ALAP 2020
IX Congreso de la Asociación Latinoamericana de Poblacion
9 a 11 diciembre
EL ROL DE LOS ESTUDIOS DE POBLACIÓN TRAS LA PANDEMIA DE COVID-19 Y EL DESAFÍO DE LA IGUALDAD EN AMÉRICA LATINA Y EL CARIBE

João Gabriel Malaguti, Escola Nacional de Ciências Estatísticas, joaogmalaguti@gmail.com
Leonardo Gravina de Faria, Universidade Federal de Juiz de Fora, leonardo.gravina.faria@gmail.com

Comparação de Métodos de Imputação para Estatística Espacial

Resumo

Para a realização de algumas estatísticas dentro da área de estatística espacial, como os índices de Moran, não se pode ter dados faltantes. Neste artigo comparam-se, por simulações Monte Carlo, dois grupos de técnicas de imputação: imputações georreferenciadas e *multiple imputation by chained equations* (MICE). Utilizando dados reais dos municípios de Minas Gerais, gera-se não resposta utilizando os diferentes mecanismos de não resposta existentes (MCAR, MAR e MNAR), imputa-se os dados faltantes antes da análise espacial e se compara os resultados com os dos dados completos. Em geral, o grupo de imputações que teve melhor performance foi o que leva em conta a distribuição geográfica dos municípios analisados, mesmo sendo mais simples que algumas das técnicas do grupo MICE.

Palavras-chave: Não resposta; Simulação Monte Carlo; Índice de Moran

Abstract

In order to produce certain statistics in the area of spatial statistics, for example Moran indexes, there can't be missing data. In this paper, we compare, via Monte Carlo simulations, two groups of imputation techniques: georeferenced imputation and multiple imputation by chained equations (MICE). Using real data from the cities in Minas Gerais, we generate nonresponse using the different existing mechanisms (MCAR, MAR and MNAR), impute the data before the spatial analysis and compare the results with those from the complete dataset. We find that the group of imputations that had better performance was the one who takes the geographic distribution of the analysed cities into account (georeferenced imputation) even being simpler than many MICE techniques.

Keywords: Nonresponse; Monte Carlo simulation; Moran Index

1. Introdução

Dados faltantes são um problema bastante comum no campo da estatística aplicada, seja por erros no preenchimento dos bancos de dados ou por não resposta. Quando se usa dados provenientes dos censos esse problema é eliminado, mas quando se utilizam os dados de pesquisas amostrais ou de registros administrativos os efeitos da não resposta têm de ser levados em conta.

Além de ser um problema prático, ele também é um problema metodológico. A literatura da área descreve três mecanismos causadores de não resposta: MCAR (não resposta completamente aleatória), MAR (não resposta aleatória) e MNAR (não resposta não aleatória (Little & Rubin, 2019). Eles podem causar a não resposta sozinhos ou em combinação, mas apenas a não resposta causada pelo mecanismo MCAR é ignorável. E ainda não existe uma maneira de determinar qual (ou quais) mecanismo(s) são responsáveis pela não resposta.

Mesmo a não reposta por MCAR é "ignorável" em parte, pois certas aplicações (como as que serão apresentadas neste artigo) dependem da existência de valores para todas as observações. Para essas, os dados faltantes precisam ser imputados.

Imputação é um método bastante flexível para o tratamento de dados faltantes, mas apresenta desvantagens, como o agrupamento de situações onde o problema é capaz de ser lidado nesta maneira e outras na qual tanto os estimadores aplicados aos dados reais e imputados apresentam viés substancial (Dempster & Rubin, 1983).

Neste artigo, comparamos dois grupos de métodos de imputação para três variáveis sujeitas aos diferentes mecanismos de não resposta, utilizando simulações Monte Carlo. O intuito sendo descobrir qual grupo (se algum) se mostra melhor para tratar os dados faltantes para estatística espacial.

Primeiramente apresentamos as medidas de estatística espacial que foram o catalisador deste estudo, depois discutimos sobre não resposta e imputação de maneira mais profunda. Na seção 3 descrevemos as simulações, e analisamos seus resultados na Seção 4.

