UNIVERSIDADE ESTADUAL DE MARINGÁ
PROGRAMA DE INICIAÇÃO CIENTÍFICA - PIC
DEPARTAMENTO DE ESTATÍSTICA
ORIENTADORA: Profª. Eniuce Menezes de Souza
ACADÊMICOS: Arthur Cesar de Moura Rocha, Paula Mitiko Heller

OBTENÇÃO, GERENCIAMENTO E VISUALIZAÇÃO DE SÉRIES TEMPORAIS DO DATASUS

Maringá, 31 de julho de 2018

UNIVERSIDADE ESTADUAL DE MARINGÁ
PROGRAMA DE INICIAÇÃO CIENTÍFICA - PIC
DEPARTAMENTO DE ESTATÍSTICA
ORIENTADORA: Profª. Eniuce Menezes de Souza
ACADÊMICOS: Arthur Cesar de Moura Rocha, Paula Mitiko Heller

OBTENÇÃO, GERENCIAMENTO E VISUALIZAÇÃO DE SÉRIES TEMPORAIS DO DATASUS

Relatório contendo os resultados finais do projeto de iniciação científica vinculado ao Programa PIC-UEM.

Maringá, 31 de julho de 2018

RESUMO

A base de dados do DATASUS é bastante rica e importante para o suporte de inúmeras pesquisas na área de saúde. Em geral, tais dados são obtidos online através do aplicativo TabWin, o qual é muito dependente do usuário, dispendioso e restrito na seleção das variáveis. Assim, este projeto se propõe a buscar métodos computacionais e estatísticos que facilitem a manipulação desses dados a fim de auxiliar pesquisas da área de saúde.

Objetiva-se desenvolver e implementar um algoritmo para facilitar a obtenção de dados do DATASUS ao longo do tempo, geração de séries temporais, além da análise descritiva e visualização de tais séries, bem como de estatísticas descritivas em plataforma interativa em formato html. Em especial, serão analisados os dados de internações por asma disponível na base de dados do DATASUS a fim de exemplificar a utilização da aplicação.

SUMÁRIO

1 INTRODUÇÃO

O Departamento de Informática do Sistema Único de Saúde (DATASUS) é um órgão do governo que tem como competência "prover os órgãos do SUS de sistemas de informação e suporte de informática, necessários ao processo de planejamento, operação e controle" (DATASUS, 2017). Dessa forma, uma de suas missões é a manutenção da base de dados necessária para o funcionamento integrado dos órgãos do SUS (DATASUS, 2017). Para o cumprimento dessa missão é gerado um banco de dados rico em informações sobre diversas condições de saúde da população brasileira, o qual está disponível para acesso ao público por meio online. Por causa de sua riqueza de dados e por ser uma fonte de coleta de dados confiável, dado que se trata de um órgão governamental, os dados disponibilizados pelo DATASUS são de fundamental importância para as pesquisas da área de saúde, tanto diretamente quanto indiretamente.

Por outro lado a dificuldade de manipulação de bancos de dados com grande volume de variáveis é uma reclamação frequente entre pesquisadores de diversas áreas do conhecimento. Essa limitação é visível em pesquisas na área da saúde como quando dito que

> "A consulta aos dados disponibilizada, diretamente, no endereço eletrônico do DATASUS, tem algumas limitações. A principal é que se encontra separada por períodos determinados, sem a possibilidade de agregar diferentes períodos para o estudo de séries temporais. Além disso, as tabelas não permitem estratificação para a análise de características adicionais, como o tipo de hospital (psiquiátrico ou geral), ou por outros grupos de idade, por exemplo" (CANDIAGO, 2007).

A partir dessa dificuldade, este projeto se propõe a buscar métodos computacionais e estatísticos que facilitem a manipulação desses dados a fim de auxiliar pesquisas da área de saúde. Para isso serão utilizados softwares como SAS e R e a base de dados do DATASUS.

Após realizada a proposta, com o intuito de verificar o cumprimento do objetivo, será realizada uma aplicação do método para um estudo descritivo sobre a asma tendo em vista que a asma é um problema de saúde pública que, conforme a Pesquisa

Nacional de Saúde (PNS) do Ministério da Saúde e Instituto Brasileiro de Geografia e Estatística (IBGE), atinge 6,4 milhões de brasileiros acima de 18 anos (BRASIL, 2015).

