Instituto Serzedello Corrêa – ISC
Pós-Graduação em Análise de Dados para o Controle

SARAH LIMA BEZERRA

Detecção de *Outliers* na Produção do SIH/SUS sob a perspectiva dos atendimentos à população dos municípios brasileiros

Brasília
2020

SARAH LIMA BEZERRA

Detecção de *Outliers* na Produção do SIH/SUS sob a perspectiva dos atendimentos à população dos municípios brasileiros

Trabalho de conclusão do curso de pós-graduação lato sensu em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista em Análise de Dados.

Orientador: Prof. MSc. Saul Campos Berardo

Brasília
2020

REFERÊNCIA BIBLIOGRÁFICA

BEZERRA, Sarah Lima. **Detecção de *Outliers* na Produção do SIH/SUS sob a perspectiva dos atendimentos à população dos municípios brasileiros**. 2020. Trabalho de Conclusão de Curso (Especialização em Análise de Dados para o Controle) – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília DF.

CESSÃO DE DIREITOS

NOME DO AUTOR: Sarah Lima Bezerra

TÍTULO: Detecção de *Outliers* na Produção do SIH/SUS sob a perspectiva dos atendimentos à população dos municípios brasileiros
GRAU/ANO: Especialista/2020

_____

Sarah Lima Bezerra
sarahlb@tcu.gov.br

Ficha catalográfica

Bezerra, Sarah Lima

Detecção de *Outliers* na Produção do SIH/SUS sob a perspectiva dos atendimentos à população dos municípios brasileiros / Sarah Lima Bezerra; orientador, Saul Campos Berardo, 2020.

97 p.

Trabalho de Conclusão de Curso (especialização) - Escola Superior do Tribunal de Contas da União, Curso de Especialização em Análise de Dados para o Controle, Brasília, 2020.

Inclui referências.

1. Análise de Dados. 2. Detecção de *Outliers*. 3. Sistema de Informação de Hospitalar. 4. Sistema Único de Saúde. 5. Metodologia CRISP-DM. I. Campos Berardo, Saul. II. Escola Superior do Tribunal de Contas da União. Especialização em Análise de Dados para o Controle. III. Título.

SARAH LIMA BEZERRA

DETECÇÃO DE *OUTLIERS* NA PRODUÇÃO DO SIH/SUS SOB A PERSPECTIVA DOS ATENDIMENTOS À POPULAÇÃO DOS MUNICÍPIOS BRASILEIROS

Trabalho de conclusão do curso de pós-graduação lato sensu em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista em Análise de Dados.

Brasília, 27 de março de 2020.

Banca examinadora:

_____

Prof. MSc. Saul Campos Berardo
Orientador
Instituto Serzedello Corrêa - TCU

_____

Prof. Dr. Edans Flávius de Oliveira Sandes
Instituto Serzedello Corrêa - TCU

AGRADECIMENTOS

Agradeço a Deus por ter me dado saúde, perseverança e coragem para superar mais um desafio.

À minha querida família por todo incentivo, amor e compreensão, em especial ao meu marido Gledson por me dar o suporte necessário para que eu conseguisse concluir o curso e este trabalho com êxito, à minha filha Laura, minha principal fonte de motivação, que, mesmo pequena, com apenas três anos, entendeu os momentos que tive que me ausentar durante esse período, e aos meus pais, que mesmo à distância, me apoiaram e torceram pelo meu sucesso.

Aos meus chefes José Renato Affonso e Maurício Ramos e Silva por todo apoio e pela autorização para que eu pudesse participar da turma de 2018 do curso de pós-graduação em Análise de Dados, promovido pelo ISC/TCU.

Ao meu orientador, colega e professor Saul Berardo por todos os ensinamentos, sugestões e ajuda, inclusive na definição do tema e do escopo do trabalho.

Aos colegas do curso de pós-graduação pelo aprendizado em conjunto e pelo convívio enriquecedor durante as aulas, em especial ao Marcelo Chaves por ter sido minha dupla em quase todos os trabalhos do curso, e a ele e ao Ricardo Santos pela troca de experiências e apoio durante a execução deste trabalho.

Ao examinador da minha banca, colega e professor Edans Sandes por todas as ideias e correções sugeridas que, com certeza, ajudaram a enriquecer o trabalho.

A todos os professores do curso pelo conhecimento transmitido.

E a todos os amigos e colegas que me ajudaram e/ou torceram para que eu lograsse êxito nessa jornada.

“A perseverança é o segredo do sucesso!”

John Lennon (1889 – 1977)

“No meio da dificuldade encontra-se a oportunidade.”

Albert Einstein (1879 - 1955)

“O homem se torna muitas vezes o que ele próprio acredita que é. Se insisto em repetir para mim mesmo que não posso fazer uma determinada coisa, é possível que acabe me tornando realmente incapaz de fazê-la. Ao contrário, se tenho a convicção de que posso fazê-la, certamente adquirirei a capacidade de realizá-la, mesmo que não a tenha no começo.”

Mahatma Gandhi (1869 – 1948)

RESUMO

A Lei Nº 8.080/1990 estabelece os princípios que regem o funcionamento do Sistema Único de Saúde (SUS), dentre eles estão a universalidade, que define que o Estado deve garantir a todos os cidadãos o acesso aos serviços de saúde oferecidos pelo SUS, independente de quaisquer características sociais ou pessoais, e a descentralização, que ocorre, especialmente, pela transferência de responsabilidades e recursos para a esfera municipal, trazendo a gestão para um nível local e tornando-a, dessa forma, mais efetiva para a população. O objetivo deste trabalho é responder, por meio de técnicas de mineração de dados aplicadas no Sistema de Informação Hospitalar do SUS (SIH/SUS), se de fato esses princípios estão sendo respeitados, ou seja, se a população de todos os municípios brasileiros está realmente tendo acesso aos serviços oferecido pelo SUS, independente de características sociais ou pessoais. O trabalho foi baseado na metodologia C*ross-Industry Standard Process for Data Mining* (CRISP-DM). Os dados do SIH/SUS foram internalizados para o ambiente do TCU, foram aplicadas técnicas estatísticas de análise exploratória e de detecção de anomalias a fim de identificar localidades com discrepâncias quanto à utilização de um dado serviço em relação às demais localidades do país, e foi desenvolvido um painel para facilitar a visualização dessas análises. Como resultado, observou-se que várias localidades foram consideradas anômalas por possuírem taxa de utilização de determinados serviços do SUS muito acima ou muito abaixo da taxa média de utilização do país. Verificou-se também que uma parcela da população com algumas características sociais ou pessoais utilizaram pouco os serviços do SUS quando comparado com o restante da população, e que algumas informações importantes para o trabalho não possuía uma qualidade adequada dos dados.

Palavras-chave: Mineração de dados. Análise de dados. Ciência de dados. Detecção de *Outliers*. Detecção de Anomalias. Teste de Normalidade. CRISP-DM. Sistema Único de Saúde. Sistema de Informação Hospitalar.

ABSTRACT

Law 8.080/1990 establishes the principles of Brazil's Unified Public Health System (SUS), among them are the universality and the decentralization. The first one defines that the State must ensure access to the services offered by SUS for all citizens, regardless of any social or personal characteristics. The second one occurs, especially, through the transfer of responsibilities and resources to the municipal sphere. The objective of this work is to answer, using data mining techniques applied in the SUS Hospital Information System (SIH/SUS), if in fact these principles are being respected, that is, if the population of all Brazilian municipalities is really having access to the services offered by SUS, regardless of social or personal characteristics. The work was based on the Cross-industry standard process for data mining methodology, known as CRISP-DM. SIH/SUS data were loaded into the TCU environment, statistical techniques of exploratory analysis and anomaly detection were applied in order to identify locations with discrepancies in the use of a service in relation to other locations in the country, and a panel was developed to facilitate the visualization of these analyzes. As a result of the work, it was observed that several locations were considered anomalous because they had a utilization rate for certain SUS services much higher or much lower than the country's average utilization rate.
It was also found that a part of the population with some social or personal characteristics used SUS services less when compared to the rest of the population, and that some important information for the work didn't have a good data quality.

Keywords: Data mining. Data analysis. Data science. Outlier detection. Anomaly Detection. Normality test. CRISP-DM. Unified Health System. Hospital Information System

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIH	Autorização de Internação Hospitalar
CNES	Cadastro Nacional de Estabelecimentos de Saúde
CNRAC	Central Nacional de Regulação de Alta Complexidade
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
DATASUS	Departamento de Informática do Sistema Único de Saúde
ER	AIH rejeitada com código de erro – Tabela do SIH
FAEC	Fundo de Ações Estratégicas e Compensação
FTP	File Transfer Protocol
HTTP	Hypertext Transfer Protocol
OMS	Organização Mundial de Saúde
OPM	Órteses, Próteses e Materiais especiais
RD	AIH reduzida – Tabela do SIH
RJ	AIH rejeitada – Tabela do SIH
SADT	Serviços Auxiliares de Diagnose e Terapia
SAS VA	SAS *Visual Analytics*
SecexSaúde	Secretaria de Controle Externo da Saúde
SIA	Sistema de Informação Ambulatorial
SIGTAP	Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM
SIH	Sistema de Informação Hospitalar
SIS	Sistema de Informação em Saúde
SP	Serviços Profissionais – Tabela do SIH
SUS	Sistema Único de Saúde
TCU	Tribunal de Contas da União
UBS	Unidades Básicas de Saúde
UPA	Unidades de Pronto Atendimento

Sumário

1 INTRODUÇÃO

Este capítulo apresenta a contextualização, o problema e a justificativa do trabalho, e como ele foi estruturado.

1.1. CONTEXTO

Em 1988, com a promulgação da Constituição Federal Brasileira, foi instituído no país o Sistema Único de Saúde (SUS), que compreende o conjunto de ações e serviços de saúde, prestados por órgãos e instituições públicas federais, estaduais e municipais, da Administração direta e indireta e das fundações mantidas pelo Poder Público (BRASIL, 1990).

Considerado um dos maiores e melhores sistemas de saúde públicos do mundo, o SUS beneficia cerca de 180 milhões de brasileiros e realiza por ano cerca de 2,8 bilhões de atendimentos, desde procedimentos ambulatoriais simples a atendimentos de alta complexidade, como transplantes de órgãos (FIOCRUZ, 20–).

A Lei Nº 8.080/1990 estabelece os princípios que regem o funcionamento do SUS. O primeiro deles é a universalidade, que define que o Estado deve garantir a todos os cidadãos o acesso aos serviços de saúde oferecidos pelo SUS, independente de quaisquer características sociais ou pessoais – gênero, raça, ocupação profissional, entre outras. Outro princípio é a descentralização das ações de saúde e o seu caráter participativo, que ocorre, especialmente, pela transferência de responsabilidades e recursos para a esfera municipal, estimulando novas competências e capacidades político-institucionais dos gestores locais, além de meios adequados à gestão de redes assistenciais de caráter regional e macrorregional, permitindo o acesso, a integralidade da atenção e a racionalização de recursos. Tal qualidade é uma conquista da rede pública de saúde, porque formaliza o reconhecimento de que o município é o principal responsável pela saúde da população.

Orientado por esses e outros princípios, o SUS parte de uma concepção ampla do direito à saúde e do papel do Estado na garantia desse direito, incorporando em sua estrutura institucional e decisória, espaços e instrumentos para democratização e compartilhamento da gestão do sistema de saúde.

A Organização Mundial de Saúde (OMS) define Sistema de Informação em Saúde (SIS) como um mecanismo de coleta, processamento, análise e transmissão da informação necessária

para se planejar, organizar, operar e avaliar os serviços de saúde, subsidiando a tomada de decisões nos níveis municipal, estadual e federal.

Um dos principais SIS que apoiam o SUS é o Sistema de Informação Hospitalar (SIH/SUS), que foi implantado pelo Ministério da Saúde por meio da Portaria GM/Ministério da Saúde n.º 896/1990.

O SIH/SUS registra informações dos atendimentos provenientes de internações hospitalares ocorridos com financiamento do SUS em hospitais públicos ou privados conveniados. Os dados são enviados por meio do formulário virtual de Autorização de Internação Hospitalar (AIH), cuja finalidade principal é a remuneração referente aos serviços realizados durante a internação hospitalar. O nível Federal recebe mensalmente uma base de dados de todas as internações autorizadas (aprovadas ou não para pagamento) para que possam ser repassados às Secretarias de Saúde os valores de Produção de Média e Alta complexidade, além dos valores de CNRAC, FAEC e de Hospitais Universitários – em suas variadas formas de contrato de gestão. O sistema sofreu diversas mudanças com vistas à melhoria, porém seu conteúdo e propósito podem ser considerados basicamente os mesmos desde sua origem (MACHADO, MARTINS e LEITE, 2016).

O SIH/SUH e os demais sistemas vinculados ao SUS são mantidos e processados pelo Departamento de Informática do Sistema Único de Saúde (DATASUS). No site do DATASUS, estão disponíveis os dados desses sistemas para *download*.

Baseado nos princípios elencados da Lei Nº 8.080/1990 e na disponibilidade dos dados do SIH/SUS, este trabalho tem como intuito verificar, a partir das informações de produção das internações hospitalares, se toda população brasileira está efetivamente tendo acesso aos serviços oferecidos pelo SUS, ou seja, se os procedimentos oferecidos estão realmente sendo utilizados pela população de todos os municípios do país de forma igualitária.

1.2. PROBLEMA E JUSTIFICATIVA

A Portaria - SECEXSAUDE 3/2019 de 10/06/2019 do TCU instituiu o Núcleo de Tratamento de Dados e Informações no âmbito da SecexSaúde, unidade técnica do Tribunal de Contas da União (TCU) responsável pelas fiscalizações e auditorias relativas à área da saúde, e atribuiu como uma de suas competências "tratar os dados dos bancos de dados do Ministério da Saúde e outros e propor o encaminhamento dos resultados obtidos".

Dentre esses bancos de dados do Ministério da Saúde, encontra-se o banco de dados do SIH/SUS, que contém as informações dos atendimentos provenientes de internações hospitalares financiadas pelo SUS. Os dados desse sistema estão disponíveis para download no portal do DATASUS.

Foi verificado, que embora tivesse uma versão desses dados no LabContas[1], ela estava desatualizada, com dados até 2016, e não existia um processo contínuo para internalização desses dados no TCU.

O SIH/SUS possui um grande volume de dados. Ele registra cerca de 12 milhões de internações hospitalares por ano. Foi constatato que após a internalização desses dados, a SecexSaúde precisaria de meios que a auxiliassem na verificação da qualidade dos dados e na identificação de possíveis irregularidades na produção das internações hospitalares do SUS.

Na mineração de dados, a detecção de anomalias é a identificação de padrões em dados que não estão em conformidade com o comportamento esperado. No contexto dos dados das internações hospitalares, os itens anômalos provavelmente se referem a uma epidemia, a uma má distribuição dos atendimentos, a indícios de fraudes ou a preenchimento incorreto dos dados.

Diante do exposto no item 1.1 CONTEXTO e nesta seção, surgem as seguintes questões a serem tratadas:

A população de todos os municípios está efetivamente tendo acesso aos serviços de internação hospitalar financiados pelo SUS? Em quais localidades são encontrados os maiores índices de utilização desses serviços? Quais localidades cuja população não tem acesso ou tem pouco acesso a esses serviços de internação hospitalar?

Quais as melhores técnicas de detecção de anomalia a serem aplicadas nos dados do SIH/SUS para verificar a efetividade dos atendimentos pelos municípios?

O princípio da universalidade que norteia o SUS está sendo cumprido em relação aos atendimentos de internações hospitalares? Toda população independente de quaisquer características sociais ou pessoais está tendo acesso aos serviços de internação hospitalar financiados pelo SUS?

[1] LabContas**:** plataforma virtual do TCU que reúne bancos de dados da Administração Pública e ferramentas de análise de conteúdo.

1.3. ESTRUTURA DO DOCUMENTO

Este documento está estruturado em nove capítulos e um apêndice.

No primeiro capítulo, encontram-se o contexto, o problema e a justificativa do trabalho.

O capítulo dois apresenta a metodologia CRISP-DM, que foi a metodologia seguida durante a elaboração do trabalho.

Nos capítulos seguintes (três a oito), tem-se o detalhamento das seis fases da metodologia CRISP-DM, em que são apresentadas as atividades e as técnicas que foram empregadas em cada uma dessas fases e os resultados obtidos nelas.

O capítulo nove consiste das considerações finais do trabalho.

Por fim, o apêndice A apresenta como foi definido o teste de normalidade utilizado nas análises do trabalho.

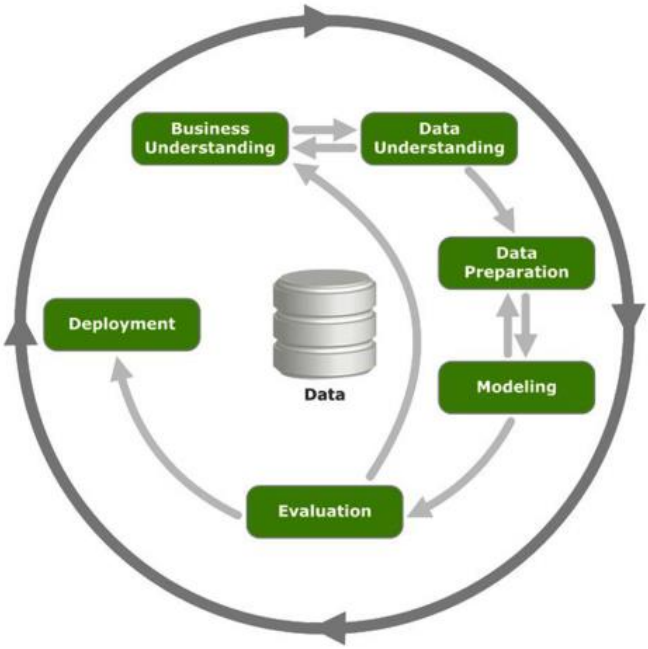
2 METODOLOGIA

O trabalho foi baseado na metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM), que é uma das metodologias mais populares e amplamente utilizadas para projetos de mineração e análise de dados.

O CRISP-DM fornece uma estrutura de natureza cíclica e iterativa, que descreve as etapas e fluxos de trabalho necessários para executar um projeto de mineração e análise de dados, desde os requisitos de negócios até as etapas finais de implantação (GHOSH, BALI e SARKAR, 2018).

Essa metodologia apresenta uma visão geral do ciclo de vida de um projeto de mineração de dados. Esse ciclo é composto por seis fases (entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação), como mostrado na figura abaixo.

Figura 1: Ciclo de Vida da Metodologia CRISP-DM



Fonte: (GHOSH, BALI e SARKAR, 2018)

Observa-se que não se trata de uma sequência de fases fixas, sendo passível de idas e vindas, que são inclusive encorajadas (HERSCU, 2017). As setas indicam as relações mais frequentes entre as fases, e não uma sequencia rígida de etapas. Elas dependem das particularidades do projeto ou do problema que está sendo resolvido.

Vasconcellos (2017) cita, como principais vantagens da metologia CRISP-DM, o fato dela poder ser aplicada a qualquer tipo de negócio e de não ter dependência de ferramenta para ser executada.

A primeira fase do ciclo consiste no entendimento do negócio, em que são determinados os objetivos de negócios e as correspondentes metas de mineração de dados. Nessa fase, as expectativas e os critérios de sucesso do projeto são definidos. Ela também avalia riscos, custos e contingências e culmina em um plano de projeto.

A fase de entendimento dos dados compreende a coleta, a descrição (quantitativa e qualitativa), a exploração (por meio de tabelas e gráficos) e a verificação da qualidade dos dados.

A fase seguinte consiste na preparação dos dados para a fase de modelagem. Ela geralmente engloba quatro etapas: seleção, limpeza, construção e integração dos dados. Segundo a IBM, essa fase consome de 50 a 70% do tempo e do esforço de um projeto.

A fase de modelagem é onde a mineração dos dados realmente acontece. É nessa fase que os algoritmos e as técnicas de mineração são empregados. Ela inclui a criação, a avaliação e o ajuste fino dos modelos, com base nos critérios de sucesso de mineração de dados estabelecidos durante a fase de entendimento do negócio (GHOSH, BALI e SARKAR, 2018).

A fase de avaliação verifica se os resultados obtidos (modelos produzidos) atendem às expectativas e aos critérios de sucesso de negócio estabelecidos durante a fase de entendimento do negócio.

A última fase do ciclo consiste na implantação, no ambiente de produção, dos modelos produzidos, validados e testados nas fases anteriores .

3 ENTENDIMENTO DO NEGÓCIO

A primeira fase do CRISP-DM consiste no entendimento do negócio, que procura identificar os objetivos e as necessidades na perspectiva de negócio, e converter este conhecimento em metas de mineração.

Nesta fase, foram realizadas as seguintes atividades:

a) Entrevistas com o responsável pelo núcleo de dados da SecexSaúde;
b) Definição dos objetivos de negócio;
c) Definição das metas de mineração;
d) Estudo bibliográfico sobre o tema do trabalho.

3.1. DEFINIÇÃO DOS OBJETIVOS DE NEGÓCIO

Diante dos problemas elencados no item 1.2 PROBLEMA E JUSTIFICATIVA, foram definidos os seguintes objetivos de negócio.

