IBGE
Instituto Brasileiro de Geografia e Estatística

Pareamento das Estatísticas do Registro Civil e dos Sistemas de Informações sobre Nascidos Vivos e Mortalidade (SINASC e SIM)

Aplicação da Técnica de Captura-Recaptura para estimativa dos totais de nascidos vivos e óbitos

2019

Nota metodológica n. 01

Versão 1 - Dezembro de 2019

Presidente da República
Jair Messias Bolsonaro

Ministro da Economia
Paulo Roberto Nunes Guedes

Secretário Especial de Fazenda
Waldery Rodrigues Junior

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Susana Cordeiro Guerra

Diretor-Executivo
Fernando José de Araújo Abrantes

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Eduardo Luiz G. Rios Neto

Diretoria de Geociências
João Bosco de Azevedo

Diretoria de Informática
David Wu Tai

Centro de Documentação e Disseminação de Informações
Marise Maria Ferreira

Escola Nacional de Ciências Estatísticas
Maysa Sacramento de Magalhães

UNIDADE RESPONSÁVEL

Diretoria de Pesquisas
Coordenação de População e Indicadores Sociais
Cristiane dos Santos Moutinho

Ministério da Economia
Instituto Brasileiro de Geografia e Estatística - IBGE
Diretoria de Pesquisas
Coordenação de População e Indicadores Sociais

Pareamento das Estatísticas do Registro Civil e dos Sistemas de Informações sobre Nascidos Vivos e Mortalidade (SINASC e SIM)

Aplicação da Técnica de Captura-Recaptura para estimativa dos totais de nascidos vivos e óbitos

2019

Nota metodológica n. 01

Versão 1 - Dezembro de 2019

IBGE
Rio de Janeiro
2019

Considerações sobre o pareamento

Com o intuito de avaliar a cobertura dos Sistemas de Estatísticas Vitais (Estatísticas do Registro Civil, do IBGE, Sistema de Informações sobre Nascidos Vivos - SINASC e Sistema de Informações sobre Mortalidade - SIM, do Ministério da Saúde) se faz necessário mensurar quantos desses eventos vitais (nascimentos e óbitos) foram alcançados pelos respectivos Sistemas. Para isso, os dados das Estatísticas do Registro Civil, SIM e SINASC passaram por um procedimento de pareamento, cujas principais etapas são descritas a seguir.

É importante ressaltar que, antes de iniciar o processo de pareamento das bases de dados, foram adotados os seguintes procedimentos:

- Aplicação de filtro na base do IBGE de modo que contenha apenas os registros de nascidos vivos e óbitos ocorridos e registrados no ano de referência, uma vez que, nessa base, há eventos que ocorreram em anos anteriores e somente foram registrados depois;

- Tratamento e padronização das variáveis, tendo em vista que uma diferença observada entre as bases se refere à construção da informação geográfica de seus registros. Na base do IBGE, existe uma variável que determina a Unidade da Federação com dois dígitos e uma outra, o município com cinco dígitos, incluindo o verificador. Nas bases do SINASC e SIM, porém, a informação geográfica de Unidade da Federação e município é observada em uma única variável contendo seis dígitos, dos quais os dois primeiros são referentes à Unidade da Federação e os quatro seguintes, ao município, sem o dígito verificador. A título de padronização, a variável do IBGE correspondente ao município foi utilizada retirando-se o dígito verificador, em conformidade com o formato adotado pelo SINASC e o SIM. Além disso, a variável sexo, tanto para nascimentos quanto para óbitos, nesses dois Sistemas, tem codificação diferente da utilizada pelo IBGE para a categoria "ignorado". Nesse caso, foi adotada a categorização do IBGE e, assim, nas bases do SINASC e SIM, os registros com sexo ignorado passaram a ter categoria 9.

- Retirada de casos da base do IBGE uma vez que eles nunca serão encontrados nas bases do SINASC e SIM. Como o IBGE levanta quatro tipos de registros que não fazem parte do âmbito desses dois Sistemas, tais registros recebem uma codificação especial na numeração da Declaração de Nascido Vivo (DN) ou da Declaração de Óbito (DO), conforme mostrado a seguir:

Quadro 1 - Registros da base do IBGE que não fazem parte do âmbito do SINASC e SIM, por temas

Nascidos vivos	Óbitos
Registro por sentença judicial	Registro realizado a partir da declaração de testemunhas qualificadas para o qual não foi emitida DO
Adoção com segundo registro	Registro por sentença judicial
Registro Administrativo de Nascimento Indígena (RANI) ou Requerimento FUNAI	Registro administrativo da FUNAI
Registro realizado a partir da declaração de testemunhas para o qual não foi emitida DN	Registro realizado a partir da declaração de testemunhas para o qual não foi emitida DO

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais.