A Organização Mundial de Saúde (OMS) estima que existam cerca de 300 milhões de casos de asma pelo mundo. No Brasil, a asma é responsável pela internação de mais de 100 mil pessoas no SUS e segundo o *Boletim Epidemiológico* n. 18, v. 47, do Ministério da Saúde, Brasil (2016), a asma foi causa de 38% das internações por doenças respiratórias crônicas (DRC) no Brasil de 2003 a 2013. Um estudo importante efetuado em várias grandes cidades do Brasil, o ISAAC (*International Study of Asthma and Allergy in Childhood*), apontou que entre 20% a 30% das crianças e adolescentes apresentam sintomas indicativos da doença (PORTARIA 054-R, 2009).

A asma é uma doença inflamatória pulmonar crônica definida pela hiperreatividade das vias aéreas inferiores e limitação do fluxo de ar (SBPT, 2002). Mesmo sendo de natureza crônica, esta patologia pode ser controlada, mas a falta de um controle apropriado pode ocasionar um custo elevado em gastos com internações, medicamentos e serviços emergenciais, além de poder privar o indivíduo com asma de suas ações rotineiras, tais como emprego e estudos (BARNES, JONSSON e KLIM, 1996).

2 JUSTIFICATIVA

Conforme mencionado na introdução, a base de dados do DATASUS é bastante rica e importante para o suporte de inúmeras pesquisas na área de saúde. Em geral, tais dados são obtidos online através do aplicativo TabWin o qual é muito dependente do usuário, dispendioso e restrito na seleção das variáveis. Neste aplicativo, o usuário pode selecionar duas variáveis, as quais serão armazenadas (uma nas linhas e outra nas colunas) em uma planilha de dados. Se o usuário quiser uma terceira variável, tal como o tempo, em meses ou anos, por exemplo, terá que repetir tal procedimento inúmeras vezes dependendo do tamanho de sua série temporal. Logo, faz-se clara a necessidade de facilitar de tal procedimento.

Outra justificativa importante é que os dados pelo TabWin são gerados de

modo agregado, mas se os arquivos brutos forem baixados via FTP, existe a possibilidade de acesso das informações de saúde por indivíduo, o qual é identificado por um código próprio de modo a preservar sua anonimidade. Embora na análise de séries temporais tais dados sejam em geral utilizados no formato agregado, a obtenção de dados não agregados é muito útil para inúmeras análises e modelagens estatísticas.

3 OBJETIVOS

A partir disso, objetiva-se neste projeto desenvolver e implementar um algoritmo para facilitar a obtenção de dados do DATASUS ao longo do tempo, geração de séries temporais, além da análise descritiva e visualização de tais séries em plataforma interativa em formato html. Especificamente, pode-se citar os seguintes objetivos específicos:

- Realizar o *download* de arquivos brutos completos via FTP e conversão para leitura em R e SAS;
- Implementar filtros para obtenção das variáveis considerando as mudanças de codificação ocorridas ao longo do tempo no registro de dados do DATASUS tais como o padrão CID9, CID10, mudança de nomes e/ou códigos de regionais de saúde, municípios, etc;
- Desenvolver uma interface html a para visualização de estatísticas descritivas e séries temporais do DATASUS, em especial, de internações por asma.

4 METODOLOGIA

A metodologia para o desenvolvimento deste projeto foi baseada na utilização de recursos computacionais mais avançados, os quais envolvem uma extensão dos conceitos aprendidos durante o Curso de Graduação em Estatística. Neste sentido, com o auxílio do software R, foi desenvolvido um programa o qual realiza o download de forma automatizada dos bancos disponibilizados pelo DATASUS via FTP. Esses dados foram obtidos em formato dbc, os quais são formas comprimidas de bancos de dados que ocupam menor espaço de memória, porém essa estrutura comprimida causa

complicações em seu uso, o que levou à conversão destes para dbf por meio do pacote "read.dbc", a qual é uma forma muito utilizada de arquivo na área de gerenciamento de dados. Este formato é lido diretamente no SAS e também no R a partir do pacote "foreign".

