Aplicações de Estimadores Bayesianos Empíricos para Análise Espacial de Taxas de Mortalidade

Alexandre E. dos Santos, Alexandre L. Rodrigues, Danilo L. Lopes

Departamento de Estatística – Universidade Federal de Minas Gerais (UFMG)
Caixa Postal 702 – 31270-901 – Belo Horizonte – MG – Brasil
{alexandrelias, alr, danilolopes}@ufmg.br

Abstract. *Maps of rates are often used for spatial dispersion analysis of certain event occurrence risk when data comes from counts by areas. However, the use of crude rates is associated to a high instability in expressing the risk of rare events in small population areas. As an alternative, will be introduced the empirical Bayesian rates; they make use either of the entire study region or the neighborhood's information in order to estimate the event occurrence risk in each area. This paper computes the rates for a real database and aims advantages on using empirical Bayes estimators instead of crude rates.*

Resumo. *Mapas de taxas são comumente utilizados para a análise da dispersão espacial do risco de ocorrência de um determinado evento quando os dados estão dispostos a partir de contagens por áreas. Um grande problema associado ao uso de taxas, porém, é a alta instabilidade que elas possuem para expressar o risco de eventos raros em regiões de população pequena. Alternativamente, são apresentadas as taxas bayesianas empíricas, que utilizam informações de toda a região ou da vizinhança para estimar o risco de ocorrência do evento em cada área. O presente artigo aplica o cálculo das taxas em um conjunto de dados reais e aponta as vantagens de utilização dos estimadores de Bayes empíricos em relação às taxas brutas.*

1. Introdução

Mapas temáticos são extremamente utilizados em diversas áreas, como saúde pública e criminologia, para visualizar a distribuição espacial de um evento, como uma doença ou um tipo de crime, e indicar possíveis áreas de alta ocorrência ou predomínio desse evento, demonstrando a necessidade de intervenção ou de investigação mais aprofundada das causas desse fato. Em muitas das aplicações, porém, em vez de observações associadas a pontos com exata posição de ocorrência do evento, os dados estão dispostos a partir de contagens por áreas, agregados por regiões geograficamente definidas, como regiões administrativas ou sanitárias. Essa agregação dos dados pode ocorrer por conveniência ou pode simplesmente refletir a forma como os dados foram disponibilizados. Em caso de dados agregados, a análise de dispersão espacial do risco de ocorrência de um acontecimento normalmente é feita via mapas de índices ou taxas de incidência, onde as áreas são sombreadas de acordo com os valores desta taxa.

A taxa bruta é o estimador mais simples para o risco de ocorrência de um evento, definindo-se como a razão entre o número de eventos ocorridos na área e o número de

pessoas expostas à ocorrência desse evento. Um grande problema associado ao uso de taxas, porém, é a alta instabilidade que elas possuem para expressar o risco de um determinado evento quando ele é raro e a população da região de ocorrência é pequena. As variações bruscas que ocorrem com estas taxas podem nada ter a ver com o fenômeno e sim com uma variabilidade associada às observações. Flutuações aleatórias casuais, como a ocorrência de um ou dois casos do evento a mais ou a menos numa localidade, causam variações substanciais nas taxas brutas se a sua população for pequena, efeito este não verificado em localidades de população grande. Além disso, para situações em que não ocorrem casos do evento em algumas regiões, a taxa bruta estima o risco de ocorrência do evento como zero, algo irreal tratando-se de dados como doenças ou crimes. Mapas de eventos baseados diretamente nessas estimativas brutas são de difícil interpretação e freqüentemente geram falsas conclusões.

A metodologia estima taxas corrigidas a partir dos valores observados utilizando-se conceitos de inferência bayesiana. O estimador Bayes empírico global calcula uma média ponderada entre a taxa bruta da localidade e a taxa global da região (razão entre o número total de casos e a população total). O estimador Bayes empírico local inclui efeitos espaciais, calculando a estimativa localmente, utilizando somente os vizinhos geográficos da área na qual se deseja estimar a taxa, convergindo em direção a uma média local em vez de uma média global. As taxas corrigidas são menos instáveis, pois levam em conta no seu cálculo não só a informação da área, mas também a informação de sua vizinhança. Mapas baseados nessas estimativas são mais interpretativos e informativos.

