Estimação de sub-registros de óbitos em pequenas áreas com os métodos bayesiano empírico e algoritmo EM

Josivan Ribeiro Justino*
Flávio Henrique Miranda de Araújo Freire**
Paulo Sérgio Lucio***

Um grande problema em estimativas demográficas no Brasil diz respeito ao nível e padrão da mortalidade. Os demógrafos que trabalham com mortalidade, no país, ainda não se sentem tão seguros sobre o real comportamento desta componente da dinâmica populacional. Por outro lado, necessita-se da disponibilidade de indicadores de mortalidade para níveis geográficos mais desagregados, sobretudo municípios. O problema é que quanto mais desagregado, mais complexo se torna o trabalho de estimar qualquer indicador social ou demográfico. Neste trabalho, objetiva-se estimar e propor correção de sub-registros de óbitos no nível municipal, segundo grupos etários, por meio de dois métodos: estimador bayesiano empírico (BE) e algoritmo EM (Expectation-Maximization). Para que os dois métodos fossem operacionalizados entre municípios semelhantes, foram realizados dois exercícios: agruparam-se os municípios segundo a mesorregião; e agruparam-se os municípios em grupos homogêneos, gerados a partir de uma análise de cluster utilizando as variáveis grau de urbanização, proporção de óbitos por causas externas e a população de cada município. Foram utilizados dados do Estado do Rio Grande do Norte, referentes a 2000. Para o total do Estado, estimou-se um sub-registro de 11% com o estimador BE e de 12,9% com o algoritmo EM. Outro resultado importante é a possibilidade de avaliar o grau de cobertura de óbitos por grupos etários em municípios e em qualquer nível de agregação a partir deste.

Palavras-chaves: Sub-registros de mortalidade. Estimador bayesiano empírico. Algoritmo EM.

Introdução

Nos últimos anos, o Ministério da Saúde tem concentrado esforços para melhorar a cobertura e qualidade dos registros de estatísticas vitais. Segundo o relatório dos indicadores sociodemográficos e de saúde (IBGE, 2009), desde a implantação dos Sistemas de Informações de Estatísticas Vitais, as bases de dados sobre nascimentos e óbitos vêm sendo ampliadas e melhoradas significativamente em todo o Brasil.

* Estatístico, mestrando em Demografia pelo Programa de Pós-Graduação em Demografia – PPGDEM do Departamento de Estatística da Universidade Federal do Rio Grande do Norte – UFRN.

** Mestre e doutor em Demografia pela Universidade Federal de Minas Gerais – UFMG. Professor do Programa de Pós-Graduação em Demografia e do Programa de Pós-Graduação em Estudos Urbanos e Regionais da Universidade Federal do Rio Grande do Norte – UFRN.

*** Mestre em Estatística pela Universidade Estadual de Campinas, doutor em Geofísica pelo Institut de Physique du Globe de Paris, e pós-doutor pelo Instituto Superior Técnico de Lisboa e pelo Centro de Geofísica da Universidade de Évora. Professor do Programa de Pós-Graduação em Demografia e do Programa de Pós-Graduação em Ciências Climáticas da Universidade Federal do Rio Grande do Norte – UFRN.

Entre outras ações desenvolvidas, em algumas regiões foram destacados consultores para treinar funcionários das Secretarias de Saúde e tentar detectar possíveis falhas no processo de levantamento dos registros vitais. Para se ter uma ideia, no Rio Grande do Norte, entre os resultados deste trabalho, destaca-se a descoberta de cemitérios clandestinos em alguns municípios. Todo esse esforço tem melhorado a qualidade dos registros de óbitos, bem como de nascimentos, nas mais variadas regiões do Brasil, ainda que persistam as diferenças regionais.

Chackiel (1987) propôs uma classificação sobre a qualidade dos registros de mortalidade, segundo o percentual de óbitos no grupo de causas mal definidas: menos de 15% – informação muito boa; entre 15% e 25% – informação relativamente boa; entre 25% e 40% – informação pouco confiável; mais de 40% – informação deficiente.

Tomando dois Estados do Nordeste brasileiro como exemplo, região caracterizada por qualidade mais precária dos registros vitais do que o Sul e o Sudeste, segundo este critério, o Rio Grande do Norte apresentava, em 2000, informação pouco confiável no que se referia aos registros de mortalidade, pois totalizava 27,6% de óbitos com *causas mal definidas*. Em 2008, esse percentual reduziu-se para apenas 3,37%. Na Paraíba, a proporção de mortes por causas mal definidas correspondeu a 7,6%, em 2008, demonstrando que os esforços para melhoria da informação de mortalidade têm apresentado bons resultados, mesmo em Estados onde os dados não eram confiáveis.

Contudo, vale ressaltar que a avaliação da qualidade dos dados não se restringe ao percentual de causas mal definidas. Baixo percentual de causas mal definidas não significa, necessariamente, que não haja sub-registro nos dados.

Quando o nível de desagregação do estudo é municipal, o efeito de subnotificações de estatísticas vitais é ainda mais impactante. No caso de informações sobre mortalidade, a ausência de registros de óbitos terá maior impacto quanto menor for a população. Nesse sentido, é justamente nos menores municípios onde mais se percebe o efeito dos sub-registros.

A preocupação com nível e padrão de mortalidade, tanto para Estados e regiões como para municípios, acarretou a implantação de um quesito no Censo Demográfico de 2010 sobre mortalidade, indagando se houve óbito naquele domicílio.

Contudo, não se pode prescindir das informações regulares e periódicas disponíveis em bases de dados como o Datasus do Ministério da Saúde e o Registro Civil disponível no Instituto Brasileiro de Geografia e Estatística (IBGE). Portanto, avaliar a qualidade das informações destas bases e corrigi-las, quando necessário, é fundamental para obtenção de indicadores confiáveis de mortalidade.

Nesse trabalho, a proposta é utilizar duas metodologias para estimar um fator de correção que, aplicado aos óbitos observados, corrija as subnotificações, gerando um número de óbitos viável para a construção de indicadores de mortalidade confiáveis para os municípios brasileiros. A proposta é estimar não um único fator de correção de sub-registro de mortalidade, mas sim, para cada município, fatores de correção por faixas etárias: crianças e jovens; adultos; e idosos.

Os métodos utilizados foram o *estimador bayesiano empírico* e o *algoritmo EM*. No exercício empírico aplicaram-se os métodos aos dados dos municípios do Rio Grande do Norte, para 2000. As informações de óbitos e população foram extraídas da base de dados do Datasus[1] do Ministério da Saúde.