As divisões das doenças de acordo com o CID9 e CID10 possuem uma estrutura de ramificação das categorias nas diversas formas de uma condição. Com isso, uma única doença que possui um código geral, pode possuir diversos códigos específicos que respondem a esse código geral. Dentro da proposta de aplicação, os códigos para a doença "asma" são 493 para o CID9 e J45 para o CID10. Porém estes são ramificados em cinco categorias, cada uma com no mínimo duas subdivisões no CID9 e quatro categorias, cada com uma ou mais subdivisões no CID10. Dessa forma, a compilação destes em um único conjunto de dados sobre "asma" é essencial para a análise. Esse trabalho foi feito com o auxílio do pacote do R "sqldf", que possibilitou o uso de funções e conceitos de Linguagem de Consulta Estruturada (*Structured Query Language* - SQL), para o agrupamento dos dados em um único conjunto macro.

A quantidade de variáveis disponíveis para análise muda com grande frequência ao longo dos anos, variando de 19 até 32 variáveis diferentes sobre uma única observação. Dessa forma, existe um grande volume de informações que não são adequadas para a realização de algumas análises, como, por exemplo, séries temporais, por conta da sua falta de continuidade. A partir disso, para a aplicação deste trabalho, optou-se por utilizar as variáveis idade, sexo, cidade de internação e cidade de residência como covariáveis para as ocorrências de asma no estado do Paraná. As três primeiras são variáveis que têm presença constante juntamente com o motivo da internação e permitem a realização de análises descritivas básicas. Cabe a observação de que a última variável não possui o mesmo comportamento, uma vez que esta possui diversos períodos entre os anos de 1994 e 1997 onde não se foram coletados esses dados. Porém, é de grande importância o conhecimento da região de origem do paciente internado, a fim de se identificar a real origem geográfica e possível foco de recorrência de uma doença. A partir disso, as informações relativas à cidade de residência do internado são utilizadas para análise a partir do ano de 1997. Além disso, é relevante frisar que a variável idade é codificada em sua forma original, sendo necessário um esforço a mais para classificá-la em uma forma mais usual, no caso

considerando as categorias como infância, jovem, adulto e idoso. É importante observar que a partir do conhecimento dos códigos utilizados pelo DATASUS para as variáveis, é possível selecioná-las com certa facilidade.

Com a utilização dos pacotes "ggplot2", "leaflet" e "dygraphs" do R, foram gerados gráficos descritivos interativos das variáveis selecionadas. Por meio do Shiny, uma ferramenta para desenvolvimento de aplicações web, foi possível a criação de um aplicativo que permite ao usuário o acesso a uma interface que apresenta três abas: na primeira, são apresentadas as informações de frequência de internações por asma por local de residência através de um mapa interativo, além disso, existe um filtro que permite ao usuário a definição de um ano de interesse; na segunda aba está apresentada a série temporal que informa a quantidade de internados conforme o passar do tempo, que pode ser apresentada levando em consideração a proporção de internados de cada sexo ou categoria de idade por meio de um filtro disponibilizado; e na terceira são apresentados dois gráficos de análise descritiva em que aparecem as proporções dos sexos e a quantidade de internados por categoria de idade, podendo ser para o estado do Paraná por inteiro ou levando em conta apenas uma cidade, sendo essa cidade a residência do interno, a qual é definida por meio de um filtro.

