ARTIGOS COMPLETOS **66**

18 a 22 de outubro de 2021
Anais do ENEPE
ISSN 1677-6321

Unoeste

ARTIGOS COMPLETOS

ANÁLISE DE CORRESPONDÊNCIA APLICADA A DADOS DE INTERNAÇÕES HOSPITALARES DO ESTADO DE SÃO PAULO..... 67

ANÁLISE DE CORRESPONDÊNCIA APLICADA A DADOS DE INTERNAÇÕES HOSPITALARES DO ESTADO DE SÃO PAULO

João Paulo de Oliveira Moraes

Universidade Estadual Paulista Júlio de Mesquita Filho – UNESP, Presidente Prudente, SP. E-mail: jp.mor@live.com

RESUMO

O Sistema Único de Saúde brasileiro é responsável por gerar um grande volume de dados, que ao serem estudados, permitem entender as demandas de saúde pública no país e assim supri-las. O Departamento de Informática do SUS é o responsável por comportar este grande volume via seus inúmeros sistemas de informação. Entre estes, o Sistema de Informações Hospitalares do SUS é o encarregado de gerenciar, armazenar e disseminar dados sobre todas as internações hospitalares realizadas no Brasil financiadas pelo SUS. O principal objetivo deste trabalho foi observar o comportamento das internações realizados no estado de São Paulo, realizadas entre janeiro e dezembro de 2019 através de uma abordagem multivariada, por meio da técnica de Análise de Correspondência. Esta técnica permite exibir a associação entre duas variáveis categóricas por meio de representações gráficas, mostrando-se uma importante ferramenta para o estudo de dados qualitativos. A base de dados em estudo foi tratada e organizada, permitindo a extração de uma amostra aleatória sistemática, onde se aplicou a técnica de Análise de Correspondência que entre os resultados apontou a relação entre doenças infeciosas e óbitos.

Palavras-chave: Sistema de Informações Hospitalares; Dados Públicos; Análise Multivariada; Análise de Correspondência.

CORRESPONDENCE ANALYSIS APPLIED IN HOSPITALIZATION DATA IN THE STATE OF SÃO PAULO.

ABSTRACT

The Brazilian Unified Health System generates a large amount of data, the study of which allows us to understand and meet the public health needs in the country. The Department of Informatics SUS is responsible for managing this large amount of data through its numerous information systems. The SUS Hospital Information System is responsible for managing, storing and disseminating data on all hospitalizations funded by SUS in Brazil. This study aims to observe the behavior of hospitalizations in the state of São Paulo, carried out between January and December 2019, through a multivariate approach using the Correspondence Analysis technique. This technique allows the representation of the association between two categorical variables through graphical representation moreover proves to be an important tool for the study of qualitative data. The database under study was prepared and organized so that a systematic random sample could be drawn, applying the Correspondence Analysis technique, which showed the association between infectious diseases and deaths, among others.

Keywords: Hospital Information System; Public Data; Multivariate Data Analysis; Correspondence Analysis.

1. INTRODUÇÃO

A demanda de saúde pública no país gera um enorme volume de dados diariamente. Nestes dados estão armazenadas as informações sobre toda ação ou qualquer procedimento realizado pelo Sistema Único de Saúde (SUS), sendo a chave principal para o desenvolvimento de pesquisas e tomada de decisões acerca de políticas públicas de saúde no território brasileiro.

Comportar esse volume de dados é uma tarefa concedida ao Departamento de Informática do SUS (DATASUS). O DATASUS busca promover a modernização através de tecnologias de informação no SUS e para que esta modernização se concretize é essencial que sistemas adequados estejam disponíveis para gerenciar seus dados

O Sistema de Informações Hospitalares do Sistema Único de Saúde (SIH-SUS) é o sistema de informação responsável por armazenar, gerenciar e disseminar dados administrativos de saúde, onde seu foco principal, destina-se a realizar pagamentos aos hospitais públicos e privados conveniados ao SUS, considerando os procedimentos prestados no regime de internação. A Autorização de Internação Hospitalar (AIH) é o documento que permite identificar e extrair tais características, pois engloba informações relacionadas ao custo, permanência e serviços oferecidos a cada internação (PEPE, 2009).