A opção por usar dados de 2000 passa pelos objetivos do trabalho. Como se pretende propor um procedimento para estimação de fator de correção de sub-registro de óbitos para os municípios brasileiros, a ideia é eliminar o máximo de viés possível na população, para que os dados de óbitos sejam corrigidos com maior precisão. Será visto mais a frente que os dois métodos propostos dependem do tamanho populacional

[1] www.datasus.gov.br.

para estimar o fator de correção de óbitos. Portanto, o último ano para o qual se têm disponíveis as duas informações necessárias é 2000, ou seja, o óbito registrado no Datasus e a população enumerada de forma censitária. Embora o IBGE já tenha disponibilizado a população total dos municípios levantada pelo censo 2010, para este ano ainda não há informação dos óbitos.

Na próxima seção, destacam-se os aspectos metodológicos, como fontes de dados, especificidades de cada método adotado, área de abrangência[2] dos exercícios empíricos, bem como a construção destas áreas de abrangência. Em seguida, são apresentados os resultados e as discussões.

Metodologia

A correção de sub-registros no nível municipal foi realizada por meio de dois métodos: estimador bayesiano empírico e o algoritmo EM (Expectation-Maximization). A ideia central é, a partir de informações de municípios próximos ou similares, obter estimativas de óbitos esperados. O quociente entre os óbitos observados e estes esperados resulta no grau de cobertura de óbitos do Sistema de Informação sobre Mortalidade (SIM) para municípios. O fator de correção é o inverso do grau de cobertura. Cabe destacar que este exercício foi feito por grupo etário quinquenal.

Estimador bayesiano empírico

O objetivo central do estimador bayesiano empírico utilizado neste trabalho parte do estimador apresentado por Marshall (1991), que propõe um estimador *contração* para taxas de mortalidade de menores de cinco anos em setores censitários de Aukland, na Nova Zelândia. Esse autor pretendia suavizar a flutuação aleatória das taxas de mortalidade, aproximando uma taxa observada em determinada área pequena a uma taxa média global, ou, ainda, a uma taxa média dos setores vizinhos, considerando o tamanho da população da área em questão. A função que operacionaliza este procedimento é a seguinte:

$$\hat{\theta}_i = m + c_i \cdot (x_i - m) \quad (1)$$

Onde θ_i corresponde à taxa suavizada; m refere-se à taxa média global ou à taxa média dos vizinhos; x_i é a taxa da área i; c_i é o fator de contração, quanto menor a população do município i menor será o peso de sua taxa x_i na taxa final suavizada, dada por θ_i.

Nesse trabalho, a aplicação do estimador bayesiano não é propriamente em taxa de incidência. A ideia é usar o estimador bayesiano empírico da fórmula anterior para estimar a razão entre os óbitos observados e o número esperado de óbitos em determinado município i (K_i).

Parte-se do princípio de que os óbitos têm distribuição Poisson $[ob_i \sim poisson(Esp_i \cdot \theta_i)]$, onde Esp_i são óbitos esperados para o município i, assumindo que o risco de um óbito ocorrer em i é o mesmo que numa área maior ou área de referência (A_r) $[Esp_i = \frac{Ob_{Ar}}{n} . n_i]$, onde Ob_{Ar} são os óbitos da área de abrangência, n_i é a população do município i e n é a população da área de referência que equivale à soma dos n_i's. θ_i, estimado inicialmente por K_i, representa o risco de um município i ter mais ou menos óbitos do que o esperado sob a hipótese de que o risco de óbitos neste município é o mesmo que o risco de óbitos na área maior de abrangência A_r. Para θ_i, assume-se uma priori não especificada com momentos constantes para todo município i $[E(\theta_i) = m$ e $V(\theta_i) = A]$. Com isso, o estimador bayesiano empírico de *contração* para estimar a razão entre óbitos observados e óbitos esperados é dado por:

$$\hat{\theta}_i = m + c_i \cdot (K_i - m) \quad (2)$$

Onde θ_i é a razão entre óbitos observados e esperados ajustado pelo método

[2] Para um melhor resultado com os métodos aplicados, tenta-se agrupar os municípios segundo critérios de similaridade, pois as duas metodologias utilizam informações do grupo de municípios como um todo para tentar corrigir as informações de um determinado município. Esse grupo de municípios similares é que se chama área de abrangência. Foi feito também outro exercício usando como área de abrangência a mesorregião do Estado em que o município esta inserido.

bayesiano empírico, ou seja, é o grau de cobertura de registros de óbitos no município i; m corresponde à razão entre óbitos observados e esperados para a área maior de abrangência, que por construção será sempre igual a 1; K_i já foi definido antes e representa a razão entre óbitos observados e esperados calculado originalmente $[K_i = \frac{Ob_i}{Esp_i} = \frac{Ob_i}{\frac{Ob_{Ar}}{n}.n_i}]$. O fator de contração ($c_i$) do valor de K_i original para o valor médio (m) é estimado por:

$$\hat{c}_i = \frac{V(\theta_i)}{V(K_i)} = \frac{s^2 - \hat{m}\sum_{i-1}^{N}\frac{n_i}{Esp_i}\frac{n}{}}{s^2 - \hat{m}\sum_{i-1}^{N}\frac{n_i/n}{Esp_i} + \frac{\hat{m}}{Esp_i}}$$

. Observe que c_i realmente funciona como um fator de contração do real valor de K_i com relação ao valor médio m, sendo que essa contração é tanto maior quanto menor for o valor esperado Esp_i, que, por sua vez, será tanto menor quanto menor for a população (n_i) do município. Portanto, em última análise, quanto menor for a população de um município maior será a variância de K_i, o que implica dizer que menos confiável será a estimativa de K_i nesta área, e isto ocorre por influência do tamanho populacional. Quanto menor for a população, mais a estimativa da relação entre óbitos observados e óbitos esperados (grau de cobertura) será influenciada pelo valor médio.

Algoritmo EM

O algoritmo EM (Expectation-Maximization) é um método de estimação que apresenta solução para diversos problemas de maximização da verossimilhança, no contexto de dados incompletos. Os conceitos fundamentais deste método foram introduzidos por Dempster, Laird e Rubin (1977). A ideia central é otimizar os parâmetros de uma função de distribuição de probabilidades, de forma que esta represente os dados da maneira mais verossímil possível. O modelo mais utilizado é aquele cuja função de distribuição de probabilidades é dada por uma mistura de Gaussianas, no entanto, neste artigo utiliza-se a adaptação desta metodologia à mistura de Poissons, uma vez que os dados deste trabalho são contagens.

A subnotificação de dados, que são elementos faltantes ou desconhecidos de variáveis latentes como os registros de óbito, é o principal objetivo de estimação do algoritmo EM. Neste artigo, tal método busca estimar a verdadeira frequência de óbitos em cada município do Rio Grande do Norte, para 2000.

