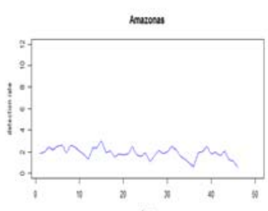
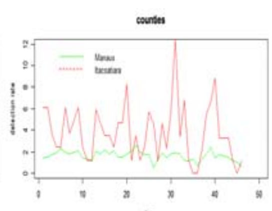
Endemic Disease Surveillance Using Bayes Factor

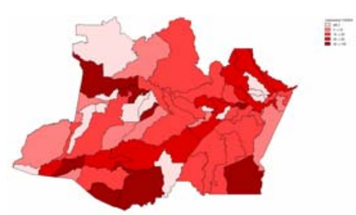
Max Sousa de Lima
maxlima@ufam.edu.br
Department of Statistics
Fed. University of Amazonas

Luiz H. Duczmal
duczmal@est.ufmg.br
Department of Statistics
Fed. University of Minas Gerais

Motivation: Disease Surveillance Maps







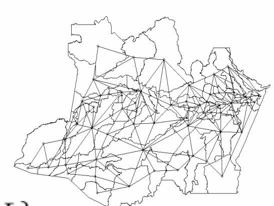
*Case Study: **New cases of Hanseniasis in the Brazilian Amazon.***

- A global pattern of decreasing number of cases is observed.
  - Are local patterns different from the global pattern?
  - Since when?
  - In which regions?

Sequential Analysis of Maps: January-October / June-October

At each time *t* we must decide if there was any change in the disease pattern of each region.

Problem Formulation: Space-time Surveillance.



- Components:

$\mathcal{S} = \{s_1, s_2, ..., s_L\}$ a map with L areas;

- Process under Surveillance:

$$\mathbf{X} = \{X_t(s_l), t = 1, 2, ... \qquad l = 1, 2, ...L\}$$

- At each time *t* we observe a vector

$$\mathbf{X}_t = (\mathrm{X}_t(s_1), \mathrm{X}_t(s_2), ..., \mathrm{X}_t(s_L))'$$

- The number of observed disease cases at the region *l*: $X_t(s_l)$

- Data set until current time *n*: $\mathcal{D}_n = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$

Objective of Space-time Surveillance:

- to discover if there is a connected set of spatial locations

$$\xi = \{s_{i_1}, ..., s_{i_L}\} \subseteq \mathcal{S}$$

- and a time interval

$$[k, n], k = k_0, k_0 + 1, ..., n$$

- such that a significant change has occurred in the usual pattern of disease cases θ_0 .

- After the changes, the pattern of disease cases in location ξ at time k is θ.

- Change parameters: (ξ, k, θ)

- What is k_0 ? When does it makes sense to define k_0 ?

Surveillance Procedure

- At current time *n*, based in $\mathcal{D}_n = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$;
- construct a stochastic process $\{\Pi_n(\xi), n \geq 1\}$;
- estimate the parameters (ξ, k, θ);
- declare that change has occurred if $\Pi_n(\xi) \geq B$.

- Here we use a posterior process, which is a function of Bayes factors.

Posterior Surveillance

- **Joint density functions for $\mathcal{D}_n = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$:**

❑ When there is no change in the process:

$$f(\mathcal{D}_n|\nu > n, \theta_0) = \prod_{t=1}^{n} f(\mathbf{X}_t|\theta_0).$$

❑ When change occurs:

$$f(\mathcal{D}_n|\nu = k, \theta, \xi) = \prod_{t=1}^{k-1} f(\mathbf{X}_t|\theta_0) \times \prod_{t=k}^{n} f(\mathbf{X}_t|\theta_0, \overline{\xi}) f(\mathbf{X}_t|\theta, \xi).$$

❑ What are the following distributions?

$$f(\mathbf{X}_t|\theta_0), \quad f(\mathbf{X}_t|\theta, \xi), \quad f(\mathbf{X}_t|\theta_0, \overline{\xi})$$

Priori distributions

❑ Joint priori for time and region of change:

$$\pi(\nu = k, \xi = \xi_j) \triangleq \pi_\nu(k) \times \pi_\xi(j).$$

Priori for time of change:

❑ *p* is the probability of occurrying a change at time *k.*

❑ k_0 is the initial moment from which we are interested in detecting change. Shiryaev (1963) used the priori $k_0=1$.