Neste trabalho, busca-se se observar as características das internações realizadas entre janeiro e dezembro de 2019 no estado de São Paulo. Considerando o grande número de internações realizadas neste período, uma amostra aleatória sistemática foi obtida da base de dados original e a técnica multivariada de análise de correspondência foi utilizada para analisar o comportamento das internações.

2. METODOLOGIA

O DATASUS disponibiliza publicamente os dados de todas as internações hospitalares financiadas pelo Sistema Único de Saúde. É possível obter arquivos referentes as autorizações aprovadas, reprovadas e serviços profissionais prestados nos âmbitos das internações hospitalares, de forma mensal para cada unidade federativa do Brasil.

No presente trabalho, optou-se por utilizar a base de dados das autorizações aprovadas, no estado de São Paulo, entre janeiro e dezembro de 2019, considerando a data de processamento da AIH no sistema. Com os arquivos em formato DBC (Database Container) obtidos do repositório, importados no software R através do pacote read.dbc, criou-se uma base de dados consolidada dos 12 arquivos referentes aos meses do período em questão, totalizando 2.606.084 observações e 113 variáveis, sendo 79 qualitativas e 34 quantitativas.

2.1. Tratamento e organização dos dados

Atentou-se em verificar incoerências apontadas na literatura sobre o SIH-SUS, como a existência de sub-registro e a falta de rigor no preenchimento de campos não relacionados ao faturamento (PEPE, 2009). Variáveis completamente zeradas ou com grande percentual das observações faltantes, variáveis de identificação das unidades prestadoras dos serviços de hospitalização, relacionadas a grupos específicos e variáveis redundantes, como código de identificação e similares foram excluídas da base de dados.

Campos como idade e dias de permanência em internação foram descartados, já que outras variáveis existentes na base permitiam obter estas observações por meio de cálculos entre datas, evitando possíveis inconsistências nas informações obtidas (ANTERO, 2015).

Um total de 4.276 AIHs apresentam custo total de R$ 0,00. De modo que não há sentido em considerar estas observações, todas foram excluídas. De modo similar, as autorizações pertencentes a um mesmo usuário foram sintetizadas, subtraindo mais de 51.000 linhas da base de dados.

O agrupamento de outros campos permitiu gerar novas variáveis sintetizadas, enquanto variáveis quantitativas continuas foram convertidas para categóricas, visando a aplicação da Análise de Correspondência. Logo, as variáveis da base de dados utilizada no presente estudo são apresentadas a seguir.

- **Desfecho da internação**: variável categórica, representa o desfecho da internação, alta, permanência/transferência ou óbito;
- **Diagnóstico**: variável categórica, representa o diagnóstico principal da patologia que levou a internação, agrupadas de acordo com os capítulos do CID-10;
- **Idade**: variável categórica, representa a idade do paciente categorizada através das faixas etárias, 0 a 15 anos, 16 a 30 anos, 31 a 45 anos, 46 a 65 anos, mais de 66 anos;
- **Mesorregião**: variável categórica, representa o local de internação categorizado de acordo com as 15 mesorregiões do estado de São Paulo;
- **Região**: variável categórica, representa o local de internação categorizado em duas classes, interior do estado e região metropolitana/litoral;
- **Sexo**: variável categórica, representa o sexo do paciente, masculino ou feminino;
- **Tempo em internação**: variável categórica, representa número de dias em hospitalização categorizada através das seguintes faixas, até uma semana, de 8 a 15 dias, 16 dias ou mais;

- **Custo total da Internação**: variável categórica, que representa a soma dos custos hospitalares categorizada de acordo com o salário mínimo vigente no ano de 2019, R$ 998,00. Custo baixo (valores até meio salário mínimo R$ 499,00), custo médio (entre meio salário mínimo e R$ 998,00), custo alto (entre um salário mínimo e R$ 1996,00) e custo elevado (mais de R$ 1996,00).