Descrição do algoritmo EM – mistura de Poissons

Nas últimas décadas, modelos de mistura têm despertado interesses por sua utilidade como um método de modelagem extremamente flexível. Trata-se de ferramenta adequada para a modelagem de dados heterogêneos, em que as observações pertencem a k populações distintas.

A distribuição de Poisson é aplicável a inúmeras situações, sendo particularmente bastante útil para modelar processos com flutuação aleatória casual, que provoca variações substanciais nas taxas brutas, se a população do município for pequena.

Passando a especificar a aplicação do algoritmo EM utilizada neste artigo, assumiram-se duas amostras aleatórias de Poissons mutualmente independentes: $X_1, ..., X_n$ com parâmetro $(\tau(i))$ e $Y_1, ... Y_n$ com parâmetro $(\beta*\tau\ (i))$, onde X_i é a incidência de mortalidade dada pelos óbitos do município i, cuja taxa de ocorrência é uma função de efeito total (β) e de um fator adicional, dado pelo parâmetro de subnotificação de óbitos na área "i" $(\tau(i))$. Esse parâmetro $\tau(i)$ representa o valor esperado de óbitos no município ou área i, que depende do tamanho populacional desta área, dado por Y_i.

Como o objetivo é estimar dados faltantes, que têm uma estimação complexa, a verossimilhança dos dados-incompletos é obtida pela soma conjunta em x_i: $\sum_{\theta}^{\infty} x_i$ $f((x_1,y_1),...,(x_n,y_n)|\ \beta,\ \tau(i),...,\ \tau(n))$, que é maximizada para obter a verossimilhança dos dados.

O algoritmo EM permite maximizar $L(\beta, \tau(1),...,\tau(n)|$dados-incompletos) consi-

derando apenas a $L(\beta, \tau(1),...,\tau(n)|$dados--completos) e a função de probabilidade condicional k(dados-aumentados|dados--incompletos, $(\beta, \tau(1),...,\tau(n))$. Desta forma:

$\log[L(\beta, \tau(1),...,\tau(n)|$dados-incompletos] = $= \log[L(\beta, \tau(1),...,\tau(n)|$dados-completos] - $- \log[k$(dados-aumentados|dados-incompletos, $(\beta, \tau(1),...,\tau(n))]$

Os Máximos Locais Estimados (MLE), por diferenciação da log-verossimilhança, são obtidos pela resolução da equação:

$$\hat{\beta}[j] = \frac{\sum_{i=1}^{n} pop_j}{\hat{\tau}[1,j-1] + \sum_{j=2}^{n} ob_j} \quad (3)$$

para $j \geq 2$. $\hat{\beta}[1]$ é o inverso da taxa de mortalidade para toda a área maior, retirando os dados de óbitos e população do município i $[\hat{\beta}[1] = \frac{\sum_{j=2}^{n} pop_j}{\sum_{j=2}^{n} ob_j}]$.

A matriz de parâmetros τ tem n linhas e m colunas. As linhas representam os municípios e as colunas correspondem ao número de simulações requisitado. O ponto de partida para estimar os parâmetros da matriz τ é a média dos óbitos de todos os municípios da área maior de abrangência, não contabilizando os óbitos do município 1:

$$\hat{\tau}[1,1] = \frac{\sum_{i=2}^{n} ob_i}{n-1}$$

Os óbitos esperados para a j-ésima simulação para o município 1 são dados por: $\hat{\tau}[1,j] = \frac{\hat{\tau}[1,j-1] + Pop[1]}{\hat{\beta}[j]+1}$. Para os demais municípios ou áreas, τ é estimado da seguinte forma: $\hat{\tau}[i,j] = \frac{Ob[i] + Pop[i]}{\hat{\beta}[j]+1}$, para $i \geq 2$. $Ob[i]$ são os óbitos observados no município i.

O procedimento de estimação dos parâmetros β e τ é iterativo. A partir da semente $\hat{\tau}[1,1]$, que é a média de óbitos, excluindo-se o município 1, o processo estima $\hat{\beta}[2]$ usando a fórmula (3). Essa estimativa de $\hat{\beta}[2]$ será usada para estimar toda a segunda coluna da matriz τ, $\hat{\tau}[i,2]$. Depois, para estimar $\hat{\beta}[3]$, serão utilizadas as estimativas da segunda coluna de $\tau(\hat{\tau}[i,2])$. Em seguida, a estimativa de $\hat{\beta}[3]$ servirá para estimar a terceira coluna (terceira simulação) de $\tau(\hat{\tau}[i,3])$, e assim por diante. E qual o critério de parada?

O critério adotado para a parada da iteração foi a distância euclidiana entre a simulação j e j-1, tendo como critério de convergência a distância menor que 0,001.

$$\sqrt{(\hat{\beta}[j] - \hat{\beta}[j-1])^2 + (\hat{\tau}[i,j] - \hat{\tau}[i-1,j])^2} < 0,001$$

Definição das áreas de abrangência

Tanto o estimador bayesiano empírico quanto o algoritmo EM usam a informação do tamanho populacional para inferir o valor esperado de óbitos em determinado município. Nesse sentido, é importante aplicar tais métodos em um conjunto homogêneo de municípios para não contaminar as estimativas com informações muito diferentes.

Os exercícios empíricos deste trabalho foram realizados com municípios do Rio Grande do Norte. Para obter uma maior homogeneização, utilizaram-se os métodos em subgrupos de municípios chamados áreas de abrangência. Num primeiro exercício, essas áreas foram as quatro mesorregiões administrativas em que se divide o Estado do Rio Grande do Norte:

- **Mesorregião Agreste**, com 43 municípios agrupados em três microrregiões: Agreste Potiguar, Baixa Verde e Borborema Potiguar. Essa mesorregião é a única em que nenhum dos seus municípios é litorâneo, tendo como cidades importantes São Paulo do Potengi, João Câmara e Santa Cruz.
- **Mesorregião Central**, com 37 municípios agrupados em cinco microrregiões: Angicos, Macau, Seridó Ocidental, Seridó Oriental e Serra de Santana. As cidades mais importantes são Angicos, Galinhos, Macau, Currais Novos, Caicó e Pedro Avelino.