2. Desenvolvimento

2.1 Estatística Espacial

Estatística espacial é um conjunto de técnicas que lida com diferentes problemas espaciais, como hidrologia, ecologia, tendências socioeconômicas, entre outros (Negreiros *et al*., 2010). Esses problemas são considerados "espaciais" porque, como Anselin e Getis (1992) colocam, os efeitos espaciais complicam qualquer entendimento simples e direto dos dados, com a presença da estrutura espacial nos dados implicando na variabilidade da similaridade com a distância entre os locais (Fortin & Dale, 2009).

O principal conceito da área é a autocorrelação espacial, que pode ser definida como a correlação intravariável através do espaço georreferenciado (Getis, 2008). Uma maneira de medir a autocorrelação espacial, ou associação espacial, é por estatísticas globais (como o I de Moran ou o C de Geary) ou por estatísticas locais (como a classe de métodos *Local Indicators of Spatial Association*, ou LISA) (Anselin, 1995). As estatísticas locais permitem a detecção

de zonas de associação espacial alta, positiva ou negativa, além de não esconder instabilidades locais.

A medida utilizada neste trabalho é o índice local de Moran (pertencente à classe LISA), que pode ser entendido como a união de duas classificações: quadrantes de Moran e suas significâncias. As zonas mencionadas anteriormente são os lugares (ou conjunto contíguo de lugares) para os quais LISA é significante. Uma análise mais profunda da estatística, incluindo suas propriedades, foi feita por Anselin (1995).

2.2 Não Resposta

Quando dados de um estudo são incompletos (faltantes), existem implicações importantes para a análise (Fitzmaurice *et al*., 2014). A perda de informação, redução da precisão e viés potencial precisam ser levados em conta, dado a complicação da análise.

Rubin (1987) formaliza a hierarquia dos mecanismos de não resposta, baseando-se nas relações entre o padrão de não resposta, as variáveis de interesse e as covariáveis.

Não resposta completamente aleatória (*missing completely at random*, ou MCAR) é quando a não resposta não é causada por qualquer variável do conjunto.

Não resposta aleatória (*missing at random*, ou MAR) ocorre quando a probabilidade que respostas são faltantes dependem do conjunto de variáveis observadas, mas independem dos valores faltantes específicos que deveriam ter sido obtidos. De maneira mais simples, MAR ocorre quando os valores faltantes da variável de interesse são faltantes devido ao efeito de covariáveis.

Não resposta não aleatória (*missing not at random*, ou MNAR) ocorre quando a probabilidade que respostas são faltantes dependem da variável de estudo.

2.3 Imputação

A ideia básica da imputação é substituir ou inserir os valores faltantes com valores imputados. O principal atrativo do método é que, após a imputação, métodos que precisam de casos completos podem ser utilizados para análise (Fitzmaurice *et al*., 2014).

No entanto, existe uma divergência sobre o uso de imputação simples contra imputação múltipla. A imputação simples falha em considerar a incerteza inerente na imputação, mas é mais simples e rápido de ser feito. A imputação múltipla contorna tal falha, mas em compensação, depende da modelagem apropriada para os dados faltantes (Kenward, 2014).

Uma das imputações mais simples e utilizadas é a imputação da média simples, que para qualquer valor faltante, imputa a média dos valores observados para a variável de interesse.

A classe de métodos aqui chamada de georreferenciados, utiliza da imputação de média condicional ou assistida (Carpenter & Kenward, 2014), usando variáveis de região. Essa imputação consiste em calcular as médias dos valores obtidos para todas as subdivisões da variável e imputar os valores faltantes de cada subdivisão com a média da mesma.

Neste artigo, são as subdivisões geográficas oficiais de mesorregião e microrregião (IBGE, 1990), e de região geográfica imediata e intermediária (IBGE, 2017). Além dessas 4, analisa-se também o que chamamos de "média da classificação", que são a média entre as imputações de microrregião e mesorregião; e região geográfica imediata e intermediária. Elas recebem o código de "MesoMicro" e "ImedInt".

Com o intuito de verificar se as definições das subdivisões não acabam por esconder certas associações mais próximas, criamos um outro método de imputação (com o código de "Viz") que utiliza a média dos municípios vizinhos. Novamente, também incluímos médias das médias, desta vez combinando "Viz" com os 4 métodos originais ("VizMeso", "VizMicro", "VizImed" e "VizInter").