- Outros casos também retirados da base do IBGE correspondem aos registros de nascidos vivos em que a mãe residia no exterior, e de óbitos em que o falecido residia no exterior.

Uma vez cumpridas essas etapas, foi criado, tanto para os registros de nascimentos, quanto para os registros de óbitos, um identificador composto pelo número da Declaração de Nascido Vivo (DN) ou da Declaração de Óbito (DO) mais um conjunto de variáveis. Esse identificador foi estruturado da seguinte maneira:

Quadro 2 - Composição dos identificadores, por temas

Nascidos vivos	Data de nascimento, sexo, Unidade da Federação de nascimento, município de nascimento, idade da mãe, Unidade da Federação de residência da mãe, município de residência da mãe, e número da DN.
Óbitos não fetais	Data do óbito, sexo, idade, Unidade da Federação de residência do falecido, município de residência do falecido, local de ocorrência do óbito, e número da DO.

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais.

A partir desse identificador, foram selecionados os casos não únicos nas bases de dados. Registros com o mesmo identificador foram analisados e considerados

repetidos, sendo um deles eliminado, de tal forma que as bases contenham apenas registros únicos.

Para tal, o primeiro passo foi separar cada uma das bases (nascidos vivos e óbitos) em duas partes: número da DN/DO único e número da DN/DO duplicado. Esse passo se fez necessário pois, por exemplo, se um registro contém número da DN/DO duplicado na base do IBGE e único nas bases do SINASC/SIM, ao proceder o *linkage* das bases, esse registro, que era único no SINASC/SIM, será pareado duas vezes, o que não é o desejado.

Tendo feito essa divisão, o primeiro pareamento foi realizado entre as bases formadas por registros com número da DN/DO único, usando a numeração da DN/DO como variável de *linkage* entre as bases. Os registros não pareados nessa fase e os registros com número da DN/DO duplicado formaram uma nova base.

Para o pareamento dessa nova base, foi criada uma variável de ligação a partir de uma chave especial formada por um conjunto de variáveis e mais uma parte da numeração da DN/DO. Antes de descrever o novo passo do pareamento, é preciso explicar como é formada a numeração da DN/DO.

A numeração da Declaração de Nascido Vivo (DN) é formada por 11 dígitos: os dois primeiros correspondem a um prefixo; o último, a um dígito verificador; e os oito restantes, à raiz do número da DN. A numeração da Declaração de Óbito (DO), por sua vez, é formada por nove dígitos: os oito primeiros são considerados a raiz do número da DO, enquanto o último é o dígito verificador.

No processo de tratamento dos dados, foram observados, em alguns casos na base do IBGE, erros de preenchimento na numeração da DN/DO. No caso da DN, esse erro ocorreu no prefixo ou no dígito verificador; e, no caso da DO, foi observado no dígito verificador. Portanto, para formar essa chave especial, foram considerados apenas os oito dígitos da raiz da numeração da DN/DO e as demais variáveis, como indicado a seguir:

Quadro 3 - Composição da chave especial para o pareamento, por temas

Nascidos vivos	Data de nascimento, sexo, Unidade da Federação de nascimento, município de nascimento, idade da mãe, Unidade da Federação de residência da mãe, município de residência da mãe, e raiz do número da DN.
Óbitos não fetais	Data do óbito, sexo, idade, Unidade da Federação de residência do falecido, município de residência do falecido, local de ocorrência do óbito, e raiz do número da DO.

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais.

Sendo assim, a base composta pelos registros com numeração da DN/DO duplicada e os registros com numeração da DN/DO única que não formaram par na primeira etapa foram pareados a partir da chave especial explicada anteriormente.

É importante destacar que foram observados casos pareados que apresentaram respostas diferentes para as variáveis. Visando a harmonização entre as respectivas bases de dados, foi criada a seguinte regra para definir de qual base seria a resposta final utilizada na modelagem:

- Para os casos em que o registro possui resposta ignorada em uma das bases e na outra há uma resposta válida, esse registro entrará no modelo com a resposta válida. Assim, por exemplo, se o registro pareado apresenta idade ignorada na base do SIM, e, na base do IBGE (Estatísticas do Registro Civil), a idade é 33 anos, a resposta final utilizada no modelo será 33 anos; e

- Para os casos em que as respostas são válidas nas duas bases de dados, serão adotadas regras de decisão, conforme mostrado a seguir:

Quadro 4 - Regras de decisão sobre qual resposta entrará no modelo, por temas

Variável	Fonte considerada
Nascidos vivos	
Data do nascimento	SINASC
Sexo	SINASC
Local de nascimento	SINASC
UF e município de nascimento	SINASC
Idade da mãe	SINASC
UF e município de residência da mãe	Estatísticas do Registro Civil
Óbitos	
Data do óbito	SIM
Sexo	SIM
Idade do falecido	SIM
UF e município de residência do falecido	Estatísticas do Registro Civil
Local de ocorrência do óbito	SIM
Natureza do óbito (natural ou não natural)	SIM

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais.