- **Mesorregião Oeste** é a segunda mais populosa, formada por 62 municípios agrupados em sete microrregiões: Chapada do Apodi, Médio Oeste, Mossoró, Pau dos Ferros, Serra de São Miguel, Umarizal e Vale do Açu. Esta mesorregião concentra as atividades salineiras e de extração de petróleo e gás natural, além de ser um polo de fruticultura irrigada. Os municípios importantes dessa mesorregião são Mossoró, Assu, Areia Branca, Apodi, Pau dos Ferros, São Rafael, Caraúbas, Patu, Tibau, São Miguel e Alexandria.
- **Mesorregião Leste** é a mais importante, formada por 25 municípios agrupados em quatro microrregiões: Litoral Nordeste, Litoral Sul, Macaíba e Natal. Esta mesorregião é a mais populosa do Estado e mais urbanizada, englobando a capital Natal e sua região metropolitana, além de concentrar o polo industrial do Estado. O turismo está praticamente todo voltado para essa mesorregião, no litoral urbano (Natal), no litoral sul (de Parnamirim até Baía Formosa) e no litoral norte (de Extremoz até Pedra Grande). As cidades importantes dessa mesorregião são Natal, Parnamirim, São Gonçalo do Amarante, Macaíba, Ceará-Mirim, Touros, São Miguel do Gostoso, Canguaretama e Tibau do Sul.

Para o segundo exercício empírico, a área maior utilizada como referência constituiu-se nos grupos homogêneos de municípios, gerados a partir de uma análise de *cluster* utilizando as variáveis grau de urbanização, proporção de óbitos por causas externas e a população de cada município.

A premissa básica para o uso dessas variáveis é que municípios maiores e mais urbanizados tendem a ter melhor infraestrutura, refletindo sistema de saúde mais bem estruturado, o que melhora a organização dos protocolos de registros de dados. Além disso, as mortes com melhor grau de cobertura de registros são aquelas classificadas por causas externas, que, em geral, requisitam ocorrências policiais e muitas vezes o óbito acontece em unidades de saúde, aumentando a chance de serem devidamente registradas. Portanto, municípios similares segundo estas variáveis tendem a formar grupos homogêneos no que se refere à qualidade de registros de óbitos e dinâmica demográfica.

A análise de *cluster* foi realizada com a distância euclidiana como critério de similaridade e, como algoritmo de agrupamento, a ligação completa. Com isso, os 167 municípios do Rio Grande do Norte foram dispostos em oito grupos homogêneos, distribuídos em todo o Estado segundo a similaridade quanto ao grau de urbanização, à proporção de óbitos por causas externas e à população de cada município.

Fontes de dados

A informação sobre grau de urbanização foi extraída do Instituto Brasileiro de Geografia e Estatística (IBGE). Os dados de população, óbitos e percentual de óbitos por causas externas foram coletados no Datasus.

Resultados

Nesta seção será exposta a análise dos resultados, para o grau de cobertura e, consequentemente, para o fator de correção de registros de óbitos, nos municípios do Estado do Rio Grande do Norte, utilizando-se o estimador bayesiano empírico (BE) e o algoritmo EM. Primeiro é feita uma análise para avaliar os métodos empregados: BE com mesorregiões como área maior; BE com grupos homogêneos como área maior; EM com mesorregiões; e EM com grupos homogêneos.

No segundo momento, analisa-se o resultado propriamente dito, do grau de cobertura de registros de óbitos nos municípios e nas mesorregiões do Rio Grande do Norte.

O grau de cobertura de óbitos nas mesorregiões foi calculado a partir dos municípios. A unidade de análise final, onde efetivamente são estimados os óbitos esperados para calcular o grau de cobertura é o município. Portanto, nas mesorregiões

calcula-se o grau de cobertura por meio do quociente entre a soma dos óbitos observados e a soma dos óbitos estimados dos municípios que pertencem a esta mesorregião.

Na terceira parte dos resultados analisa-se o grau de cobertura de óbitos segundo grupos etários. Conforme mencionado anteriormente, tanto o método BE quanto o EM foram aplicados nos municípios aos grupos quinquenais de idade. Aqui nos resultados, para avaliar o grau de cobertura, agruparam-se os resultados em faixas etárias maiores para tornar os resultados mais robustos, menos sujeitos a flutuações aleatórias.

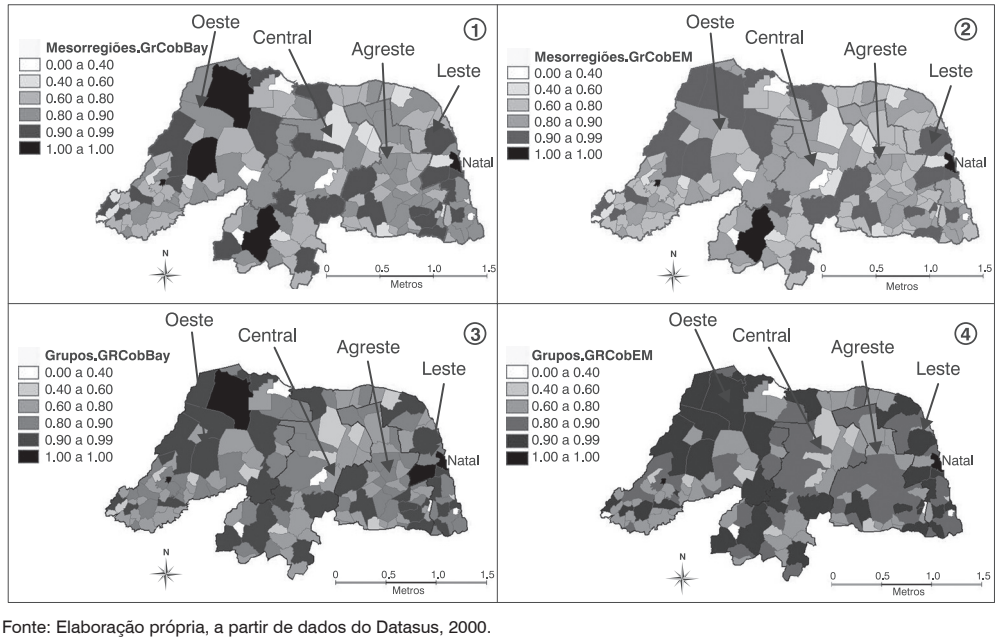
Cabe ressaltar ainda que, quando o algoritmo EM ou o estimador bayesiano empírico estimou valor menor que o óbito observado, esta estimativa foi desconsiderada, permanecendo o valor observado. Nesse caso, o grau de cobertura foi considerado 1 (100%).

Avaliando os métodos

Os mapas da Figura 1 apresentam os resultados do grau de cobertura de óbitos dos municípios do Rio Grande do Norte para todos os exercícios realizados, permitindo avaliar diferenças nos métodos aplicados. O primeiro mapa mostra a distribuição espacial do grau de cobertura dos óbitos estimados a partir do BE operacionalizado nos municípios, tendo como área maior a mesorregião. O segundo mapa usa o algoritmo EM com mesorregiões, o terceiro apresenta os resultados do BE por grupos homogêneos e o quarto mostra a distribuição espacial do grau de cobertura encontrado a partir do algoritmo EM implementado por grupos homogêneos de municípios.