3.1.1 Objetivo geral

Verificar se de fato a população brasileira está tendo acesso, de forma igualitária, aos serviços oferecidos pelo SUS, por meio da análise dos dados de internações hospitalares, a fim de auxiliar o TCU (SecexSaúde) nas suas atividades de controle, fornecendo indícios a serem usados em suas auditorias, e detectando possíveis deficiências no preenchimento dos dados do sistema de informações hospitalares.

3.1.2 Objetivos específicos

1. Internalizar os dados do SIH/SUS e demais bases de dados que sejam necessárias para o trabalho;
2. Utilizar técnicas de mineração de dados para detectar de forma automática anomalias estatísticas nos registros de internações hospitalares do SUS;
3. Construir um painel apresentando os dados e as anomalias identificadas nos registros do SIH/SUS

3.2. DEFINIÇÃO DAS METAS DE MINERAÇÃO

Para os objetivos específicos definidos no item anterior, foram definidas metas de mineração a serem alcançada durante o trabalho. São elas:

a) Criar script, usando a linguagem de programação *Python*, para internalizar de forma automática os dados do SIH/SUS disponíveis no site do DATASUS – objetivo específico 1;

b) Criar script, usando a linguagem de programação *Python*, para verificar de forma automática a distribuição estatística da amostra de dados a ser analisada e transformar essa distribuição em uma distribuição gaussiana (normal), quando necessário – objetivo específico 2;

c) Criar script, usando a linguagem de programação *Python*, para detectar de forma automática anomalias estatísticas nos registros de internações hospitalares do SUS – objetivo específico 2;

d) Construir um painel, usando a ferramenta *SAS Visual Analytics* (SAS VA), para apresentar os dados e as anomalias identificadas durante as análises nos registros do SIH/SUS – objetivo específico 3.

3.3. ESTUDO BIBLIOGRÁFICO SOBRE O TEMA

3.3.1 **Descentralização**

A descentralização da gestão e das políticas da saúde no país é um dos princípios organizativos do SUS, que foi estabelecido a partir da Constituição Federal de 1988 e regulamentada pelas Leis 8.080/90 (Lei Orgânica da Saúde) e 8.142/90. A partir desse princípio, a gestão passa a ser feita de forma integrada entre a União, os estados e os municípios. O poder e a responsabilidade sobre o setor passam a ser distribuídos entre os três níveis de governo, objetivando uma prestação de serviços com mais eficiência e qualidade e também a fiscalização e o controle por parte da sociedade (FIOCRUZ, 20–).

Segundo o Ministério da Saúde (2017), as responsabilidades de cada ente são:

- Governo federal

A gestão da saúde no nível federal é de responsabilidade do Ministério da Saúde. Ele formula as políticas nacionais de saúde, mas não realiza as ações. Para isso, depende de seus parceiros (estados, municípios, ONGs, fundações, empresas, etc.).

O governo federal é o principal financiador da rede pública de saúde. Historicamente, o Ministério da Saúde aplica metade dos recursos gastos no país em saúde pública em todo o Brasil, e estados e municípios, em geral, contribuem com a outra metade dos recursos.

- Governo estadual

O gestor estadual é responsável pela organização do atendimento à saúde em seu território.

Ele deve aplicar recursos próprios, inclusive nos municípios, e os repassados pela União.

Os estados possuem secretarias específicas para a gestão de saúde.

Além de ser um dos parceiros para a aplicação de políticas nacionais de saúde, o estado formula suas próprias políticas de saúde. Ele coordena e planeja o SUS em nível estadual, respeitando a normatização federal.

- Governo municipal

A partir do Pacto pela Saúde, de 2006, o gestor municipal assina um termo de compromisso para assumir integralmente as ações e serviços de seu território, tornando-se, assim, o principal responsável pela saúde de sua população.

Ele deve aplicar recursos próprios e os repassados pela União e pelo estado.

Os municípios possuem secretarias específicas para a gestão de saúde.

O município formula suas próprias políticas de saúde e também é um dos parceiros para a aplicação de políticas nacionais e estaduais de saúde. Ele coordena e planeja o SUS em nível municipal, respeitando a normatização federal e o planejamento estadual.

Pode estabelecer parcerias com outros municípios para garantir o atendimento pleno de sua população, para procedimentos de complexidade que estejam acima daqueles que pode oferecer.

O Decreto 7.508 de 2011, que regulamenta a Lei 8.080/90, estabelece um novo arranjo para a descentralização: a região de saúde, que é um espaço geográfico contínuo constituído por agrupamentos de municípios limítrofes, delimitado a partir de identidades culturais, econômicas e sociais e de redes de comunicação e infraestrutura de transportes compartilhados,

com a finalidade de integrar a organização, o planejamento e a execução de ações e serviços de saúde.

Cada região formada nos estados deverá garantir a integralidade no atendimento através da parceria entre os municípios componentes.

3.3.2 Tipos de unidades de atendimento da rede pública de saúde municipal

O atendimento à população na rede pública de saúde municipal pode existir em Unidades Básicas de Saúde (UBS), em Unidades de Pronto Atendimento (UPA) e em hospitais públicos (MERELES, 2016).

As UBS foram criadas para ser o ambiente primário de atendimento ao cidadão. Elas são conhecidas popularmente como centros ou postos de saúde. Nelas, são realizados os procedimentos menos complexos, como vacinação e curativos. O objetivo é que elas consigam atender a maior parte dos problemas de saúde da população local, evitando, assim, encaminhamento para outras unidades, como emergências e hospitais. Para isso, o ideal é que cada bairro de um município tenha pelo menos uma UBS.

As UPAs foram criadas para concentrar os atendimentos de saúde de média complexidade. Elas possuem uma estrutura intermediária entre os postos de saúde e os hospitais. Geralmente, possuem aparelhos para alguns exames (como raio-X e eletrocardiografia) e leitos de observação. Segundo o Ministério da Saúde, nas localidades que contam com UPA, 97% dos casos são solucionados na própria unidade.

Os hospitais públicos são estruturas que contam com o maior número de aparelhos de exames diversos, onde normalmente trabalham os especialistas de diversas áreas e onde ocorrem as cirurgias e o atendimento de casos mais complexos. Muitos deles possuem prontos-socorros para atendimentos de urgência de casos mais graves.

3.3.3 Infosas

Em 2013, procurando incorporar ferramentas de mineração de dados no sistema de controle do SUS, o Ministério da Saúde contratou uma equipe do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais para desenvolver um sistema interativo e automático de detecção de anomalias na produção do SUS para posterior auditoria e verificação. Assim, surgiu o InfoSAS.

O InfoSAS utiliza diversos algoritmos que procuram capturar discrepâncias, produzindo escores que permitem ordenação e priorização do resultado.

As anomalias estatísticas encontradas pelo InfoSAS podem ser originadas por fraudes, mas também serem resultados de processos corretos de ocorrência pouco frequente, como mutirões ou transferências de serviços entre prestadores, ou de informação incorreta nos registros de atendimentos, como o lançamento do endereço do hospital ao invés do endereço do paciente. Taxas de cobertura com valores muito acima do esperado pela taxa brasileira podem também ser resultado de uma má distribuição do atendimento no país, com os municípios com atendimento anômalo sendo aqueles poucos com o atendimento correto (ASSUNÇÃO, CARVALHO, *et al.*, 2016)

O InfoSAS mostrou-se uma importante ferramenta para detecção de casos anômalos que podem servir de indícios para diversas atividades de auditoria e controle.

O trabalho se baseou nessa ferramenta, mas com escopo e objetivos diferentes. No InfoSAS, busca-se detectar as taxas de atendimentos por habitante muito superiores à média nacional e os estabelecimentos considerados discrepantes em relação ao valor e à quantidade dos atendimentos. Seu principal foco é a detecção de possíveis fraudes. O foco deste trabalho está mais voltando para a verificação da efetividade de políticas públicas na área da saúde, analisando se de fato toda população brasileira está tendo acesso aos serviços de internação hospitalar financiados pelo SUS. Neste trabalho, busca-se detectar não apenas as localidades (municípios ou regiões de saúde) cuja população foi muito atendida (em relação às demais localidades do país), mas também as localidades que não tiveram acesso ou tiveram pouco acesso a esses serviços.

O InfoSAS trabalha com os dados dos sistemas SIA e SIH. Este trabalho usa apenas os dados do SIH.

Outro ponto de diferença é que o InfoSAS analisa os procedimentos do SUS no nível da sua forma (Tabela 3: Composição do número do procedimento e Figura 10: Exemplo de Número de Procedimento), e este trabalho faz a análise nos quatro níveis de abstração do serviço (procedimento, forma, subgrupo e grupo).

E, por fim, o código do InfoSAS não está disponível para a comunidade e a ideia deste trabalho é que seu código seja aberto.

4 ENTENDIMENTO DOS DADOS

A fase de entendimento dos dados compreende a coleta, a descrição (quantitativa e qualitativa), a exploração (por meio de tabelas e gráficos) e a verificação da qualidade dos dados.

Nesta fase, foram realizadas as seguintes atividades:

a) Levantamento das bases de dados disponibilizadas no LabContas necessárias para o trabalho;
b) Internalização dos dados ainda não disponibilizados no LabContas necessários para o trabalho;
c) Estudo, análise e exploração dos dados disponíveis para o trabalho.

4.1. LEVANTAMENTO DAS INFORMAÇÕES DISPONIBILIZADAS NO LABCONTAS NECESSÁRIAS PARA O TRABALHO

Iniciou-se esta fase fazendo um levantamento dos dados que seriam necessários para a execução do trabalho, a saber: SIH/SUH, CNES, IBGE e Regiões de Saúde.

Primeiramente, verificou-se que as informações do CNES, IBGE e Regiões de Saúde já haviam sido internalizadas no ambiente LabContas nas bases de dados BDU_SECEXSAUDE_CNES, BD_IBGE e BDU_SECEXSAUDE_AUX, respectivamente.

4.1.1 IBGE

A base de dados BD_IBGE disponível no ambiente LabContas possui a tabela POPULACAO_MUNICIPIO, em que podem ser encontradas informações sobre as populações de todos os 5570 (cinco mil, quinhentos e setenta) municípios do país.

4.1.2 Regiões de Saúde

A base de dados BDU_SECEXSAUDE_AUX disponível no ambiente LabContas possui a tabela BR_REGSAUD, que possui a relação código município x região de saúde. Um município pertence apenas a uma região de saúde.

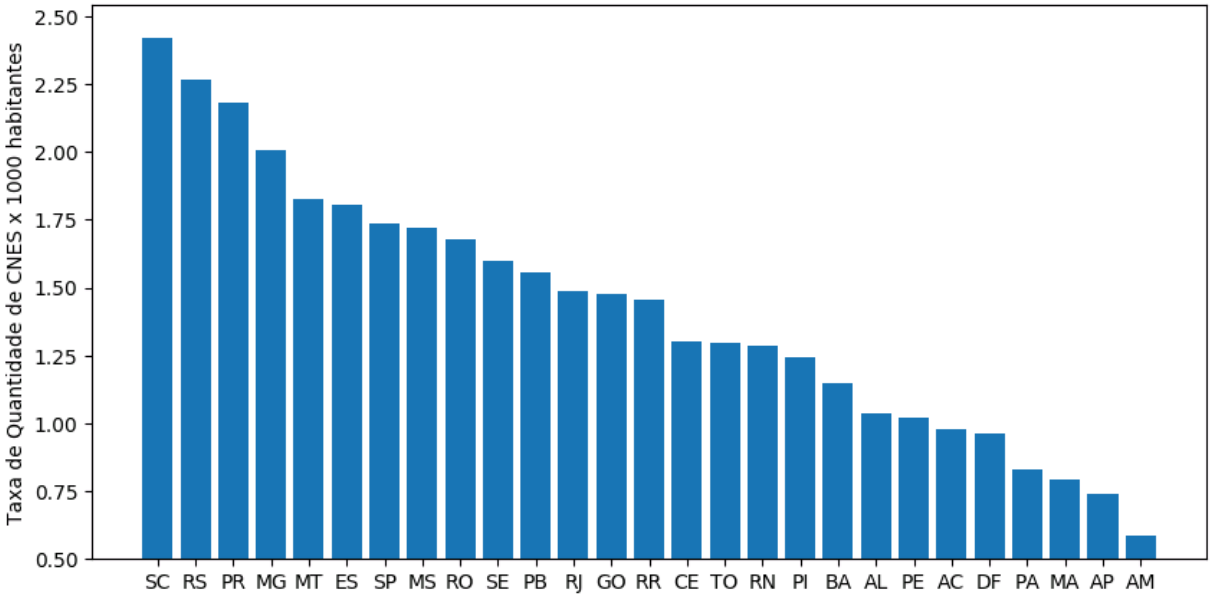
Existem 464 regiões de saúde cadastradas. Cada região de saúde pertence a um único estado federativo.

4.1.3 CNES

A base de dados BDU_SECEXSAUDE_CNES disponível no ambiente LabContas possui a tabela ST (Estabelecimentos), em que podem ser encontradas informações de todas as unidades de saúde públicas ou privadas do país, ou seja, todas as instituições cadastradas no CNES.

A figura abaixo representa a situação da distribuição das instituições de saúde nos estados em dezembro de 2018 (mês/ano utilizado no decorrer do trabalho para os dados dessa tabela).

Figura 2: Taxa de Quantidade de CNES (a cada 1000 habitantes) por UF



Fonte: Elaborada pela autora (2020).

Na figura, o estado do Brasil que possui a maior taxa de quantidade de estabelecimentos do CNES por habitante é Santa Catarina, com o valor de 2,42 estabelecimentos a cada mil habitantes. O estado com a menor taxa é Amazonas, com o valor de 0,58 estabelecimentos por mil habitantes. A taxa média dos estados é de 1,42 estabelecimentos por mil habitantes.

4.2. INTERNALIZAÇÃO DOS DADOS NÃO DISPONIBILIZADOS NO LABCONTAS NECESSÁRIOS PARA O TRABALHO

4.2.1 **Fonte dos dados do SIH/SUS: DATASUS**

O DATASUS disponibiliza os dados do SIH/SUS para download via dois protocolos da internet: HTTP e FTP.

Usando o protocolo HTTP, os dados das tabelas principais podem ser encontrados nos endereços:

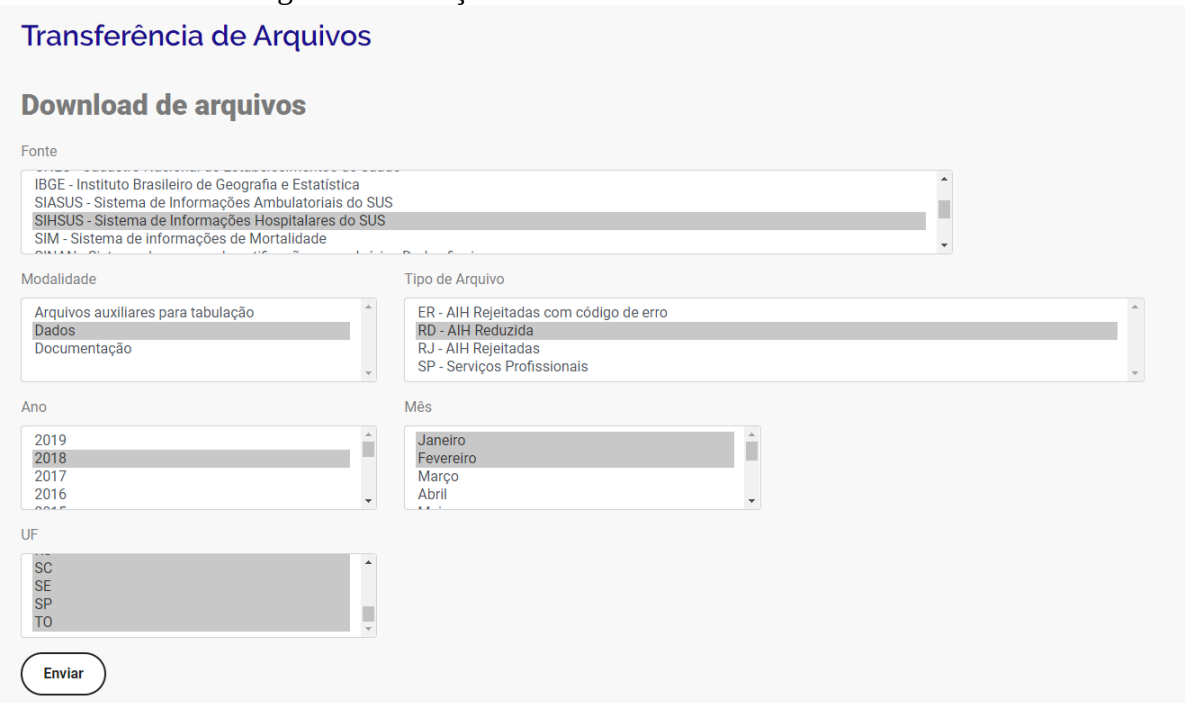
a) http://datasus.saude.gov.br/transferencia-de-arquivos/
b) http://www.datasus.gov.br/DATASUS/index.php?area=0901&item=1&acao=25

Para ter acesso aos dados, basta escolher, em um dos dois endereços, 'SIHSUS – Sistema de Informações Hospitalares do SUS' como Fonte, 'Dados' como Modalidade, um ou mais tipos de arquivo (existem quatro disponíveis: ER, RD, RJ e SP), um ou mais anos, um ou mais meses, e uma ou mais UF, conforme Figura 3: Obtenção dos dados do SIH/SUS via HTTP. Como resultado, obtém-se a lista dos arquivos de dados referentes à pesquisa realizada.

Os tipos de arquivos se referem a AIH Rejeitadas com código de erro (ER), AIH Reduzida (RD), AIH Rejeitadas (RJ) e Serviços Profissionais (SP).

É importante destacar que, para cada tipo de arquivo, existe um arquivo de dados por mês para cada UF (incluindo o Distrito Federal), ou seja, existem 27 arquivos por mês, totalizando 324 arquivos de dados por ano por tipo de arquivo.

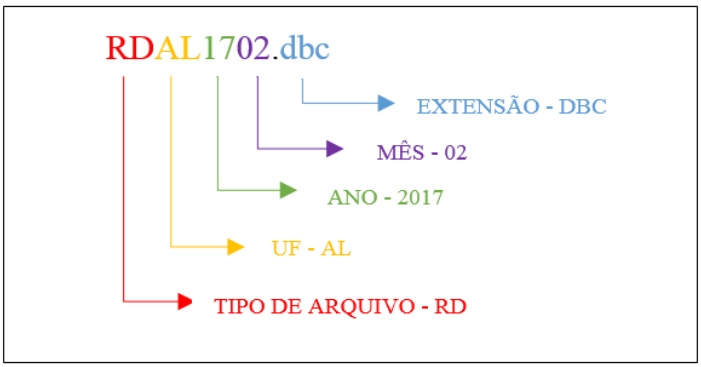
Figura 3: Obtenção dos dados do SIH/SUS via HTTP



Fonte: Elaborada pela autora - *print screen* de (Transferência de Arquivos DATASUS, 2020).

A nomenclatura do arquivo de dados é composta pelo tipo de arquivo (dois primeiros dígitos), UF (dígitos 3 e 4), ano (dígitos 5 e 6) e mês (últimos dois dígitos), seguido da extensão do arquivo [2], conforme exemplificado na Figura 4: Exemplo de Nome de Arquivo do SIH.

Figura 4: Exemplo de Nome de Arquivo do SIH/SUS

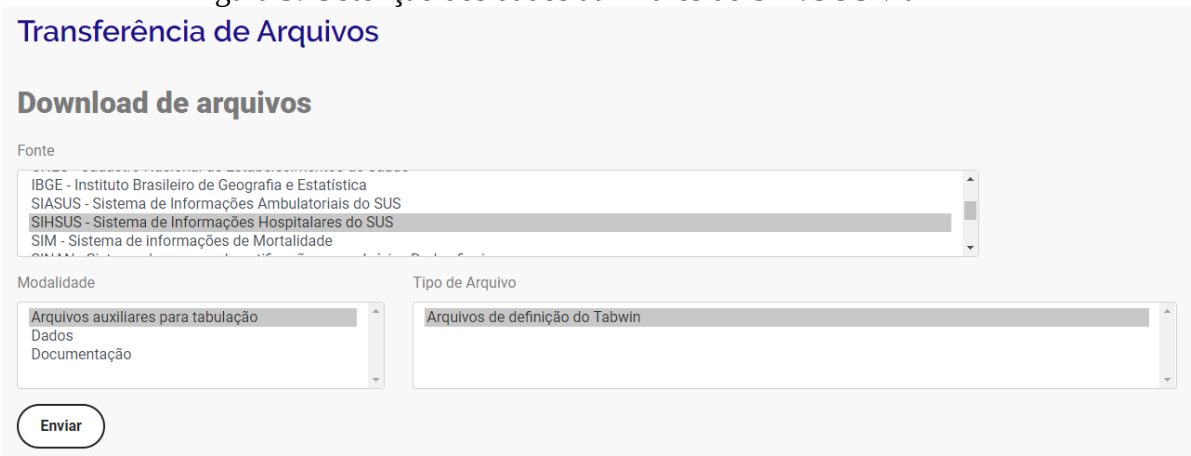


Fonte: Elaborada pela autora (2020).