Podemos descrever o processo da imputação por "Viz" de maneira algorítmica. A partir da matriz de vizinhança presente em arquivos *shapefile*, o algoritmo verifica todos os municípios com dados faltantes que não possuem vizinhos com dados faltantes e imputa o valor da média dos vizinhos. Feito isso, ele faz o mesmo para os municípios com no máximo um vizinho com dados faltantes e faz o mesmo, então atualizando o banco de dados.

É feita uma verificação para saber se algum novo município entre na categoria de apenas um vizinho (por exemplo, se um município tem dois vizinhos com dados faltantes e um deles foi imputado, ele passa a ter apenas um vizinho com dado faltante), se sim, o programa continua imputando para aqueles com apenas um vizinho. Quando não existem mais, ele aumenta o valor para dois vizinhos faltantes e segue o processo até todos os dados serem imputados.

Além da imputação pela média simples descrita no início desta seção, outros 8 métodos pertencentes à classe MICE são analisados neste estudo (Van Buuren & Groothuis-Oudshoorn, 2010; Van Buuren, 2018). Todos os métodos desta classe aqui descritos podem ser utilizados na linguagem de programação R, pelo pacote *mice*.

O método "sample" imputa uma amostra aleatória dos valores observados, sendo junto com a média marginal os métodos MICE mais simples.

Três dos métodos são baseados em regressão linear normal, "norm.predict", "norm.nob" e "norm" (Van Buuren, 2018). O primeiro destes, também chamado de imputação por regressão, utiliza o modelo linear sem considerar o erro (Equação 1). Enquanto que "norm.nob" (ou imputação por regressão estocástica) o considera (Equação 2). O último desses, "norm", usa imputação múltipla Bayesiana (Equação 3).

$$\dot{y} = \beta_0 + \beta_1 X_{Faltante} \quad \text{(Eq.1)}$$

$$\dot{y} = \beta_0 + \beta_1 X_{Faltante} + \epsilon \quad \text{(Eq.2)}$$

$$\dot{y} = \dot{\beta_0} + \dot{\beta_1} X_{Faltante} + \dot{\epsilon} \quad \text{(Eq.3)}$$

Nos quais, $\dot{y}$ são os valores imputados, $X_{Faltante}$ são o conjunto de covariáveis dos dados faltantes, β_0 e β_1 são os estimadores de mínimos quadrados para os dados observados, ϵ é retirado aleatoriamente de uma $N(0, \sigma^2)$; e $\dot{\beta_0}$, $\dot{\beta_1}$ e $\dot{\epsilon}$ são retirados aleatoriamente da distribuição a posteriori dados os valores observados.

Também Bayesiano é o método "2l.lmer", que imputa usando um modelo normal de dois níveis.

Árvores de classificação e regressão (*classification and regression trees*, ou CART) são bastante populares na área de aprendizagem de máquina (Breiman *et al*, 1984) e podem ser aplicadas em imputação. Tais modelos buscam preditores e pontos de corte que subdividem a amostra em subamostras mais homogêneas. O processo se repete até a definição de uma árvore binária (Van Buuren, 2018).

O método "pmm" é *predictive mean matching*, que cria subamostras dos dados observados. Ele calcula o valor predito da variável de interesse de acordo com o modelo de imputação escolhido. Para cada valor faltante, forma-se um conjunto (com 5 candidatos, nesta aplicação) dos casos completos que possuem valores preditos próximos ao valor predito para o faltante. Seleciona-se aleatoriamente entre os candidatos e o valor observado do sorteado é o utilizado para a imputação.

Finalmente, o último método analisado da classe MICE é o método do indicador aleatório (*random indicator*, ou RI). Ele estima um offset entre a distribuição dos dados observados e dos dados faltantes, iterando algoritmicamente os modelos de resposta e imputação, assumindo que os dois têm os mesmos preditores.

3. Estudo de Simulação

3.1 Dados Utilizados

Antes de chegar na simulação Monte Carlo propriamente dita, é preciso discutir os dados e a métrica de comparação utilizada neste estudo.

Primeiro foi definido o limite territorial. Escolhemos realizar este estudo com os municípios de Minas Gerais por duas razões: primeiro, é a UF com maior número de municípios (853) e suas subdivisões apresentam tanto grupos homogêneos quanto heterogêneos.