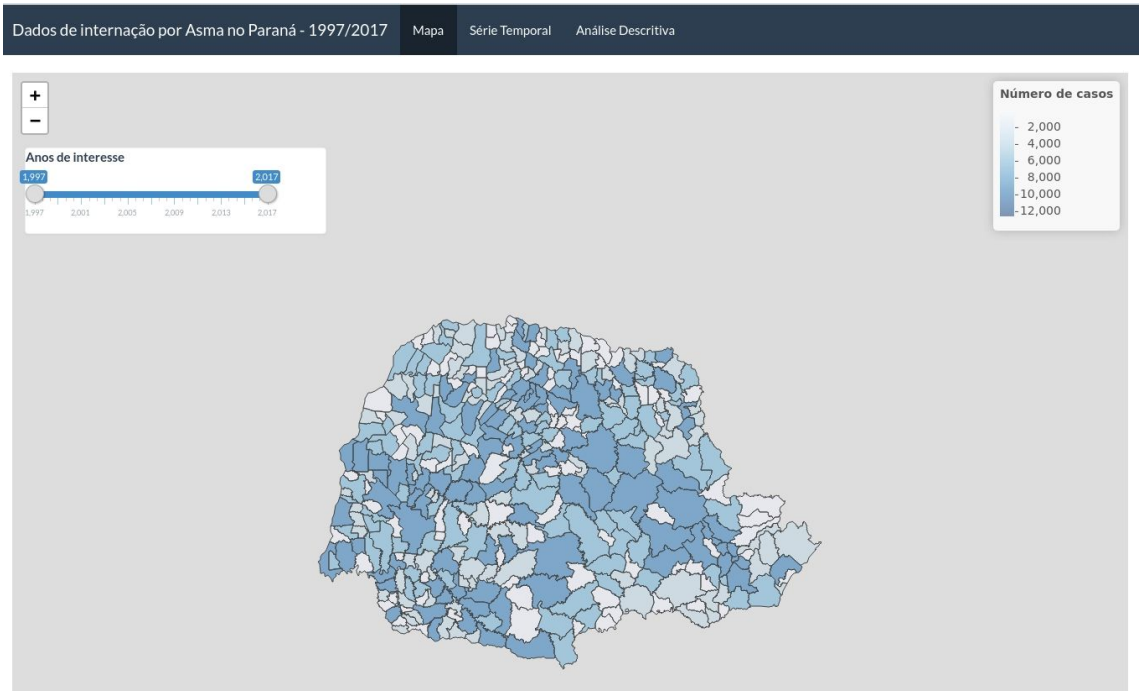
Para que fosse possível o acesso remoto à essa aplicação foi utilizado o Shiny Server, o qual é um servidor desenvolvido especificamente para aplicativos do Shiny. Dessa forma, através de um endereço html, o usuário pode acessar e interagir com os gráficos sem ter de processar todo o programa desenvolvido, uma vez que todas as análises e criação dos gráficos e mapas são feitos no servidor. Por fim, a utilização do aplicativo será apresentada no próximo tópico por meio de uma análise descritiva dos dados.

4 RESULTADO E DISCUSSÃO

Como resultado final do trabalho desenvolvido, foram gerados mapas, gráficos descritivos e séries temporais de dados referentes a internações causadas por asma no estado do Paraná desde o ano de 1997 até o final do ano de 2017. Esses mapas e gráficos podem ser acessados através de um endereço html.

A primeira imagem disponibilizada pelo aplicativo é um mapa que apresenta a frequência de internações por região, o qual pode ser visto na Figura 1.

Figura 1 - Mapa de internações



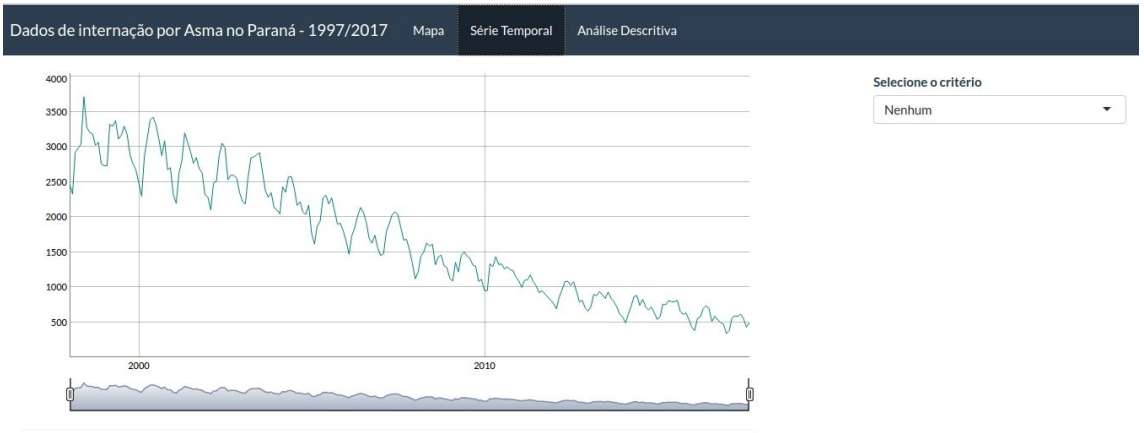
Conforme apresentado na legenda da Figura 1, as áreas em tons mais escuros de azul são aquelas que possuem maior quantidade de registros de internações por causa de asma. Em contraponto, as regiões com tons mais claros de azul, se aproximando do branco, são aquelas com menor frequência. Observa-se na Figura 1 a flexibilidade que o usuário tem para escolher um ou mais anos específicos, bem como períodos com possibilidade de rápida escolha do ano inicial, final e tamanho do intervalo de tempo.