O método de Bayes empírico foi aplicado para a correção das taxas de óbitos por neoplasia maligna do esôfago por microrregiões dos estados de região Sul e São Paulo, entre os anos de 1996 e 2002. O presente artigo também demonstra as vantagens de utilização de taxas bayesianas empíricas em vez de taxas brutas na estimação do risco de morte por neoplasia maligna do esôfago na região analisada.

Rotinas para o cálculo de taxas bayesianas global e local estão disponíveis dentro do pacote "spdep" do ambiente estatístico livre R (http://www.r-project.org/) e também dentro do software livre de geoprocessamento TerraViewPlus 3.0.3, disponível no endereço eletrônico http://www.dpi.inpe.br/terraview/index.php; este último foi aqui utilizado para o cálculo de ambas as estimativas bayesianas empíricas pois apresenta a vantagem de se poder acessar as suas funções através de uma interface amigável, enquanto no primeiro as funções são acessadas através de janela de comandos.

2. Metodologia

Em uma abordagem bayesiana, assumimos que os conhecimentos e as incertezas acerca do real valor do risco de ocorrência de um determinado evento em cada área dentro de uma determinada região podem ser representados por uma distribuição de probabilidade. Mais especificamente, os valores desconhecidos e fixos das taxas seriam realizações de variáveis aleatórias com uma certa distribuição conjunta. O objetivo é atualizar nosso conhecimento acerca destas quantidades desconhecidas após a observação dos dados.

O conhecimento prévio sobre os riscos θ_i de ocorrência do evento é expresso na *distribuição de probabilidade a priori* ($p(\theta)$); as observações dos dados possuem uma distribuição de probabilidade que depende dos reais riscos de ocorrência do evento em

cada área, os parâmetros que se desejam estimar. Essa distribuição das observações é conhecida como *função de verossimilhança* (p(x|θ)); a partir da distribuição a *priori* e da função de verossimilhança é possível determinar, via teorema de Bayes, a *distribuição de probabilidade a posteriori* (p(θ|x)), uma atualização de seus conhecimentos anteriores. A partir da distribuição *a posteriori* pode-se derivar estimativas pontuais para os reais riscos de ocorrência do evento em cada região. Para maiores informações sobre métodos de análise bayesiana, veja Gelman (1995).

Tipicamente a distribuição a posteriori é de forma muito complexa, o que impossibilita o cálculo analítico de quantidades de interesse, tais como a média a posteriori dos parâmetros e os seus desvios padrão a posteriori. Este problema é contornado utilizando-se simulações via MCMC [Gelman 1995]. Os métodos que se utilizam dessa estimação, denominados completamente bayesianos, são preferíveis por poderem ser generalizados para modelos mais complexos, apesar do grande esforço computacional que eles requerem; alternativamente existem os métodos bayesianos empíricos, que se destacam por utilizar os dados observados para estimar os parâmetros da priori, não estando associados às dificuldades de determinação da distribuição, seja por uma insuficiência de conhecimentos sobre a variável em estudo ou por uma certa subjetividade envolvida na escolha da mesma. Bernardinelli e Montomoli (1992) apresentam um interessante estudo comparando métodos completamente bayesianos e bayesianos empíricos. Para o problema em estudo os métodos bayesianos empíricos apresentam resultados semelhantes àqueles apresentados pelos métodos completamente bayesianos e possuem a vantagem de serem de fácil integração a ambientes de geoprocessameto estando disponíveis em diversos deles, como TerraView, GeoDa, dentre outros.

Marshall (1991) propõe um método extremamente simples de ser implementado para o cálculo das estimativas bayesianas empíricas $\hat{\theta}_i$ e que não supõe nenhuma distribuição específica para os θ_i:

$$\hat{\theta}_i = C_i r_i + (1 - C_i)\hat{m}$$

em que $C_i = \dfrac{s^2 - \hat{m}/\overline{n}}{s^2 - \hat{m}/\overline{n} + \hat{m}/n_i}$, $\hat{m}$ é a taxa global dos eventos, $\overline{n}$ é o número médio de pessoas em risco, n_i é o número de pessoas observadas na área i, $s^2 = \sum_i \dfrac{n_i (r_i - \hat{m})^2}{n}$, n é o número de pessoas observadas em todas as áreas juntas e r_i é a taxa observada na área i.