O DATASUS também disponibiliza o dicionário de dados das duas principais tabelas cujos dados estão disponíveis para download: RD e SP. Trata-se do arquivo IT_SIHSUS_1603.pdf, que pode ser baixado escolhendo a opção 'Documentação' como Modalidade na busca.

Os dados auxiliares podem ser acessados ao escolher a opção 'Arquivos auxiliares para tabulação' como Modalidade e 'Arquivos de definição do TabWin' como Tipo de Arquivo, conforme Figura 5: Obtenção dos dados auxiliares do SIH/SUS via HTTP.

Figura 5: Obtenção dos dados auxiliares do SIH/SUS via HTTP



Fonte: Elaborada pela autora – *print screen* de (Transferência de Arquivos DATASUS, 2020).

[2] O DATASUS disponibiliza os dados dos arquivos para download em duas extensões: dbf ou dbc.

Dbf é uma extensão de arquivo de banco de dados desenvolvida pela empresa Ashton-Tate, que é a empresa que originou o banco de dados dBase, e foi introduzida pela primeira vez no dBase II. É acrônimo para dBase database file. (WHAT IS FILE EXTENSION TEAM, 20–)

Dbc representa um arquivo dbf compactado.

Usando o protocolo FTP, os dados das tabelas principais podem ser acessados por meio dos endereços:

a) ftp://ftp.datasus.gov.br/dissemin/publicos/SIHSUS/199201_200712/Dados/ (para dados de janeiro de 1992 até dezembro de 2007)
b) ftp://ftp.datasus.gov.br/dissemin/publicos/SIHSUS/200801_/Dados/ (para dados a partir de janeiro de 2008)

Os dados auxiliares podem ser acessados por meio do endereço ftp://ftp.datasus.gov.br/dissemin/publicos/SIHSUS/200801_/Auxiliar/.

4.2.2 Internalização dos dados das tabelas principais do SIH/SUS

As principais tabelas do SIH cujo os dados estão disponíveis pelo DATASUS para download são RD e SP, que representam respectivamente AIH Reduzida e Serviços Profissionais.

A tabela de AIH reduzida possui dados referentes à internação, ao paciente e ao estabelecimento em que a internação ocorreu. A tabela de Serviços Profissionais possui informações sobre os profissionais que atuaram durante a internação.

Foi verificado que existem dados disponíveis da tabela RD desde janeiro de 1992 e da tabela SP desde junho de 1998. A SecexSaúde solicitou que fossem disponibilizados no LabContas os dados de janeiro de 2008 em diante.

Para isso, foi criado um script de carga dos dados usando a linguagem de programação *Python* (Python, 2001), por meio da ferramenta Jupyter Notebook (PROJECT JUPYTER, c2020), que verifica os arquivos disponíveis no FTP do DATASUS que ainda não foram internalizados no LabContas - usando a biblioteca *ftplib* (PYTHON SOFTWARE FOUNDATION, c2001-2020). Para cada arquivo encontrado, carrega os dados do arquivo em um *dataframe* da biblioteca *pandas* (Pandas documentation, 2008) - usando as bibliotecas *PySUS* (COELHO, 201-?) ou *dbfread* (Read DBF Files with Python, 20–), dependendo da extensão do arquivo - e faz a carga dos dados do *dataframe* no banco de dados - usando a biblioteca *pyodbc* (MICHAEL KLEEHAMMER, 201-?).

Durante a criação e execução do script, surgiram algumas dificuldades que precisaram ser contornadas. As principais delas foram:

a) Descompactação de arquivos com a extensão dbc para dbf.

Para descompactar os arquivos com extensão dbc para dbf, o DATASUS disponibiliza a ferramenta TabWin (TabWin, 2008) ou o executável dbf2dbc.exe, que integra o TabWin e pode ser chamado via linha de comando. Ambos são executados no sistema operacional Windows. Entretanto, o ambiente do TCU destinado ao desenvolvimento na linguagem de programação *Python* é Linux.

No início do trabalho, optou-se por fazer essa descompactação em uma máquina com o sistema operacional Windows, em um passo anterior à execução do script de carga que estava sendo construído.

Após um período de pesquisa e tentativa de inserção dessa etapa no script, foi encontrada a biblioteca *pysus* do *Python* que possui uma função que recebe um arquivo com extensão dbc e retorna um *dataframe* da biblioteca *pandas* do *Python*, tornando a solução independente do sistema operacional que o script estava sendo executado.

b) Lentidão na inserção dos dados das tabelas no banco de dados.

Na primeira versão do script, a inserção dos dados no banco de dados estava sendo feita por meio do método to_sql da biblioteca *pandas*. Entretanto, a carga de cada arquivo estava demorando em média 4h para ser finalizada, tornando o processo de carga inviável, pelo fato de terem sido carregados 7776 arquivos no total (arquivos correspondentes ao período de janeiro de 2008 a dezembro de 2019).

A solução foi usar o método *bulk insert* (MICROSOFT, 2020), que insere os dados numa tabela de dados a partir de um arquivo em um formato especificado pelo usuário (no trabalho, foram usados arquivos com extensão csv). Para isso, foi utilizada a biblioteca *pyodbc*. O tempo médio para finalizar a carga de um arquivo passou para menos 20s.

c) Dados já carregados no banco de dados estavam desatualizados.

Quando o trabalho foi iniciado, foi verificado que já existiam as tabelas RD e SP no banco de dados com dados de janeiro de 2008 a dezembro de 2016. Entretanto, foi verificado que esses dados estavam desatualizados em relação aos arquivos disponíveis no FTP do DATASUS.

Essa checagem foi realizada comparando o quantitativo de registros dos arquivos carregados no banco de dados com o quantitativo de registros dos arquivos disponíveis no FTP.

A solução para contornar esse problema foi recarregar, no LabContas, os dados de todos os arquivos das tabelas SP e RD de janeiro de 2008 em diante.

4.2.3 **Internalização dos dados das tabelas auxiliares do SIH/SUS**

Existem quatro arquivos com os dados auxiliares do SIH/SUS disponíveis para download pelo DATASUS, conforme tabela abaixo.

Tabela 1: Arquivos das tabelas auxiliares do SIH/SUS

ARQUIVO	CONTEÚDO
TAB_SIH_199201-199712.zip	Tabelas auxiliares referentes aos dados de 01/1992 a 12/1997.
TAB_SIH_199801-200307.zip	Tabelas auxiliares referentes aos dados de 01/1998 a 07/2003.
TAB_SIH_200308-200712.zip	Tabelas auxiliares referentes aos dados de 08/2003 a 12/2007.
TAB_SIH.zip	Tabelas auxiliares referentes aos dados de 02/2008 em diante.

Fonte: Elaborada pela autora (2020).

Optou-se por internalizar os dados do arquivo TAB_SIH.zip por serem compatíveis com os dados das tabelas principais que foram inseridos no LabContas, ou seja, dados a partir de janeiro de 2008.

O arquivo TAB_SIH.zip contém 246 arquivos, dos quais 237 são arquivos que representam as tabelas de referência (tabelas *lookup*) das principais tabelas dos sistemas do SUS. Esses arquivos possuem a extensão dbf ou a extensão cnv (arquivos de conversão).

A SecexSaúde solicitou que fossem disponibilizados no LabContas os dados auxiliares referenciados diretamente pelas tabelas RD e SP. Foi então realizado um estudo para identificar quais arquivos deveriam ser carregados e chegou-se na lista a seguir com 55 arquivos: BR_MUNICGESTOR, BR_REGIAO, BR_UF, CARATEND, CID10GRUPO, CID10_3D, CNAE, COMPLEX2, CONTRAC, DIARIASUTI, ETNIA, FAECTP, FINANC, GESTAO, IDADEBAS, IDADEDET, IDENT, IDENTIFIC, INSCPN, INSTRU, LEITOS, MARCAUTI, MESES, MORTES, MUNICBR, MUNIDB, NACION3D, NATJUR, NATUREZA, NUMFILH, PERM, REGCT, SAIDAPERM, SEQAIH, SEXO, SIMNAO, TP_DIAGSEC, TP_VAL, UF, VINCPREV, BR_MUNICIP, LIBGESTOR, ANO, CBO, CID10, S_CLASSEN, TB_FORMA, TB_GRUPO, TB_SIGTAP, TB_SUBGR, TCHBR, TCNESBR, TMANTBR, TPROCUNI, RACACOR.

Para a carga dos dados auxiliares, foi criado um script usando a linguagem *Python*, por meio da ferramenta *Jupyter Notebook*, que, para cada arquivo da lista acima, carrega os dados do arquivo em um *dataframe* da biblioteca *pandas* – usando a biblioteca *dbfread* para arquivos com extensão dbf e usando a função open (W3SCHOOLS, c1999-2020) para arquivos com extensão cnv – e carrega os dados do *dataframe* no banco de dados usando a biblioteca *pandas*.

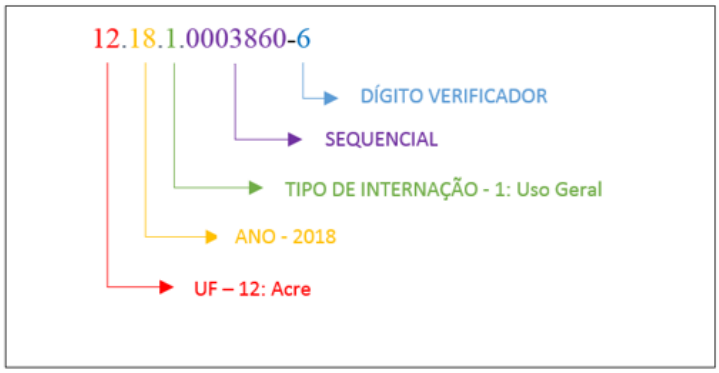
4.3. ENTENDIMENTO DOS PRINCIPAIS CAMPOS DO SUS USADOS NO TRABALHO

Nesta seção, foram detalhados alguns campos do SUS que tiveram relevância na execução do trabalho.

4.3.1 **Número da AIH**

Segundo Ministério da Saúde (2017), a numeração da AIH (campo N_AIH da tabela RD) constitui-se de 13 dígitos, incluído o dígito verificador, de acordo com a composição exemplificada na **Erro! Fonte de referência não encontrada.** e detalhada na Tabela 2: Composição da Numeração da AIH.

Figura 6: Exemplo de Numeração da AIH



Fonte: Elaborada pela autora (2020).

Tabela 2: Composição da Numeração da AIH

DÍGITO	SIGNIFICADO
1° e 2°	Unidade da Federação, de acordo com o código do Instituto Brasileiro de Geografia e Estatística / IBGE (ex: 25 - Paraíba, 31 - Minas Gerais), exceto nos casos das séries numéricas específicas da Central Nacional de Regulação de Alta Complexidade (CNRAC), que iniciam com o número 99 para todo Brasil, sem divisão por UF
3° e 4°	Dois últimos algarismos do ano de referência (Ex.: 16 para 2016)
5º	Tipo de internação; Valores possíveis: 1, 3 ou 5. • 1 (um) para identificar que a autorização é de Internação (AIH) - uso geral; • 3 (três) para identificar que a numeração é de internação (AIH) específica da CNRAC; • 5 (cinco) para identificar que a autorização é de internação (AIH) específica para a estratégia de aumento do acesso aos Procedimentos Cirúrgicos Eletivos no âmbito do Sistema Único de Saúde (SUS), seja componente I, II ou III, definidos pela Portaria GM/Ministério da Saúde nº 1.340, de 29 de junho de 2012 e legislação correlata.
6° ao 12°	Sequencial (valores: 0.000.001 até 9.999.999)
13º	Dígito verificador (valores: 0 a 9)

Fonte: Elaborada pela autora (2020).

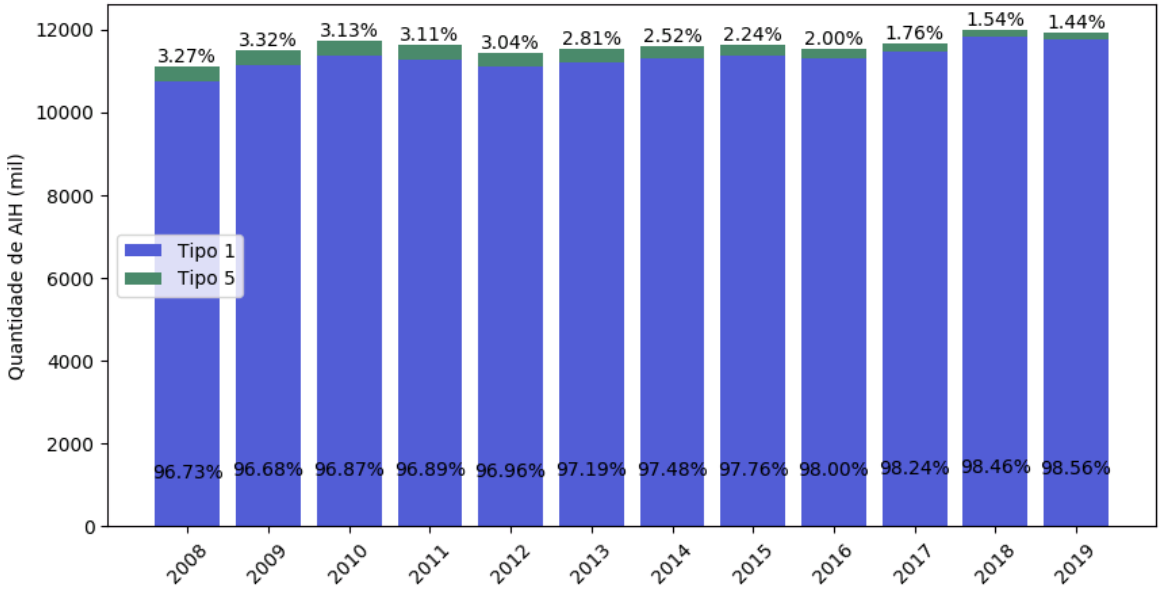
4.3.2 **Tipo de AIH**

O campo IDENT da tabela RD indica o tipo de AIH. Os valores possíveis para esse campo são: 1 (indica nova AIH) e 5 (indica AIH para continuidade de tratamento em procedimentos que admitem longa permanência, como, por exemplo, na psiquiatria e no tratamento de tuberculose).

A data de internação na AIH 5 permanece a mesma da AIH 1, mesmo que a internação se prolongue por meses (ou anos), representando uma única internação.

A Figura 7: Gráfico da Quantidade de Tipo de AIH x Ano mostra uma comparação do quantitativo de AIHs dos tipos 1 e 5 no decorrer dos anos. Pode-se observar que o percentual de AIHs do tipo 5 tem diminuído no decorrer dos anos. A média percentual de AIHs desse tipo é 2.5, variando entre 1.44 a 3.32%.

Figura 7: Gráfico da Quantidade de Tipo de AIH x Ano



Fonte: Elaborada pela autora (2020).

4.3.3 **Localidade**

Durante o trabalho, foram realizadas análises em dois tipos de localidade: município ou região de saúde.

O município usado como base para as análises foi o de residência do paciente, já que o foco do trabalho era destacar as localidades cuja população estava com acesso anômalo (tanto alto quanto baixo) em relação aos serviços de internação hospitalar financiados pelo SUS. E a

região de saúde usada também foi a de residência do paciente, isto é, a região de saúde ao qual o município de residência do paciente pertence.

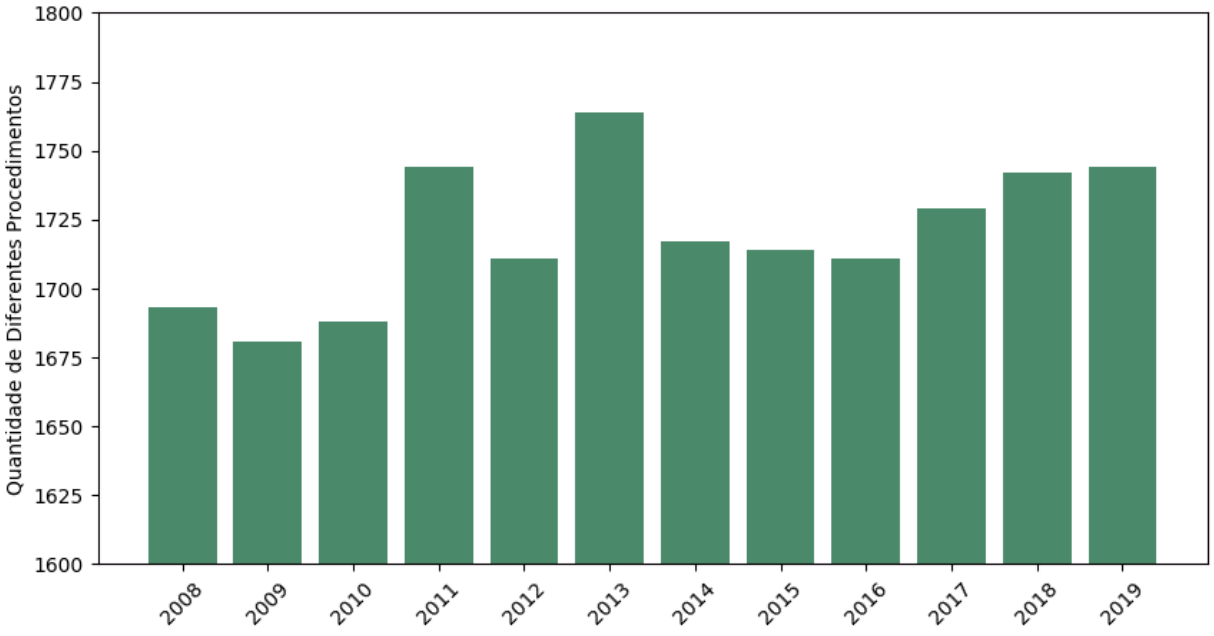
A tabela RD possui a informação do município de residência do paciente (campo MUNIC_RES) e a tabela BR_REGSAUD possui a relação das regiões de saúde e dos municípios que pertencem a cada uma delas.

4.3.4 **Procedimento Realizado**

Existem dois campos de procedimento na tabela RD: PROC_REA e PROC_SOLIC.

O campo PROC_REA indica o procedimento realizado que deu origem a AIH. A figura abaixo mostra o quantitativo de diferentes procedimentos que deram origem a AIH no decorrer dos anos.

Figura 8: Quantitativo de Diferentes Procedimentos Realizados x Ano



Fonte: Elaborada pela autora (2020).

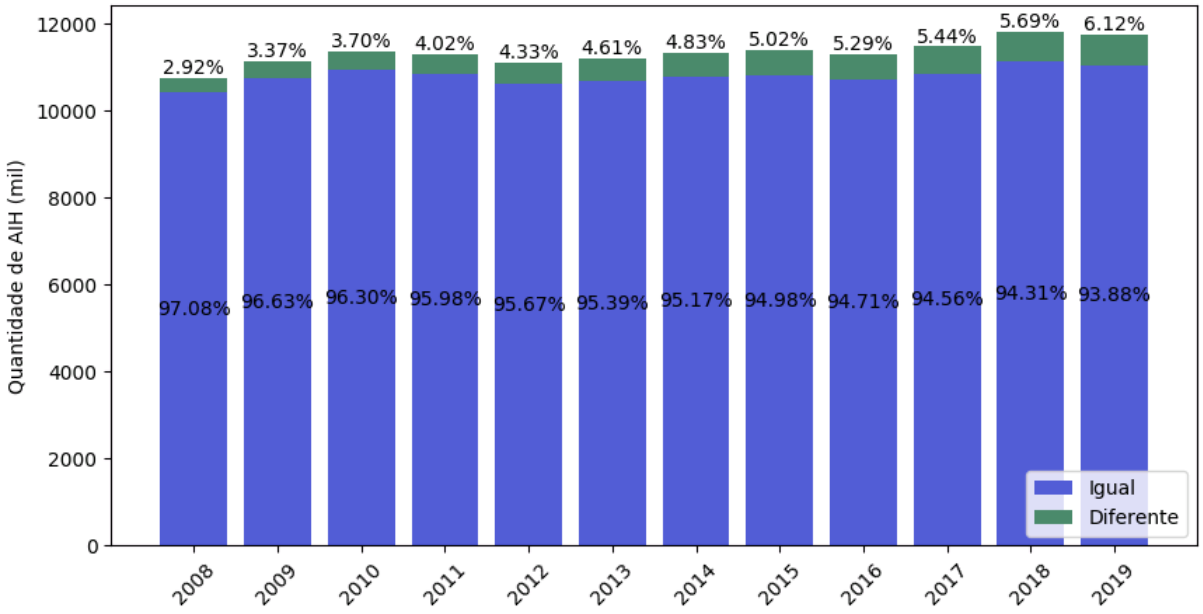
O campo PROC_SOLIC indica o procedimento solicitado na AIH. Geralmente esse campo possui o mesmo valor do campo PROC_REA que foi realmente realizado (PROC_REA), mas em alguns casos isso não acontece, conforme apresentado na Figura 9: Comparação de Procedimento Realizado x Solicitado por Ano.