Aplicação da Técnica de Captura-Recaptura para estimativa dos totais de nascidos vivos e óbitos

A aplicação da Técnica de Captura-Recaptura visa estimar os totais de nascidos vivos e óbitos, sendo essencial para o estudo da cobertura dos Sistemas de Estatísticas Vitais (Estatísticas do Registro Civil, do IBGE, Sistema de Informações sobre Nascidos Vivos - SINASC e Sistema de Informações sobre Mortalidade - SIM, do Ministério da Saúde). Para isso, foi aplicada uma modelagem, descrita a seguir, com o intuito de estimar esses totais.

A modelagem adotada foi o Modelo Linear Generalizado (Generalized Linear Model - GLM). O GLM é implementado após um pareamento bem-sucedido de dois bancos de dados. Huggins (1989, 1991) apresenta uma abordagem que visa modelar a probabilidade de o indivíduo ser capturado usando variáveis presentes nas respectivas bases de dados, como, por exemplo, sexo, idade, local de residência, nível educacional, densidade populacional do local de residência etc. O GLM trabalha com a probabilidade condicional, permitindo a modelagem das probabilidades de captura em termos de covariáveis observáveis. Essa abordagem é útil quando as fontes de dados observadas têm dependência e probabilidades de captura heterogêneas. Deve ser ressaltado que intervalos de confiança para o tamanho da população também são calculados nessa abordagem.

Uma função logística é usada para executar o GLM, de acordo com as características individuais e/ou municipais. Considere-se a seguinte equação do modelo:

$$ln\left(\frac{p_{ib}}{1-p_{ib}}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

Onde:

p_{ib} é a probabilidade de o indivíduo i ser capturado em cada uma das fontes de dados b (SIM e IBGE);

β_0 é o intercepto;

β_j é o parâmetro para a $j-ésima$ variável, $j = 1, 2, .., \mathrm{k}$;

x_j é a $j-ésima$ variável, $j = 1, 2, .., \mathrm{k}$;

$i = 1, 2, 3, \ldots, \mathrm{N}$ é o número de registros;

$b = 1, 2$ é o número de fontes de dados; e

k é o número de covariáveis no modelo.

Segundo Huggins (1989, 1991), para estimar a população, precisa-se usar a probabilidade estimada de captura, calculada pelo modelo acima, usando as seguintes etapas:

$$1)\ \alpha = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j$$

$$2)\ p_{ib}^{*} = \frac{e^{\alpha}}{e^{\alpha} + 1}$$

$$3)\ \hat{p}_i = 1 - \prod_{i=1}^{N} (1 - p_{ib}^{*})$$

$$4)\ \widehat{N} = \sum_{i=1}^{N} \frac{1}{\hat{p}_i}$$

Onde:

$\hat{\beta}_0$ e $\hat{\beta}_j$ são as estimativas dos parâmetros obtidas por meio do citado modelo;

p_{ib}^{*} é a probabilidade estimada de um indivíduo *i* ser capturado pela fonte *b*;

$\hat{p}_i$ é a probabilidade do indivíduo *i* ser capturado pelo menos por uma das fontes; e

$\widehat{N}$ é o total estimado de nascimentos/óbitos (se houver uma baixa probabilidade de um nascimento/morte reportado ser capturado em qualquer uma das fontes, haverá um número maior de nascimentos/mortes estimados por nascimento/morte individual relatado).

Também se pode calcular a variância do estimador $\widehat{N}$ da seguinte maneira:

$$Var(\widehat{N}) = \sum_{i=1}^{N} \frac{1 - \hat{p}_i}{(\hat{p}_i)^2}$$

E, com essas informações, pode-se ter um intervalo de confiança de 95%:

$$IC = \widehat{N} \pm 1{,}96\sqrt{Var(\widehat{N})}$$

Uma vantagem significativa do Método GLM é que ele ajuda a contornar dois aspectos presentes nesse tipo de estudo: o viés introduzido pela dependência entre as duas fontes e a captura heterogênea. Nesse sentido, a modelagem permite que a probabilidade de captura em determinada fonte seja prevista de acordo com cada combinação de variáveis no modelo (por exemplo, para uma morte masculina de um recém-nascido no Estado do Rio de Janeiro, ocorrida em uma unidade de saúde etc.). A aplicação faz isso para cada nascimento/morte relatado, possibilitando estimar a probabilidade de captura em cada fonte (ou seja, a probabilidade conjunta de captura)

e o total de nascimentos/mortes. Essa forma de abordagem é consistente com a sugestão de Chandrasekar e Deming (1949) para reduzir o viés da dependência.