We modify the Shiryaev priori :

$$\pi_\nu(k) = p(1-p)^{k-k_0} \quad \text{for} \quad k = k_0, k_0+1, ...$$

$$p \sim U(0,1)$$

Priori Distributions

Priori for the region of change (Gangnon and Kayton, 2000)**:**

$$\pi_\xi(j) = \pi(\xi = \xi_{l_j}) = E^{-1} \times exp\{-p_1 E_1(\varsigma_j) - p_2 E_2(\varphi_j) - p_3 E_3(\varrho_j)\}$$

$$E = \sum_{\xi_{l_j} \in \Xi} exp\{-p_1 E_1(\varsigma_j) - p_2 E_2(\varphi_j) - p_3 E_3(\varrho_j)\}.$$

- Parameters of shape, size and orientation of the cluster $(\varphi_j, \varsigma_j, \varrho_j)$
  - E_1, E_2 e E_3 are scores for those parameters;
  - p_1, p_2 e p_3 are weights for those parameters.

Posterior Process

Posterior probability that change has not occurred until the current time:

$$\tilde{\Pi}_n(0) = \mathbb{P}(\nu > n | \mathcal{D}_n)$$

$$\tilde{\Pi}_n(0) = \frac{Beta(1, n - k_0 + 2)}{\sum_{i=1}^{J} \sum_{k=k_0}^{n} Beta(2, k - k_0 + 1)\pi_\xi(i) BF_n(k, \xi_i) + Beta(1, n - k_0 + 2)}.$$

Posterior probability that change has occurred at some location until the current time:

$$\tilde{\Pi}_n(\xi_j) = \mathbb{P}(k_0 \leq \nu \leq n, \xi = \xi_j | \mathcal{D}_n)$$

$$\tilde{\Pi}_n(\xi_j) = \frac{\sum_{k=k_0}^{n} Beta(2, k - k_0 + 1)\pi_\xi(j) BF_n(k, \xi_j)}{\sum_{i=1}^{J} \sum_{k=k_0}^{n} Beta(2, k - k_0 + 1)\pi_\xi(i) BF_n(k, \xi_i) + Beta(1, n - k_0 + 2)}$$

Bayes Factor:

$$BF_n(k, \xi_j) = \int_\Theta \left(\prod_{t=k}^{n} \frac{f(\mathbf{X}_t | \theta, \xi_j)}{f(\mathbf{X}_t | \theta_0, \xi_j)} \right) \pi(\theta | \xi_j) d\theta, \qquad \Theta \in \mathcal{R}^d.$$

Posterior Process

- Relation between the probabilities:

$$\tilde{\Pi}_n(\xi_j) = \frac{\Psi_n(\xi_j)}{1 + (\tilde{\Pi}_n^{-1}(0) - 1)^{-1}} = \frac{\Psi_n(\xi_j)}{(1 - \tilde{\Pi}_n(0))^{-1}}$$

$$\Psi_n(\xi_j) = \frac{\sum_{k=k_0}^{n} Beta(2, k - k_0 + 1)\pi_\xi(j) BF_n(k, \xi_j)}{\sum_{i=1}^{J}\sum_{k=k_0}^{n} Beta(2, k - k_0 + 1)\pi_\xi(i) BF_n(k, \xi_i)}.$$

$\tilde{\Pi}_n(\xi_j)$ is a crescent function of cumulative BFs.