Usando esses mapas apenas com o objetivo de cotejar os exercícios realizados, observa-se que, aparentemente, não há grandes diferenças, nem por método, nem por área maior de abrangência usada. Em

FIGURA 1
Grau de cobertura de registros de óbitos para os municípios, empregando os métodos bayesiano empírico e algoritmo EM, tendo como área maior mesorregiões e grupos homogêneos
Estado do Rio Grande do Norte – 2000



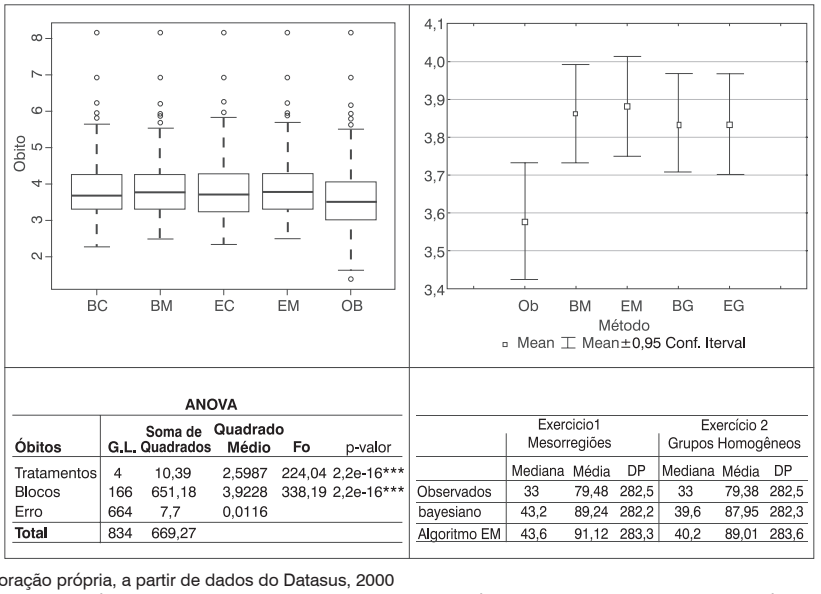
Fonte: Elaboração própria, a partir de dados do Datasus, 2000.

linhas gerais, alguns municípios, sobretudo os maiores como Natal e Mossoró, têm grau de cobertura bem alto, chegando a 1 em algumas situações. Numa visão mais geral, os mapas demonstram que a porção mais central do Estado, nas mesorregiões central e agreste, é a que apresenta o maior número de municípios com menor grau de cobertura. Mas essas análises regionais e municipais serão detalhadas mais a frente.

Novamente avaliando as possíveis diferenças entre os métodos e exercícios realizados, a Figura 2 mostra os box-plot dos óbitos observados, estimados pelo BE e pelo EM, para cada área maior de abrangência utilizada. Esses gráficos ratificam a primeira impressão constatada nos mapas anteriores: não há grande diferença na distribuição dos óbitos estimados segundo o método empregado, bem como parece não haver diferença quando se utilizam mesorregião ou grupos homogêneos de municípios como área maior. O que esses gráficos revelam é que, de fato, os dois estimadores testados aumentam os valores dos óbitos quando comparados aos óbitos observados. De acordo com a Figura 2, no exercício com as mesorregiões como área maior, o estimador BE estima em média 9,6 óbitos por municípios a mais que os óbitos observados. Já o algoritmo EM acrescenta 11,4 óbitos em média por município com relação aos observados. Quando os estimadores foram aplicados segundo grupos homogêneos, esses valores foram 8,6 e 9,3 para estimador BE e algoritmo EM, respectivamente. Se tomarmos a mediana, também os resultados apontam para o mesmo cenário, em que os óbitos estimados, seja pelo BE seja pelo EM, têm mediana por volta de 10 óbitos a mais que a mediana dos óbitos observados.

Quando a comparação é segundo a área de abrangência, os valores de mediana e média apontam para as mesorregiões como modo de operacionalização dos métodos que estimam maiores valores de óbitos, ainda que a diferença com relação

FIGURA 2
Box Plot, Gráfico de médias com intervalo de 95% de confiança para médias, tabela de análise de variância e sumário estatístico de todos os exercícios realizados (1)
Municípios do Rio Grande do Norte – 2000



ANOVA

Óbitos	G.L.	Soma de Quadrados	Quadrado Médio	Fo	p-valor
Tratamentos	4	10,39	2,5987	224,04	2,2e-16***
Blocos	166	651,18	3,9228	338,19	2,2e-16***
Erro	664	7,7	0,0116		
Total	834	669,27			

	Exercicio1 Mesorregiões			Exercício 2 Grupos Homogêneos		
	Mediana	Média	DP	Mediana	Média	DP
Observados	33	79,48	282,5	33	79,38	282,5
bayesiano	43,2	89,24	282,2	39,6	87,95	282,3
Algoritmo EM	43,6	91,12	283,3	40,2	89,01	283,6

Fonte: Elaboração própria, a partir de dados do Datasus, 2000
(1) Óbitos observados, óbitos estimados pelo estimador bayesiano empírico usando as mesorregiões como área maior, óbitos estimados pelo estimador bayesiano empírico usando os grupos homogêneos de municípios como área maior, óbitos estimados pelo algoritmo EM usando as mesorregiões como área maior e óbitos estimados pelo algoritmo EM usando os grupos homogêneos como área maior.

aos resultados por grupo homogêneo não seja significante. Na verdade, não há diferença significante entre os quatro exercícios de estimação de óbitos utilizados neste trabalho. A diferença existente é entre os óbitos observados e todas as tentativas de estimação, mostrando que, de fato, os métodos estão elevando os óbitos originalmente registrados. É o que revela a tabela de análise de variância da Figura 2, e está ilustrado também com o gráfico de médias de óbitos e o respectivo intervalo de confiança, segundo método de estimação e área maior de abrangência.

Apesar das diferenças não significantes, no que se refere à forma de agregação dos municípios, o que nesse trabalho convencionou-se a chamar de área maior de abrangência, os resultados quando os métodos foram aplicados por mesorregião são um pouco melhores, pois apresentam valores médios e medianos maiores. Além disso, o desvio padrão não é diferente de uma forma de agregação para outra. As maiores divergências nas estimativas ocorrem quando os dois métodos – BE e EM – foram aplicados aos municípios reunidos nos grupos homogêneos, sobretudo naqueles com menor população.

Diante disso, a seguir serão analisados os resultados por mesorregião, por municípios e por faixa etária. Essas análises estarão baseadas nos óbitos estimados pelos métodos aplicados, tendo como área maior de abrangência as mesorregiões.