2.2. Determinação da Amostra

Uma amostra aleatória sistemática foi extraída da base da dados. Como descrito em Cochran (1977), esta técnica garante que cada unidade tenha igual probabilidade de seleção para a amostra.

Suponha que os N indivíduos da população sejam numeradas de 1 a N em determinada ordem. Suponha ainda que N possa ser descrito como um produto de dois inteiros n e k, de modo que $k = N = n$. Para determinar uma amostra de tamanho n é necessário seguir os seguintes passos:

1. selecionar um número i, contido entre 1 e k, aleatoriamente;
2. selecionar cada $k - ésimo$ elemento após i de modo que a amostra contenha os elementos, $i;\ i + k;\ i\ + 2k; \ldots;\ i\ (n - 1)k$.

Assim, o primeiro elemento i é obtido aleatoriamente e os elementos restantes são selecionados sistematicamente através de k intervalos.

Para o presente trabalho, optou-se por extrair uma amostra de 10% de cada uma das 15 mesorregiões do estado de São Paulo, totalizando ao fim $n\ = 255082$ indivíduos selecionados.

2.3. Análise de Correspondência

A análise de correspondência, também chamada de análise de correspondência simples, é uma técnica exploratória desenvolvida por Benzécri (1976). O método pode ser considerado como uma forma de exibir o padrão de associação entre variáveis categóricas graficamente, sendo uma valiosa ferramenta para dados qualitativos que normalmente não são o foco de grande parte das técnicas multivariadas.

Esta técnica tem a capacidade de reduzir a dimensão de uma matriz de dados não negativos e visualizá-la em um subespaço de menor dimensionalidade. Os dados de interesse são dispostos em uma tabela de contingência, onde as entradas das I linhas e J colunas representam categorias de duas variáveis qualitativas e os valores nas células da tabela são as respectivas frequências. Desta maneira, pode-se explorar a relação entre variáveis categóricas por meio de uma representação bidimensional, onde categorias são representadas pelas coordenadas dos pontos traçados e uma medida chamada de inércia determina a variação total de uma tabela de contingência, assim como sua decomposição determina a inércia associada a cada dimensão. Considera-se um resultado significativo aquele onde duas dimensões (componentes) conseguem explicar juntas mais de 70% da variabilidade total.

O desenvolvimento algébrico e aplicações da técnica podem ser encontrados com mais detalhes na literatura sobre análise multivariada (JOHNSON; WICHERN, 2007; MINGOTI, 2005; RENCHER, 2002; GREENACRE, 1984).

Todas as análises aqui apresentadas foram obtidas por meio do software R, com o auxílio dos pacotes FactoMineR e factoextra.

3. RESULTADOS E DISCUSSÃO

A amostra foi composta de 255082 observações no período de janeiro a dezembro de 2019, onde 56,8% pertencem ao sexo feminino e 43,2% ao masculino. A Tabela 1 apresenta os resultados descritivos para as variáveis na amostra

Tabela 1. Caracterização das internações do estado de São Paulo em 2019.

Variável (Categorias)	N= 255082 (%)
Sexo	
Masculino	110231 (43,2%)
Feminino	144851 (56,8%)
Faixa etária	
De 0 a 14 anos	34333 (13,5%)
De 15 a 30 anos	56748 (22,2%)
De 31 a 65 anos	111047 (43,5%)
Mais de 66 anos	52954 (20,8%)
Região	
Interior	132792 (52,1%)
Metropolitana e Litoral	122290 (47,9%)
Permanência	
Até 7 dias	211104 (82,8%)
De 8 a 15 dias	27043 (10,6%)
De 16 dias ou mais	16935 (6,6%)
Tempo em internação	
Alta	223962 (87,8%)
Permanência/Trans.	18366 (7,2%)
Óbitos	12754 (5,0%)
Diagnóstico (Cap.CID-10)	
Infecciosas	11316 (4,4%)
Tumores	20833 (8,2%)
Sangue	2446 (1,0%)
Endócrinas	4490 (1,8%)
Mentais	7175 (2,8%)
Sistema nervoso	5011 (2,0%)
Olhos	5037 (2,0%)
Ouvido	591 (0,2%)
Aparelho circulatório	27497 (10,8%)
Aparelho respiratório	22870 (9,0%)
Diagnóstico (Cap.CID-10)	
Aparelho digestivo	27811 (10,9%)
Pele	5821 (2,3%)
Osteomuscular	5022 (2,0%)
Aparelho geniturinário	19664 (7,7%)
Gravidez	44635(17,5%)
Perinatal	6345 (2,5%)
Malformações	2563 (1,0%)
Achados anormais	4591 (1,8%)
Causas externas	24562 (9,6%)
Fatores/Contatos	6802 (2,7%)
Custo total da internação	
Baixo	99152 (38,9%)
Médio	92911 (36,4%)
Alto	27914 (10,9%)
Elevado	35105 (13,8%)