A taxa bayesiana empírica global é, portanto, uma média ponderada entre a taxa bruta da localidade e a taxa global da região. Se a localidade apresentar uma população considerável, sua taxa apresentará pequena variabilidade e ela permanecerá praticamente inalterada. Se, por outro lado, a localidade apresentar uma população pequena, a estimativa da taxa bruta terá grande variância e pouco peso será atribuído a essa taxa

instável, tornando a taxa bayesiana mais próxima do valor esperado de uma área escolhida ao acaso naquela região.

A distribuição a priori para θ_i em todas as estimativas bayesianas discutidas acima são não-espaciais; isto é, a média e a variância a priori são estabelecidas como constantes para todas as áreas. O estimador bayesiano empírico pode ser generalizado para incluir efeitos espaciais, ao se exigir que a estimativa ajustada para uma área se aproxime de uma média da "vizinhança" em vez de uma média global (considera-se como vizinhança da área *i* todas as demais áreas que compartilham fronteira com a i-ésima área). Com isso adiciona-se uma suavidade espacial ao modelo, pois as estimativas bayesianas globais não variam segundo a configuração espacial das áreas, o que não parece razoável na maioria das situações. Para o cálculo de estimativas bayesianas empíricas locais modifica-se a distribuição a *priori* de θ_i para permitir que a média e a variância sejam relacionadas à vizinhança de *i*, em vez de permanecerem constantes para todas as áreas; então a taxa observada em pequenas populações irá convergir para uma média local em vez da global. A estimativa bayesiana local consiste em uma pequena alteração do método proposto por Marshall (1991): em vez de $\hat{m}$ e $\overline{n}$ utilizam-se $\hat{m}_i$ e $\overline{n}_i$, representando, respectivamente, a taxa local na vizinhança da área *i* e o número médio de eventos nesta vizinhança.

3. Aplicações do método

Como uma aplicação do método de Bayes empírico, foram utilizados dados de óbitos de indivíduos do sexo masculino entre 50 e 59 anos por neoplasia (tumor) maligna do esôfago para as 157 microrregiões dos estados de São Paulo, Paraná, Santa Catarina e Rio Grande do Sul entre os anos de 1996 e 2002, bem como os seus respectivos totais populacionais, ambos obtidos a partir do endereço eletrônico do Departamento de Informação e Informática do SUS – DATASUS (http://www.datasus.gov.br/). O software TerraViewPlus 3.0.3 foi utilizado para o cálculo das taxas brutas e das taxas bayesianas global e local.

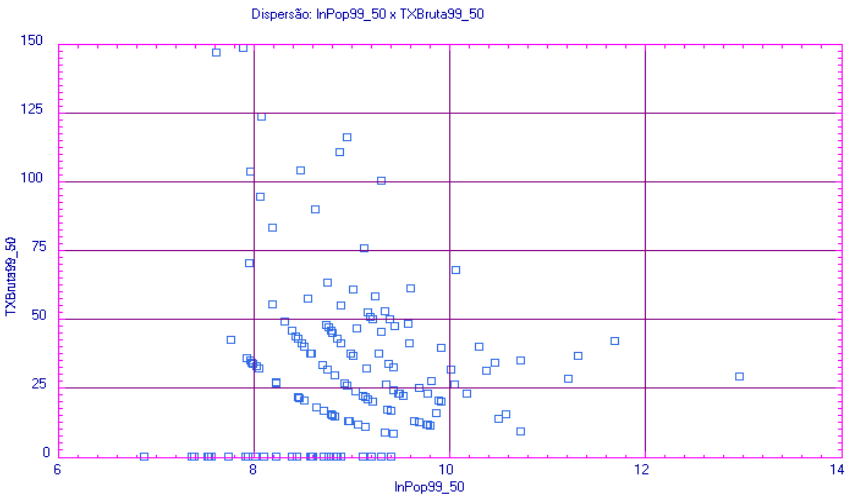

Figura 1. Gráfico de dispersão do logaritmo populacional versus taxa bruta de ocorrência de neoplasia maligna do esôfago na população masculina com idade entre 50 e 59 anos no ano de 1999.

