ISSN: 2318-0854

A QUALIDADE DO PREENCHIMENTO DAS VARIÁVEIS SOCIODEMOGRÁFICAS E OPERACIONAIS DAS DECLARAÇÕES DE ÓBITOS DOS CENTENÁRIOS NO SEMIÁRIDO BRASILEIRO

Juliana Barbosa Medeiros [1]
Neir Antunes Paes [2]

RESUMO

Um problema comum em investigações científicas é a ocorrência de dados faltantes. Sendo esse volume moderado ou alto pode comprometer seriamente a qualidade e a confiabilidade dos indicadores deles derivados. Nesse sentido, foram desenvolvidos métodos estatísticos para a imputação de dados, cujo problema está muito presente nas Declarações de Óbitos (DO), particularmente da população centenária que é o contingente que mais cresce no Brasil. Seguindo esta tendência, desponta o Semiárido, com estudos ausentes sobre essa abordagem. Traçou-se como objetivo avaliar a completude dos registros de óbitos e imputar os dados faltantes das variáveis sociodemográficas e operacionais das DO dos centenários do Semiárido brasileiro. Para os triênios 2000/02, 2009/11 e 2018/20 foram avaliadas as seguintes variáveis: sexo, idade, raça/cor, situação conjugal, escolaridade, caracterização municipal, local de ocorrência e assistência médica. Foi adotada a imputação múltipla sendo avaliados o mecanismo e o padrão dos dados ausentes a serem imputados. Considerou-se o mecanismo de não-resposta como completamente aleatório e para o padrão como monotônico. Houve uma melhora da completude das variáveis ao longo dos anos para a maioria das variáveis e categorias estudadas. De acordo com os critérios para a imputação de dados as seguintes categorizações de incompletude foram encontradas para as variáveis em ambos os sexos em todo o período: ?5% (sexo, idade, caracterização municipal e local de ocorrência); 5-15% (raça/cor, situação conjugal); ?15% (escolaridade, assistência médica). A assistência médica atingiu >50% inviabilizando a imputação. Sexo, idade e caracterização do município foram praticamente preenchidos. A imputação realizada com as demais variáveis permite a elaboração de avaliações mais assertivas e confiáveis sobre a dinâmica do preenchimento de variáveis da DO, contribuindo para um melhor esclarecimento do processo de envelhecimento no Brasil, em especial nas regiões com maior carência de estudos científicos e acadêmicos.

Palavras-chave: Centenários, Semiárido, Estatísticas Vitais, Qualidade dos Registros, Imputação de Dados

INTRODUÇÃO

A evolução na disponibilidade das bases de dados dos eventos vitais, principalmente através do SIS, torna o uso das análises quantitativas cada vez mais inerentes ao meio dos estudos epidemiológicos, que utilizam a estatística como principal ferramenta para trabalhar com as informações contidas nas bases (OLIVEIRA et al., 2020).

[1] Doutora em Modelos de Decisão e Saúde da Universidade Federal da Paraíba – UFPB, julianabcnet@hotmail.com;
[2] Professor Orientador: Doutor, Programa de Pós Graduação em Modelos de Decisão e Saúde da Universidade Federal da Paraíba – UFPB, neirpaes@yahoo.com.br

ISSN: 2318-0854

Nesta busca pelo conhecimento empírico, a qualidade da informação é imprescindível, uma vez que, várias técnicas e aplicações ficam limitadas a completude das informações. A completude é um dos pilares de avaliação da qualidade de uma base de dados, uma vez que, a incompletude refere-se as informações em branco contidas nas variáveis o que compromete a consistência dos dados (ROMERO; CUNHA, 2006).

A perda de informações dentro de um formulário, é conhecido como problema de dados faltantes ou missing-values, que em alguns casos provocam a exclusão total do sujeito. Essa exclusão pode provocar um viés nos resultados, uma vez que, os sujeitos excluídos, podem pertencer a determinadas categorias que não terão representatividade na amostra final. Em outros casos, os valores não declarados inviabilizam o uso de determinada variável, pois a quantidade de dados faltantes na mesma é muito alta (NUNES, L. N.; KLÜCK; FACHEL, 2010).