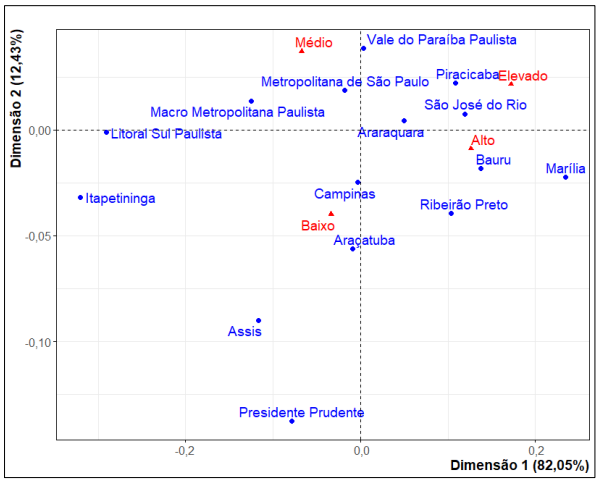
Fonte: Elaborado pelo autor.

Para investigar as características das internações hospitalares, foi aplicado a técnica de análise de correspondência, explorando as características de duas variáveis dispostas em uma tabela de contingência.

O primeiro par de variáveis analisados foram custo total e mesorregião. A inércia total obtida em cada uma das 3 dimensões ($v = min(I - 1; J - 1) = 3$) foi igual a $\varphi^2 = 0{,}008 + 0{,}001 + 0{,}001 = 0{,}0100$. A primeira dimensão explica 82,05% da inércia total, a segunda 12,42%. Logo, a representação bidimensional das duas primeiras dimensões é adequada, já que juntas explicam 94,47% da variação total dos dados.

A representação gráfica deste cruzamento é mostrada na Figura 1. As mesorregiões de Araçatuba, Ribeirão Preto e Campinas estão próximas do ponto que representa a faixa de custo baixo, indicando que grande parte das suas internações tem valores totais dentro desta faixa. Vale do Paraíba Paulista, Metropolitana de São Paulo e Macro Metropolitana Paulista correspondem mais ao custo médio. Bauru, Marília, Araraquara, São José do Rio Preto e Piracicaba são as regiões mais próximas das faixas de custo alto e elevado. A mesorregião de Prudente está mais próxima da faixa de custo mais baixo do que das outras, o mesmo ocorre com Itapetininga. O Litoral Sul Paulista de certo modo está em um ponto intermediário entre as categorias baixo e médio, indicando que os custos para esta mesorregião variam entre estes intervalos.

Figura 1. Mapa de correspondência: Custo e Mesorregião.