Essa diferença ocorre, porque, durante a internação, a hipótese diagnóstica inicial pode não ser confirmada ou pode surgir uma condição clínica superveniente, ou ainda, ser identificada outra patologia de maior gravidade, complexidade ou intercorrência que implique na necessidade de mudança de procedimento. Assim, o código do procedimento original fica

registrado no campo PROC_SOLIC e o código do novo procedimento fica registrado no campo PROC_REA (MINISTÉRIO DA SAÚDE, 2017).

A figura abaixo mostra a comparação, por ano, entre o quantitativo de AIHs com procedimento realizado igual ao procedimento solicitado em relação ao quantitativo de AIHs que esses procedimentos são diferentes. Pode-se observar que o percentual desses procedimentos sendo diferentes tem aumentado no decorrer dos anos.

Figura 9: Comparação de Procedimento Realizado x Solicitado por Ano



Fonte: Elaborada pela autora (2020).

Nas análises realizadas durante o trabalho, foi utilizada a informação do procedimento realizado (PROC_REA).

4.3.5 **Serviço**

Em 2008, a partir da Portaria SAS n° 3848/07, foi implantada a Tabela Unificada de Procedimentos, Medicamentos, Órteses e Próteses e Materiais Especiais do SUS. Essa tabela é gerada pelo Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM do SUS (SIGTAP) e tem a função de unir os procedimentos médicos cobrados pelos sistemas SIH/SUS e SIA/SUS3 servindo como mais um esforço para a integração das bases de dados nacionais (SANTOS, 2009).

[3] SIA/SUS: Sistema de Informação Ambulatorial do SUS

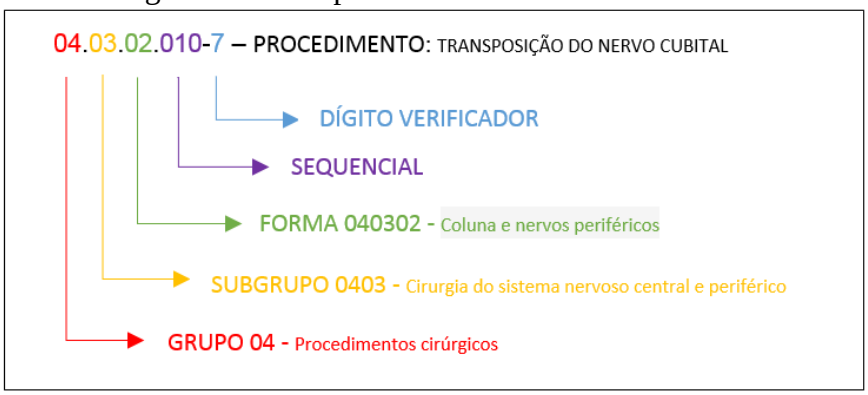
O número do procedimento no SIGTAP é composto por dez dígitos e contém as informações detalhadas na **Erro! Fonte de referência não encontrada.** e exemplificada na Figura 10: Exemplo de Número de Procedimento.

Tabela 3: Composição do número do procedimento

DÍGITO	SIGNIFICADO	QUANTIDADE NO SIGTAP	EXEMPLOS
1° e 2°	Grupo do SIGTAP ao qual o procedimento pertence.	8 grupos	‘Procedimentos cirúrgicos’, ‘Transplantes de órgãos’ e ‘Órteses, próteses e materiais especiais’
1° ao 4°	Subgrupo do SIGTAP ao qual o procedimento pertence	59 subgrupos	'Cirurgia do aparelho da visão', 'Tratamento de lesões, envenenamentos e outros, decorrentes de causas externas', 'Diagnóstico em vigilância epidemiológica e ambiental' e 'Hemoterapia'
1° ao 6°	Forma do SIGTAP ao qual o procedimento pertence	382 formas	'Anestesiologia', 'Cirurgia em nefrologia', 'Coleta e exames para fins de doação de orgãos, tecidos e células e de transplante' e 'Avaliação de morte encefálica'
1° ao 10°	Procedimento do SIGTAP	5182 procedimentos	‘BIOPSIA CIRURGICA DE TIREOIDE’, ‘AVALIACAO VOCAL’, ‘MAMOPLASTIA PÓS-CIRURGIA BARIÁTRICA’ e ‘ARTROPLASTIA PARCIAL DE QUADRIL’

Fonte: Elaborada pela autora (2020).

Figura 10: Exemplo de Número de Procedimento



Fonte: Elaborada pela autora (2020).

No exemplo da figura acima, o procedimento ‘Transposição do nervo cubital’ (número 0403020107) pertence à forma ‘Coluna e nervos periféricos’ (número 040302), que pertence ao subgrupo ‘Cirurgia do sistema nervoso central e periférico’ (número 0403), que, por sua vez, pertence ao grupo ‘Procedimentos cirúrgicos’ (número 04).

No trabalho, foi realizada análise em todos os níveis de abstração de serviço acima, ou seja, o procedimento 0403020107 foi analisado, todos os procedimentos da forma 040302 (incluindo o procedimento 0403020107) foram analisados, todos os procedimentos do subgrupo 0403 (incluindo os procedimentos da forma 040302) foram analisados e todos os procedimentos do grupo 04 (incluindo os procedimentos do subgrupo 0403) foram analisados.

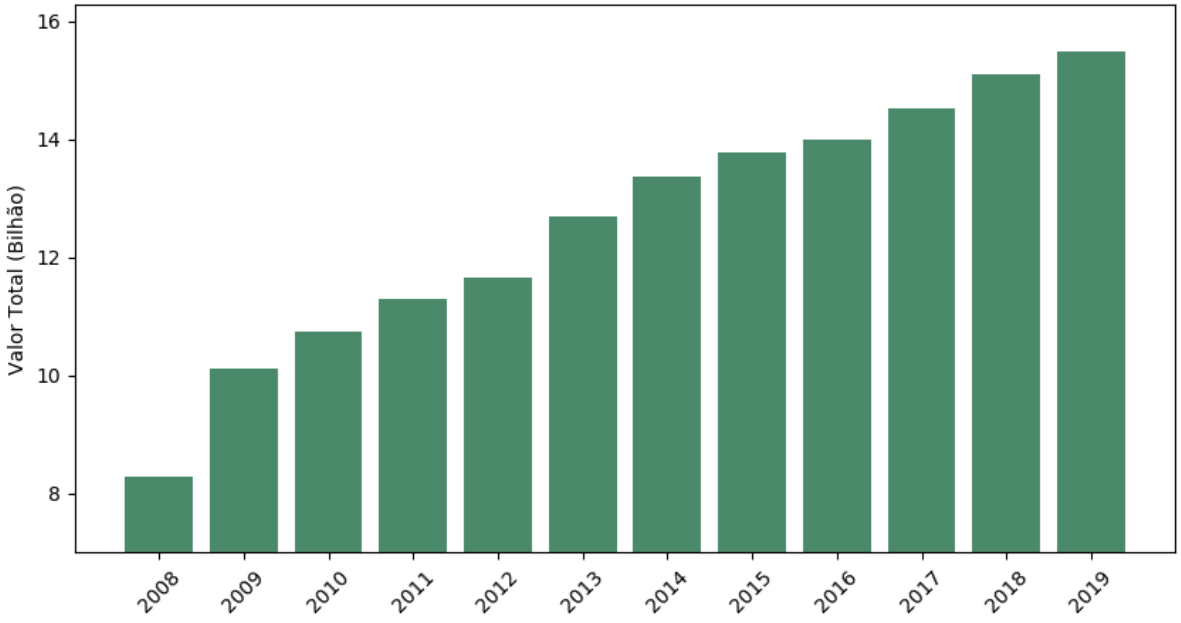
Cada procedimento, forma, subgrupo ou grupo analisado foi considerado um serviço no trabalho.

4.3.6 **Valor Total da AIH**

O campo VAL_TOT da tabela RD indica a soma total dos valores gastos com a AIH. Esse valor engloba os valores com serviços profissionais (médicos e cirurgiões dentistas) e serviços hospitalares - diárias, taxas de salas, alimentação, higiene pessoal, de apoio ao paciente no leito, materiais hospitalares, medicamentos, Serviços Auxiliares de Diagnose e Terapia (SADT), exceto medicamentos especiais e SADT especiais, e serviços profissionais, exceto médicos e cirurgiões dentistas.

A Figura 11: Gráfico do Valor Total das AIHs x Ano mostra a evolução dos valores totais gastos com internações hospitalares no decorrer dos anos.

Figura 11: Gráfico do Valor Total das AIHs x Ano



Fonte: Elaborada pela autora (2020).

5 PREPARAÇÃO DOS DADOS

A fase de preparação dos dados consiste em deixar os dados prontos para a fase de modelagem. É a construção de um conjunto de dados obtidos dos dados brutos iniciais, porém que passaram pela limpeza e transformação necessárias para a próxima etapa.

Nesta fase, foram realizadas as seguintes atividades:

a) Seleção dos dados;

b) Limpeza de dados;

c) Criação de coluna derivadas.

5.1. IBGE

Os dados do IBGE referentes à população dos municípios foram carregados para um *dataframe* da biblioteca *pandas*, que consistia inicialmente das colunas COD_MUNICIPIO (formada pelos seis primeiros dígitos da coluna COD_IBGE, ou seja, o código do município sem o dígito verificador, para ficar compatível com a informação de município contida no SIH/SUS), NM_MUNICIPIO, UF, LATITUDE, LONGITUDE e POPULACAO.

A partir dessas informações, foram adicionadas ao *dataframe* as colunas POPULACAO_UF e POPULACAO_BRASIL, calculadas a partir da soma da população dos municípios que compõe cada ente.

Em um novo ciclo da metodologia, acrescentou-se ao *dataframe* a região de saúde ao qual o município pertencia e a coluna POPULACAO_REGSAUD, calculada a partir da soma da população dos municípios que compõe a região de saúde.

5.2. TABELAS AUXILIARES

Criou-se um *dataframe* da biblioteca *pandas* com a junção de 4 (quatro) tabelas auxiliares: TB_SIGTAP, TB_FORMA, TB_GRUPO, TB_SUBGR, que contêm, respectivamente, a descrição dos procedimentos, das formas, dos grupos e dos subgrupos do SIGTAP.

Criou-se outro *dataframe* da biblioteca *pandas* com a relação dos CNES de todos os hospitais do país com seu respectivo município e sua região de saúde.

5.3. TABELA RD DO SIH/SUS

A tabela RD é formada pelas informações das internações hospitalares financiadas pelo SUS. Cada linha corresponde a uma AIH.

Criou-se um *dataframe* da biblioteca *pandas* a partir dos dados dessa tabela, filtrados pela coluna IDENT=1 (a fim de excluir as AIHs do Tipo 5 para evitar duplicação de informação), agrupados por ANO_CMPT (ano de competência da AIH), PROC_REA (procedimento realizado que deu origem à internação) e MUNIC_RES (município de residência do paciente), e com as colunas agregadas QTD (quantidade de AIHs dos pacientes do município MUNIC_RES que o procedimento PROC_REA deu causa à internação no ano ANO_CMPT) e VL_TOTAL (soma dos valores totais das AIHs dos pacientes do município MUNIC_RES que o procedimento PROC_REA deu causa à internação no ano ANO_CMPT).

Para o trabalho, em acordo com a SecexSaúde, foram selecionadas as linhas referentes ao ano de 2018, totalizando um *dataframe* com 1.071.918 linhas.

Realizou-se uma junção desse *dataframe* com os dados do IBGE (descritos no item 5.1 IBGE).

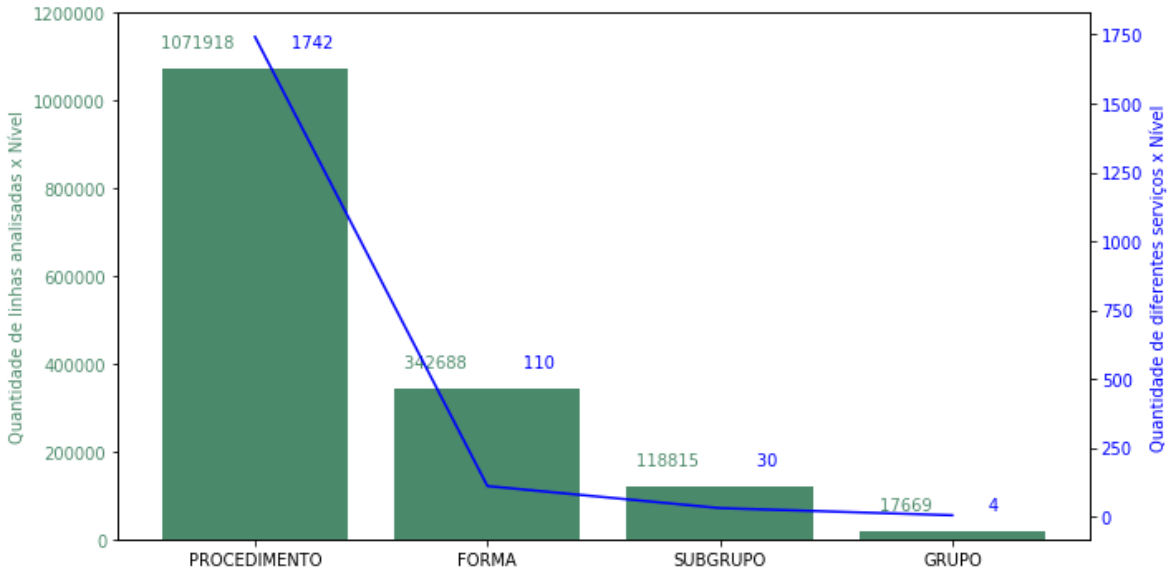
No trabalho, além da análise no nível dos procedimentos que deram causa às internações, foram analisados também outros níveis de abstração de serviço: forma, grupo e subgrupo.

Por esse motivo, foram adicionadas ao *dataframe* linhas referentes a esses três níveis, totalizando um *dataframe* com 1.551.090 linhas.

Então, por exemplo, para o procedimento de número 0403020107 (Figura 10: Exemplo de Número de Procedimento), além de uma linha por município que utilizou esse procedimento em 2018, no *dataframe*, foram adicionadas também uma linha para cada município que utilizou um procedimento da forma de número 040302 (que engloba o procedimento 0403020107), uma linha para cada município que utilizou um procedimento do subgrupo de número 0403 (que engloba a forma 040302) e uma linha para cada município que utilizou um procedimento do grupo de número 04 (que engloba o subgrupo 0403).

A figura abaixo mostra o número de linhas do *dataframe* e a quantidade de diferentes serviços que foram utilizados no SIH/SUS em 2018 por nível de abstração de serviço (procedimento, forma, subgrupo ou grupo).

Figura 12: *Dataframe* detalhado x Nível de abstração de serviço



Fonte: Elaborada pela autora (2020).

Foram adicionadas ao *dataframe* as colunas derivadas listadas na tabela a seguir.

Tabela 4: Colunas derivadas

COLUNA	SIGNIFICADO	CÁLCULO
COD_FORMA	Forma SIGTAP do Procedimento	6 (seis) primeiros dígitos da coluna PROC_REA
COD_SUBGRUPO	Subgrupo SIGTAP do Procedimento	4 (quatro) primeiros dígitos da coluna PROC_REA
COD_GRUPO	Grupo SIGTAP do Procedimento	2 (dois) primeiros dígitos da coluna PROC_REA
NIVEL	Nível de abstração do serviço, que indica se a linha do *dataframe* se refere a um procedimento, uma forma, um subgrupo ou um grupo	Valores = {'PROCEDIMENTO', 'FORMA', 'SUBGRUPO', 'GRUPO'}
QTD_UF	Quantidade de atendimentos cujo	Soma da quantidade de AIHs cujo procedimento / forma / subgrupo / grupo

	procedimento / forma / subgrupo / grupo (dependendo da coluna NIVEL) deu origem à internação na UF	(dependendo da coluna NIVEL) deu origem à internação na UF
QTD_BRASIL	Quantidade de atendimentos cujo procedimento / forma / subgrupo / grupo (dependendo da coluna NIVEL) deu origem à internação no Brasil	Soma da quantidade de AIHs cujo procedimento / forma / subgrupo / grupo (dependendo da coluna NIVEL) deu origem à internação no Brasil
TX	Taxa de atendimentos por habitante do serviço na localidade (município ou região de saúde)	TX = QTD * Fator de habitantes / POPULACAO, onde QTD = quantidade de AIHs cujo procedimento / forma / subgrupo / grupo (dependendo da coluna NIVEL) deu origem à internação na localidade; Fator de habitantes = 100 e POPULACAO = população da localidade
TX_UF	Taxa de atendimentos por habitante do serviço na UF	TX_UF = QTD_UF * Fator de habitantes / POPULACAO_UF, onde QTD_UF = quantidade de AIHs cujo procedimento / forma / subgrupo / grupo (dependendo da coluna NIVEL) deu origem à internação na UF; Fator de habitantes = 100 e POPULACAO_UF = população da UF
TX_BRASIL	Taxa de atendimentos por habitante do serviço no Brasil	TX_UF = QTD_BRASIL * Fator de habitantes /POPULACAO_BRASIL, onde QTD_BRASIL = quantidade de AIHs cujo procedimento / forma / subgrupo / grupo (dependendo da coluna NIVEL) deu origem à internação no Brasil; Fator de habitantes = 100 e POPULACAO_BRASIL = população do Brasil

DESCRICAO	Descrição do serviço	Descrição referente ao procedimento, forma, subgrupo ou grupo (dependendo da coluna NIVEL), encontrada no *dataframe* das descrições do SIGTAP (item 5.2) Exemplo: Nível = 'PROCEDIMENTO', procedimento = '0308020030': DESCRICAO = 'Procedimento: 0308020030: TRATAMENTO DE INTOXICACAO OU ENVENENAMENTO POR EXPOSICAO A MEDICAMENTO E SUBSTANCIAS DE USO NAO MEDI'
DESCRICAO_COMPLETA	Descrição do serviço levando em conta toda a hierarquia dos níveis de serviço (GRUPO > SUBGRUPO > FORMA > PROCEDIMENTO)	Descrições referentes ao procedimento, forma, subgrupo ou grupo (dependendo da coluna NIVEL) e dos níveis de abstração acima, encontradas no *dataframe* das descrições do SIGTAP (item 5.2) Exemplo: Nível = 'SUBGRUPO', subgrupo = '0401': DESCRICAO_COMPLETA = 'Grupo: 04: Procedimentos cirúrgicos - SubGrupo: 0401: Pequenas cirurgias e cirurgias de pele, tecido subcutâneo e mucosa'

Fonte: Elaborada pela autora (2020).

6 MODELAGEM

Na fase de modelagem são selecionadas e aplicadas as técnicas de Data Mining mais apropriadas, com base nos objetivos identificados na fase de entendimento do negócio.

Nesta fase foram realizadas as seguintes atividades:

a) Seleção dos algoritmos a serem usados no teste de normalidade dos dados;
b) Verificação da distribuição estatística dos dados analisados;
c) Seleção dos métodos de detecção de *outliers* usados na análise;
d) Detecção de *outliers* usando métodos estatísticos e baseados em proximidade;
e) Avaliação das análises de detecção de *outliers*.

6.1. TESTE DE NORMALIDADE

De acordo com Griffiths (2008), a distribuição normal é uma curva simétrica, na forma de uma curva de sino, em que a maioria das observações é encontrada no centro da curva e a densidade de probabilidade diminui à medida que se distancia da média. Tanto a média quanto a mediana estão no centro e têm a maior densidade de probabilidade. Ela possui a forma de uma curva de sino.

Os testes de normalidade são utilizados para verificar se a distribuição de probabilidade associada a um conjunto de dados pode ser aproximada pela distribuição normal. (ESTATCAMP, 20–)

No trabalho, o teste de normalidade foi feito usando duas funções do módulo de funções estatísticas *stats* da biblioteca *scipy* (SCIPY DEVELOPERS, c2020) do *Python*: *shapiro* – para os casos que o tamanho da amostra era menor de 5000 (THE SCIPY COMMUNITY, 2019) – e *anderson* – para todos os casos. Esses testes representam, respectivamente, os testes de normalidade Shapiro-Wilk e Anderson-Darling.

Em ambas as funções, foi considerado que os dados seguiam uma distribuição normal quando o *p-value*[4] calculado era maior que 0,05, indicando que a hipótese de que os dados seguiam uma distribuição normal não podia ser rejeitada. Esse valor de p-value é tipicamente

[4] *p-value*: é a probabilidade de se obter uma estatística de teste igual ou maior que aquela observada em uma amostra, sob a hipótese nula.

utilizado e referenciado em vários livros e artigos, como Alaudeen, England e Chopra (2019), Yu-Wei e Bhatia (2017) e Laurae (2016).

Foi considerado que os dados de uma amostra seguiam uma distribuição normal quando passavam nos testes das duas funções simultaneamente (no caso de amostras com tamanho entre 48 e 5000) e quando passavam no teste de Anderson-Darling (no caso de amostras maiores do que 5000), conforme apresentado no APÊNDICE A – DEFINIÇÃO DO TESTE DE NORMALIDADE deste trabalho.