Utilizamos três variáveis de estudo, retiradas de dados censitários do ano de 2010: Indice de Desenvolvimento Humano Municipal (IDHM), PIB *per capta* e número de óbitos. Um dos fatores que levaram a essa escolha foi o fato de suas médias serem bastante díspares (respectivamente 0,668; 141,38; e 11221,78) o que ajuda na generalização da comparação.

Outro fator que influenciou na decisão das variáveis foi a classificação do IDHM em faixas (Pinto, Costa & Marques, 2013), que foram utilizadas para a geração de não resposta no segundo cenário analisado (descrito em mais detalhes na seção 3.3).

3.2 Métrica de Comparação

Como o intuito do estudo é comparar dentre os métodos de imputação qual gera os resultados mais próximos dos valores reais, foi necessário desenvolver uma medida única para isto, uma que leva em conta tanto a classificação nos diferentes quadrantes de Moran quanto a significância encontrada nas estatísticas LISA.

Baseado nos valores imputados, realizam-se os cálculos para as estatísticas de interesse. Estas são comparadas com as estatísticas reais (provindas da população, sem qualquer não resposta) e são pontuados: 0, se os valores diferem e 1, caso sejam iguais. Calcula-se a média entre as duas métricas (quadrantes de Moran e significâncias LISA) para cada município e, ao final, a média para todo o território analisado.

A medida criada para a comparação dos métodos foi denominada razão de acertos de Moran. De modo geral, quanto mais próximo o valor é de 1, melhor é a imputação para a criação dessas estatísticas.

3.3 Simulação

Estudos de simulação Monte Carlo são métodos estatísticos computacionais baseados na geração de diferentes amostras independentes para se obter respostas aproximadas para problemas probabilísticos (Brandimarte, 2014; Lewis & Orav, 1989).

A precisão deste método está relacionada ao número de amostras geradas (iterações ou rodadas do estudo), seguindo a lei dos grandes números que descreve que com um grande número de ensaios, o valor esperado dos resultados observados converge para a média verdadeira.

Como para este estudo é preciso comparar os métodos sabendo qual é o mecanismo gerador de não resposta, partimos de dados populacionais e em cada rodada geramos aleatoriamente não resposta para ser imputada.

Criamos dois cenários de análise, baseados nos mecanismos de não resposta citados anteriormente. Um na qual a não resposta é completamente aleatória com probabilidade de 0,25, chamado de "MCAR".

O segundo cenário tem a não resposta gerada com base nas faixas da variável IDHM, assumindo que municípios com IDH menor tem probabilidade maior de não resposta. Este cenário funciona como MNAR para a variável de IDHM, e como MAR para as demais. A probabilidade média de não resposta é aproximadamente 0,18.

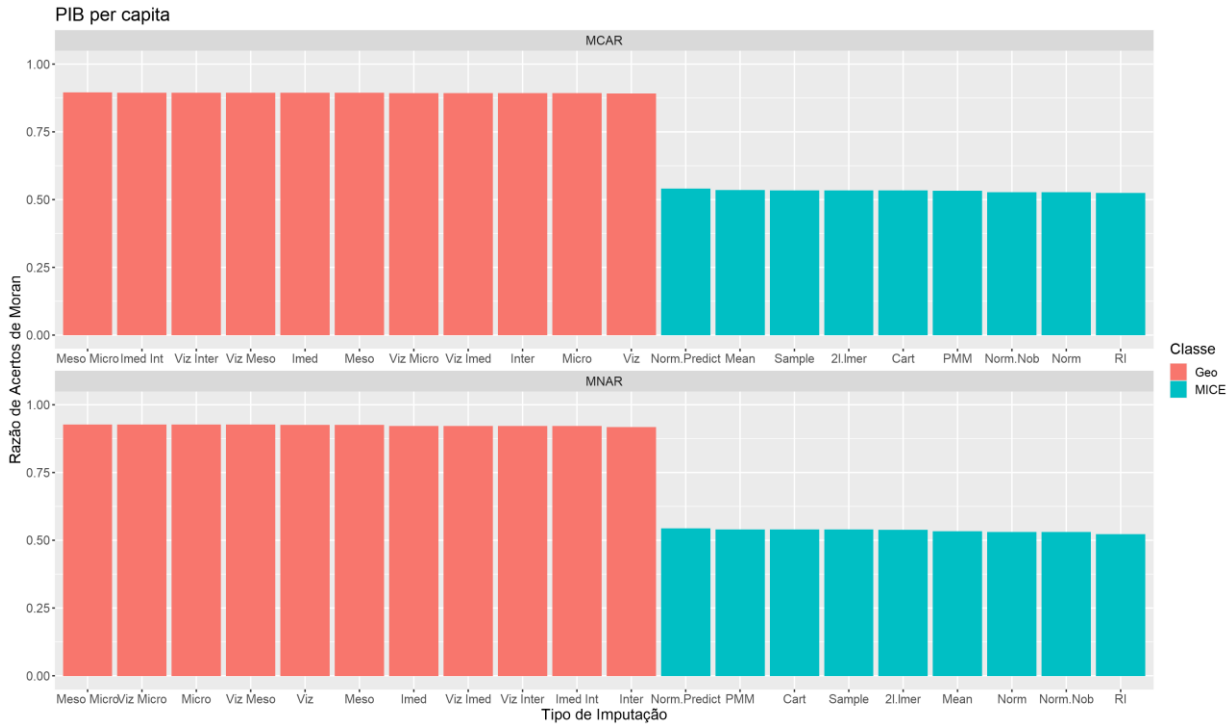
Usando o *software* estatístico R (R Core Team, 2020), criamos um código que gera não resposta, imputa os valores faltantes e calcula as estatísticas de estudo e a métrica de comparação. Para cada combinação de cenário, variável e método de imputação foram realizadas 10.000 iterações Monte Carlo.