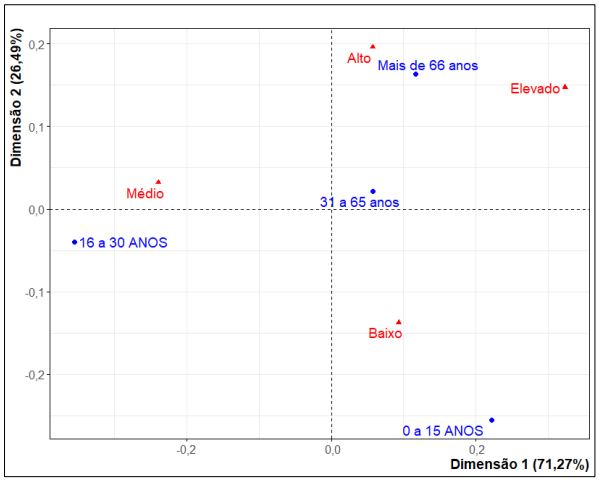


Fonte: Elaborado pelo próprio autor.

A próxima característica investigada, juntamente com o custo de internação foi a variável idade. A inércia total neste caso é igual a $\varphi^2 = 0,0651$. A primeira componente explica 71,27% da inércia total e a segunda 26,49%. Logo, as duas primeiras dimensões conseguem explicar juntas mais de 97,76% da inércia.

Pela Figura 2, observa-se claramente que as faixas etárias de indivíduos mais velhos tendem a corresponder a categorias de maiores gastos com internação. A faixa etária de indivíduos com mais de 66 anos está próxima da categoria de custo alto, assim como a faixa etária entre 46 a 65 anos está mais próxima da categoria elevado. Indivíduos com idade entre 16 e 30 anos, correspondem a gastos da modalidade médio e os mais jovens (0 a 15 anos) a de valores mais baixos

Figura 2. Mapa de correspondência: Custo e Idade.

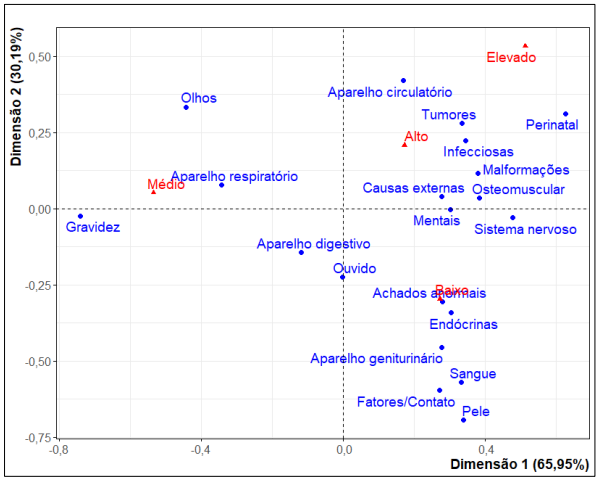


Fonte: Elaborado pelo próprio autor.

O diagnóstico principal da causa de internação, está categorizado de acordo com os 22 capítulos do CID-10. Aplicando uma análise conjunta com a variável custo total da internação, obteve-se uma inércia total de $\varphi^2 = 0{,}2611$, onde 96,14% desta está contida nas duas primeiras dimensões.

Da representação bidimensional, mostrada na Figura 3, nota-se que de certo modo quatro grupos de diagnósticos se formam em relação ao custo total. As causas de internação mais correspondentes a categoria de custo baixo são as doenças do aparelho digestivo, ouvido, achados anormais em exame, doenças endócrinas, doenças do aparelho geniturinário, doenças do sangue, fatores e contatos que influenciam o estado de saúde e doenças de pele. Em relação ao custo médio, as categorias de diagnóstico mais próximos são respectivamente, gravidez, doença do aparelho respiratório e doença dos olhos. As de custo alto, tumores, doenças infecciosas, malformações, causas externas, doenças osteomusculares, transtornos mentais e doenças do sistema nervoso e um pouco distante as doenças do aparelho circulatório. Por fim, custo elevado se associa as afecções perinatais.

Figura 3. Mapa de correspondência: Custo e Diagnóstico Principal.