Ao analisar a distribuição das tonalidades de azul, é possível verificar que as ocorrências são bem distribuídas pelo estado, uma vez que, apenas a região próxima à capital do estado, a qual possui grande volume de internações em sua região metropolitana, se apresenta uma sequência maior regiões com pequena quantidade de

internações, o que pode indicar a migração das internações das cidades da região para a capital.

Na segunda aba do aplicativo html, é possível encontrar o gráfico da série temporal da quantidade de internações por asma no estado do Paraná (Figura 2).

Figura 2 - Série temporal de internações



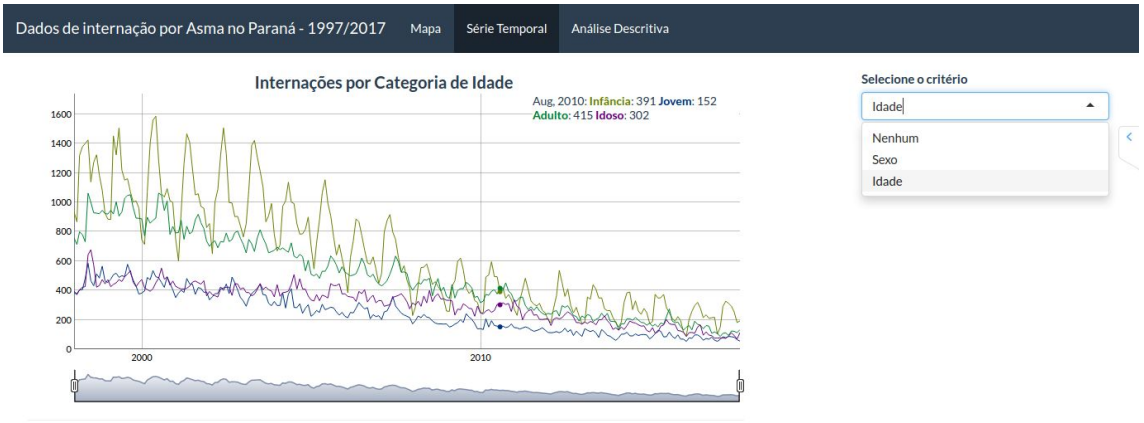
Conforme apresentado na Figura 2, é possível afirmar que a quantidade de internações por conta desta doença vem reduzindo de forma gradativa com o passar dos anos. Apesar disso, é importante destacar que a doença apresenta um comportamento cíclico de picos de internações nos primeiros meses do ano. Isso pode indicar a presença de interferência de algum fator climático na incidência das internações, o que pode ser investigado em trabalhos futuros.

Ao utilizar o filtro disponível, tornou-se visível que o montante de mulheres internadas por causa de asma é superior ao de homens, conforme apresentado na Figura 3. Além disso, a partir da análise dos gráficos apresentados na Figura 4, é possível afirmar que a incidência de internações de crianças com asma é muito superior às outras categorias e reforçando o comportamento cíclico da doença. O segundo grupo com maior quantidade de internações é aquele composto por adultos, seguido pelos idosos e por fim os jovens. Para esta análise foram consideradas crianças aquelas que possuem menos de 10 anos de idade, jovens aqueles que possuem entre 10 e 24 anos de

idade, adultos aqueles que possuem entre 25 e 60 anos de idade e, por fim, idosos aqueles que possuem mais de 60 anos de idade.