Como critérios de avaliação da completude pode ser classificada em: excelente (menos de 5%); boa (5 a 9%); regular (10 a 19%); ruim (20 a 49%); e muito ruim (50% ou mais) (PAES; GOUVEIA, 2010). Na tentativa de obter o maior número de informações, técnicas são empregadas para resgatar esses valores por meio de re-checagem, autópsia verbal, linkage ou imputação dos valores, resultando esta última alternativa em uma completude total das informações finais.

O linkage e a imputação de dados têm sido cada vez mais utilizados por pesquisadores e órgãos governamentais por possuírem baixo custo operacional e eficiência no resgate dos dados. Contudo, a maioria das aplicações utilizam apenas uma dessas técnicas, cuja escolha depende das variáveis estudadas. Utiliza-se o linkage para realizar o pareamento com outras bases de dados, ou seja, o linkage é uma técnica de relacionamento de bases de dados que busca identificar de forma precisa se o registro de um indivíduo também está contido em outra base de dados. Contudo, as variáveis de estudo dessa tese, contidas na DO, não estão contidas em outras bases de dados de interesse, como no caso do SIH, portanto para corrigir os problemas de incompletude das variáveis o foco traçado neste trabalho é concentrado na imputação de dados.

Diante do exposto, o presente estudo tem como objetivo avaliar a completude dos registros de óbitos e imputar os dados faltantes das variáveis sociodemográficas e operacionais das DO dos centenários do Semiárido brasileiro.

ISSN: 2318-0854

METODOLOGIA

O presente estudo foi do tipo ecológico, transversal, referente aos óbitos de centenários do Semiárido brasileiro, com abordagem quantitativa e que fez uso de informações sociodemográficas presentes nas Declarações de Óbitos, no período de 2000 a 2020.

Para o estudo da mortalidade teve-se como foco todos os casos de óbitos de indivíduos com 100 anos ou mais registrados na base de microdados do Sistema de Informações de Mortalidade (SIM). Visando minimizar as flutuações aleatórias no número de óbitos anuais, onde pequenos números geram altas variabilidades nas taxas, optou-se por trabalhar com os triênios: 2000, 2001 e 2002; 2009, 2010 e 2011; 2018, 2019 e 2020.

A unidade de análise observacional do estudo foram os centenários dos municípios pertencentes ao Semiárido brasileiro, que por sua vez, foram agregados em espaços regionalizados, denominados de acordo com a nova tipologia de caracterização municipal (Urbano, Intermediário e Rural), com o propósito de se obter melhor representatividade quantitativa.

Para o presente estudo foram utilizadas seis variáveis da DO: faixa etária, raça/cor, situação conjugal, escolaridade, assistência médica e local de ocorrência do óbito. Essas variáveis foram analisadas separamente para cada sexo (masculino e feminino).

Como critérios de avaliação da completude das variáveis neste estudo utilizou-se a classificação proposta por Romero e Cunha (2006), que define a completude das variáveis de acordo com o percentual de missing-values : excelente (menos de 5%); boa (5 a 9%); regular (10 a 19%); ruim (20 a 49%); e muito ruim (50% ou mais).

Para corrigir os problemas de incompletude das variáveis utilizou-se a técnica de imputação de valores para completar as informações ausentes das variáveis contidas na DO selecionadas neste estudo. Para aplicação desta técnica foi necessário avaliar o mecanismo e o padrão de dados ausentes dos dados a serem imputados. Os três mecanismos de não-resposta segundo Little e Rubin (2019) são:

i) ausentes de forma completamente aleatória;

ii) ausentes de forma aleatória e

iii) ausentes de forma não-aleatória.

Os padrões de não-resposta se referem à forma com que os valores ausentes estão distribuídos em uma base de dados e podem ser classificados em:

i) padrão univariado,

ISSN: 2318-0854

ii) padrão monotônico

iii) padrão não-monotônico.

Após essas verificações foram imputadas as informações ausentes das variáveis sociodemográficas e local de ocorrência do óbito.