A grande variabilidade das taxa brutas ao expressar o risco de morte por neoplasia do esôfago em localidades de população pequena está representada na Figura 1. No gráfico, o eixo horizontal é o logaritmo do total populacional e cada ponto representa uma microrregião. Observa-se um formato de funil com a variação das taxas sendo muito maior nas microrregiões com menor contingente populacional. Essa grande variabilidade em microrregiões pequenas está associada à pura flutuação aleatória. Por exemplo, a microrregião de Bananal (SP), que não registrou nenhum caso de morte por tumor no esôfago em homens entre 50 e 59 anos no ano de 1999, teria a sua taxa aumentada de 0 para 103 óbitos por neoplasia do esôfago por 100.000 indivíduos caso ocorresse apenas um único caso de morte pela doença entre os seus 969 homens entre 50 e 59 anos. Caso isso ocorresse o risco de morte pela doença em Bananal em vez de nulo seria estimado como o oitavo maior dentro da região em estudo.

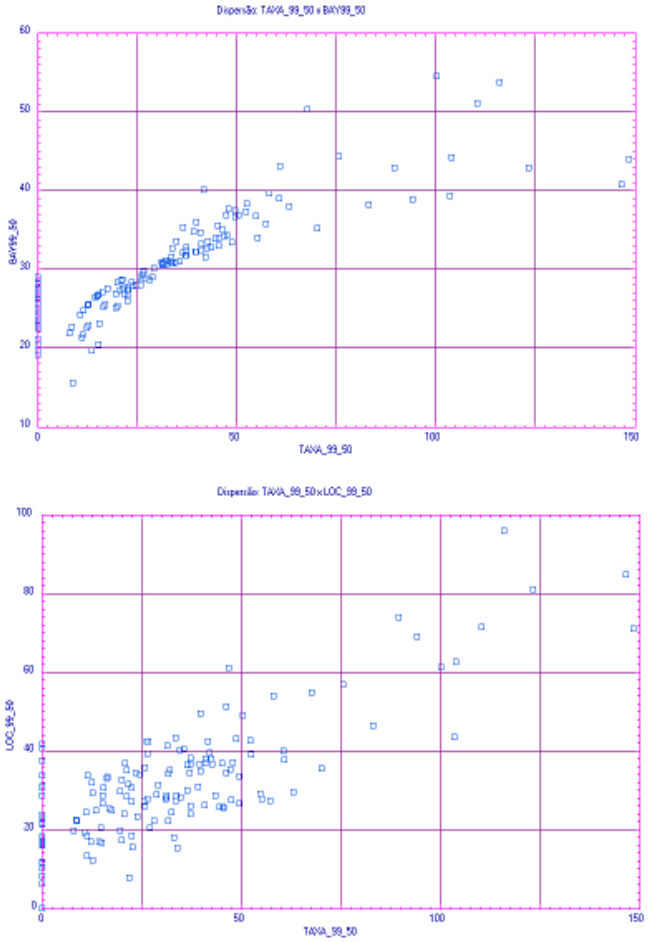

Figura 2. Gráficos de dispersão da taxa bruta versus taxas bayesianas globais e locais de óbitos por neoplasia maligna do esôfago para população masculina com idade entre 50 e 59 anos no ano de 1999.

A Figura 2 apresenta gráficos de dispersão entre as taxas brutas e as taxas bayesianas empíricas das microrregiões analisadas. Verifica-se um grande número de microrregiões cujos riscos de mortalidade por neoplasia do esôfago são estimados pela taxa bruta como nulos, o que significaria dizer que a população masculina entre 50 e 59 anos daquela microrregião não está sujeita ao risco de morte por neoplasia do esôfago,

algo que obviamente não correspondente à realidade. A Figura 3 apresenta gráficos de dispersão das taxas bayesianas contra o logaritmo do total populacional. A taxa bayesiana global apresenta um formato de funil inverso daquele observado para a taxa bruta, pois o método de Bayes empírico global aproxima a taxa bayesiana da taxa global para as pequenas populações, enquanto para as populações grandes a taxa bayesiana global aproxima-se da taxa bruta e incorpora a sua variabilidade. Já para a taxa local, quando a população é pequena a taxa se aproxima da taxa de eventos de cada vizinhança, e por isso o padrão verificado no gráfico de dispersão é uma nuvem de pontos.

