obtainable by combinations of $H_{j,k}^{(0)}$ ’s and $H_{j,k}^{(1)}$ ’s, with the obvious restriction that only one of $H_{j,k}^{(0)}$ and $H_{j,k}^{(1)}$ may be included in any given model, for each (j, k) . Define $H^{(0)}$ to be the model in which $\gamma \equiv \mathbf{0}$. Detecting deviations from uniformity will be equated with declaring in favor of models $H \in \mathcal{H} \setminus H^{(0)}$ under an appropriate posterior. The precise form of our posterior will follow from the Poisson sampling model and the specification of (i) a conditional prior density $p(\{\omega_{j,k}\} | \gamma)$, and (ii) a probability mass function $\Pr(\gamma)$.

We model the $\gamma_{j,k}$ ’s as independent Bernoulli random variables i.e., $\Pr(\gamma) = \prod_{j,k} \alpha_{j,k}^{\gamma_{j,k}} (1 - \alpha_{j,k})^{1 - \gamma_{j,k}}$, for values $\{\alpha_{j,k}\}$ in $[0, 1]$. Then, conditional on γ , we specify that

$$\omega_{j,k} | \gamma_{j,k} \sim \begin{cases} \delta_{e_{j+1, ch(k)}/e_{j,k}}, & \text{if } \gamma_{j,k} = 0 \\ \text{Dirichlet}(c_j T e_{j+1, ch(k)}), & \text{if } \gamma_{j,k} = 1, \end{cases} \quad (5)$$

where $\delta_{e_{j+1, ch(k)}/e_{j,k}}$ indicates a point mass at $e_{j+1, ch(k)}/e_{j,k}$, T is a constant meant to capture an overall level of relative risk in D , and the c_j are scale-dependent hyperparameters whose effect is to influence the relative variation within each scale. Finally, given γ and the $\omega_{j,k}$ ’s, sampling from the conditional distributions $\mathbf{Y}_{j+1, ch(k)} | \omega_{j,k}, Y_{j,k}$, according to the multinomial distributions defined in (3) yields the observations in \mathbf{Y}_J .

Using standard calculations, it is easy to show that under this model the posterior evidence in favor of a hypothesis $H \in \mathcal{H}$ i.e., $\Pr(H | \mathbf{Y}_J) = \Pr(\gamma | \mathbf{Y}_J)$, for $\gamma \equiv \gamma(H)$, is given by

$$\Pr(\gamma | \mathbf{Y}_J) = \prod_{j=0}^{J-1} \prod_{k=1}^{N_j} \rho_{j,k}^{\gamma_{j,k}} (1 - \rho_{j,k})^{1 - \gamma_{j,k}}, \quad (6)$$

where

$$\rho_{j,k} = \frac{O_{j,k}}{1 + O_{j,k}} \quad (7)$$

and

$$\begin{aligned} O_{j,k} &= \frac{\Pr(\gamma_{j,k} = 1 | \mathbf{Y}_{j+1, ch(k)}, Y_{j,k})}{\Pr(\gamma_{j,k} = 0 | \mathbf{Y}_{j+1, ch(k)}, Y_{j,k})} \\ &= \frac{\alpha_{j,k}}{1 - \alpha_{j,k}} \times \frac{\Pr(\mathbf{Y}_{j+1, ch(k)} | Y_{j,k}, \gamma_{j,k} = 1)}{\Pr(\mathbf{Y}_{j+1, ch(k)} | Y_{j,k}, \gamma_{j,k} = 0)}, \end{aligned} \quad (8)$$

with

$$\begin{aligned} \Pr(\mathbf{Y}_{j+1, ch(k)} | Y_{j,k}, \gamma_{j,k} = i) &= \int \left(\Pr(\omega_{j,k} | \gamma_{j,k} = i) \times \right. \\ &\left. \Pr(\mathbf{Y}_{j+1, ch(k)} | Y_{j,k}, \omega_{j,k}, \gamma_{j,k} = i) \right) d\omega_{j,k}, \quad i = 0, 1. \end{aligned} \quad (9)$$

From the above expressions we see that the posterior is itself

a product of independent Bernoulli random variables, with probability of success $\rho_{j,k}$ for the (j, k) -th variable, where the $\rho_{j,k}$ are defined in terms of the posterior odds $O_{j,k}$, and the marginal data likelihoods arising in the latter have the standard, closed-form expressions under a multinomial-Dirichlet model. Therefore, from expression (6) we see that selecting the most likely model in \mathcal{H} , which corresponds to selecting the optimal combination of 0’s and 1’s, reduces to deciding whether $\rho_{j,k}$ or $1 - \rho_{j,k}$ is larger, for each (j, k) . But each $\rho_{j,k}$ is a monotone increasing function of the posterior odds $O_{j,k}$. So this problem is equivalent to choosing the hypothesis $H_{j,k}^{(i)}$, $i = 0, 1$, that maximizes $\Pr(H_{j,k}^{(i)} | \mathbf{Y}_J)$, for each (j, k) .

4. APPLICATIONS

We applied our framework to simulated and real data. Here we provide brief descriptions of the simulation design and mortality data.

4.1 Simulation

We conducted a simulation study, aimed at illustrating the potential of our multiscale detection framework, as viewed from a handful of simple scenarios. The initial data space D was taken to be a square region, and the nested hierarchy $\mathcal{B}^{(J)}$ was defined through the generic quad-tree structure. Specifically, D was partitioned into $2^j \times 2^j = 2^{2j}$ identical subregions, for $j = 0, 1, \dots, 4$, so that each square subregion at scales $j < 4$ consists of four sub-subregions at scale $j + 1$. At the finest scale $J = 4$, there were a total of $16 \times 16 = 256$ subregions. Simulations were conducted under a total of five different scenarios of landscape design. The first landscape has uniform relative risk; that is, each subregion has a relative risk of one. The rest of the landscapes each contain a single area of elevated relative risk. The locations of these areas were chosen so as to allow for examination of the effect on our detection framework from different degrees of nesting of the elevated area within the elements of the nested hierarchy.

4.2 Application to Tuscany Gastric Cancer Mortality Data

We applied our multiscale method to male gastric cancer mortality data obtained from the Tuscan region of Italy during the period 1980-1989 [2, 1]. For our hierarchy of nested partitions we took the set of nested subregions corresponding to three of the five geo-political units used by the European community. Data were obtained for males over 35 years of age at the finest level. Italian age specific rates for the same calendar period were used to obtain expected numbers of deaths [1].

5. REFERENCES

- [1] E. Dreassi and A. Biggeri. Edge effect in disease mapping. *Journal of the Italian Statistical Society*, 7(3):267–283, 1998.
- [2] A. B. Lawson. *Statistical Methods in Spatial Epidemiology*. John Wiley Sons, Chichester, 2001.
- [3] M. M. Louie and E. D. Kolaczyk. A multiscale method for disease mapping in spatial epidemiology. *Statistics in Medicine*, 25(8):1287–1306, 2006.