De acordo com os percentuais de dados ausentes nas variáveis foi possível determinar o método de imputação a ser utilizado. Segundo Harrell Jr (2001) é possível serem definidas linhas gerais para a escolha entre os métodos de imputação de acordo com a proporção de dados faltantes em qualquer uma das variáveis.

- Proporção ≤ 0,05 → Neste caso pode ser usada imputação única ou analisar somente os dados completos.
- Proporção entre 0,05 e 0,15 → Imputação única pode ser usada aqui provavelmente sem problemas, entretanto o uso da imputação múltipla é indicado.
- Proporção ≥ 0,15 → A imputação múltipla é indicada na maior parte dos casos.

Optou-se pela imputação múltipla, pois essa técnica produz resultados não viesados e com erros padrão apropriados.

Uma questão relevante na imputação múltipla é a escolha da quantidade de imputações m a serem realizadas. Alguns autores argumentam que um m entre 3 e 5 já é suficiente para gerar resultados satisfatórios (RUBIN, 1996; SCHAFER, 1999). A decisão sobre a quantidade de imputações se baseia em um indicador denominado por Rubin (RUBIN, 1996) de Eficiência Relativa (ER), expresso como função da quantidade de imputações (m) e do percentual de dados ausentes da variável (λ). O resultado do indicador aponta o percentual de eficiência dos valores imputados de cada variável. Assim, o cálculo para obtenção da Eficiência Relativa é dado por:

$$ER = \left(1 + \frac{\lambda}{m}\right)^{-1}$$

De acordo com os dados observados e o percentual de dados faltantes, foi calculado a eficiência relativa para cada variável e após esse passo foi optado o número de imputações para os dados faltantes das variáveis do estudo.

A técnica de imputação múltipla cria m cópias da base de dados onde os valores ausentes são substituídos por valores plausíveis imputados através de técnicas adequadas de estimação. Os valores imputados para o padrão monotônico são obtidos por meio de métodos estatísticos inferenciais como Método da Regressão Linear Bayesiana ou Método da Média Preditiva e para o padrão não-monotônico tem-se o método de Monte Carlo baseado em Cadeias

ISSN: 2318-0854

de Markov. Um número m de bancos distintos e completos são gerados, e cada um deles deve ser analisado (RUBIN, 1996; SCHAFER, 1999). Para a combinação entre todas as m estimativas individuais de todas as imputações realizadas, recorreu-se às Regras de Rubin (RUBIN, 1987) que se utiliza de estimativas da média e da variância entre as imputações.

Para cada análise das m bases de dados completas, obtém-se uma estimativa para um parâmetro escalar de interesse Q, ou seja, Q_j, j=1,2,…,m. Segundo Schafer (1999), Q pode ser qualquer medida escalar a ser estimada, tal como média, correlação, coeficiente de regressão ou razão de chances. Então a estimativa combinada $\bar{Q}$ será a média das estimativas individuais ($\widehat{Q_J}$):

$$\bar{Q}=\frac{1}{m}\sum_{j=1}^{m}\widehat{Q_J}$$

Para a variância combinada, primeiramente calcula-se a variância dentro das imputações:

$$\bar{U}=\frac{1}{m}\sum_{j=1}^{m}U_j$$

e a variância entre imputações:

$$B=\frac{1}{m-1}\sum_{j=1}^{m}(\widehat{Q_J}-\hat{Q})^2$$

Então, a variância total, que é a variância combinada, será:

$$T=\bar{U}+\left(1+\frac{1}{m}\right)B$$

Para este estudo o Mecanismo de não resposta foi o Completamente Não Aleatório (MCAR) e o padrão de não resposta foi o monotônico, para a imputação múltipla foi utilizado o m = 5.

O banco de dados foi elaborado utilizando-se o aplicativo Microsoft Office Excel. Para a análise dos dados, foi utilizado o software SPSS 20.0 (Statistical Package for the Social Sciences, Chicago, EUA). Por se tratar de um estudo que emprega apenas dados secundários provenientes de bancos de dados de domínio público, disponibilizados online, justificou-se o não encaminhamento deste estudo para aprovação por comitê de ética em pesquisa, segundo o que estabelece o inciso III, artigo primeiro da Resolução nº 510/16 do Conselho Nacional de SaúdeA metodologia do artigo deverá apresentar os caminhos metodológicos e uso de

ISSN: 2318-0854

ferramentas, técnicas de pesquisa e de instrumentos para coleta de dados, informar, quando for pertinente, sobre a aprovação em comissões de ética ou equivalente, e, sobre o direito de uso de imagens.