Figura 3 - Séries temporais de internações por sexo

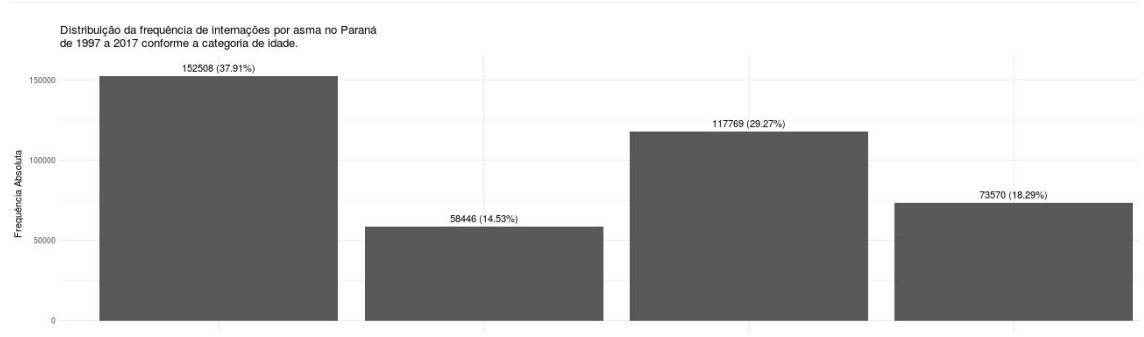
Figura 4 - Séries temporais de internações por classificação de idade



Na última aba disponível são apresentados dois gráficos descritivos. O primeiro (Figura 5) apresenta que, no estado do Paraná, a proporção de internações de mulheres é quase 10% maior que a de homens. O que vai de encontro com o encontrado na série temporal. O segundo (Figura 6) apresenta o gráfico de frequência de internação por categoria de idade, o qual indica que as crianças e os adultos são as categorias que mais são internadas por asma no Paraná, os quais representam mais de dois terços da população analisada. Esse resultado é coerente com o que foi apresentado anteriormente. As causas dessas proporções é tópico que deve ser abordado em pesquisas futuras.

Figura 5 - Proporção de internação por sexo

Figura 6 - Frequência de internações por ano



5 CONCLUSÃO

A base de dados do DATASUS é fonte de informação importante para o desenvolvimento de inúmeras pesquisas na área de saúde. Apesar disso existe uma grande dificuldade na manipulação de seus dados, muitas vezes por conta do grande volume de variáveis. Dessa forma, a fim de facilitar a utilização desses dados, o presente trabalho teve por objetivo desenvolver e implementar um algoritmo para facilitar a obtenção de dados do DATASUS ao longo do tempo, geração de séries temporais, além da análise descritiva e visualização de tais séries em plataforma interativa em formato html. Para isso, foram utilizados o R, o Shiny e o Shiny Server em conjunto.

Uma vez desenvolvido o programa que permite tal atuação, para avaliar a usabilidade do mesmo, foi realizada uma análise descritiva a partir das informações fornecidas pelo aplicativo. Nessa análise foi encontrado que, ao longo dos anos, o volume de internações por causa da asma no Paraná vem diminuindo de forma quase constante. Apesar dessa redução, é visível existe um comportamento cíclico nas incidências de internação, onde há um pico nos primeiros meses do ano.

Além disso, foi identificado que as mulheres são a maior parte dos internados pela doença no estado. Em acréscimo a isso, foi encontrado que as crianças representam mais de um terço das internações. As causas desses comportamentos e proporções não compõem o escopo do presente trabalho. Dessa forma, há a sugestão de realização de pesquisas futuras que respondam aos questionamentos sobre a motivação do comportamento cíclico das internações e a maior proporção de mulheres e crianças.