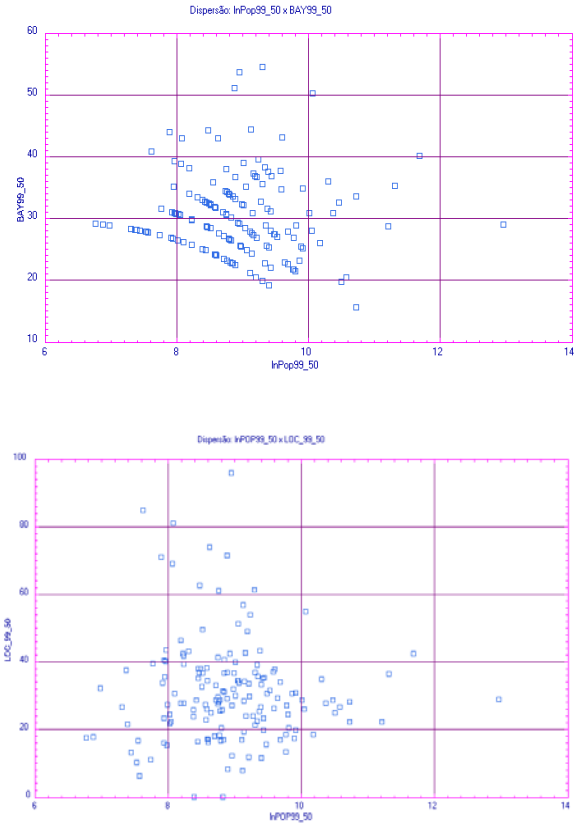

Figura 3. Gráficos de dispersão obtidos a partir das taxas bayesiana global (acima) e local (abaixo) de neoplasia maligna do esôfago versus o logaritmo populacional. Ambos obtidos com dados da população masculina, com idade entre 50 e 59 anos no ano de 1999.

A Figura 4 apresenta mapas temáticos para as três taxas: bruta, bayesiana global e bayesiana local. No mapa de taxas brutas observa-se um predomínio das cores azuis, que representam valores baixos de taxas, e a existência de algumas microrregiões isoladas em vermelho, que representa valores altos de taxas. Para o mapa de taxas bayesianas globais o padrão observado é de predomínio de valores em torno de uma média global, com algumas microrregiões apresentando taxas bem menores e outras

apresentando taxas bem maiores. Isso se deve à tendência do método de Bayes Empírico Global possui de aproximar as estimativas da média global. Já para o mapa de taxas bayesianas locais observa-se um padrão mais suave de cores, com microrregiões próximas no espaço possuindo proximidade de tons, permitindo que se visualize mais perfeitamente grupos de microrregiões com altos ou baixos índices de mortalidade pela doença.