RESULTADOS E DISCUSSÃO

A completude representa a magnitude ou nível de declaração de uma determinada variável informada quando o registro foi realizado no sistema. Problemas no preenchimento das variáveis são decorrentes de declarações errôneas ou omissões de informações referentes a essas variáveis (PAES, 2018). A completude é traduzida pela proporção de campos preenchidos com resposta não-nulas e a incompletude pelo percentual de informações ignoradas ou não preenchidas, também denominadas de missings. Portanto, nesta primeira etapa avaliou-se a qualidade das variáveis contidas nos registros de óbitos dos centenários, ao longo dos anos, verificando o percentual de incompletude das variáveis que foram utilizadas nesta investigação. A Tabela 1 apresenta um panorama da situação em termos do percentual de informações ignoradas ou não preenchidas das variáveis selecionadas na Declaração de Óbito dos centenários, por sexo, para os triênios de 2000 a 2020.

Tabela 1 Número e percentual de informações ignoradas ou não preenchidas segundo variáveis selecionadas da Declaração de Óbito dos centenários, por sexo, nos espaços regionalizados do Semiárido brasileiro, nos triênios de 2000 a 2020.

Ano do óbito		**2000-2002**						**2009-2011**						**2018-2020**					
Variável	Sexo	Rural		Intermediário		Urbano		Rural		Intermediário		Urbano		Rural		Intermediário		Urbano	
		n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Faixa Etária	H	0	0,0	0	0,0	0	0,0	0	0,0	0	0,0	0	0,0	0	0,0	0	0,0	0	0,0
	M	0	0,0	0	0,0	0	0,0	0	0,0	0	0,0	0	0,0	0	0,0	0	0,0	0	0,0
Raça/cor	H	21	17,8	8	19,5	32	31,7	88	13,5	27	11,2	76	15,2	60	4,0	20	4,1	45	5,6
	M	41	19,0	12	17,4	34	22,2	126	11,5	52	12,9	125	12,2	96	4,0	23	2,8	65	3,5
Situação Conjugal	H	12	10,2	2	4,8	6	5,9	50	7,7	19	7,9	44	8,8	133	8,8	39	8,0	95	10,1
	M	11	5,1	6	8,7	6	3,9	97	8,9	35	8,6	113	10,0	237	9,9	79	9,5	177	9,7
Escolaridade	H	36	30,5	10	24,4	45	44,6	150	23,0	64	26,0	145	29,1	251	16,6	91	18,7	178	19,0
	M	73	33,7	28	40,6	52	34,0	247	22,5	93	23,0	266	26,0	405	16,9	143	17,3	397	21,7
Assistência médica?	H	96	81,4	39	95,1	75	74,3	554	85,0	198	82,0	418	83,8	830	54,8	271	55,6	527	56,3
	M	172	79,6	52	75,4	119	77,8	902	82,3	345	85,0	830	81,1	1362	56,8	464	56	992	54,2
Local de ocorrência	H	4	3,4	0	0,0	3	3,0	1	0,2	0	0,0	1	0,2	1	0,1	0	0,0	0	0,0
	M	2	0,9	0	0,0	1	0,6	1	0,1	0	0,0	1	0,1	1	0,0	2	0,2	0	0,0

Fonte: Sistema de Informações sobre Mortalidade – Ministério da Saúde.
Legenda: H: Homens M: Mulheres.

ISSN: 2318-0854

Observa-se na Tabela 1 que a variável *Faixa Etária* não apresentou dados ausentes ou ignorado, o que se deve ao fato que durante a coleta de microdados do SIM foi utilizado como filtro os registros de óbitos dos indivíduos com 100 anos ou mais declarados ou calculados pelo sistema após verificação ou rechecagem.