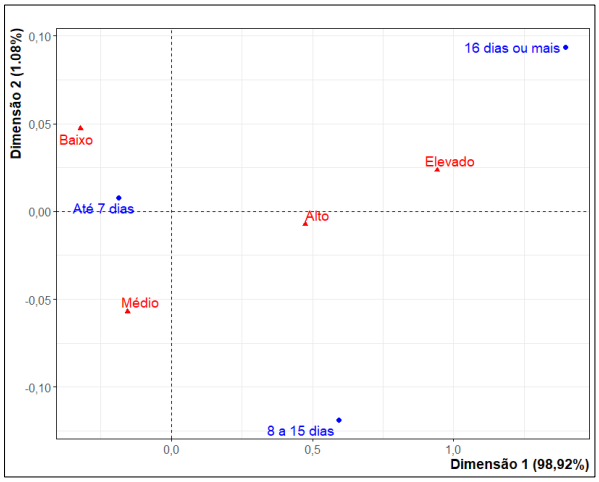
Fonte: Elaborado pelo próprio autor.

Considerando a variável tempo em internação e a variável custo, obteve-se a inércia total igual a $\varphi^2 = 0{,}198$. onde a primeira dimensão explica 92,92% e a segunda 1,08%.

A representação desta análise, apresentada na Figura 4 não apresentou nenhuma proximidade acentuada entre as categorias no plano, mas é possível observar algumas características. Nas hospitalizações que duraram até 7 dias os custos totais permaneceram nas faixas baixo e médio. As que duraram entre 8 e 15 dias nas faixas de médio e alta, e de 16 dias ou mais os custos foram mais elevados.

De forma generalizada, observa-se que, o tempo de internação está diretamente ligado ao custo da internação, isto é, quanto mais tempo um paciente fica hospitalizado, maior tende a ser o custo total da sua internação

Figura 4. Mapa de correspondência: Custo e Tempo em Internação.

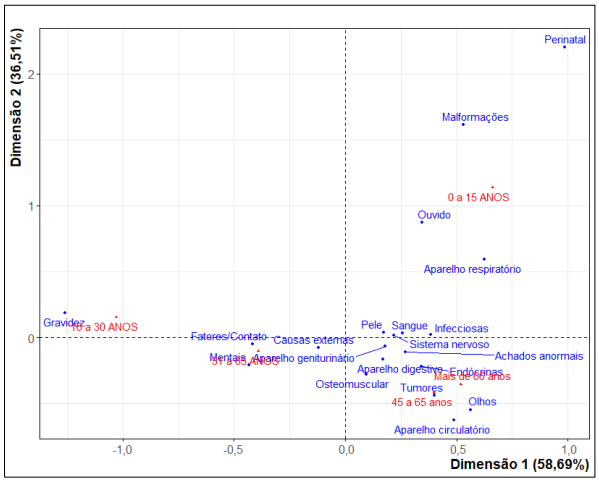


Fonte: Elaborado pelo próprio autor.

A análise de correspondência entre as variáveis diagnóstico e idade, apresentou inercia total de $\varphi^2 = 0{,}707$. As duas primeiras dimensões explicam juntas, 95,20% da variação total.

Na Figura 5, nota-se que grande parte das causas principais se concentram próximas a categoria de mais de 66 anos. Pessoas na faixa etária de 45 a 65 tendem a ser acometidas pôr tumores, doenças dos olhos e do aparelho circulatório. Nota-se que transtornos mentais, fatores que influenciam o estado de saúde ou contato com os serviços de saúde e causas externas acometem pessoas com idade entre 31 e 65 anos. Pessoas entre 0 e 15 anos tendem a ser hospitalizadas por doenças de ouvido, aparelho respiratório, malformações congênitas e afecções do período perinatal. Por fim, a faixa etária entre 16 e 30 anos, corresponde a gravidez

Figura 5. Mapa de correspondência: Idade e Diagnóstico Principal.