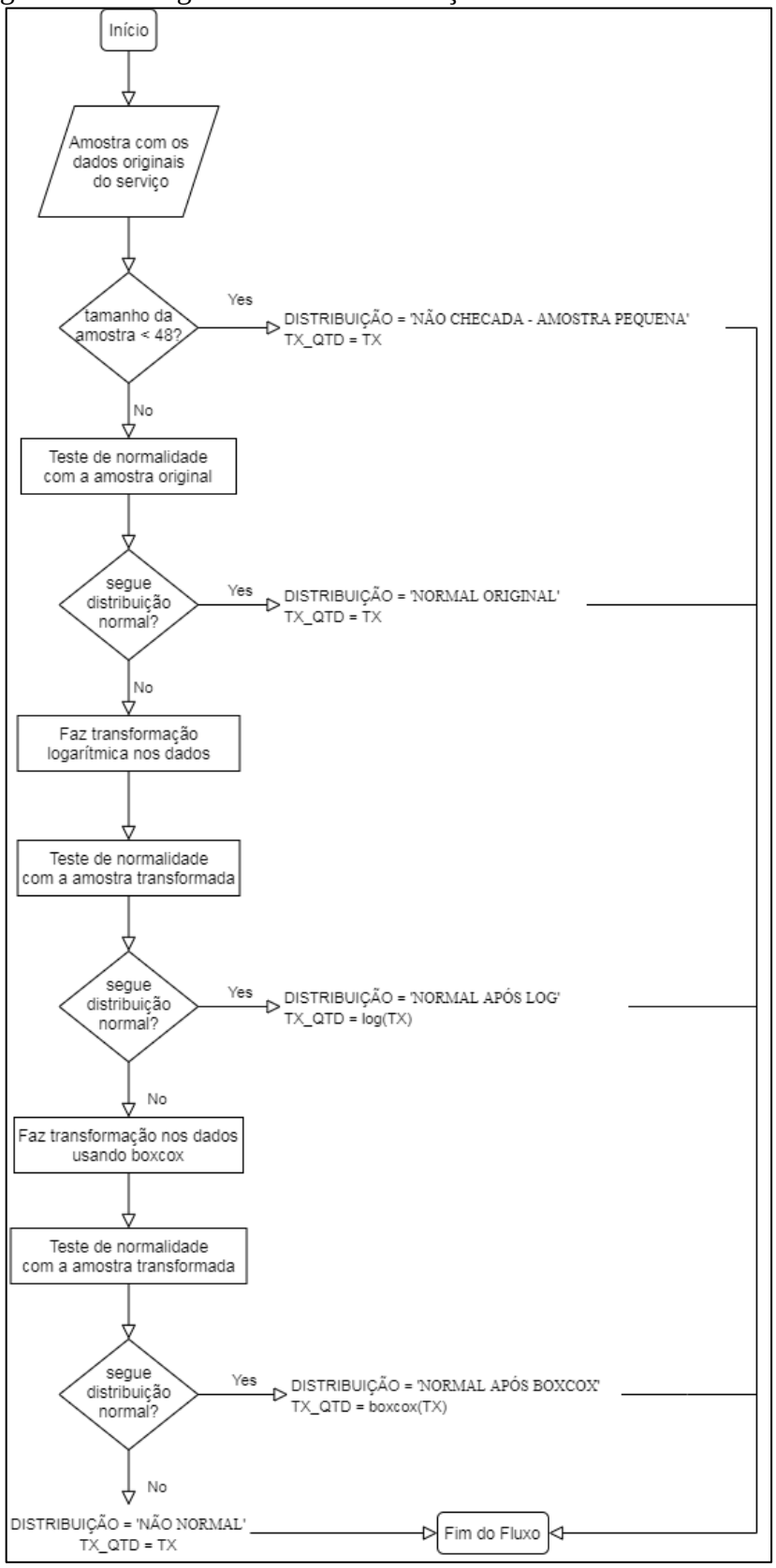
6.2. DISTRIBUIÇÃO DOS DADOS DA ANÁLISE

Alguns métodos usados na análise de detecção de *outliers*, chamados de métodos estatísticos (descritos no item 6.3 DETECÇÃO DE OUTLIERS), pressupõe um certo tipo de distribuição dos dados.

Dessa forma, foi necessário realizar testes de normalidade e/ou transformações nos dados, conforme descrito na Figura 13: Fluxograma das Transformações e Testes de Normalidade, para tentar chegar numa amostra de dados que seguia uma distribuição de dados previamente conhecida (no caso do trabalho, uma distribuição normal) para poder utilizar esses métodos.

Tan, Steinbach e Kumar (2009) reforçam a importância de identificar o tipo de distribuição de dados ao utilizar técnicas estatísticas para detecção de *outliers*, pois embora a maioria dos dados possam ser representados pelas distribuições mais comuns, como a Gaussiana, de Poisson ou binominal, alguns conjuntos de dados seguem uma distribuição atípica, e se uma distribuição errada for considerada, a detecção de *outlier* pode se tornar imprecisa.

Figura 13: Fluxograma das Transformações e Testes de Normalidade



Fonte: Elaborada pela autora (2020).

O fluxo acima foi realizado para cada serviço (procedimento, forma, subgrupo ou grupo) que deu origem a uma internação hospitalar financiada pelo SUS. Foram criadas as colunas ‘DISTRIBUICAO’ e ‘TX_QTD’ para receberem, respectivamente, o valor do tipo de distribuição da amostra dos dados e o valor da taxa a ser usado na análise de detecção de *outliers*.

Para cada serviço, foram selecionados os dados da coluna ‘TX’, que representavam a amostra original dos dados do serviço.

7. Primeiramente, era verificado se o tamanho dessa amostra era menor do que 48 (conforme APÊNDICE A – DEFINIÇÃO DO TESTE DE NORMALIDADE). Em caso positivo, a coluna ‘DISTRIBUICAO’ era preenchida com o valor ‘NÃO CHECADA - AMOSTRA PEQUENA’ e a coluna ‘TX_QTD’ era preenchida com os valores originais da coluna ‘TX’. Em caso negativo, o teste de normalidade descrito no item 6 TESTE DE NORMALIDADE era realizado com a amostra original dos dados do serviço.

Caso esses dados seguissem uma distribuição normal, a coluna ‘DISTRIBUICAO’ era preenchida com o valor ‘NORMAL ORIGINAL’ e a coluna ‘TX_QTD’ era preenchida com os valores da coluna ‘TX’.

No caso de a distribuição não ser normal, verificava-se se os dados seguiam uma distribuição LOG-NORMAL. Para isso, era feita uma transformação logarítmica nos dados da amostra e aplicado o teste de normalidade sobre eles, já que uma distribuição log-normal transformada se torna uma distribuição normal (GARNER, 2015).

Em caso positivo (dos dados seguirem uma distribuição normal após a transformação logarítmica), a coluna ‘DISTRIBUICAO’ era preenchida com o valor ‘NORMAL APÓS LOG’ e a coluna ‘TX_QTD’ era preenchida com os valores de log da coluna ‘TX’.

No caso de não passar em nenhum dos testes de normalidade acima (normal e log-normal), usava-se a função *boxcox* (Boxcox) do módulo de funções estatísticas *stats* da biblioteca *scipy* do *Python* para tentar transformar a distribuição dos dados da coluna ‘TX’ em uma distribuição normal. Após passar pela transformação usando a função *boxcox*, os dados eram submetidos ao teste de normalidade para ver se estavam realmente seguindo uma distribuição normal. É importante destacar que nem sempre a transformação usando *boxcox* consegue converter os dados em uma distribuição normal (MCNEESE, 2016)

Em caso positivo (dos dados seguirem uma distribuição normal após a transformação usando *boxcox*), a coluna 'DISTRIBUICAO' era preenchida com o valor 'NORMAL APÓS BOXCOX' e a coluna 'TX_QTD' era preenchida com os valores de retorno da coluna 'TX' após transformação *boxcox*.

No caso de não conseguir transformar os dados para que eles seguissem uma distribuição normal, a coluna 'DISTRIBUICAO' era preenchida com o valor 'NÃO NORMAL' e a coluna 'TX_QTD' era preenchida com os valores da coluna 'TX'.

7.1. DETECÇÃO DE *OUTLIERS*

A detecção de *outlier* é o processo de encontrar objetos de dados com comportamento muito diferente do esperado. Esses objetos são chamados *outliers*.

Um *outlier* também pode ser entendido como um objeto que se desvia de forma significativa dos demais, como se fosse produzido por um mecanismo diferente, ou tenha recebido interferência externa ao processo normal de seu domínio (KAMBER, PEI e HAN, 2011). Ele também é conhecido como ponto fora da curva, valor atípico ou anomalia.

Existem algumas formas de categorizar os métodos de detecção de *outliers*.

Uma delas divide os métodos em: supervisionados (em que os dados usados para treino são rotulados indicando se é um dado normal ou uma anomalia), não supervisionados (em que os dados não possuem rótulo identificando se o dado é normal ou é uma anomalia) ou semi-supervisionados (em que os dados usados para treino estão todos rotulados como dados normais) (HALDER e OZDEMIR, 2018).

No trabalho, foram usados métodos não supervisionados, já que os dados não possuíam rótulo indicando se o dado era normal ou era uma anomalia. Esses métodos assumem que os dados normais são mais frequentes do que os dados anômalos.

Outra forma de categorizar os métodos de detecção de *outliers* é classificá-los em: estatísticos, baseados em proximidade e baseados em cluster.

Os métodos estatísticos (também conhecidos como métodos baseados em modelo) fazem suposições sobre a normalidade dos dados. Eles assumem que os objetos de dados considerados normais são gerados por um modelo estatístico (estocástico) e que os dados que não seguem o modelo são *outliers*. Dessa forma, uma estratégia para detecção de *outlier* seria analisar a probabilidade de objetos pertencerem ou se adaptarem ao modelo. Objetos com baixa

probabilidade são candidatos a *outliers*. Uma vantagem de usar métodos estatísticos é que a detecção de *outliers* pode ser estatisticamente justificada (KAMBER, PEI e HAN, 2011).

Os métodos estatísticos para a detecção de *outliers* podem ser divididos em duas categorias principais: métodos paramétricos e métodos não paramétricos. As técnicas paramétricas assumem uma distribuição conhecida das observações e estimam os parâmetros dos dados (ESKIN, 2000 APUD FREITAS; IGOR, 2019), como, por exemplo, os métodos boxplot e Z-Score. As técnicas não paramétricas não assumem um modelo estatístico a priori. Elas tentam determinar o modelo a partir dos dados de entrada (KAMBER, PEI e HAN, 2011). O histograma é um exemplo de método estatístico não paramétrico.

Os métodos baseados em proximidade entendem que os dados têm uma relação de vizinhança. Eles assumem que determinado objeto é anômalo, caso seus vizinhos mais próximos estejam suficientemente distantes, ou seja, quando esse objeto se distância consideravelmente de toda sua vizinhança.

Esses métodos são divididos em duas categorias: métodos baseados em distância e métodos baseados em densidade. Os métodos baseados em distância observam os vizinhos mais próximos em um determinado raio. Um objeto é considerado anômalo quando sua vizinhança tem um número baixo de instâncias. Os métodos baseados em densidade observam a densidade de um objeto em relação aos seus vizinhos. Objetos anômalos têm uma densidade menor se comparados com sua vizinhança.

Os métodos baseados em agrupamento assumem que os objetos considerados normais são aqueles que fazem parte de grandes clusters, enquanto que os objetos que compõem os pequenos clusters são os *outliers*.

No trabalho, foram utilizados métodos estatísticos e métodos baseados em proximidade (baseados em densidade), detalhados nos itens a seguir.

6.3.1 **Z-Score**

Um método estatístico simples para a detecção de *outliers* univariados é o Z-Score, também conhecido como Teste Z ou Escore Padronizado. Este método se baseia na propriedade de que os dados seguem uma distribuição normal (TOCCI e JAIME, 2018).

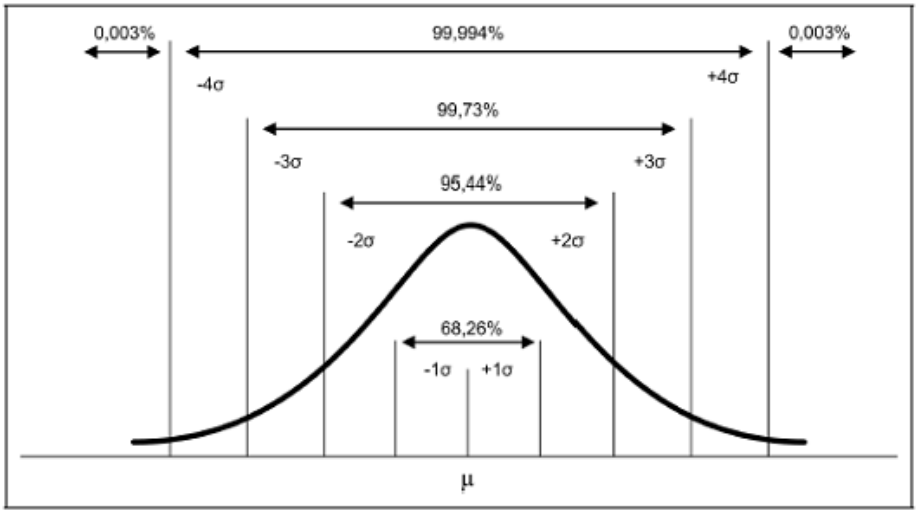
O Z-Score mede exatamente quantos desvios padrões um ponto está acima ou abaixo da média, ou seja, especifica a localização precisa do ponto na distribuição. Para calcular o z-score de cada objeto do conjunto de dados analisado, a fórmula abaixo foi aplicada:

z-score do objeto = (objeto - média) / desvio

Dela, pode-se inferir que um z-score com valor positivo significa que o objeto está acima da média, um z-score com valor negativo significa que o objeto está abaixo da média e um z-score com valor próximo a zero significa que o objeto tem um valor próximo à média.

Neste método, um limite é definido para classificar o dado como normal ou *outlier*, ou seja, para um objeto do conjunto de dados analisado ser considerado anômalo, o valor absoluto de seu z-score deve ser maior do que esse limite. Geralmente, usa-se o valor de três desvios padrões da média para esse limite (DESHPANDE e KOTU, 2018) (Z-scores review). A figura abaixo mostra que usando esse limite de três, que foi o valor escolhido para execução deste trabalho, 99,73% dos dados são considerados normais.

Figura 14: Porcentagem da população x Quantidade de desvio padrão da média



Fonte: (Distribuição Normal)

Segundo Tocci e Jaime (2018), existem dois agravantes que tornam o método não tão robusto:

- A média e o desvio padrão são altamente influenciados pelos valore dos *outliers.*
- O método tem um comportamento impreciso em bases de dados pequenas. Ele nunca detecta um *outlier* para um conjunto de dados com menos de 11 pontos.

6.3.2 **Z-Score modificado**

O segundo método utilizado foi uma variação do z-score descrito no item acima. Este método estatístico usa a mediana e o desvio absoluto da mediana (MAD – *Median of the absolute deviation*) no lugar da média e do desvio padrão, respectivamente (TOCCI e JAIME,

2018). Dessa forma, o valor calculado para um ponto não sofre interferência dos valores dos *outliers* como no Z-Score.

Para calcular o z-score modificado de cada objeto do conjunto de dados analisado, a fórmula abaixo foi aplicada:

MAD = mediana (|objeto – mediana|) para todos os objetos do conjunto de dados

z-score modificado do objeto= 0.6745*(objeto - mediana) / MAD

No trabalho, o limite para classificar o dado como normal ou *outlier* para este método continuou sendo três.

6.3.3 **IQR**

O terceiro método utilizado é um método estatístico conhecido como IQR, que se baseia na amplitude interquartil (IQR – *Interquartile Range*).

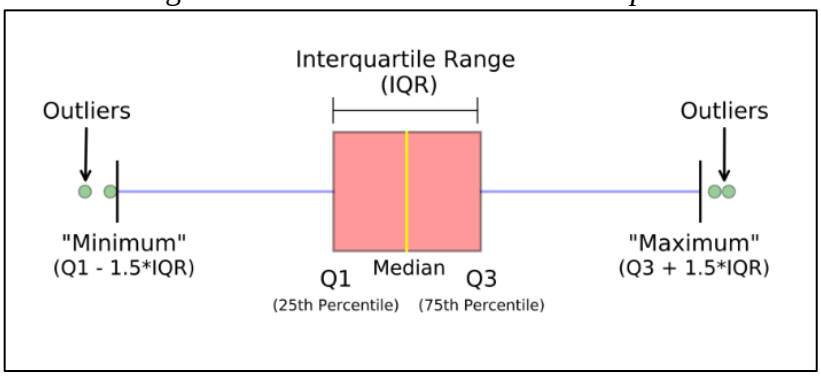
Ele utiliza o resumo de cinco números: menor número do conjunto de dados (Min); mediana entre o menor número e a mediana geral (Q1 – quartil inferior); mediana de todo conjunto de dados (Q2); mediana entre o maior número e a mediana geral (Q3 – quartil superior); e maior do conjunto de dados.

O IQR é definido pelo valor Q3 – Q1. Esse intervalo corresponde a 50% dos números do conjunto de dados analisado.

Qualquer objeto que seja maior do que Q3 + 1.5 x IQR ou menor do que Q1 – 1.5 x IQR é considerado *outlier*. Os demais objetos são considerados normais. O intervalo entre Q1 – 1.5 x IQR e Q3 + 1.5 x IQR equivale a uma distância de três desvios padrões, que foi o limite escolhido no trabalho para o método Z-Score.

Esse é o método usado para plotar o gráfico *boxplot*, conforme figura abaixo.

Figura 15: Medidas do Gráfico de *Boxplot*



Fonte: (CARMONA)

6.3.4 **Fator *Outlier* Local (LOF)**

LOF é um método de detecção de *outlier* baseado em densidade, que atribui para cada objeto um fator de anormalidade local, resultado da diferença da densidade do objeto com a densidade da sua vizinhança. É local, pois a pontuação da anomalia depende de quão isolado o objeto está em relação à sua vizinhança. Mais precisamente, a localidade é dada pelos k-vizinhos mais próximos, cuja distância é usada para estimar a densidade local

A ideia principal deste método é comparar a densidade em torno de um objeto com a densidade em torno de seus vizinhos locais. A suposição básica dos métodos de detecção de *outlier* baseados em densidade é que a densidade em torno de um objeto normal é semelhante à densidade em torno de seus vizinhos, enquanto a densidade em torno de um objeto *outlier* é significativamente diferente da densidade em torno de seus vizinhos (KAMBER, PEI e HAN, 2011).

No trabalho, foi usada a classe *LocalOutlierFactor* do módulo *neighbors* da biblioteca *sklearn* (API Sklearn) do Python, com os seguintes valores de entrada setados: n_neighbors=20 (valor *default*), contamination=0.1 (valor *default*).

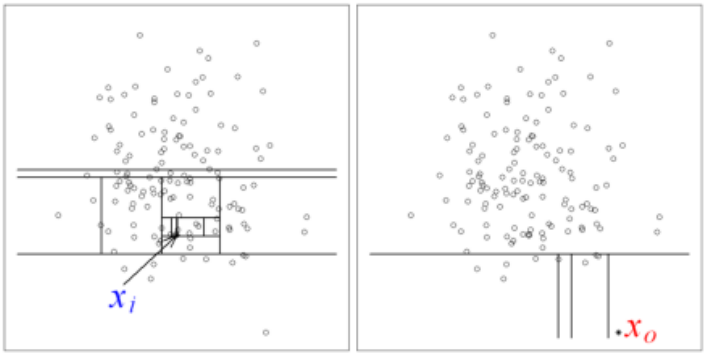
6.3.5 ***Isolation Forest***

O método *Isolation Forest* é um método de detecção de *outlier* ensemble, que se baseia na ideia de que é mais fácil isolar um elemento anômalo do que descrever um ponto de dados normal.

O objetivo deste método é medir o isolamento de um ponto. Para isso ele faz a seleção de uma característica dos dados e, em seguida, constrói uma árvore escolhendo aleatoriamente um valor para dividir os dados entre um conjunto de valores máximo e mínimo do recurso selecionado.

Os *outliers* são detectados quando apresentam um caminho curto entre a raiz da árvore e a folha. Um caminho curto, representa uma separação simples, onde foi utilizada poucas partições para isolar determinado ponto. Os dados normais geralmente precisam de mais partições, como mostra a Figura 16: Isolamento dos pontos com *Isolation Forest,* com dado normal em (a) e um *outlier* em (b). O método considera o caminho médio para determinar se determinado ponto é um *outlier*.

Figura 16: Isolamento dos pontos com *Isolation Forest*



Fonte: (NETO, 2018)

No trabalho, foi usada a classe *IsolationForest* do módulo *ensemble* da biblioteca *sklearn* (API Sklearn) do Python, com o seguinte valor de entrada setado: random_state=np.random.RandomState(123).

É importante destacar que o método LOF, descrito no item anterior, costuma indicar menos *outliers* que o *Isolation Forest* e nem sempre eles indicam os mesmos. Os dois métodos costumam concordar com apenas 1.3% das indicações de anomalias, reforçando que não é uma tarefa fácil a detecção de *outliers* (MOLIN, 2019).

7.2. ANÁLISES REALIZADAS

Durante o trabalho, foram realizadas oito rodadas de análises de detecção de *outliers*.

Cada rodada significa uma combinação de atributos que foram testados, conforme tabela a seguir.