Além dessas funcionalidades, o Método GLM também é útil em dois aspectos: para prever a probabilidade de captura de cada fonte de acordo com diferentes combinações de variáveis e, assim, entender melhor as operações de cada sistema; e por permitir precisão na identificação de quais subgrupos populacionais representam a maior porcentagem de mortes perdidas por qualquer sistema, podendo informar esforços adicionais para capturar esses nascimentos/mortes.

Por fim, é possível se ter vários modelos de acordo com a combinação das variáveis que se decide usar. Uma maneira de ajudar a escolher a melhor modelagem é usar o menor AIC (Akaike Information Criteria), que relaciona a probabilidade condicional do modelo ao número de parâmetros estimados. Sendo assim, as variáveis selecionadas para aplicação do método em cada base foram:

Nascidos vivos

- Unidade da Federação e município de residência da mãe;
- Grupo de idade da mãe;
- Local de nascimento;
- Percentual da população municipal de mulheres de 25 a 39 anos de idade que completaram o ensino médio com base no Censo Demográfico 2010; e
- Densidade populacional do município com base no referido Censo.

Óbitos

- Unidade da Federação e município de residência do falecido;
- Sexo do falecido;
- Grupo de idade do falecido;
- Local de ocorrência do óbito;
- Causa da morte (natural/não natural);
- Percentual da população municipal de 25 a 39 anos de idade que completou o ensino médio com base no Censo Demográfico 2010; e
- Densidade populacional do município com base no referido Censo.

Essa modelagem requer variáveis compatíveis, o que significa que ambas as fontes têm a mesma codificação de categorias de variáveis. Além disso, é importante destacar que houve casos pareados que apresentaram respostas diferentes para as variáveis. Para esses casos, foi criada uma regra para definir de qual base seria a resposta final utilizada na modelagem, e esse passo está descrito na primeira parte da presente nota metodológica.

Uma vez aplicado o Modelo GLM, é possível estimar o volume dos eventos vitais (nascimentos e óbitos) e, consequentemente, os seus respectivos sub-registros.

As estimativas de sub-registro de nascimentos e óbitos levarão em consideração os nascimentos e óbitos ocorridos e não registrados até o 1º trimestre do ano subsequente ao ano de nascimento (atendendo o prazo legal para efetivação do registro). Esse indicador estará associado a uma cobertura das Estatísticas do Registro Civil sem a incorporação dos registros tardios.

Para estimar o total de eventos, é necessário incorporar todos os nascidos vivos e óbitos captados, incluindo aqueles presentes na base das Estatísticas do Registro Civil e que não são objeto de pareamento.

Assim, no caso dos **nascidos vivos**, são incorporados os seguintes registros:

- Registro por sentença judicial;
- Registro Administrativo de Nascimento Indígena (RANI) ou Requerimento FUNAI; e
- Registro realizado a partir da declaração de testemunhas para o qual não foi emitida DN.

No caso dos **óbitos**, são incorporados os seguintes registros:

- Registro realizado a partir da declaração de testemunhas qualificadas para o qual não foi emitida DO;
- Registro por sentença judicial;
- Registro administrativo da Funai; e
- Registro realizado a partir da declaração de testemunhas para o qual não foi emitida DO.

Referências

CHANDRASEKAR, C.; DEMING, W. E. On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistics Association*, Alexandria: ASA, v. 44, n. 245, p. 101-115, 1949. Disponível em: http://unstats.un.org/unsd/vitalstatkb/Attachment1057.aspx?AttachmentType=1. Acesso em: nov. 2019.

HUGGINS, R. M. Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, Washington, DC: International Biometric Society - IBS, v. 47, n. 2, p. 725-732, Jun. 1991. Disponível em: https://www.jstor.org/stable/2532158. Acesso em: nov. 2019.

HUGGINS, R. M. On the statistical analysis of capture experiments. *Biometrika*, Oxford: Oxford University; London: Biometrika Trust, v. 76, n. 1, p. 133-140, Mar. 1989. Disponível em: https://www.jstor.org/stable/2336377. Acesso em: nov. 2019.

04 de dezembro de 2019

Diretoria de Pesquisas