4. Resultados

Para uma melhor visualização dos resultados, podemos dispor os dados de maneira gráfica (Figuras 1, 2 e 3). Estes gráficos de barras tem o valor da razão de acertos de Moran por método de imputação, com as classes sendo diferenciadas pela cor.

Uma leitura rápida mostra que a classe de imputações georreferenciadas se mostra melhor do que a classe MICE para todas as variáveis e cenários. Os valores da classe são próximos de 0,9 comparados com 0,5 (classe MICE para IDHM e PIB per capta) e 0,7 (classe MICE para o número de óbitos).

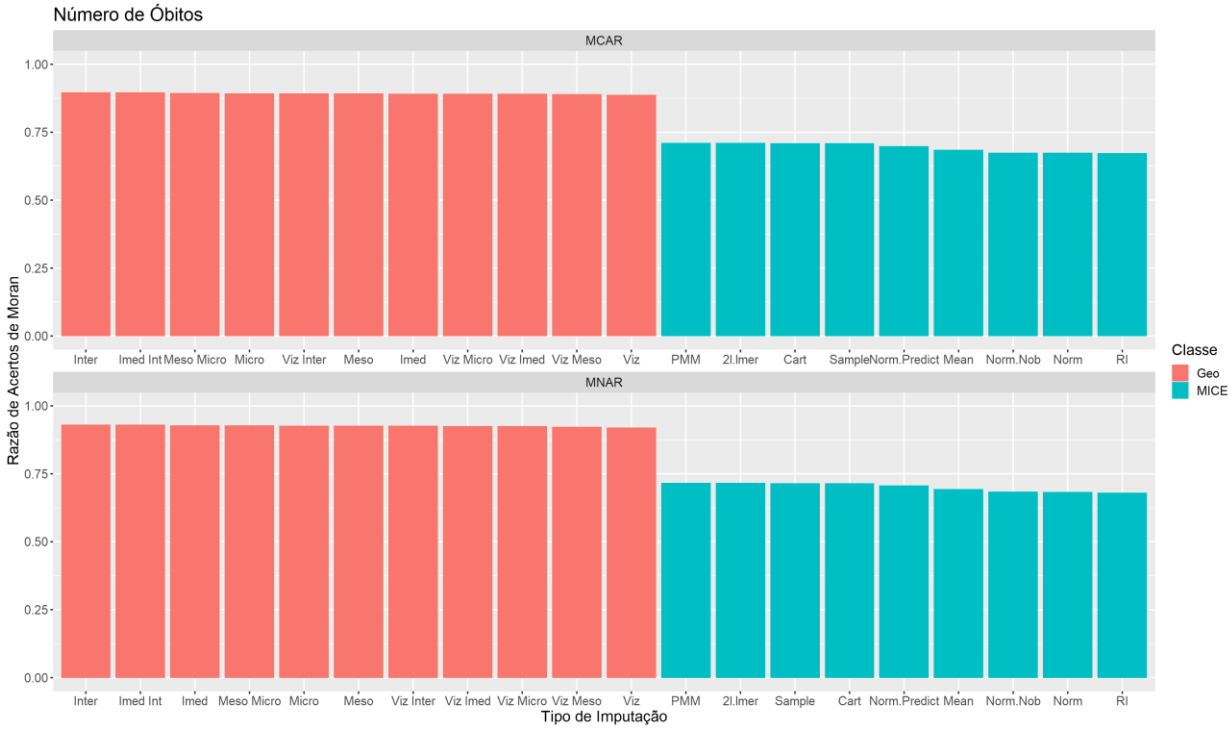
Figura 1 – Razão de Acertos de Moran por Tipo de Imputação para PIB per capita