Tabela 5: Resumo dos atributos das rodadas de análise

RODADA	LOCALIDADE	NÍVEL INICIAL DE SERVIÇO	LOCALIDADES ANALISADAS POR SERVIÇO
1	Município	Procedimento	Municípios cuja população usou o serviço
2	Município	Procedimento	Todos os municípios
3	Município	Forma	Municípios cuja população usou o serviço
4	Município	Forma	Todos os municípios

5	Região de Saúde	Procedimento	Regiões de Saúde cuja população usou o serviço
6	Região de Saúde	Procedimento	Todas as regiões de saúde
7	Região de Saúde	Forma	Regiões de Saúde cuja população usou o serviço
8	Região de Saúde	Forma	Todas as regiões de saúde

Fonte: Elaborada pela autora (2020).

A coluna LOCALIDADE indica se as análises da rodada foram feitas no nível do município ou da região de saúde de residência do paciente.

A coluna NÍVEL INICIAL DE SERVIÇO indica se as análises foram a partir do nível do procedimento, ou seja, foram analisados os quatro níveis de abstração de serviço (procedimento, forma, subgrupo e grupo) ou se foram a partir do nível da forma, ou seja, foram analisados apenas os três últimos níveis de serviço (forma, subgrupo e grupo).

A coluna LOCALIDADES ANALISADAS POR SERVIÇO indica se foram analisadas na rodada apenas as localidades cuja população teve acesso ao serviço ou se todas as localidades independentemente de a população ter tido acesso ou não ao serviço.

Nas rodadas em que todas as localidades do país são analisadas, podem-se encontrar localidades que foram consideradas *outliers* por não terem nenhum, terem pouco ou terem muito acesso a um determinado serviço de internação hospitalar. Essas são as rodadas pares: dois, quatro, seis e oito.

Nas demais rodadas, podem-se encontrar localidades que foram consideradas *outliers* por terem pouco ou muito acesso a um determinado serviço de internação hospitalar, mas não são encontradas localidades que não tiveram nenhum acesso a um determinado serviço, porque só são analisadas as localidades cuja população (mesmo que tenha sido apenas um paciente) teve acesso ao serviço.

Todas as análises foram baseadas na coluna TX_QTD, derivada da coluna TX (taxa de atendimentos por habitante do serviço na localidade) conforme descrito no item 6.2 DISTRIBUIÇÃO DOS DADOS DA ANÁLISE.

A taxa de atendimentos por habitante do serviço na localidade é calculada a partir da quantidade de AIHs que um determinado serviço deu origem à internação em uma localidade pela população dessa localidade (TX = QTD * Fator de habitantes / POPULACAO, onde QTD = quantidade

de AIHs cujo serviço deu origem à internação na localidade; Fator de habitantes = 100 e POPULACAO = população da localidade).

A ideia era encontrar a localidade (município ou região de saúde) que tinha uma TX_QTD considerada anômala em relação às demais localidades do país quanto à utilização de um serviço (procedimento, forma, subgrupo ou grupo) que originou uma internação hospitalar financiada pelo SUS.

Uma localidade considerada *outlier* em relação a um serviço era classificada como *outlier* alto (quando sua TX_QTD era maior que o valor médio da TX_QTD de todas as localidades que utilizaram esse serviço) ou como *outlier* baixo (quando sua TX_QTD era menor que o valor médio da TX_QTD de todas as localidades que utilizaram esse serviço).

A Tabela 6: Resumo do Quantitativo de Linhas por Rodada – Análise e Resultado apresenta a quantidade de linhas (total, relativas a procedimentos, formas, subgrupos e grupos) usadas durante cada rodada de análises, a quantidade de linhas que foram consideradas *outliers* (total, altos e baixos) e a quantidade de linhas que não foram consideradas *outliers*. Cada linha analisada representa a TX_QTD de um serviço em uma dada localidade no ano de 2018.

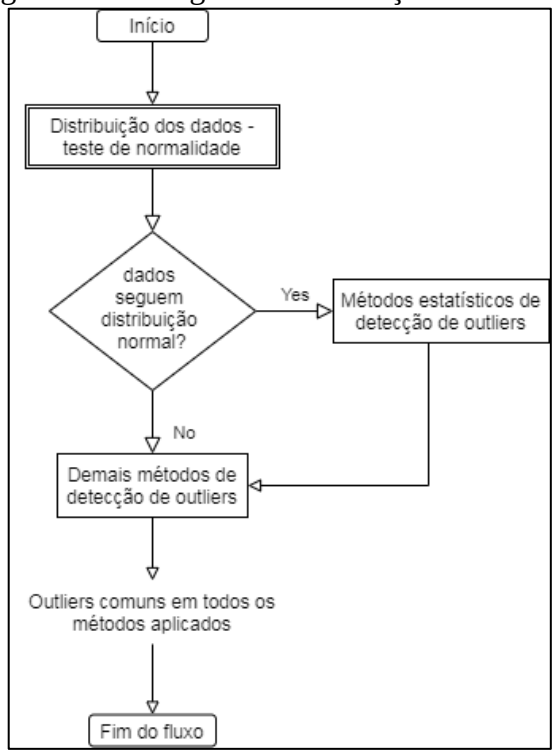
Tabela 6: Resumo do Quantitativo de Linhas por Rodada – Análise e Resultado

QUANTIDADE DE LINHAS	RODADA							
	1	2	3	4	5	6	7	8
TOTAL	1551090	10505020	479172	802080	342581	824182	55972	62928
PROCEDIMENTOS	1071918	9702940	N/A	N/A	286609	761254	N/A	N/A
FORMAS	342688	612700	342688	612700	42042	48070	42042	48070
SUBGRUPOS	118815	167100	118815	167100	12189	13110	12189	13110
GRUPOS	17669	22280	17669	22280	1741	1748	1741	1748
OUTLIERS TOTAL	17837	7219605	5797	85572	5564	241163	974	3615
OUTLIERS ALTO	14307	257622	4292	19219	4543	44969	722	2211
OUTLIERS BAIXO	3530	6961983	1505	66353	1021	196194	252	1404
NÃO SÃO *OUTLIERS*	1533253	3285415	473375	716508	337017	583019	54998	59313

Fonte: Elaborada pela autora (2020).

Em todas as rodadas, o fluxograma definido na Figura 17: Fluxograma de Detecção de *Outliers* foi executado para todos os serviços disponíveis na rodada.

Figura 17: Fluxograma de Detecção de *Outliers*



Fonte: Elaborada pela autora (2020).

O primeiro passo do fluxo consistia na execução do fluxograma definido na Figura 13: Fluxograma das Transformações e Testes de Normalidade. Os dados para a análise (coluna TX_QTD) e a informação de que os dados seguiam ou não uma distribuição normal (coluna DISTRIBUIÇÃO) eram as saídas desse passo.

A Tabela 7: Resumo dos Tipos de Distribuição por Rodada apresenta o resumo da classificação do tipo da distribuição dos dados de todos os serviços analisados por rodada.

Tabela 7: Resumo dos Tipos de Distribuição por Rodada

QUANTIDADE DE SERVIÇOS (POR TIPO DE DISTRIBUIÇÃO)	RODADA							
	1	2	3	4	5	6	7	8
TOTAL ANALISADOS	1886	1886	144	144	1886	1886	144	144
NORMAL ORIGINAL	0	0	0	0	0	0	0	0
NORMAL APÓS LOG	747	0	32	0	791	26	38	14
NORMAL APÓS BOXCOX	378	0	45	0	408	68	71	34
NÃO NORMAL	350	1886	65	144	206	1792	32	96
NÃO CHECADA (AMOSTRA PEQUENA)	411	0	2	0	481	0	3	0

Fonte: Elaborada pela autora (2020).

O segundo passo do fluxo consistia em verificar o tipo de distribuição que a amostra dos dados do serviço que estava sendo analisado seguia. Caso ela seguisse uma distribuição normal, eram realizados os métodos estatísticos (definidos nos itens 6.3.1 Z-Score, 6.3.2 Z-Score modificado e 6.3.3 IQR) e em seguida os demais métodos de detecção de *outliers* (definidos nos itens 6.3.4 Fator *Outlier* Local (LOF) e 6.3.5 *Isolation Forest*). Caso contrário, apenas os demais métodos eram executados.

O último passo era encontrar as localidades que foram consideradas *outliers* em todos os métodos de detecção de *outliers* executados para aquele serviço naquela rodada, ou seja, encontrar as localidades pertencentes a interseção dos resultados de todos os métodos de detecção de *outliers* executados.

6.4.1 **Rodada 1**

Nesta rodada, foram considerados os procedimentos que deram origem às AIHs do SIH/SUS nos municípios do país, totalizando 1886 serviços do SIGTAP, sendo quatro grupos, 30 subgrupos, 110 formas e 1742 procedimentos.

Para cada serviço, foram analisados somente os municípios de residência dos pacientes que tiveram pelo menos uma AIH com o serviço realizado (PROC_REA).

Dos 1742 procedimentos, 22 foram utilizados por apenas um município (Quantidade mínima de municípios que usaram um procedimento), e um procedimento foi utilizado por 5554 municípios (Quantidade máxima de municípios que usaram um procedimento). A tabela abaixo mostra um resumo desses quantitativos pelos quatro níveis analisados: Procedimento, Forma, Subgrupo e Grupo.

Tabela 8: Resumo da utilização do serviço por Nível – Rodada 1

MEDIDA / NIVEL	PROCEDIMENTO	FORMA	SUBGRUPO	GRUPO
TOTAL DE SERVIÇOS ANALISADOS	1742	110	30	4
QUANTIDADE MÍNIMA DE MUNICÍPIOS QUE USARAM UM SERVIÇO	1	34	389	2867
NÚMERO DE SERVIÇOS QUE FORAM USADOS PELA	22	1	1	1

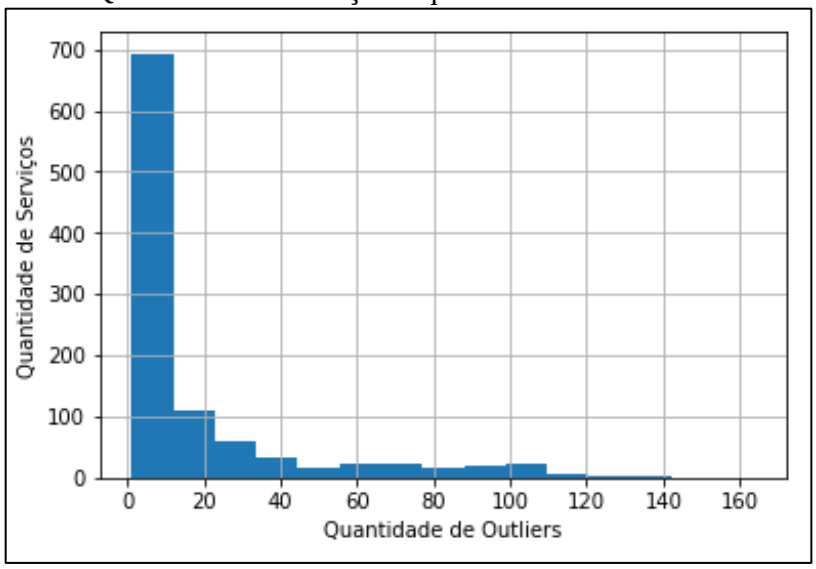
MENOR QUANTIDADE DE MUNICÍPIOS				
EXEMPLOS DE SERVIÇOS QUE FORAM USADOS PELA MENOR QUANTIDADE DE MUNICÍPIOS	0403080061 (Nucleotractomia trigeminal e/ ou espinal); 0403080096 (Tratamento de movimento anormal por estereotaxia com micro-registro);	030109 (Atendimento/ Acompanhamento em saúde do idoso)	0501 (Coleta e exames para fins de doação de orgãos, tecidos e células e de transplante)	02 (Procedimentos com finalidade diagnóstica)
QUANTIDADE MÁXIMA DE MUNICÍPIOS QUE USARAM UM SERVIÇO	5554	5554	5554	5554
NÚMERO DE SERVIÇOS QUE FORAM USADOS PELA MAIOR QUANTIDADE DE MUNICÍPIOS	1	1	1	2
EXEMPLOS DE SERVIÇOS QUE FORAM USADOS PELA MAIOR QUANTIDADE DE MUNICÍPIOS	0310010039 (Parto Normal)	030314 (Tratamento de doenças do ouvido / Apófise Mastóide e Vias Aéreas)	0303 (Tratamentos Clínicos – Outras Especialidades)	03 (Procedimentos Clínicos); 04 (Procedimentos cirúrgicos)
QUANTIDADE MÉDIA DE MUNICÍPIOS QUE USARAM UM SERVIÇO	615,33	3115,34	3960,5	4417,25

Fonte: Elaborada pela autora (2020).

Em relação aos *outliers* encontrados nesta rodada, foram encontradas anomalias em 1026 serviços (dentre os 1886 analisados), sendo que 449 serviços tiveram *outliers* altos (taxas acima da média) e baixos (taxas abaixo da média), 137 serviços tiveram apenas *outliers* baixos e 440 tiveram apenas *outliers* altos.

O histograma a seguir representa a distribuição da quantidade de municípios considerados *outliers* (Quantidade de *outliers*) pela frequência de serviços que tiveram cada uma das faixas de quantidade de *outliers* (Quantidade de serviços). Foram analisadas as quantidades de *outliers* encontrados nos 1026 serviços.

Figura 18: Quantidade de serviços x quantidade de *Outliers* – Rodada 1



Fonte: Elaborada pela autora (2020).

Percebe-se que a maioria desses serviços tiveram poucos *outliers*, ou seja, quase 700 serviços tiveram menos de dez municípios considerados *outliers*

Nesta rodada, a média de *outliers* por serviço foi 17,38, o mínimo de *outliers* por serviço foi 1 e o máximo de *outliers* por serviço foi 164.

6.4.2 **Rodada 2**

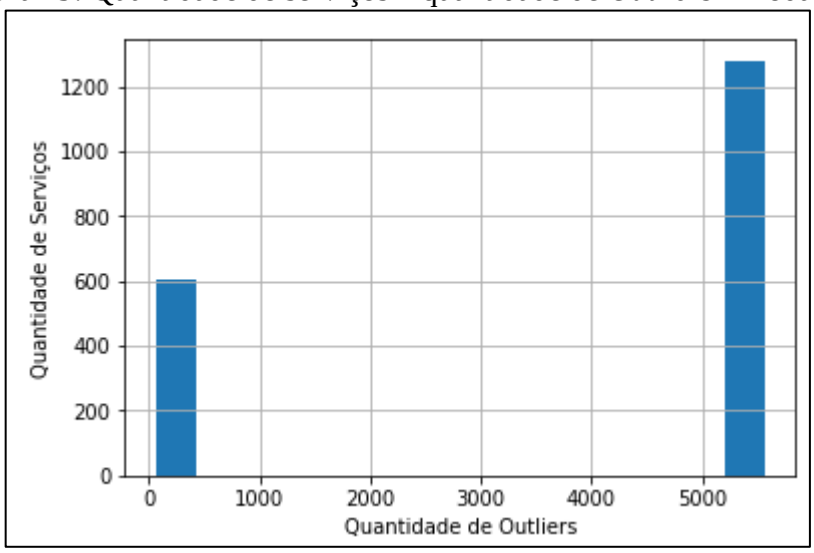
Nesta rodada, para a análise de um serviço, foram também acrescentados os municípios cuja população não teve acesso a esse serviço nas internações hospitalares financiadas pelo SUS. Dessa forma, para cada serviço, foram analisadas as taxas de quantidade para todos os 5570 municípios do país. Os municípios que não utilizaram o serviço da análise, ficaram com taxa zero. Esta rodada foi pensada tentando encontrar os municípios que estavam tendo menos acesso aos serviços de internação hospitalar financiados pelo SUS.

Foram analisados os mesmos 1886 serviços da rodada anterior. Entretanto de 1.551.090 linhas totais analisadas na rodada anterior, passou-se a ter 10.505.020, em que 85,2% estavam com valor zero na quantidade do serviço utilizado pelo município (coluna QTD), e, por consequência, estavam com valor zero na taxa de atendimentos desse serviço no município (coluna TX).

Em relação aos *outliers* desta rodada, foram encontradas anomalias em todos os 1886 serviços analisados, sendo que 1755 serviços tiveram *outliers* altos (taxas acima da média) e baixos (taxas abaixo da média), nenhum serviço teve apenas *outliers* baixos e 131 tiveram apenas *outliers* altos.

O histograma a seguir representa a distribuição da quantidade de municípios considerados *outliers* (Quantidade de *outliers*) pela frequência de serviços que tiveram cada uma das faixas de quantidade de *outliers* (Quantidade de serviços). Foram analisadas as quantidades de *outliers* dos 1886 serviços.

Figura 19: Quantidade de serviços x quantidade de *Outliers* – Rodada 2



Fonte: Elaborada pela autora (2020).

Percebe-se que a maioria dos serviços tiveram muitos *outliers* detectados. Mais de 1200 serviços tiveram mais de 5000 municípios considerados *outliers*.

Nesta rodada, a média de *outliers* por serviço foi 3827,99, o mínimo de *outliers* por serviço foi 60 e o máximo de *outliers* por serviço foi 5570 (que representa todos os municípios).

6.4.3 **Rodada 3**

Nesta rodada, pensou-se em subir um grau de abstração em relação ao nível dos serviços analisados, ou seja, as análises foram feitas a partir do nível da forma, para tentar contornar alguma possível inconsistência no preenchimento dos dados pelos municípios. Por exemplo, um município poderia estar preenchendo o procedimento ‘TIREOIDECTOMIA TOTAL C/ ESVAZIAMENTO GANGLIONAR’ de código ‘0402010051’ com o código correto, enquanto outro município poderia estar preenchendo a informação do mesmo procedimento com o código ‘0402010043’ (que se refere ao procedimento ‘TIREOIDECTOMIA TOTAL’). Subindo o nível para a forma ‘040201’, que significa ‘Cirurgia de tireóide e paratireóide’, acreditou-se que esse tipo de erro seria mais difícil de acontecer.

Assim, nesta rodada, foram analisados 144 serviços, sendo quatro grupos, 30 subgrupos e 110 formas.

Das 110 formas, 34 foram utilizadas por apenas um município (Quantidade mínima de municípios que usaram um procedimento), e uma forma foi utilizada por 5563 municípios (Quantidade máxima de municípios que usaram um procedimento). A tabela abaixo mostra um resumo desses quantitativos pelos três níveis analisados: Forma, Subgrupo e Grupo.

Tabela 9: Resumo da utilização do serviço por Nível – Rodada 3

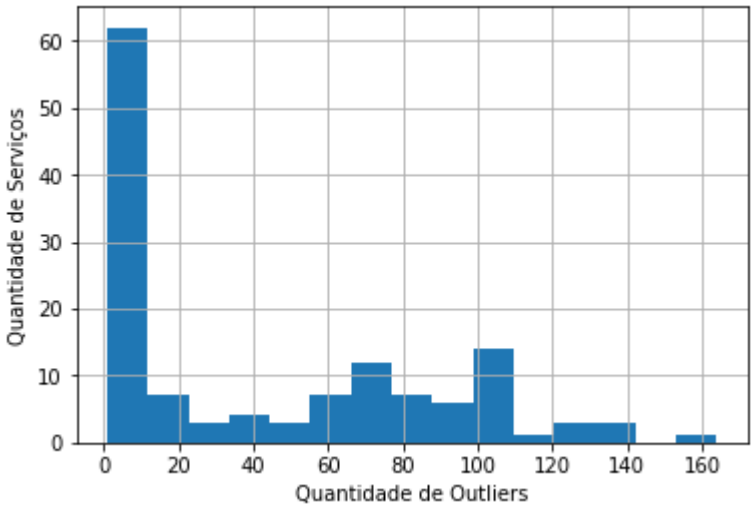
MEDIDA / NIVEL	FORMA	SUBGRUPO	GRUPO
TOTAL DE SERVIÇOS ANALISADOS	110	30	4
QUANTIDADE MÍNIMA DE MUNICÍPIOS QUE USARAM UM SERVIÇO	34	389	2867
NÚMERO DE SERVIÇOS QUE FORAM USADOS PELA MENOR QUANTIDADE DE MUNICÍPIOS	1	1	1
EXEMPLOS DE SERVIÇOS QUE FORAM USADOS PELA MENOR QUANTIDADE DE MUNICÍPIOS	030109 (Atendimento/ Acompanhamento em saúde do idoso)	0501 (Coleta e exames par a fins de doação de orgãos, tecidos e células e de transplante)	02 (Procedimentos com finalidade diagnóstica)
QUANTIDADE MÁXIMA DE MUNICÍPIOS QUE USARAM UM SERVIÇO	5563	5570	5570
NÚMERO DE SERVIÇOS QUE FORAM USADOS PELA MAIOR QUANTIDADE DE MUNICÍPIOS	1	1	2
EXEMPLOS DE SERVIÇOS QUE FORAM USADOS PELA MAIOR QUANTIDADE DE MUNICÍPIOS	030314 (Tratamento de doen ças do ouvido / apófise mastóide e vias aéreas)	0303 (Tratamentos clínicos - outras especialidades)	03 (Procedimentos Clínicos); 04 (Procedimentos cirúrgicos)
QUANTIDADE MÉDIA DE MUNICÍPIOS QUE USARAM UM SERVIÇO	3115,34	3960,5	4417,25

Fonte: Elaborada pela autora (2020).