Resultaldos agregados por mesorregiões

A Tabela 1 traz o grau de cobertura dos óbitos e o fator de correção, segundo o método de estimação, apresentando, para cada mesorregião e metodologia utilizada, o maior e o menor grau de cobertura encontrado para um município, além do grau de cobertura de óbitos para o total da mesorregião e o respectivo fator de correção de sub-registro de óbito.

Observa-se que o grau de cobertura para o total do Estado do Rio Grande do Norte, com dados de 2000, foi de 88,95% e 87,11%, respectivamente com os métodos BE e EM. Isso implica sub-registro de 11,05% e 12,89%. A mesorregião com maior cobertura é a Leste, a única com valor acima de 90% para os dois métodos em análise. Essa é a região com maior percentual de população urbana, composta por 25 municípios, onde estão situados a capital do Estado, sua região metropolitana e dois dos três municípios mais populosos do Rio Grande do Norte – Natal e Parnamirim. Por outro lado, a região Agreste é a que apresenta o pior grau de cobertura de registros de óbitos.

Resultaldos por municípios

Os mapas da Figura 1 indicam o grau de cobertura dos municípios, em que se verifica certa heterogeneidade. Há municípios que apresentaram grau de cobertura acima de

TABELA 1
Estatísticas do grau de cobertura dos óbitos nos municípios, por modelo de estimador utilizado, tendo as mesorregiões como área maior
Estado do Rio Grande do Norte – 2000

Mesorregiões	Municípios	Grau de cobertura bayesiano (%)			Fator de correção bayesiano	Grau de cobertura EM (%)			Fator de correção EM
		Máximo	Mínimo	Geral		Máximo	Mínimo	Geral	
Agreste	43	0,9781	0,4343	0,8400	1,1905	0,9780	0,4201	0,8114	1,2324
Central	37	0,9922	0,3614	0,8534	1,1718	0,9869	0,3542	0,8335	1,1997
Leste	25	0,9998	0,3185	0,9317	1,0733	0,9990	0,3193	0,9128	1,0955
Oeste	62	0,9936	0,3068	0,8724	1,1463	0,9884	0,3336	0,8616	1,1606
Total	**167**	**0,9998**	**0,3068**	**0,8895**	**1,1243**	**0,9990**	**0,3193**	**0,8711**	1,1480

Fonte: Elaboração própria, a partir de dados do Datasus, 2000.

95% (ou seja, de sub-registros inferiores a 5%), como Açu, Apodi, Caicó, Caraúbas, Ceará-Mirim, Macaiba, Mossoró, Natal, Riacho de Santana, Santa Cruz, Serrinha, Touros e Viçosa. A maioria destes municípios pertence às mesorregiões Leste e Oeste, as mais urbanizadas do Estado (Figura 3). Por outro lado, existem municípios com grau de cobertura de óbitos abaixo de 40%, tais como Bodó, Jundiá, Porto do Mangue, Timbaúba dos Batistas e Vila Flôr.

O grau de cobertura dos óbitos nos municípios está diretamente correlacionado com o grau de urbanização. Testes de correlação mostraram significância entre estas duas variáveis, seja quando o grau de cobertura foi estimado a partir do BE (p-valor=0,00011), seja quando estimado pelo EM (p-valor=0,00006), no exercício com a mesorregião como área maior. Parece que municípios mais urbanizados tendem a ter maior estrutura organizacional, implicando melhores rotinas gerenciais, o que impacta na melhoria das informações prestadas à Secretaria de Estado de Saúde.

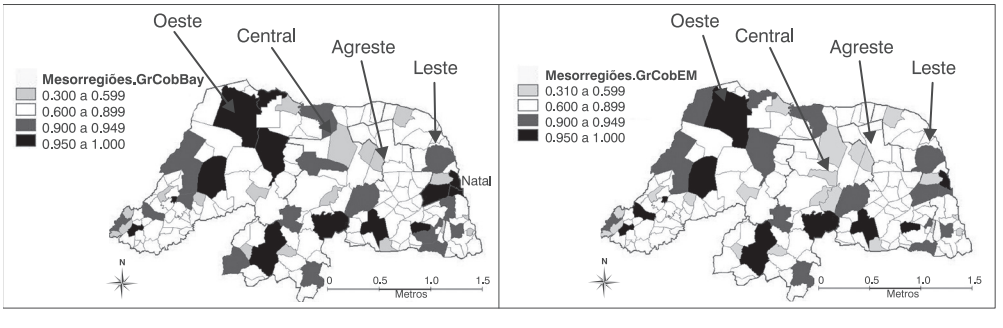
Outra avaliação que se fez foi verificar se há correlação espacial dos sub-registros de óbitos com o espaço no território do Rio Grande do Norte. O índice de Moran evidenciou que não há tal relação (p-valor=0,327 para sub-registros estimados por BE e p--valor=0,365 quando a estimativa foi gerada pelo EM, ambos com mesorregiões como área maior), de maneira que não se pode dizer que existe uma configuração espacial bem definida dos sub-registros de óbitos no Estado.

Quando o enfoque é a Região Metropolitana de Natal, o grau de cobertura de registros de óbitos estimado pelo método BE foi de 97,5% (sub-registro de 2,5%). Já para o EM, o grau de cobertura correspondeu a 95,5% (sub-registro de 4,5%). Este resultado é fortemente influenciado por Natal, visto que, quando se retira o município desta análise, o conjunto dos demais municípios metropolitanos apresentou sub-registro de 7%, pelo método BE, e de 11%, pelo Algoritmo EM.

Avaliando o sub-registro de óbitos por faixa etária

Conforme mencionado anteriormente, a operacionalização das metodologias neste trabalho foi por grupos etários quinquenais. Os resultados de grau de cobertura por municípios e mesorregiões analisados nas seções anteriores foram obtidos a partir do quociente da soma dos óbitos observados e a soma dos óbitos estimados em cada grupo etário quinquenal. Portanto, é possível avaliar a qualidade dos registros de óbitos segundo o grupo etário. Com o objetivo de tornar as estimativas de grau de cobertura mais robustas, os grupos quinquenais foram agregados em quatro grupos maiores, conforme apresentado na Tabela 2.

FIGURA 3
Grau de cobertura de registros de óbitos para os municípios, utilizando os métodos bayesiano empírico e algoritmo EM, tendo como área maior as mesorregiões
Estado do Rio Grande do Norte – 2000



Fonte: Elaboração própria, a partir de dados do Datasus, 2000.

Os resultados por faixa etária, nos exercícios metodológicos tendo as mesorregiões como área maior, apresentaram valores muito próximos para os dois estimadores (Tabela 2). Na mesorregião Agreste, a faixa etária de 0 a 14 anos registrou os mesmos graus de cobertura, para os dois estimadores e consequentemente o mesmo valor de sub-registro de óbito. Esta semelhança também ocorreu na mesorregião Oeste, na faixa etária de 35 a 59 anos e na de mais de 60 anos.