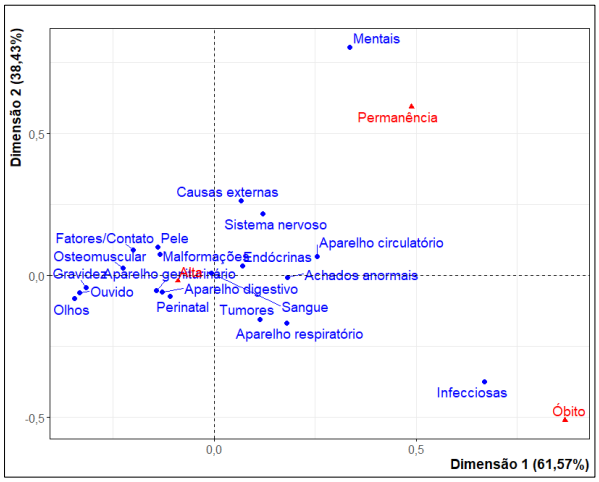
Fonte: Elaborado pelo próprio autor.

Considerando os campos de diagnostico principal e do desfecho da internação, a inércia total é de $\varphi^2 = 0{,}1007$, explicada em 61,67% pela primeira dimensão e 38,43% pela segunda.

Observando a representação das duas primeiras dimensões na Figura 6, nota-se que grande parte dos diagnósticos se encontram próximos ao desfecho alta. De modo geral, pacientes com estes tipos de comorbidades acabam recebendo alta após um período de internação. Em relação aos transtornos mentais, é possível observar a proximidade com a categoria permanência, indicando que pacientes acometidos com

esta causa tendem a permanecer hospitalizados por períodos mais longos. Já as doenças infecciosas apresentam como desfecho da internação o óbito.

Figura 6. Mapa de correspondência: Desfecho da internação e Diagnóstico Principal.



Fonte: Elaborado pelo próprio autor.

5. CONSIDERAÇÕES FINAIS

A análise de correspondência permitiu através das suas representações gráficas, mostrar pontos interessantes sobre a dinâmica do Sistema de Informações Hospitalares e das internações no ano de 2019.

Através dela, pode-se observar as relações entre variáveis desfecho de internação e diagnostico principal, onde apontou-se que doenças infecciosas de modo geral implicaram em óbitos dos usuários hospitalizados.

Outro ponto importante a se destacar, está na relação entre custo e idade dos usuários. Já que a representação gráfica apontou que pacientes idosos tendem a estar associados com custos mais elevados de internação comparado a pacientes de faixa etárias mais jovens.

Apesar das carências apresentadas, dado a ausência de variáveis que auxiliariam no levantamento de outras informações relevantes, o Sistema de Informações Hospitalares se mostrou importante quando se trata de mostrar a importância do Sistema Único de Saúde para a sociedade. O número de internações realizadas no período e o custo expedido nas mesmas são pontos fortes que comprovam esta ideia.

Deste modo, o estudo de dados públicos de saúde, sejam oriundos do SIH ou de um dos inúmeros sistemas do DATASUS deve ser difundido. Muitas informações estão presentes nestes sistemas, destaca-las auxiliam no processo de conhecimento dos mecanismos e transmissão das informações para a população, além de servir como uma forma de valorização do SUS.

REFERÊNCIAS

ANTERO, V. H. d. M. **Análise de Dados Públicos de Saúde com Mapas Auto-Organizáveis de Kohonen (Som)**. 61 p. Monografia (Trabalho de Graduação) — Universidade Federal de Pernambuco, Centro de Informática, Graduação em Ciência da Computação, Recife, PE, 2015.

COCHRAN, W. G. **Sampling Techniques**. 3. ed. New York: John Wiley, 1977.

GREENACRE, M. J. **Theory and Applications of Correspondence Analysis**. Great Britain: Academic Press, 1984.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6. ed. New Jersey: Pearson Prentice Hall, 2007.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte, MG: Editora UFMG, 2005.

PEPE, V. E. **Sistema de Informações Hospitalares do Sistema único de Saúde (SIH-SUS)**. In:. Ministério da Saúde Brasília, 2009. v. 2. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/experiencia_brasileira_sistemas_saude_volume2.pdf. Acesso em: 13 mar 2019.

RENCHER, A. C. **Methods of Multivariate Analysis**. 2nd ed. ed. New York: J. Wiley, 2002. (Wiley Series in Probability and Mathematical Statistics). https://doi.org/10.1002/0471271357