Em relação aos *outliers* desta rodada, foram encontradas anomalias em 133 serviços (dentre os 144 analisados), sendo que 116 serviços tiveram *outliers* altos (taxas acima da média) e baixos (taxas abaixo da média), 10 serviços tiveram apenas *outliers* baixos e 7 tiveram apenas *outliers* altos.

O histograma a seguir representa a distribuição da quantidade de municípios considerados *outliers* (Quantidade de *outliers*) pela frequência de serviços que tiveram cada uma das faixas de quantidade de *outliers* (Quantidade de serviços). Foram analisadas as quantidades de *outliers* encontrados nos 133 serviços.

Figura 20: Quantidade de serviços x quantidade de *Outliers* – Rodada 3



Fonte: Elaborada pela autora (2020).

Percebe-se que a maioria desses serviços tiveram poucos *outliers*. Mais de 60 serviços tiveram menos de dez municípios considerados *outliers*.

Nesta rodada, a média de *outliers* por serviço foi 43,58, o mínimo de *outliers* por serviço foi 1 e o máximo de *outliers* por serviço foi 164.

6.4.4 **Rodada 4**

Nesta rodada, pensou-se em continuar a análise a partir do nível da forma, mas completando com quantidade zero, os municípios que não utilizaram o serviço que está sendo analisado.

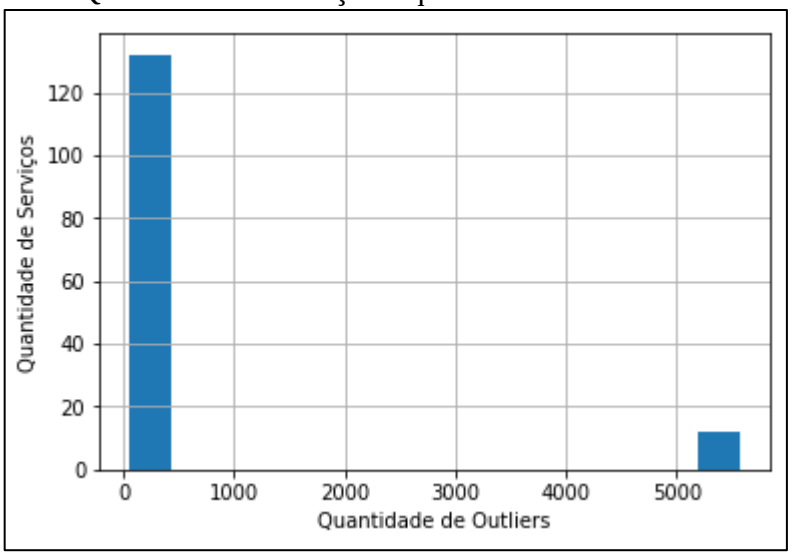
Assim, nesta rodada, continuaram sendo analisados os 144 serviços da rodada anterior. Entretanto de 479.172 linhas totais analisadas na rodada anterior, passou-se a ter 802.080, em que 40,25% estavam com valor zero na quantidade do serviço utilizado pelo município (coluna QTD), e, por consequência, estavam com valor zero na taxa de quantidades do município para o serviço (coluna TX).

Em relação aos *outliers* desta rodada, foram encontradas anomalias em todos os 144 serviços analisados, sendo que 115 serviços tiveram *outliers* altos (taxas acima da média) e

baixos (taxas abaixo da média), nenhum serviço teve apenas *outliers* baixos e 29 tiveram apenas *outliers* altos.

O histograma a seguir representa a distribuição da quantidade de municípios considerados *outliers* (Quantidade de *outliers*) pela frequência de serviços que tiveram cada uma das faixas de quantidade de *outliers* (Quantidade de serviços). Foram analisadas as quantidades de *outliers* encontrados nos 144 serviços.

Figura 21: Quantidade de serviços x quantidade de *Outliers* – Rodada 4



Fonte: Elaborada pela autora (2020).

Percebe-se que a maioria dos serviços tiveram poucos *outliers* detectados.

Nesta rodada, a média de *outliers* por serviço foi 594,25, o mínimo de *outliers* por serviço foi 60 e o máximo de *outliers* por serviço foi 5570 (que representa todos os municípios).

As ações e os serviços de saúde não são estruturados apenas na escala dos municípios. Existem no Brasil milhares de pequenas municipalidades que não possuem em seus territórios condições de oferecer serviços de alta e média complexidade; por outro lado, existem municípios que se tornam referência e garantem o atendimento da sua população e de municípios vizinhos (MINISTÉRIO DA SAÚDE, 2009).

Muitos pacientes informam como município de residência aquele no qual foram atendidos por temer não receber o atendimento no local escolhido, e essa informação equivocada acaba influenciando no resultado das análises realizadas (AGUIAR, MELO, *et al.*, 2013).

Segundo a ANS (2020), a região de saúde é um espaço geográfico contínuo constituído por agrupamentos de municípios limítrofes, delimitado a partir de identidades culturais,

econômicas e sociais e de redes de comunicação e infraestrutura de transportes compartilhados, com a finalidade de integrar a organização, o planejamento e a execução de ações e serviços de saúde. O SUS possui 437 regiões de saúde cadastradas.

Pensou-se em repetir as análises feitas nas primeiras quatro rodadas, usando-se as regiões de saúde no lugar de municípios. O intuito principal foi diminuir esse problema da informação equivocada do município de residência, e, por consequência, reduzir o número de serviços que não são utilizados por nenhuma localidade.

6.4.5 **Rodada 5**

Nesta rodada, foram considerados os procedimentos que deram origem às AIHs do SIH/SUS nas regiões de saúde do país, totalizando 1886 serviços do SIGTAP, sendo quatro grupos, 30 subgrupos, 110 formas e 1742 procedimentos.

Para cada serviço, foram analisados somente as regiões de saúde que tiveram pelo menos uma AIH com o serviço realizado (PROC_REA).

Dos 1742 procedimentos, 23 foram utilizados por apenas uma região de saúde (Quantidade mínima de regiões de saúde que usaram um procedimento), e 38 procedimentos foram utilizados por 437 regiões de saúde (Quantidade máxima de regiões de saúde que usaram um procedimento). A tabela abaixo mostra um resumo desses quantitativos pelos quatro níveis analisados: Procedimento, Forma, Subgrupo e Grupo.

Tabela 10: Resumo da utilização do serviço por Nível – Rodada 5

MEDIDA / NIVEL	PROCEDIMENTO	FORMA	SUBGRUPO	GRUPO
TOTAL DE SERVIÇOS ANALISADOS	1742	110	30	4
QUANTIDADE MÍNIMA DE REGIÕES DE SAÚDE QUE USARAM UM SERVIÇO	1	13	117	432
NÚMERO DE SERVIÇOS QUE FORAM USADOS PELA MENOR QUANTIDADE DE REGIÕES DE SAÚDE	23	1	1	1
EXEMPLOS DE SERVIÇOS QUE FORAM USADOS PELA MENOR	0304010162 (Moldagem em colo e/ou corpo do utero);	030109 (Atendimento/ Acompanhamento em saúde do	0504 (Processamento de tecidos para transplante)	02 (Procedimentos com finalidade diagnóstica)

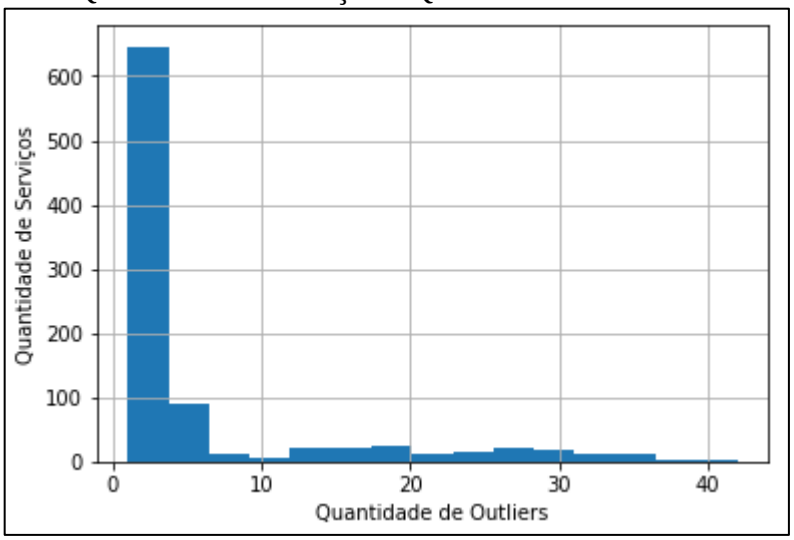
QUANTIDADE DE REGIÕES DE SAÚDE	0403080061 (nucleotractomia trigeminal e/ou espinal)	idoso)		
QUANTIDADE MÁXIMA DE REGIÕES DE SAÚDE QUE USARAM UM SERVIÇO	437	437	437	437
NÚMERO DE SERVIÇOS QUE FORAM USADOS PELA MAIOR QUANTIDADE DE REGIÕES DE SAÚDE	38	35	16	2
EXEMPLOS DE SERVIÇOS QUE FORAM USADOS PELA MAIOR QUANTIDADE DE REGIÕES DE SAÚDE	0303020032 (Tratamento de Anemia Aplástica e outras anemias); 0303030020 (Tratamento de desnutrição)	030301 (Tratamento de doenças infecciosas e parasitárias); 040704 (Parede e cavidade abdominal)	0304 (Tratamento em oncologia); 0407 (Cirurgia do aparelho digestivo, órgãos anexos e parede abdominal)	03 (Procedimentos clínicos); 04 (Procedimentos Cirúrgicos)
QUANTIDADE MÉDIA DE REGIÕES DE SAÚDE QUE USARAM UM SERVIÇO	164,52	382,2	406,3	435,25

Fonte: Elaborada pela autora (2020).

Em relação aos *outliers* encontrados nesta rodada, foram encontradas anomalias em 920 serviços (dentre os 1886 analisados), sendo que 243 serviços tiveram *outliers* altos (taxas acima da média) e baixos (taxas abaixo da média), 179 serviços tiveram apenas *outliers* baixos e 498 tiveram apenas *outliers* altos.

O histograma a seguir representa a distribuição da quantidade de regiões de saúde consideradas *outliers* (Quantidade de *outliers*) pela frequência de serviços que tiveram cada uma das faixas de quantidade de *outliers* (Quantidade de serviços). Foram analisadas as quantidades de *outliers* encontrados nos 920 serviços.

Figura 22: Quantidade de Serviços x Quantidade de *Outliers* – Rodada 5



Fonte: Elaborada pela autora (2020).

Percebe-se que a maioria desses serviços tiveram poucos *outliers*. Mais de 600 serviços tiveram até cinco regiões de saúde consideradas *outliers*.

Nesta rodada, a média de *outliers* por serviço foi 6.04, o mínimo de *outliers* por serviço foi 1 e o máximo de *outliers* por serviço foi 42.

6.4.6 **Rodada 6**

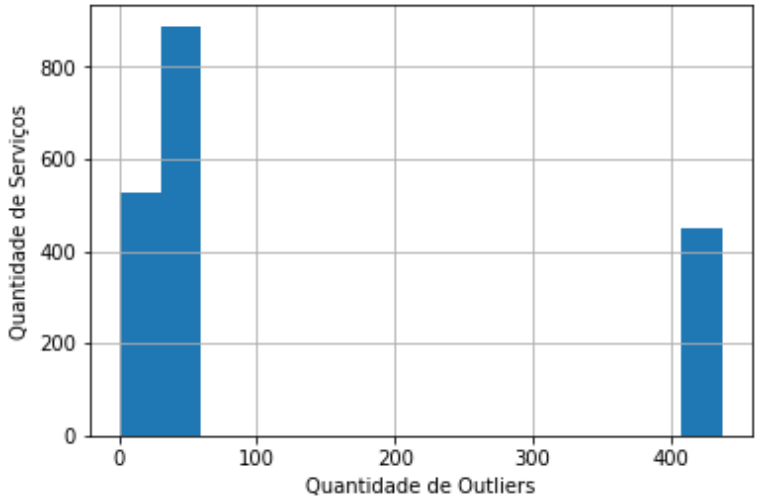
Nesta rodada, para a análise de um serviço, foram também acrescentadas as regiões de saúde cuja população não teve acesso a esse serviço nas internações hospitalares financiadas pelo SUS. Dessa forma, para cada serviço, foram analisadas as taxas de quantidade para todas as 437 regiões de saúde do país. As regiões de saúde que não utilizaram o serviço da análise, ficaram com taxa zero. Esta rodada foi pensada tentando encontrar as regiões de saúde que estavam tendo menos acesso aos serviços de internação hospitalar financiados pelo SUS.

Foram analisados os mesmos 1886 serviços da rodada anterior. Entretanto de 342.581 linhas totais analisadas na rodada anterior, passou-se a ter 824.182, em que 58,43% estavam com valor zero na quantidade do serviço utilizado pela região de saúde (coluna QTD), e, por consequência, estavam com valor zero na taxa de quantidades da região de saúde para o serviço (coluna TX).

Em relação aos *outliers* desta rodada, foram encontradas anomalias em 1864 serviços, sendo que 1635 serviços tiveram *outliers* altos (taxas acima da média) e baixos (taxas abaixo da média), 30 serviços tiveram apenas *outliers* baixos e 199 tiveram apenas *outliers* altos.

O histograma a seguir representa a distribuição da quantidade de regiões de saúde consideradas *outliers* (Quantidade de *outliers*) pela frequência de serviços que tiveram cada uma das faixas de quantidade de *outliers* (Quantidade de serviços). Foram analisadas as quantidades de *outliers* encontrados nos 1864 serviços.

Figura 23: Quantidade de Serviços x Quantidade de *Outliers* – Rodada 6



Fonte: Elaborada pela autora (2020).

Percebe-se que a maioria dos serviços tiveram poucos *outliers* detectados.

Nesta rodada, a média de *outliers* por serviço foi 129,37, o mínimo de *outliers* por serviço foi 1 e o máximo de *outliers* por serviço foi 437 (que representa todas as regiões de saúde).

6.4.7 **Rodada 7**

Nesta rodada, as análises foram feitas a partir do nível da forma, totalizando 144 serviços, sendo quatro grupos, 30 subgrupos e 110 formas.

Das 110 formas, uma forma foi utilizada por 13 regiões de saúde (Quantidade mínima de regiões de saúde que usaram uma forma), e 35 formas foram utilizadas por 437 regiões de saúde (Quantidade máxima de regiões de saúde que usaram uma forma). A tabela abaixo mostra um resumo desses quantitativos pelos três níveis analisados: Forma, Subgrupo e Grupo.

Tabela 11: Resumo da utilização do serviço por Nível – Rodada 7

MEDIDA / NIVEL	FORMA	SUBGRUPO	GRUPO
TOTAL DE SERVIÇOS ANALISADOS	110	30	4
QUANTIDADE MÍNIMA DE REGIÕES DE SAÚDE QUE USARAM UM SERVIÇO	13	117	432

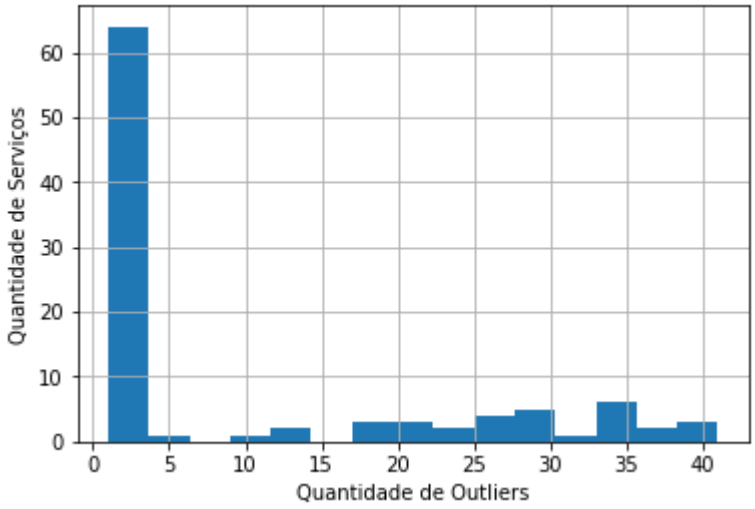
NÚMERO DE SERVIÇOS QUE FORAM USADOS PELA MENOR QUANTIDADE DE REGIÕES DE SAÚDE	1	1	1
EXEMPLOS DE SERVIÇOS QUE FORAM USADOS PELA MENOR QUANTIDADE DE REGIÕES DE SAÚDE	030109 Atendimento/ Acompanhamento em saúde do idoso)	0504 (Processamento de tecidos para transplante)	02 (Procedimentos com finalidade diagnóstica)
QUANTIDADE MÁXIMA DE REGIÕES DE SAÚDE QUE USARAM UM SERVIÇO	437	437	437
NÚMERO DE SERVIÇOS QUE FORAM USADOS PELA MAIOR QUANTIDADE DE REGIÕES DE SAÚDE	35	16	2
EXEMPLOS DE SERVIÇOS QUE FORAM USADOS PELA MAIOR QUANTIDADE DE REGIÕES DE SAÚDE	030106 (Consulta/ Atendimento ás urgências (em geral)); 030302 (Tratamento de doenças do sangue, órgãos hematopoiéticos e alguns transtornos imunitários)	0305 (Tratamento em nefrologia); 0308 (Tratamento de lesões, envenenamentos e outros, decorrentes de causas externas)	03 (Procedimentos clínicos); 04 (Procedimentos Cirúrgicos)
QUANTIDADE MÉDIA DE REGIÕES DE SAÚDE QUE USARAM UM SERVIÇO	382,2	406,3	435,25

Fonte: Elaborada pela autora (2020).

Em relação aos *outliers* desta rodada, foram encontradas anomalias em 97 serviços (dentre os 144 analisados), sendo que 41 serviços tiveram *outliers* altos (taxas acima da média) e baixos (taxas abaixo da média), 23 serviços tiveram apenas *outliers* baixos e 33 tiveram apenas *outliers* altos.

O histograma a seguir representa a distribuição da quantidade de regiões de saúde consideradas *outliers* (Quantidade de *outliers*) pela frequência de serviços que tiveram cada uma das faixas de quantidade de *outliers* (Quantidade de serviços). Foram analisadas as quantidades de *outliers* encontrados nos 97 serviços.

Figura 24: Quantidade de Serviços x Quantidade de *Outliers* – Rodada 7



Fonte: Elaborada pela autora (2020).

Percebe-se que a maioria desses serviços tiveram poucos *outliers*. Mais de 60 serviços tiveram menos de cinco regiões de saúde consideradas *outliers*.

Nesta rodada, a média de *outliers* por serviço foi 10,04, o mínimo de *outliers* por serviço foi 1 e o máximo de *outliers* por serviço foi 41.

6.4.8 **Rodada 8**

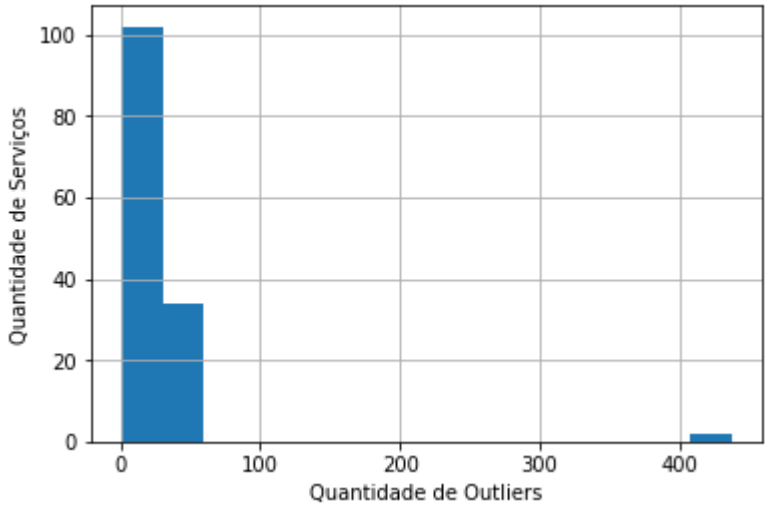
Nesta rodada, continuou-se a análise a partir do nível da forma, mas completando com quantidade zero, as regiões de saúde que não utilizaram o serviço que está sendo analisado.

Assim, continuaram sendo analisados os 144 serviços da rodada anterior. Entretanto de 55.972 linhas totais analisadas na rodada anterior, passou-se a ter 62.928, em que 11,05% estavam com valor zero na quantidade do serviço utilizado pela região de saúde (coluna QTD), e, por consequência, estavam com valor zero na taxa de quantidades da região de saúde para o serviço (coluna TX).

Em relação aos *outliers* desta rodada, foram encontradas anomalias em 138 serviços (dos 144 analisados), sendo que 79 serviços tiveram *outliers* altos (taxas acima da média) e baixos (taxas abaixo da média), 21 serviços tiveram apenas *outliers* baixos e 38 tiveram apenas *outliers* altos.