De forma geral, nos dois métodos utilizados, a faixa etária com maior sub-registro foi a dos adultos jovens (15 a 34 anos): 16% no BE e 20% no EM. A Tabela 2 mostra também que a mortalidade daqueles com mais de 60 anos é a mais bem registrada no Rio Grande do Norte, apresentando os menores sub-registros de óbitos. Os outros dois grupos etários (0 a 14 e 35 a 59 anos) têm praticamente a mesma cobertura de registros de óbitos.

Discussão

Este artigo apresentou uma metodologia de correção de sub-registro de óbitos, que permite avaliar o grau de cobertura de óbitos por grupos etários em municípios e em qualquer nível de agregação a partir deste, tais como mesorregiões, conforme o exercício apresentado para o Estado do Rio Grande do Norte, usando dados do Censo 2000. Para esses dados, foram encontrados 11% de sub-registro, empregando a metodologia do estimador bayesiano empírico, o que representa a falta de 1.648 óbitos não registrados, e 12,8% usando o algoritmo EM, significando 1.962 óbitos não enumerados em 2000 no Rio Grande do Norte.

Segundo Cavalini e Leon (2008), a estimativa oficial de sub-registro do SIM, para o Brasil em geral, é de 17,7%. Contudo, em seus estudos também utilizando o estimador bayesiano empírico, estes autores encontraram um percentual de correção de sub-registros de mortalidade de 5,85% para todo o Brasil. Estimativas do IBGE para 2005 apresentam índices de sub-registro de óbitos para a Região Nordeste acima de 26%, sendo 30% no RN. A diferença nas estimativas mostra a necessidade de uma melhor avaliação dos métodos de estimação.

Com base nos dados trabalhados neste artigo, observou-se que o grau de urbaniza-

TABELA 2
Grau de cobertura e estimativa de sub-registro de óbitos para os municípios, por faixa etária e modelo de estimador, tendo as mesorregiões como área maior
Estado do Rio Grande do Norte – 2000

Mesorregiões	Índice	Bayesiano				Algoritmo EM			
		0 a 14 anos	15 a 34 anos	35 a 59 anos	60 anos e mais	0 a 14 anos	15 a 34 anos	35 a 59 anos	60 anos e mais
Agreste	GrCob (%)	0,768	0,734	0,776	0,885	0,765	0,674	0,748	0,856
	Sub-registro	0,232	0,266	0,224	0,115	0,235	0,326	0,252	0,144
Central	GrCob (%)	0,799	0,754	0,802	0,889	0,767	0,719	0,799	0,869
	Sub-registro	0,201	0,246	0,198	0,111	0,233	0,281	0,201	0,131
Leste	GrCob (%)	0,917	0,896	0,921	0,945	0,878	0,878	0,902	0,931
	Sub-registro	0,083	0,104	0,079	0,055	0,122	0,122	0,098	0,069
Oeste	GrCob (%)	0,858	0,825	0,836	0,896	0,819	0,773	0,831	0,896
	Sub-registro	0,142	0,175	0,164	0,104	0,181	0,227	0,169	0,104
Total	**GrCob (%)**	**0,863**	**0,836**	**0,861**	**0,913**	**0,832**	**0,798**	**0,847**	**0,899**
	Sub-registro	**0,137**	**0,164**	**0,139**	**0,087**	**0,168**	**0,202**	**0,153**	**0,101**

Fonte: Elaboração própria com base nos dados do IBGE

ção é um fator determinante na qualidade das informações. A região metropolitana de Natal apresentou apenas 2,4% de sub-registros, com estimador bayesiano empírico, e 4% com algoritmo EM.

Os exercícios mostraram que, mesmo em regiões com grau de urbanização mais baixo, como as mesorregiões Central e Agreste, e em áreas com piores indicadores sociais, é possível encontrar municípios com elevado grau de cobertura de registros de mortalidade, indicando que boa gerência administrativa pode ser universalizada na qualidade dos registros vitais.

Espera-se uma melhora constante na qualidade dos sistemas de informação brasileiros, para que métodos de estimação passem a ser utilizados com o objetivo de avaliar e atestar a qualidade dos registros oficiais, diminuindo sua função de correção de sub-registros.

Os resultados encontrados no trabalho apontam não haver diferença significativa nas estimativas de sub-registro entre os métodos utilizados: bayesiano empírico ou algoritmo EM. Esperava-se que, criando grupos homogêneos de municípios, as duas metodologias pudessem operar resultados mais robustos, pois estariam usando informações de áreas realmente similares. Talvez em exercícios futuros seja o caso de delimitar melhor os grupos homogêneos com a inclusão de novas variáveis.

Outro ponto importante a ser ressaltado, e projetando futuras tentativas nessa temática, é a pergunta: o que estamos encontrando aqui é, de fato, sub-registro? Na opinião dos autores, pela construção dos métodos, acredita-se que estamos diante de uma metodologia capaz de mensurar quantos óbitos estão faltando em determinado município ou grupo etário. Veja por exemplo a forma de construção do estimador bayesiano empírico. Com ele, estima-se de forma suavizada o quociente entre o valor observado de óbito e um número esperado de óbito num determinado município, sob a hipótese de taxa de mortalidade constante para todos os municípios que compreendem certa área maior. Em outras palavras, o que se faz é suavizar este quociente (fator k descrito logo abaixo da fórmula 2 da seção metodológica). Depois do fator k suavizado, por meio do método bayesiano empírico, multiplica-se pelos óbitos esperados estimados inicialmente, sob aquela hipótese de que os óbitos eram constantes para toda a área maior. Assim, se obtêm os novos óbitos estimados, tendo sido usado o método bayesiano.

O grau de cobertura então é o resultado dos óbitos observados divididos por estes últimos óbitos estimados. O desafio daqui em diante consiste em combinar esses métodos, inicialmente usados para correção de flutuações aleatórias, com métodos demográficos de correção de sub-registros.

Referências

CARVALHO, J. A. M. de.; SAWYER, D. O.; RODRIGUES, R. do N. **Introdução a alguns conceitos básicos e medidas em demografia**. Belo Horizonte: Abep, 1994 (Série Textos didáticos n. 1).

CARVALHO, J. A. de. **Crescimento populacional e estrutura demográfica no Brasil**. Belo Horizonte: UFMG/Cedeplar, 2004.

CAVALINI, L. T.; PONCE DE LEON, A. C. M. Sistemas de informação em saúde do Brasil: informações incompletas e estratégias de correção. In: XI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE. **Anais...** Campos do Jordão, 2008. Disponível em: <http://www.sbis.org.br/cbis11/arquivos/828.PDF>. Acesso em: 05 out. 2011.