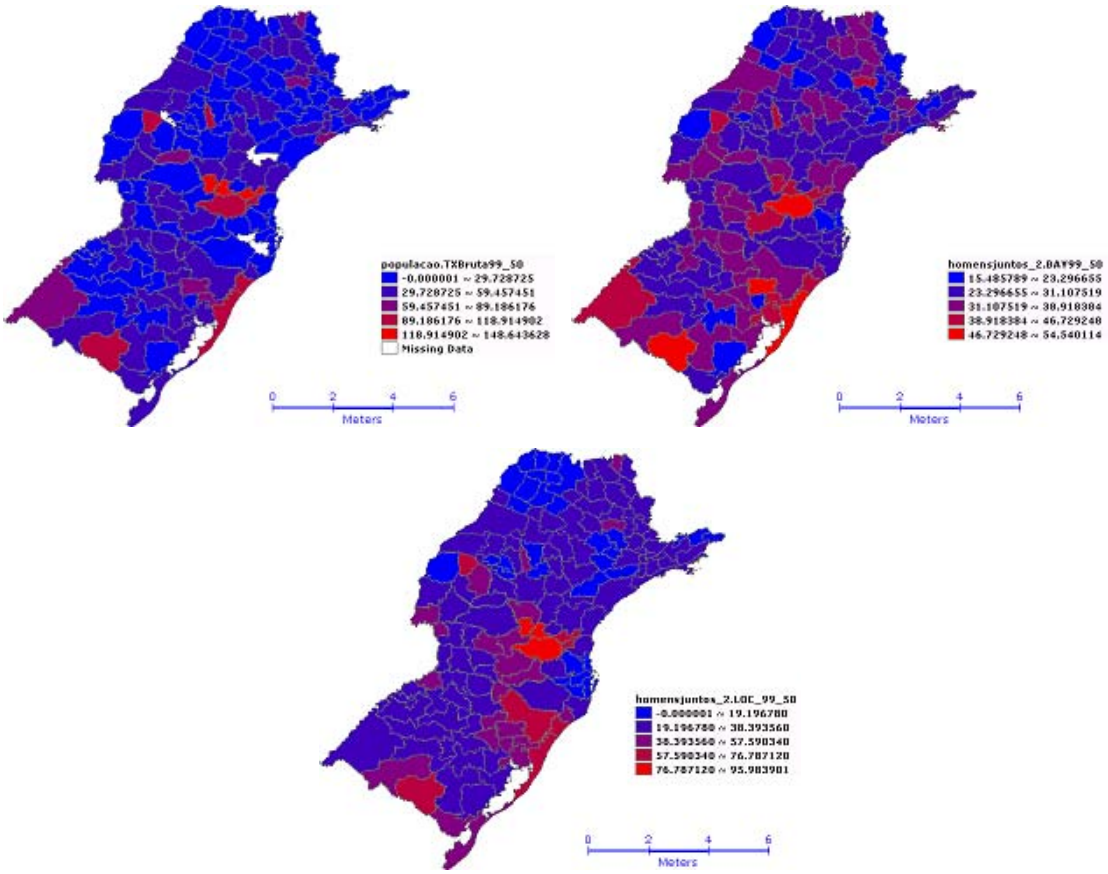

Figura 4. Mapas temáticos para as três taxas obtidas de óbitos por neoplasia maligna do esôfago para população masculina com idade entre 50 e 59 anos no ano de 1999: taxa bruta (esquerda); taxa bayesiana global (direita) e taxa bayesiana local (abaixo).