Fonte: Elaborado pelos autores.

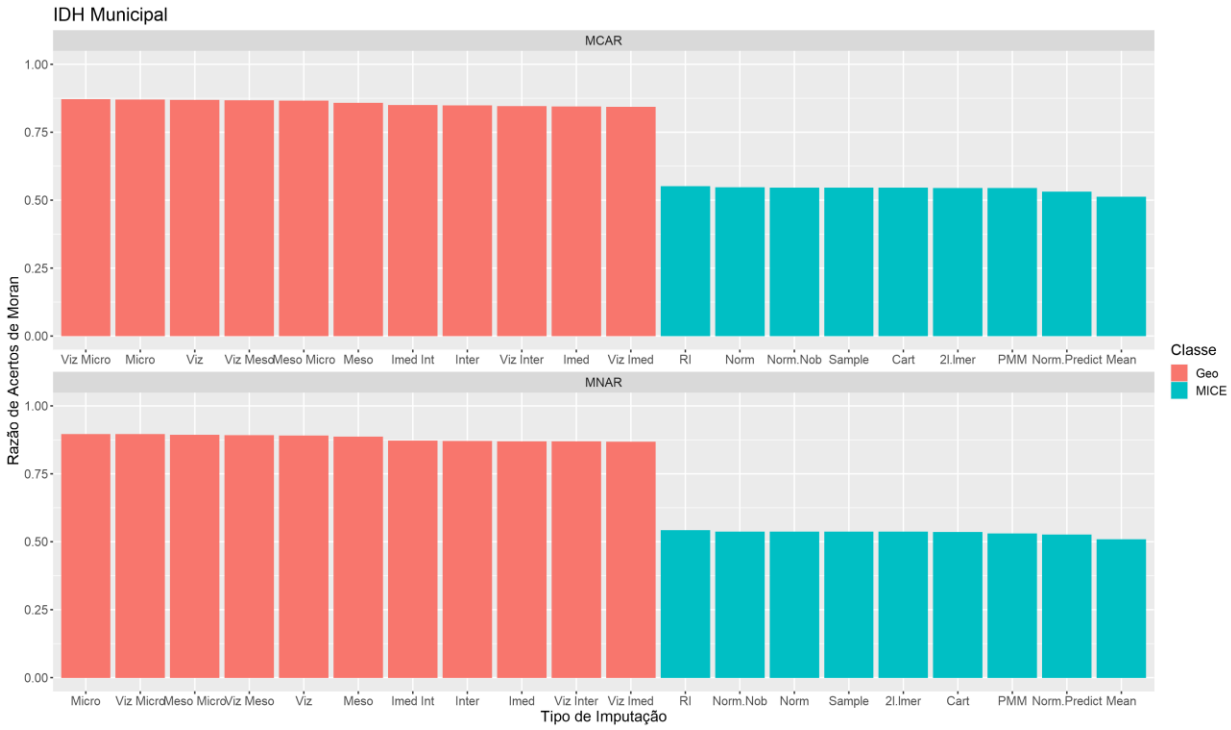
Uma análise mais atenta também nota que inexiste método ótimo, isto é, nenhum dos métodos é consistentemente o melhor, tanto de um modo geral quanto dentro de cada classe.

Figura 2 – Razão de Acertos de Moran por Tipo de Imputação para Número de Óbitos



Fonte: Elaborado pelos autores.