Bayes Factor

$$BF_n(k, \xi_j) = \int_\Theta \left(\prod_{t=k}^{n} \frac{f(\mathbf{X}_t|\theta, \xi_j)}{f(\mathbf{X}_t|\theta_0, \xi_j)} \right) \pi(\theta|\xi_j) d\theta, \qquad \Theta \in \mathcal{R}^d.$$

If $\pi(\theta|\xi)$ is difficult to obtain, substitute it by the *intrinsic priori* (Moreno et al., 1998).

$$\pi(\theta|\xi, \mathcal{D}_{k_0-1}) = \pi^I(\theta|\xi)\mathbb{E}_{\mathbf{X}_t|\theta,\xi}\left[BF_0^I(k_0, \xi)\right]$$

where

$$BF_0^I(k_0, \xi) = \frac{\prod_{t=1}^{k_0-1} f(\mathbf{X}_t|\theta_0, \xi)}{\int_\Theta \prod_{t=1}^{k_0-1} f(\mathbf{X}_t|\theta, \xi)\pi^I(\theta|\xi)}, \qquad \Theta \in \mathcal{R}^d.$$

$$\mathcal{D}_{k_0-1}\{\mathbf{X}_{k_0-1}, \ldots \mathbf{X}_1\}$$

$\pi^I(\theta|\xi)$ is an improper conventional priori

(for example, the Jefrey priori for the problem).

Surveillance Procedure

- Surveillance function

$$\tilde{\Pi}_n = \underset{\xi_j \in \Xi}{Sup} \left[\tilde{\Pi}_n(\xi_j)\right].$$ (posterior mode)

- Instant of alarm

$$\tilde{\tau} = inf\{n \geq k_0 : \tilde{\Pi}_n > \tilde{B}\} \quad (0 < \tilde{B} < 1)$$

- Cluster location

$$\tilde{\xi} = \underset{\xi_j \in \Xi}{argmax} \left[\tilde{\Pi}_n(\xi_j)\right]$$

- Instant of change

$$\tilde{k} = \underset{k_0 \leq k \leq n}{argmax} \left[\Pi_n(k, \tilde{\xi})\right]$$

Application: New cases of Hanseniasis in the Amazon

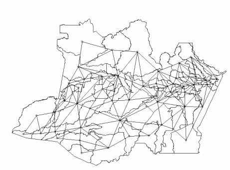
- Statistics Model

$$t = 1, 2, ..., 46 \quad \text{and} \quad L = 62$$
$$\mathbf{X}_t = (X_t(s_1), ..., X_t(s_L))'$$ (Poisson Process)

- Before change occurs:

$$X_t(s_l) \sim Poi(\rho_{k_0-1}(\xi) N_t(s_l)) \quad \forall \quad s_l \in \xi$$

$\rho_{k_0-1}(\xi)$ (Probability of a new case)



- After change occurs:

$$\nu = k,\ k_0 \leq k \leq n = 46$$ (unknown time)

$$\rho(\xi) \sim \text{Uniform } (\rho_{k_0-1}(\xi), 1)$$

Bayes Factor

$$BF_n(k,\xi) = \frac{\rho_{k_0-1}(\xi)}{1-\rho_{k_0-1}(\xi)} \frac{\Gamma(x_{n,k}(\xi)+1)}{\mu_{n,k}(\xi)^{x_{n,k}(\xi)+1}} exp(\mu_{n,k}(\xi)) \left(\mathbb{F}_z(1/\rho_{k_0-1}) - \mathbb{F}_z(1)\right).$$

$$x_{n,k}(\xi) = \sum_{t=k}^{n} \sum_{s_l \in \xi} x_t(s_l) \quad \text{and} \quad \mu_{n,k}(\xi) = \sum_{t=k}^{n} \sum_{s_l \in \xi} \mu_{0,t}(s_l)$$

$\mathrm{Z} \sim \mathrm{Gama}(x_{n,k}(\xi)+1, \mu_{n,k}(\xi))$ with D.F. $\mathbb{F}_z(.)$

Surveillance Algorithm

- Elliptic clusters (Kulldorff et al., 2006)

$\varsigma = a$ (a= Semi major axis)

$\varphi = a/b$ (b=Semi minor axis)