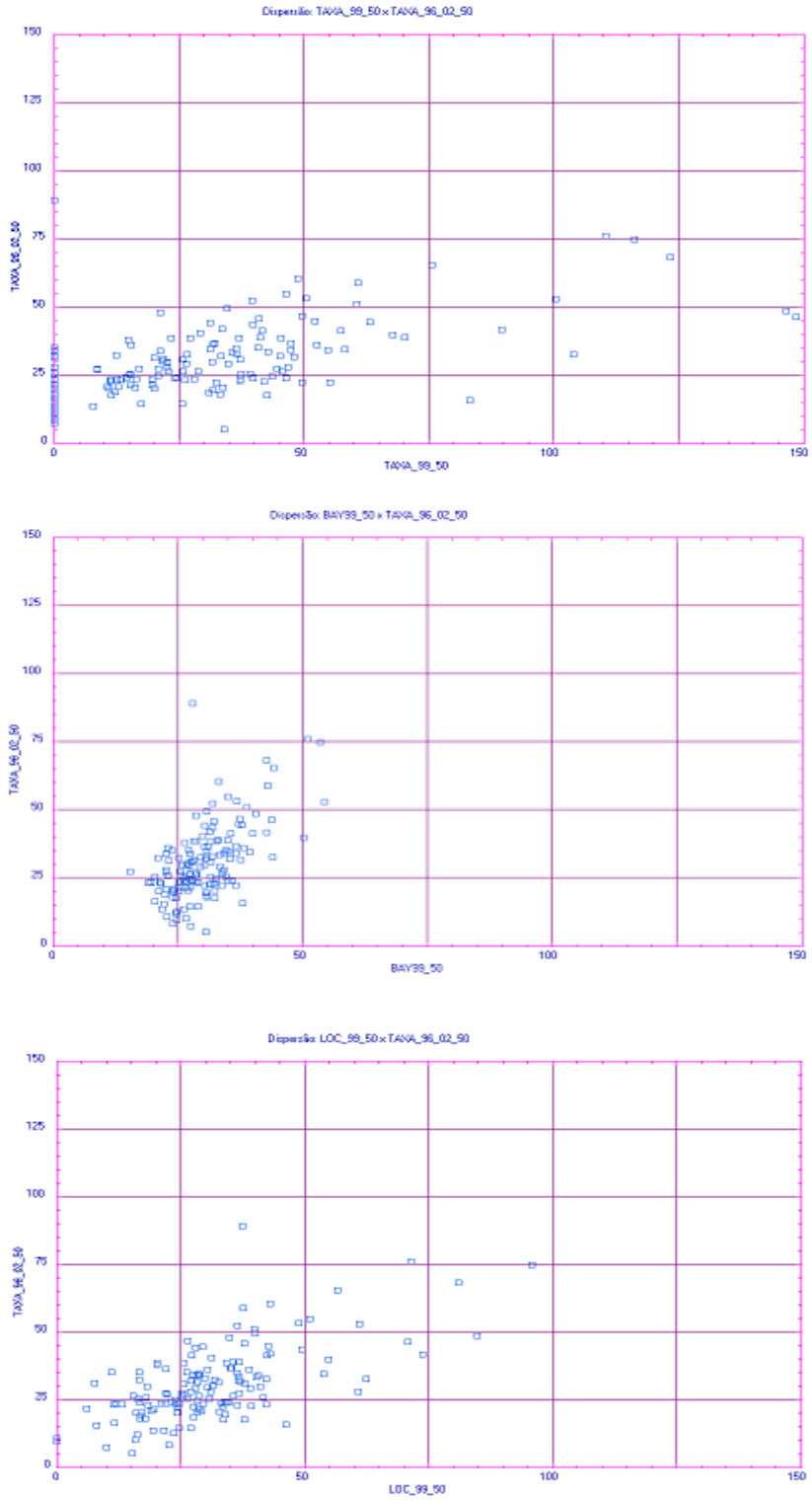

Figura 5. Gráficos de dispersão obtidos com a taxa bruta do período de 1996 a 2002 versus taxas brutas (alto), bayesiana global (centro) e bayesiana local (abaixo) do ano de 1999.

Para se testar o poder de estimação do risco de morte por neoplasia do esôfago das taxas bayesianas frente à estimação feita pela taxa bruta foi utilizado o seguinte procedimento: inicialmente calculou-se uma taxa do período de 1996 a 2002 como a

razão entre o número total de óbitos de indivíduos do sexo masculino entre 50 e 59 anos por neoplasia maligna do esôfago e a soma das populações correspondentes de cada ano do período analisado. Sob a suposição de que o risco de morte pela doença possui um comportamento linear dentro do período de 1996 a 2002, esta taxa total do período pode ser considerada uma boa estimativa para o risco de morte pela doença no meio do período, que, no caso, corresponde ao ano de 1999. Como o período em estudo não é muito extenso, essa suposição pode ser facilmente adotada. Assim, espera-se que as melhores estimativas do risco de morte geradas somente a partir dos dados de 1999 estejam próximas da taxa calculada a partir de todo o período de 1996 a 2002. A Figura 5 apresenta gráficos de dispersão entre a taxa total do período e cada uma das taxas utilizadas no presente artigo: bruta, bayesiana global e bayesiana local. Observa-se que, para as taxas bayesianas o gráfico se aproxima da reta com inclinação 1 e intercepto 0, indicando uma proximidade entre a taxa total do período e as taxas bayesianas do ano de 1999, enquanto no gráfico de taxa bruta versus a taxa do período existe uma nuvem de pontos que não indica nenhuma relação aparente. Essa análise aponta evidências do quão vantajosas são as taxas bayesianas para estimação do risco de ocorrência do evento.