Figura 3 – Razão de Acertos de Moran por Tipo de Imputação para IDH Municipal



Fonte: Elaborado pelos autores.

É necessário apontar que, para a classe MICE, as funções dispunham do banco de dados com todas as variáveis de subdivisões geográficas, dispondo, portanto, de mais variáveis que qualquer dos métodos georreferenciados. Mesmo a imputação georreferenciada mais grosseira (em mesorregiões, com apenas 12 categorias) se mostra melhor que o melhor método MICE.

5. Considerações Finais

O objetivo deste trabalho, ao perceber que os cálculos das estatísticas de Moran obrigatoriamente precisam de dados completos, foi comparar diferentes métodos de imputação para se determinar qual seria o melhor para se possibilitar a análise espacial.

Usando dados censitários reais construímos um estudo de simulação que compara métodos de imputação de duas classes (georreferenciada e MICE) para diferentes mecanismos de não resposta.

Por simulações Monte Carlo mostramos que os métodos pertencentes à classe georreferenciada são mais precisos que os pertencentes à classe MICE, mesmo quando o mecanismo de não resposta é não ignorável e com os métodos MICE capazes de utilizar todas as informações geográficas presentes no banco de dados.

6. Referências Bibliográficas

Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical analysis, 27(2), 93-115.

Anselin, L., & Getis, A. (1992). Spatial statistical analysis and geographic information systems. The Annals of Regional Science, 26(1), 19-33.

Brandimarte, P. (2014). *Handbook in Monte Carlo Simulation – Applications in Financial Engineering, Risk Management, and Economics* (1a ed.). Hoboken: John Wiley & Sons.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Carpenter, J. R., & Kenward, M. G. (2014). Developments of methods and critique of ad hoc methods. In: Fitzmaurice, G. M., Kenward, M. G., Molenberghs, G., Verbeke, G., & Tsiatis, A. A. (eds.) *Handbook of Missing Data Methodology* (1a ed., páginas 23-46). Chapman and Hall/CRC Press.

Dempster, A. P., & Rubin, D. B. (1983). Overview. Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography.

Fitzmaurice, G. M., Kenward, M. G., Molenberghs, G., Verbeke, G., & Tsiatis, A. A. (2014). Missing data: Introduction and statistical preliminaries. In: Fitzmaurice, G. M., Kenward, M. G., Molenberghs, G., Verbeke, G., & Tsiatis, A. A. (eds.) *Handbook of Missing Data Methodology* (1a ed., páginas 3-22). Chapman and Hall/CRC Press.

Fortin, M. J., & Dale, M. R. (2009). Spatial autocorrelation. In: Fotheringham, A. S., & Rogerson, P. A. (eds.) *The SAGE handbook of spatial analysis* (1a ed., páginas 89-103). Sage.

Getis, A. (2008). A history of the concept of spatial autocorrelation: A geographer's perspective. Geographical analysis, 40(3), 297-309.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). (1990). Divisão regional do Brasil em mesorregiões e microrregiões geográficas.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). (2017). Divisão do Brasil em Regiões Geográficas Intermediárias e Imediatas.

Kenward, M. G. (2014). Multiple Imputation: Introduction. In: Fitzmaurice, G. M., Kenward, M. G., Molenberghs, G., Verbeke, G., & Tsiatis, A. A. (eds.) *Handbook of Missing Data Methodology* (1a ed., páginas 235-238). Chapman and Hall/CRC Press.

Lewis, P. A.W. & Orav, E. J. (1989) *Simulation Methodology for Statisticians, Operations Analysts and Engineers, Volume 1*. (1a ed.) Belmont: Wadsworth.

Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (3a ed.). John Wiley & Sons.

Negreiros, J. G., Painho, M. T., Aguilar, F. J., & Aguilar, M. A. (2010). A comprehensive framework for exploratory spatial data analysis: Moran location and variance scatterplots. International Journal of Digital Earth, 3(2), 157-186.

Pinto, D. G. C., Costa, M. A. C., & Marques, M. L. D. A. C. (2013). O Índice de Desenvolvimento Humano Municipal Brasileiro.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2020.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons.

Van Buuren, S. (2018). Flexible Imputation of Missing Data. CRC Press.

Van Buuren, V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. Journal of statistical software, 1-68.