CAVALINI, L. T.; PONCE DE LEON, A. C. M. Correção de sub-registros de óbitos e proporção de internações por causas mal definidas. **Rev. Saúde Pública**, v. 41, n. 1, p. 85-93, 2007. Disponível em: <http://www.scielo.br/pdf/rsp/v41n1/13.pdf>. Acesso em: 05 out. 2011.

CHAKIEL, J. La investigacion sobre causas de morte en la América Latina. **Notas de Poblacion**, Santiago, Chile, n. 44, p.9-30, ago. 1987.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society B**, v. 39, n. 1, p. 1-38, 1977.

FREIRE, F. H. M. A. **Projeção populacional para pequenas áreas pelo método das componentes demográficas usando estimadores bayesianos espaciais**. Tese (Doutorado em Demografia) – Centro de Desenvolvimento e Planejamento Regional, Universidade Federal de Minas Gerais, Belo Horizonte, 2001. Disponível em: <http:// www.cedeplar.ufmg.br/>. Acesso em: 14 jul. 2010.

FREIRE, F. H. M. A.; ASSUNÇÃO, R. M. Intervalo de confiança para a taxa de fecundidade total de pequenas áreas. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS, 11. **Anais...** Belo Horizonte: Abep, 1998. Disponível em: <http:// www.abep.org.br/>. Acesso em: 14 out. 2010.

GELMAN, A. **Bayesian data analysis**. London; New York: Chapman & Hall, 1995. HASSELBLAD, V. Estimation of parameters for a mixture of normal distributions. **Technometrics**, v. 8, n. 3, p. 431-444, 1966.

HORTA, M. M. Modelos de mistura de distribuições na segmentação de imagens SAR polarimétricas multi-look. São Carlos, 2009.

IBGE. **Indicadores sociodemográficos e de saúde no Brasil**, p. 41-78, 2009. Disponível em: <http://www.ibge.gov.br/home/estatistica / população / indic_ sociosaude / 2009 / indicsaude.pdf>. Acesso em: 24 mar. 2010.

JANNUZZI. P de M. **Indicadores sociais no Brasil**. Campinas: Editora Alínea, 2001.

JEFFREYS, H. An alternative to the rejection of observations. **Proceedings of the Royal Society of London A**, v. 137, n. 831, p. 78-87, 1932.

MARSHALL, R. M. Mapping disease and mortality rates using Empirical Bayes Estimators. **Journal of the Royal Statistical Society**, Series C: Applied Statistics, v. 40, n. 2, p. 283-294, 1991.

MCLACHLAN, G. J.; KRISHNAN, T. **The EM algorithm and extensions**. New York: John Wiley & Sons, 1997.

MCLACHLAN, G. J.; PEEL, D. **Finite mixture models**. New York: John Wiley & Sons, 2000.

PAES, N. A.; ALBUQUERQUE, M. E. E. Avaliação da qualidade dos dados populacionais e cobertura dos registros de óbitos para as regiões brasileiras. **Revista de Saúde Pública**, USP, v. 33, n. 1, p. 33-43, 1999. Disponível em:<http://www.scielosp.org/pdf/rsp/v33n1/0021.pdf/>. Acesso em: 2 nov. 2010.

PAES, N. A. Avaliação da cobertura dos registros de óbitos dos Estados brasileiros em 2000. **Revista de Saúde Pública**, USP, v. 39 n.6, p. 882-90, 2005. Disponível em:<http://www.scielosp.org /pdf/rsp/v29n6/06.pdf/>. Acesso em: 16 jul. 2010.

REDNER, R. A.; WALKER, H. F. Mixture densities, maximum likelihood and the EM algorithm. **SIAM Review**, v. 26, n. 2, p. 195-239, 1984.

Resumen

Estimación de subregistros de fallecimientos en pequeñas áreas con los métodos bayesiano empírico y algoritmo EM

Un gran problema, en lo que se refiere a estimativas demográficas en Brasil, está relacionado con el nivel y patrón de la mortalidad. Los demógrafos que trabajan con mortalidad en el país todavía no se sienten muy seguros sobre el comportamiento real de este componente de la dinámica poblacional. Por otro lado, es necesario que se disponga de indicadores de mortalidad para niveles geográficos más desagregados, sobre todo municipios. El problema es que cuanto

más desagregado, más complejo se hace el trabajo de estimar cualquier indicador social o demográfico. Este trabajo tiene por objetivo estimar y proponer una corrección de subregistros de fallecimientos en el nivel municipal, según grupos de edad, por medio de dos métodos: estimador bayesiano empírico (BE) y algoritmo EM (Expectation-Maximization). Con el objeto de que los dos métodos fueran puestos en funcionamiento entre municipios semejantes, se realizaron dos ejercicios: se agruparon los municipios según la mesorregión; y se agruparon los municipios en grupos homogéneos, generados a partir de un análisis de *cluster,* utilizando las variables grado de urbanización, proporción de óbitos por causas externas y la población de cada municipio. Se utilizaron datos del Estado de Río Grande do Norte, referentes al año 2000. Para el total del Estado, se estimó un subregistro de un 11% con el estimador BE y de un 12,9% con el algoritmo EM. Otro resultado importante es la posibilidad de evaluar el grado de cobertura de óbitos por grupos de edad en municipios y en cualquier nivel de agregación a partir de este nivel.

Palabras-claves: Subregistros de mortalidad. Estimador bayesiano empírico. Algoritmo EM.

Abstract

Estimation of death underreporting at small areas using empiric Bayesian and EM algorithm methods

Level and standard of mortality are major demographic estimation problems in Brazil. Demographists dealing with mortality in Brazil still do not fell assured of the real behavior of this population dynamics component. On the other hand, there is a need for mortality indicators available for more disaggregated geographic levels, mostly municipalities. The difficulty is that the more disaggregated, the more complex is the task for estimating any social or demographic indicator. In this study, we aimed to estimate and to propose the correction of death underreporting at the municipal level, according to age, using two methods: the empiric Bayesian estimator (BE) and the EM (Expectation-Maximization) algorithm. For the two methods to be operational within comparable municipalities, two steps were performed: we grouped the municipalities according to a mesoregion; and we grouped them into two homogeneous groups, created from a cluster analysis using the variables level of urbanization, proportion of death from external causes and the population of each municipality. We used data collected in 2000 from the State of Rio Grande do Norte. For the entire State, we estimated underreporting to be 11% using the BE estimator, and 12.9 % using the EM algorithm. Another important finding was the capability to assess the level of death coverage by age groups in the municipalities and, at any level of aggregation.

Keywords: Death underreporting. Empiric Bayesian estimator. EM algorithm.

Recebido para publicação em 01/11/2011
Aceito para publicação em 07/01/2012