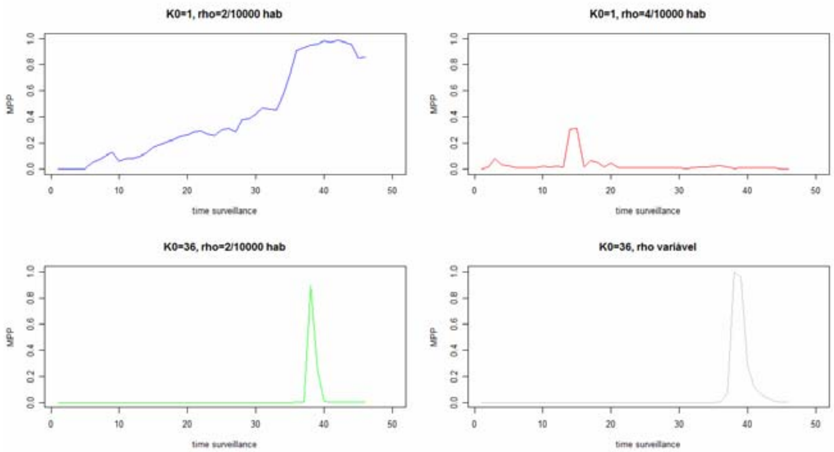
$$\pi_\xi(j) \propto exp\{-E_1(\varsigma_j) - E_2(\varphi_j)\}$$

$E_1(\varsigma_j) = d(s_l, s_{l'})/Max(s_l)$ (normalized distances)

$a = d(s_l, s_l')$

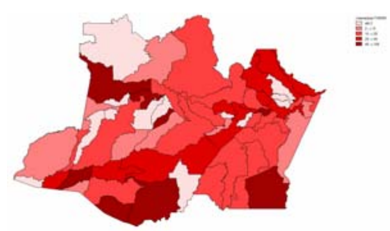
$E_2(\varphi_j) = a/b.$ (after normalizion)

Results



MPP- Maximum Posterior Probability

Results



- Cluster detected up to time n=34:

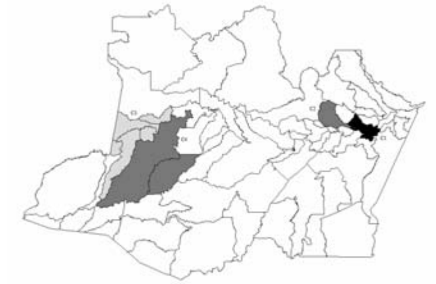
$$\pi(\xi = \tilde{\xi}) = 0.000289$$

$$\Pi_n(\tilde{\xi}) = \mathbb{P}\{\xi = \tilde{\xi}, 1 \leq \tilde{k} \leq 34\} = 0.571$$

- This result is *not* interesting from the surveillance viewpoint.

- We are only interested in “alive” clusters.

Results



- Locations of MPPs in the last four months, with k_0=36.

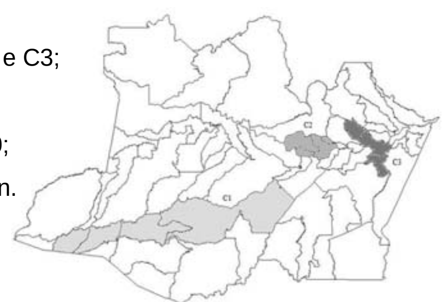
$$\tilde{\Pi}_{43} = 0.0349, \quad \tilde{\xi}_1 = C1 \qquad \tilde{\Pi}_{44} = 0.0056, \quad \tilde{\xi}_2 = C2$$

$$\tilde{\Pi}_{45} = 0.0014, \quad \tilde{\xi}_3 = C3 \qquad \tilde{\Pi}_{46} = 0.0007, \quad \tilde{\xi}_4 = C4$$

- At this time there is no evidence of change of pattern of cases since the beggining of this year which persists to current time.

Simulated data study

- Injected Clusters in the map: C1, C2 e C3;
- Surveillance period: t=1,2,...n=50;
- Cases distributed over the map: 3000;
- Cases follow a multinomial distribution.



$$X_t(s_l) \sim \text{Poisson}(\rho_t(s_l)N_t(s_l))$$

$$(X_t(s_1), ..., X_t(s_L)|X^t = x) \sim \mathbf{M}(x, \phi_t(s_1), ..., \phi_t(s_l))$$

$$\phi_t(s_l') = \rho_t(s_l')N_t(s_l')/\sum_{l=1}^{L} \rho_t(s_l)N_t(s_l)$$

Simulated Cluster Model

$$X_t(s_l) \sim \mu_{0,t}(s_l) = \rho_{k_0-1} N_t(s_l) \quad \text{for all} \quad s_l \in \mathcal{S}.$$