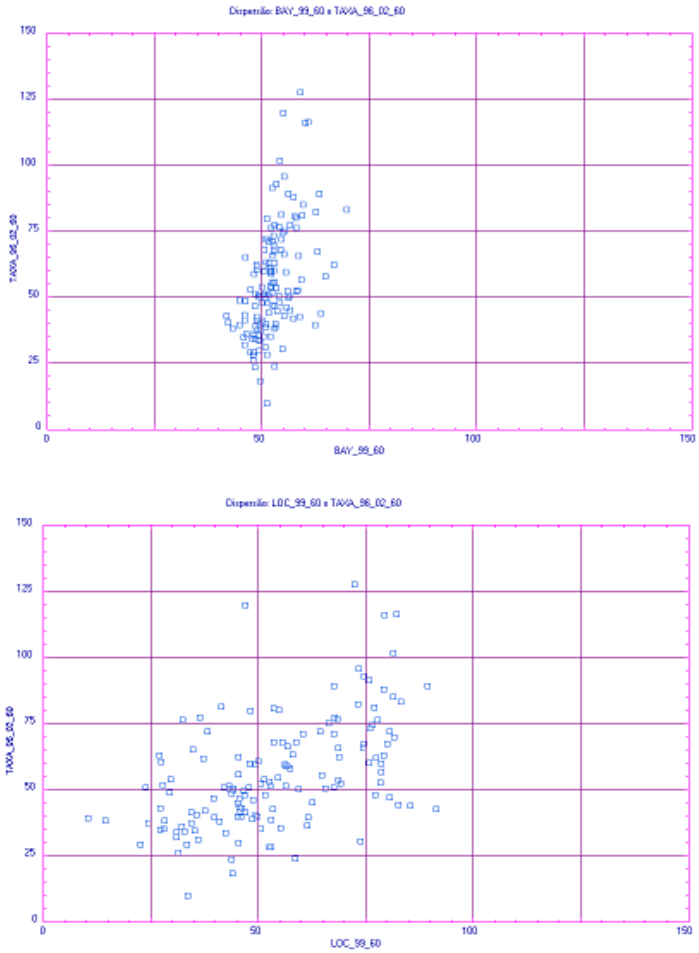

Figura 6. Gráficos de dispersão obtidos a partir de dados de neoplasia maligna do esôfago para a população masculina de idade entre 60 a 69 anos do ano de 1999. Taxas bayesiana global (acima) e local (abaixo) versus taxa bruta do período de 1996 a 2002.

É importante ressaltar que nem mesmo os métodos bayesianos conseguem estimar com exatidão o risco de ocorrência em situações em que o fenômeno em estudo é muito raro ou as populações são extremamente baixas. Por exemplo, para a faixa etária de 60 a 69 anos, por exemplo, os valores estimados pela taxa bayesiana global convergiram em grande parte para a média global, como demonstrado na Figura 6; isso se deve principalmente aos baixos valores de população observados para essa faixa etária, o que transmite uma grande incerteza às estimativas da taxa bruta. Para as faixas etárias mais jovens as estimativas bayesianas são muito instáveis devido à extrema raridade dessa causa de morte em idades mais jovens.

4. Discussões e Conclusões

A partir dos dados utilizados para análise foram apontadas vantagens da aplicação de taxas bayesianas empíricas em relação às taxas brutas. As taxas bayesianas demonstraram apresentar menor variabilidade e uma maior adequação aos reais riscos de ocorrência do evento em cada área da região em estudo. Para aquelas áreas em que a taxa bruta apresentaria grande variabilidade, o método de Bayes Empírico Global estas taxas em direção a média global da região. O método de Bayes Empírico Local adiciona um efeito espacial às estimativas, tornando-as próximas de uma média local, o que gera uma certa suavidade no mapa temático.

No ano em estudo, os métodos de Bayes Empíricos estimaram de forma extremamente satisfatória o real risco de ocorrência do evento em cada área para aquelas faixas etárias em que o número de casos e os totais populacionais são suficientes. Faixas etárias em que as contagens são pequenas os estimadores bayesianos empíricos perdem em grande parte a qualidade de estimação do risco real, mas ainda se apresentam melhor do que as taxas brutas. Para o conjunto de dados analisado as taxas bayesianas empíricas locais apresentaram melhores estimativas do que as taxas bayesianas globais, demonstrando que uma variação do risco suave é adequada para o conjunto de dados.

Bibliografia

Assunção, R. M. (2003). "Estimadores Bayesianos Empíricos Espaciais de Taxas", http://www.est.ufmg.br/~assuncao/cursos/espacial/aplicado/aulas/empbayes.pdf, September.

Bailey T. C., Gatrell A. C. (1995) "Interactive spatial data analysis". Harlow Essex, England: Longman Scientific & Technical; New York, NY: J. Wiley, p. 303-306.

Bernardinelli, L., Montomoli, C. (1992) "Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk". *Statistics in Medicine,* 11, 983-1007.

Gelman, A. (1995) "Bayesian data analysis". London; New York: Chapman & Hall.

Marshall, R. M. (1991) "Mapping disease and mortality rates using Empirical Bayes Estimators", In *Journal of the Royal Statistical Society, Series C: Applied Statistics*, Vol. 40, No. 2, pages 283-294. London, England.