Avaliando a variável *Raça/cor* foi possível observar que os valores de dados ausentes ou ignorados variaram entre 4,0-31,7% entre os homens e 2,8-22,2% entre as mulheres, portanto a incompletude dessa variável foi maior entre os homens, principalmente aqueles que residiam em áreas urbanas, porém para ambos os sexos houve um importante descenso da incompletude ao longo dos anos.

A variável S*ituação Conjugal* também apresentou maior percentual de *missings* entre os homens com variação de 4,8-10,2 e 3,9-10,0% entre as mulheres. Um fato preocupante e que merece destaque é o aumento do percentual de registros ausentes dessa variável ao longo dos anos, para ambos os sexos, exceto os homens residentes em área rural entre os dois primeiros triênios quando houve uma diminuição no percentual. Em todas as outras situações o acréscimo de pontos percentuais foi marcante no decorrer do período, sinalizando um piora na captação dessa informação, sendo a única, de um modo geral, que andou na contramão da tendência das demais.

O registro da *Escolaridade*, apesar de elevados percentuais sobretudo nos primeiros triênios, melhorou de forma expressiva ao longo dos anos. O percentual de dados ausentes também foi maior entre os homens, com uma variação de 16,6-44,6%, já entre as mulheres a variação foi 16,9-40,6%.

Ainda avaliando a variável *Escolaridade* os municípios classificados como Urbanos, apresentaram maiores percentuais de dados ausentes, situação análoga encontrada nas variáveis *Raça/cor* e *Situação Conjugal*. A presença de *missings* foi superior ou próximo a 30% nos dois primeiros triênios, em ambos os sexos, o que dificulta uma análise com mais consistência, originada pela negligência com as informações socioeconômicas, em especial o nível de instrução quando do preenchimento da DO pelo médico, profissional responsável por esse registro.

As deficiências no preenchimento do formulário são essencialmente um fenômeno sociocultural, que depende do contexto onde o médico está inserido e trabalha (MENDONÇA; DRUMOND; CARDOSO, 2010). Nesse sentido, a escolaridade é um dos itens mais negligenciados no seu preenchimento e de difícil recuperação (BARBUSCIA; RODRIGUES-

ISSN: 2318-0854

JÚNIOR, 2011). Estudos evidenciam que os patologistas negligenciam mais esse preenchimento que o clínico geral (STUQUE; CORDEIRO; CURY, 2003).

A variável *Assistência Médica* (durante a doença que ocasionou o óbito) apresentou os percentuais mais elevados de incompletude durante o período analisado e em comparação a todas as outras variáveis estudadas. Os percentuais de *missings* foram acima de 50% em todas as situações analisadas. O maior percentual encontrado foi entre os homens que residiam em municípios caraterizados como Intermediário (95,1%) no primeiro triênio, já os menores valores foram encontrados entre as mulheres, residentes em área urbana, no último triênio (54,2%). Apesar dos elevados percentuais, houve melhora do registro dessa variável ao longo dos anos. Sem embargo, os percentuais de dados ausentes nessa variável, no triênio mais recente, foram tão elevados, acima de 50%, que inviabiliza qualquer tentativa de recuperação desses dados. Sendo assim, essa variável ficou excluída do processo de imputação dos dados na etapa seguinte e não será introduzida no método de imputação múltipla.

Por fim, uma variável bem declarada, com valores quase equivalentes a 0 “zero” e com percentuais menores comparadas as demais variáveis sociodemográficas, foi *Local de ocorrência* do óbito que apresentou poucos dados ausentes ou ignorados. O percentual diminuiu ao longo dos anos, tendo as mulheres menores percentuais de dados ausentes, principalmente aquelas que residiam em área urbana ou intermediária.