After an unknown time: $\nu = k, k_0 \leq k \leq n$

$$X_t(s_l) \sim \mu_{1,t}(s_l) = \rho_t(\xi) N_t(s_l) \quad \text{for all} \quad s_l \in \xi = C2.$$

with

$$\rho_t(\xi) > \rho_{k_0-1}(\xi) = 0.05$$

- Simulation of the process:

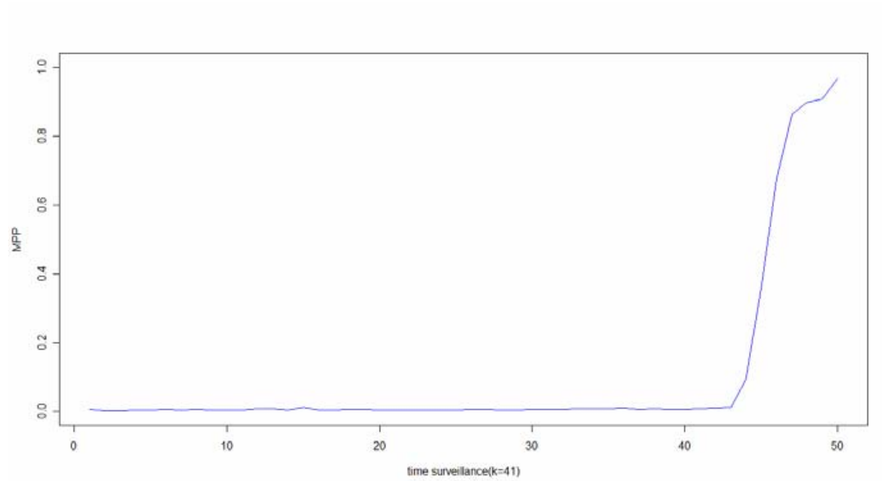
$$\rho_t(\xi) \sim U(0.05, 0.1); \quad k = 41, \quad k_0 = 1.$$

- In this model we use the same distributions, but now with real data.

$$\pi_\xi(j)$$

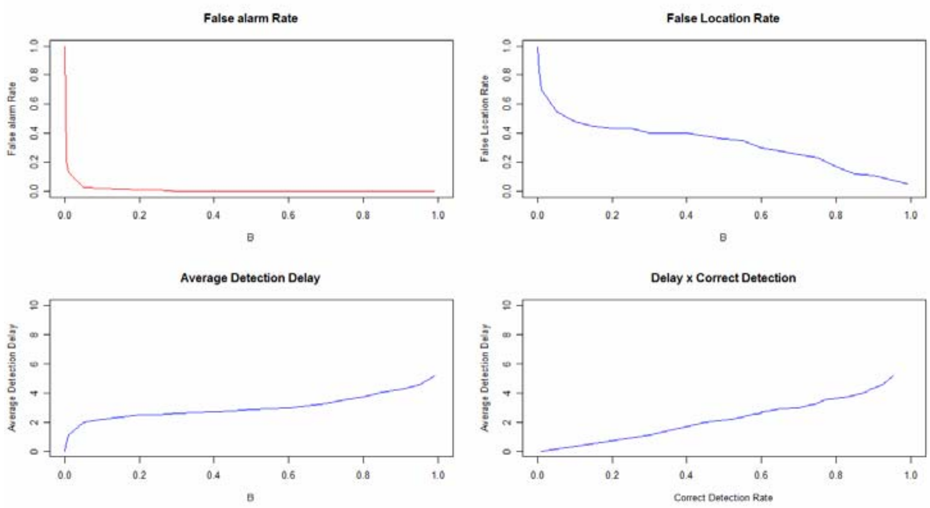
$$\rho_t(\xi) \sim U(\rho_{k_0-1}, 1).$$

Results



- MPP for the simulated process

Results



Performance measurements obtained in the simulation

Conclusions

Different from other adaptive methods, which rely simply on changing data, our method is based on changing models derived from the same data.

The Cumulative Bayes Factor method is effective for the early detection of endemic diseases, where the baseline of number of cases has non constant average.

Thanks!