Este trabalho foi um passo inicial no sentido de aproximar os resultados de pesquisas realizadas na Universidade da sociedade, tanto em nível de graduação, como de pós graduação. Pretende-se que resultados de modelos desenvolvidos, predições e classificações, que são tão importantes para tomadas de decisão tanto de gestores quanto membros da comunidade de um modo geral, fiquem disponíveis de modo fácil. agradável e interativo. A partir do momento que se tem um servidor e todos os relatórios interativos são gerados diretamente do código em R, tem-se um facilitador que reduz o tempo do pesquisador e atende as demandas mais diversificadas da sociedade.

Além disso, o contato direto dos alunos com o download automático de dados, gerenciamento e geração de aplicativos interativos diretamente da linguagem R, vem de encontro com a atual tendência e expectativa do mercado de trabalho. Atualmente, as grandes empresas em nível nacional e internacional, usam o R/Shiny para acompanhamento contínuo e análise de dados, tais como, Google, Amazon, Boticário, Avon, Uber, Airbnb, dentre muitas outras. A busca pelo R/Shiny tem sido cada vez maior. Tais plataformas poderia ser desenvolvidas em SAS, entretanto, seria necessária a compra de diversos módulos extras, o que implicaria em altos custos para as empresas. Assim, o desenvolvimento do projeto trata-se de uma experiência importante aos alunos de graduação tanto para iniciação científica quanto para o desenvolvimento de habilidades para enfrentar o mercado de trabalho nesta era

tecnológica.

REFERÊNCIAS

BARNES, P.J.; JONSSON, B.; KLIM, J.B. The costs of asthma. **European Respiratory Journal**, Copenhagen, v. 9, n. 4, p. 636-642, Abril 1996.

BRASIL. DATASUS. **Histórico / Apresentação.** Disponível em: <http://datasus.saude.gov.br/datasus>. Acesso em: 26 maio 2017.

BRASIL. MINISTÉRIO DA SAÚDE. **Boletim Epidemiológico n. 18, v. 47.** 2016. Disponível em: <http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/svs/noticiassvs/23618-boletim-epidemiologico-apresenta-perfil-da-morbimortalidade-por-doencasrespiratorias-cronicas>. Acesso em: 27 maio 2017.

BRASIL. PORTAL BRASIL. **Asma atinge 6,4 milhões de brasileiros.** 2015. Disponível em: <http://www.brasil.gov.br/saude/2015/01/asma-atinge-6-4-milhoes-de-brasileiros>. Acesso em: 27 maio 2017.

CANDIAGO, R. H. **Uso do DATASUS para avaliação de mudanças nos padrões das internações psiquiátricas no Brasil.** 2007. 109 f. Dissertação (Mestrado) - Curso de Pós-graduação em Ciências Médicas: Psiquiatria, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2007. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/11406/000612639.pdf?sequence=1>. Acesso em: 26 maio 2017.

KLEINMAN, K; HORTON, N. J. SAS and R: Data Management, Statistical Analysis, and Graphics, 2nd Edition, 468p. 2014.

PORTARIA 054-R. Diretrizes Terapêuticas para o Manejo da Asma não Controlada. Diário Oficial dos poderes do estado, 2009.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. 2013.

SAS 9.3 SQL Procedure, User's Guide, 2012.

SBPT. Sociedade Brasileira de Pneumologia e Tisiologia. III Consenso brasileiro no manejo da asma. **Jornal de pneumologia**, São Paulo, v. 28.

ANEXOS

I Encontro de Modelagem Estatística

ESTATÍSTICA

Certificado

Certificamos que o trabalho **"Obtenção de Dados Não Agregados no Datasus: Análise das internações por asma no estado do Paraná de 1992 a 2017"**, título de autoria de **Arthur C.M. Rocha, Paula M. Heller, Eniuce M. Souza,** foi apresentado no formato pôster durante o **"I Encontro de Modelagem Estatística"**, realizado no Anfiteatro Yoshiaki Fukushigue (Térreo do Bloco E90) da Universidade Estadual de Maringá - UEM, nos dias 04 e 05 de dezembro de 2017.

Maringá, 05 de dezembro de 2017.

Clédina Regina Lonardan Acorsi
Chefe do Departamento de Estatística

Eniuce Menezes de Souza
Coordenadora do Evento e do
Programa de Pós-Graduação em Bioestatística

UEM PARANÁ FUNDAÇÃO ARAUCÁRIA