Estudo realizado com idosos longevos no Nordeste brasileiro, entre 2001 e 2015, também evidenciou maior incompletude das variáveis sociodemográficas entre os homens, assim como a variável *Escolaridade* apresentou maiores percentuais de dados ausentes, comparadas às variáveis *Raça/cor* e *Estado Civil* (MENDES, 2018). Medeiros (2015) realizou uma análise da mortalidade em idosos longevos no Brasil, entre 2000 e 2010, cujos resultados corroboram com os percentuais de dados ausentes tanto para a variável *Assistência Médica*, com percentual de incompletude também acima de 50%, como a variável *Local de ocorrência* do óbito com os menores percentuais analisados. Em ambos os estudos, houve melhora da completude das variáveis ao longo dos anos, assim como foi evidenciado na maioria das variáveis estudadas entre os centenários.

CONSIDERAÇÕES FINAIS

A imputação realizada com as variáveis, exceto assistência médica que atingiu um percentual de incompletude acima de 50%, permitiu a elaboração de avaliações mais assertivas e confiáveis sobre a dinâmica do preenchimento de variáveis da DO, contribuindo para um

ISSN: 2318-0854

melhor esclarecimento do processo de envelhecimento no Brasil, em especial nas regiões com maior carência de estudos científicos e acadêmicos.

REFERÊNCIAS

BARBUSCIA, D. M.; RODRIGUES-JÚNIOR, A. L. Completude da informação nas Declarações de Nascido Vivo e nas Declarações de Óbito, neonatal precoce e fetal, da região de Ribeirão Preto, São Paulo, Brasil, 2000-2007. **Cad. Saúde Pública**, v. 27, n. 6, p. 1192-1200, 2011.

HARRELL JR, F.E. **Regression modeling strategies: with applications to linear models, logistic regression and survival analysis**. New York: Springer-Verlag, 2001.

LITTLE, R. JA; RUBIN, D. B. **Statistical analysis with missing data**. John Wiley & Sons, 2019.

MEDEIROS, W. R. **Mortalidade em idosos longevos e "mais jovens" no Brasil.** 2015. 108f. Tese (Doutorado em Saúde Coletiva) - Centro de Ciências da Saúde, Universidade Federal do Rio Grande do Norte, Natal, 2015.

MENDES, T. C. O. **Perfis da mortalidade de idosos no Nordeste: estudo comparativo entre três faixas etárias e seus fatores contextuais relacionados.** 2018. 160f. Tese (Doutorado em Saúde Coletiva) - Centro de Ciências da Saúde, Universidade Federal do Rio Grande do Norte, Natal, 2018.

MENDONÇA, F. M.; DRUMOND, E.; CARDOSO, A. M. P. Problemas no preenchimento da Declaração de Óbito: estudo exploratório. **Revista Brasileira de Estudos de População**, v. 27, p. 285-295, 2010.

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. **Revista Brasileira de Epidemiologia**, v. 13, p. 596-606, 2010.

OLIVEIRA, G. M. M. et al. Estatística Cardiovascular–Brasil 2020. **Arquivos brasileiros de Cardiologia**, v. 115, p. 308-439, 2020.

PAES, N.A. **Demografia estatística dos eventos vitais: com exemplos baseados na experiência brasileira**. João Pessoa: Editora do CCTA, 2018. 215p.

PAES, N. A.; GOUVEIA, J. F. Recuperação das principais causas de morte do Nordeste do Brasil: impacto na expectativa de vida. **Revista de Saúde Pública**, v. 44, p. 301-309, 2010.

ROMERO, D. E.; CUNHA, C. B. Avaliação da qualidade das variáveis sócioeconômicas e demográficas dos óbitos de crianças menores de um ano registrados no Sistema de Informações sobre Mortalidade do Brasil (1996/2001). **Cadernos de Saúde Pública**, v. 22, p. 673-681, 2006.

ISSN: 2318-0854

RUBIN, D. B. Multiple imputation after 18+ years. **Journal of the American statistical Association**, v. 91, n. 434, p. 473-489, 1996.

SCHAFER, J. L. Multiple imputation: a primer. **Statistical methods in medical research**, v. 8, n. 1, p. 3-15, 1999.

STUQUE, C. O.; CORDEIRO, J. A.; CURY, P. M. Avaliação dos erros ou falhas de preenchimento dos atestados de óbito feitos pelos clínicos e pelos patologistas. **J. Bras. Patol. Med. Lab.**, v. 39, n. 4, p. 361-364, 2003.