O histograma a seguir representa a distribuição da quantidade de regiões de saúde consideradas *outliers* (Quantidade de *outliers*) pela frequência de serviços que tiveram cada uma das faixas de quantidade de *outliers* (Quantidade de serviços). Foram analisadas as quantidades de *outliers* encontrados nos 138 serviços.

Figura 25: Quantidade de Serviços x Quantidade de *Outliers* – Rodada 8



Fonte: Elaborada pela autora (2020).

Percebe-se que a maioria dos serviços possuem poucos *outliers* detectados.

Nesta rodada, a média de *outliers* por serviço foi 26,19, o mínimo de *outliers* por serviço foi 1 e o máximo de *outliers* por serviço foi 437.

A tabela a seguir apresenta um resumo das análises realizadas nas oito rodadas.

Tabela 12: Resumo das Rodadas de Análise – Fase de Modelagem

MEDIDA / RODADA	1	2	3	4	5	6	7	8
TOTAL DE LOCALIDADES ANALISADAS	5570	5570	5570	5570	437	437	437	437
TOTAL DE SERVIÇOS ANALISADOS	1886	1886	144	144	1886	1886	144	144
SERVIÇOS COM *OUTLIERS*	1026	1886	133	144	920	1864	97	138
QTD MIN DE *OUTLIERS* POR SERVIÇO	1	60	1	60	1	1	1	1
QTD DE SERVIÇO COM QTD MIN DE *OUTLIERS*	230	1	7	1	373	35	40	23
EXEMPLOS DE SERVIÇOS COM QTD MIN DE *OUTLIERS*	041205; 0303010177; 0412050072	0416	030305; 041201; 041611	0416	0301; 040201; 0416060102	040603; 0303010037; 0409050083	0501; 040504; 050103	0408; 030314; 041502
QTD MAX DE *OUTLIERS* POR SERVIÇO	164	5570	164	5570	42	437	41	437
QTD DE SERVIÇO COM QTD MAX DE *OUTLIERS*	1	628	1	2	1	448	2	2
EXEMPLOS DE SERVIÇOS COM QTD MAX DE *OUTLIERS*	0407	041303; 0303160012; 0505020084	0407;	030109; 041303	0407040099;	041303; 0303010096; 0403010136	030315; 040402	030109; 041303
QTD MEDIA DE *OUTLIERS* POR SERVIÇO	17,38	3827,99	43,58	594,25	6,04	129,37	10,04	26,19

Fonte: Elaborada pela autora (2020).

7 AVALIAÇÃO

Na fase de avaliação, analisam-se os resultados obtidos na fase de modelagem para verificar se de fato atendem aos objetivos de negócio definidos na fase de entendimento do negócio.

Nesta fase, foram realizadas as seguintes atividades:

a) Avaliação dos resultados das rodadas realizadas na fase de modelagem sob uma perspectiva das localidades que foram consideradas *outliers*;
b) Construção de um painel para ajudar na visualização dos resultados de detecção de *outliers* da fase de modelagem e na análise dos dados do SIH/SUS;
c) Determinação dos próximos passos.

8.1. ANÁLISE DO RESULTADO DA FASE DE MODELAGEM

Esta análise foi feita sob a perspectiva das localidades (municípios ou regiões de saúde) que foram consideradas anômalas.

Na primeira rodada, 4454 localidades (no caso municípios) foram consideradas *outliers* em pelo menos um dos 1886 serviços analisados.

1147 municípios tiveram comportamento anômalo em apenas um serviço, como, por exemplo, o município de Machadinho D'Oeste (RO), que foi considerado *outlier* apenas no serviço Tratamento de outras doenças do aparelho respiratório (procedimento: 0303140135).

O município que foi considerado *outlier* mais vezes nos serviços analisados foi Belém (PA), tendo comportamento anômalo em 39 serviços, dentre eles: Parto cesariano c/ laqueadura tubaria (procedimento: 0411010042), Biopsias múltiplas intra-abdominais em oncologia (procedimento: 0416040209) e Tratamento de Pacientes sob cuidados prolongados (forma: 030313).

Em média, uma localidade foi considerada *outlier* em quatro serviços.

1240 municípios foram considerados *outliers* baixos (com taxas abaixo da média) em pelo menos um serviço.

Dentre eles, 623 foram *outliers* baixo em apenas um serviço, como, por exemplo, o município de Jaguariúna (SP), que foi *outlier* baixo no serviço Tratamento de crises epiléticas não controladas (procedimento: 0303040165).

Belém (PA) foi o município considerado mais vezes *outlier* baixo. Em todas as 39 vezes que foi considerado *outlier*, sua taxa estava abaixo da média.

A tabela a seguir apresenta alguns desses casos. Pode-se verificar que o procedimento Exérese de cisto sacrococcígeo estava a mais de quatro desvios padrão abaixo da média nacional (coluna QTD Desvio Padrão).

Tabela 13: Exemplos de procedimentos em que Belém(PA) foi considerado *outlier* baixo

Serviço	Distribuição	TX	TX_QTD	Média Nacional	Desvio Padrão	QTD Desvio Padrão	Qtd Serviços Realizados
0401020088: exérese de cisto sacrococcígeo	Normal após boxcox	0,00007	-7,22804	-4,29408	0,71458	4,10586	1
0303170093: tratamento em psiquiatria (por dia)	Normal após boxcox	0,00007	-5,54469	-3,05225	0,65958	3,77881	1
0303060140: tratamento de embolia pulmonar	Normal após log	0,00020	-8,51230	-4,82686	0,99397	3,70781	3
0408050160: reconstrução ligamentar intra-articular do joelho (cruzado anterior)	Normal após log	0,00013	-8,91774	-4,70975	1,17909	3,56883	2
0408050039: artrodese de medias / grandes articulações de membro inferior	Normal após log	0,00007	-9,61081	-5,67292	1,10766	3,55515	1

Fonte: Elaborada pela autora (2020).

3799 municípios foram considerados *outliers* altos (com taxas acima da média) em pelo menos um serviço analisado.

1119 municípios foram *outliers* alto em apenas um serviço, como, por exemplo, o município de Rio Branco (AC), que foi *outlier* alto no serviço Curetagem pós-abortamento / puerperal (procedimento: 0411020013).

O município que foi considerado *outlier* alto mais vezes nos serviços analisados foi Tigrinhos (SC), tendo comportamento anômalo em 28 serviços.

A tabela a seguir apresenta alguns desses exemplos. Pode-se verificar que a população de Tigrinhos teve acesso aos procedimentos da forma Gerais mais de oito desvios padrão acima da média nacional (coluna QTD Desvio Padrão).

Tabela 14: Exemplos de procedimentos em que Tigrinhos (SC) foi considerado *outlier* alto

Serviço	Distribuição	TX	TX_QTD	Média Nacional	Desvio Padrão	QTD Desvio Padrão	Qtd Serviços Realizados
040806: gerais	Não normal	0,79608	0,79608	0,10374	0,08086	8,56267	13
0415010012: tratamento c/ cirurgias múltiplas	Não normal	0,97979	0,97979	0,12639	0,13230	6,45041	16
04: Procedimentos cirúrgicos	Não normal	8,57318	8,57318	2,46800	0,97053	6,29058	140
0303160047: tratamento de transtornos hemorrágicos e hematológicos do feto e do recém-nascido	Não normal	0,18371	0,18371	0,02837	0,03072	5,05726	3
0409: Cirurgia do aparelho geniturinário	Normal após boxcox	1,22474	0,20906	-1,18951	0,38692	3,61464	20

Fonte: Elaborada pela autora (2020).

As dez maiores taxas de municípios consideradas *outliers* são apresentadas na Tabela 15: *Outliers* Mais Significativos. Todas elas se referem a *outliers* altos, ou seja, que a taxa está acima da média.

Pode-se verificar que metade dessas taxas pertencem ao município Coração de Maria (BA).

Tabela 15: *Outliers* Mais Significativos

Serviço	Distribuição	TX	TX_QTD	Média	STD	Qtd STD	Qtd Serviços	UF	Município
040704 : Parede e cavidade abdominal	Não normal	8,781	8,781	0,182	0,154	55,744	1985	BA	Coração de Maria
0407040064 : HERNIOPLASTIA EPIGASTRICA	Não normal	2,614	2,614	0,019	0,048	53,675	591	BA	Coração de Maria
0407040129 : HERNIOPLASTIA UMBILICAL	Não normal	2,654	2,654	0,045	0,053	49,468	600	BA	Coração de Maria
0407 : Cirurgia do aparelho digestivo, orgãos anexos e parede abdominal	Não normal	10,891	10,891	0,433	0,252	41,564	2462	BA	Coração de Maria
0303150025 : TRATAMENTO DE DOENCAS GLOMERULARES	Não normal	2,142	2,142	0,018	0,053	40,443	249	PI	Fronteiras
0407040102 : HERNIOPLASTIA INGUINAL / CRURAL (UNILATERAL)	Não normal	2,659	2,659	0,078	0,064	40,053	601	BA	Coração de Maria
0301060088 : DIAGNOSTICO E/OU ATENDIMENTO DE URGENCIA EM CLINICA MEDICA	Não normal	12,479	12,479	0,135	0,320	38,584	443	SC	Arroio Trinta
0406010935 : REVASCULARIZACAO MIOCARDICA C/ USO DE EXTRACORPOREA (C/ 2 OU MAIS ENXERTOS)	Não normal	0,949	0,949	0,018	0,025	36,561	411	PR	Campina Grande do Sul
0303070102 : TRATAMENTO DE OUTRAS DOENCAS DO APARELHO DIGESTIVO	Não normal	5,509	5,509	0,081	0,157	34,615	748	RN	Alexandria
030308 : Tratamento de doenças da pele e do tecido subcutâneo	Não normal	4,895	4,895	0,117	0,144	33,192	931	MA	Passagem Franca

Fonte: Elaborada pela autora (2020)

A tabela a seguir apresenta o resumo do resultado das oito rodadas de análise realizadas na fase de modelagem.

Tabela 16: Resumo das Rodadas de Análise da Fase de Modelagem sob a Perspectiva das Localidades

MEDIDAS / RODADAS	1	2	3	4	5	6	7	8
TOTAL DE LOCALIDADES ANALISADAS	5570	5570	5570	5570	437	437	437	437
TOTAL DE SERVIÇOS ANALISADOS	1886	1886	144	144	1886	1886	144	144
LOCALIDADES CONSIDERADAS *OUTLIERS* (EM PELO MENOS 1 SERVIÇO)	4454	5570	2679	5570	437	437	349	437
QTD MIN DE SERVIÇOS EM QUE UMA LOCALIDADE FOI CONSIDERADA *OUTLIER*	1	1163	1	9	1	478	1	2
QTD DE LOCALIDADE COM QTD MIN	1147	1	1321	3	1	1	117	14
EXEMPLOS DE LOCALIDADE COM QTD MIN	Machadinho D'Oeste (RO); Marapanim	Porto Alegre (RS)	Cacoal (RO); Rorainópolis	Altos (PI); Araras (SP); Taubaté (SP)	Circ.Fé Vale/ Histórico (SP)	Ribeira Do	Ilha Do Bananal (TO);	Zé Doca (MA); Viçosa

	(PA); Piranhas (GO)		(RR); Augusto Corrêa (PA)			Pombal (BA)	Timon (MA); Caicó (RN)	(MG); Araras (SP)
QTD MAX DE SERVIÇOS EM QUE UMA LOCALIDADE FOI CONSIDERADA *OUTLIER*	39	1352	18	36	48	815	19	33
QTD DE LOCALIDADE COM QTD MAX	1	1	1	1	1	1	1	1
EXEMPLOS DE LOCALIDADE COM QTD MAX	Belém (PA)	Balsa Nova (PR)	Benjamin Constant (AM)	Balsa Nova (PR)	Planalto (RS)	Recife (PE)	Alto Solimões (AM)	Santa Fé Do Sul (SP)
QTD MÉDIA DE SERVIÇOS EM QUE UMA LOCALIDADE FOI CONSIDERADA *OUTLIER*	4,00	1296,15	2,16	15,36	12,73	55,86	2,79	8,27
LOCALIDADES CONSIDERADAS (EM PELO MENOS 1 SERVIÇO) *OUTLIERS* BAIXOS	1240	5570	849	5570	273	437	123	433
QTD MIN DE SERVIÇOS EM QUE UMA LOCALIDADE FOI CONSIDERADA *OUTLIER* BAIXO	1	686	1	4	1	396	1	1
QTD DE LOCALIDADE COM QTD MIN	623	1	536	2	110	1	71	14
EXEMPLOS DE LOCALIDADE COM QTD MIN	Vilhena (RO); Novo Gama (GO); Jaguariúna (SP)	Belo Horizonte (MG)	Brasiléia (AC); Joviânia (GO); Brasília (DF)	Paulista (PE); São Paulo (SP)	18ª Região Iguatú (CE); Extremo Oeste (SC); Uva Vale (RS)	Planalto (RS)	Petrolina(PE); Café (RO); Nordeste I (GO)	Alto Do Tietê (SP); Carbonífera Costa Doce (RS); Uberlândia Araguari (MG)

QTD MAX DE SERVIÇOS EM QUE UMA LOCALIDADE FOI CONSIDERADA *OUTLIER* BAIXO	39	1298	18	30	41	614	19	28
QTD DE LOCALIDADE COM QTD MAX	1	1	1	1	1	1	1	1
EXEMPLOS DE LOCALIDADE COM QTD MAX	Belém (PA)	Tonantins (AM)	Benjamin Constant (AM)	Benjamin Constant (AM)	Metropolitana I (RJ)	Metropolitana I (RJ)	Alto Solimões (AM)	Alto Solimões (AM)
LOCALIDADES CONSIDERADAS (EM PELO MENOS 1 SERVIÇO) *OUTLIERS* ALTOS	3799	5570	2019	4756	434	437	302	394
QTD MIN DE SERVIÇOS EM QUE UMA LOCALIDADE FOI CONSIDERADA *OUTLIER* ALTO	1	3	1	1	1	10	1	1
QTD DE LOCALIDADE COM QTD MIN	1119	7	1024	956	7	1	129	66
EXEMPLOS DE LOCALIDADE COM QTD MIN	Rio Branco (AC); Urutaí (GO); Brasília (DF)	Tonantins (AM); Sebastião Barros (PI); Serra Nova Dourada (MT)	Brejo (MA); Varzedo (BA); Ponto Belo (ES)	Panamá (GO); Itiquira (MT); Baturité (CE)	Aracaju (SE); Circ.Fé Vale/Histórico (SP); Dourados (MS)	Alto Solimões (AM)	Rio Caetés (PA); Vale Do Canindé (PI); Pampa (RS)	Rio Madeira (AM); Cocais (PI); 10ª Região (PB)
QTD MAX DE SERVIÇOS EM QUE UMA LOCALIDADE FOI CONSIDERADA *OUTLIER* ALTO	28	508	16	24	48	356	11	31
QTD DE LOCALIDADE COM QTD MAX	1	1	1	1	1	1	1	1

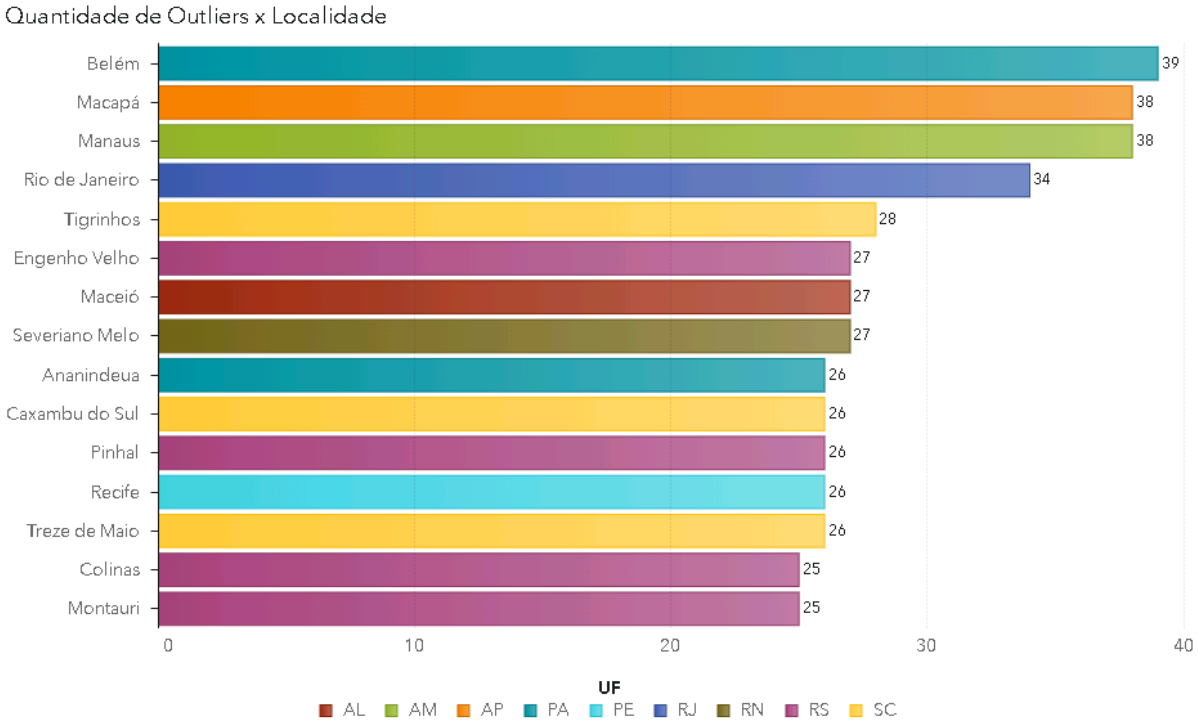
EXEMPLOS DE LOCALIDADE COM QTD MAX	Tigrinhos (SC)	Belo Horizonte (MG)	Tigrinhos (SC)	Balsa Nova (PR)	Planalto (RS)	2ª Rs Metropolitana (PR)	Santa Fé do Sul (SP)	Santa Fé do Sul (SP)

Fonte: Elaborada pela autora (2020)

8.2. PAINEL PARA VISUALIZAÇÃO DOS DADOS

Foi criado um painel, usando a ferramenta SAS VA (SAS VISUAL ANALYTICS, 2019), para ajudar na visualização dos resultados de detecção de *outliers* da fase de modelagem e na análise dos dados do SIH/SUS. Abaixo seguem algumas telas desse painel.

Figura 26: Quantidade de *Oultiers* x Localidade

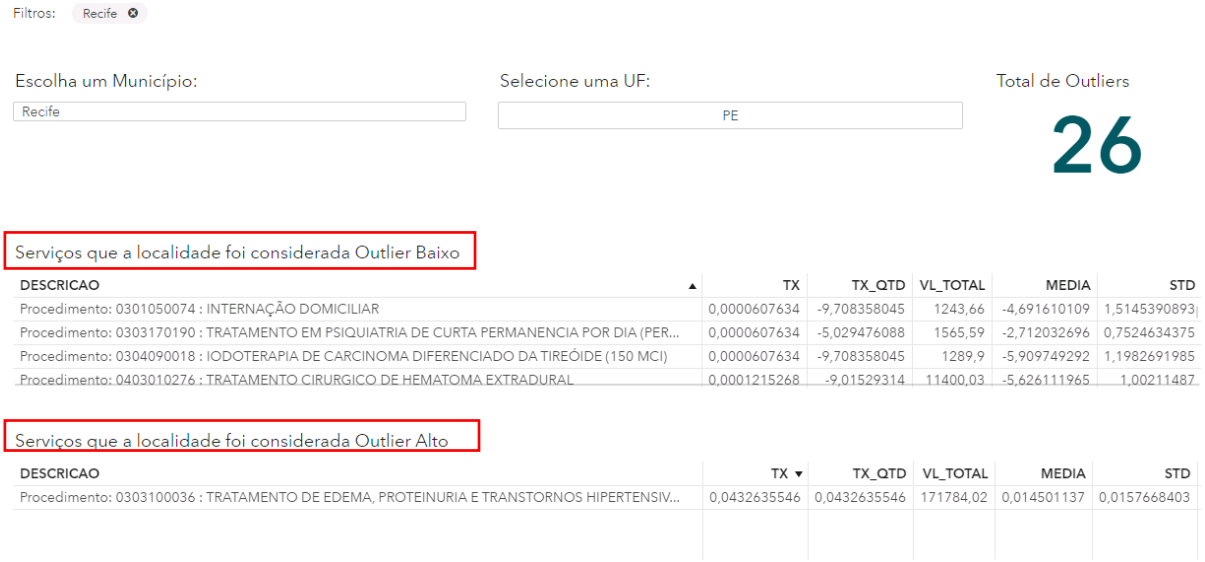


Fonte: Elaborada pela autora (2020).

Nessa primeira tela, obtém-se a lista das localidades que foram consideradas mais vezes *outliers* na análise dos serviços. As cores das barras, indicam a UF que a localidade pertence. Dessa forma, pode-se perceber que o estado Rio Grande do Sul (RS) aparece quatro vezes e o estado Santa Catarina aparece com três localidades.

Na tela a seguir, são listados os serviços que a localidade selecionada foi considerada *outier* (tanto *outlier* alto quanto *outiler* baixo). Para cada serviço, é mostrado a taxa de atendimentos por habitante do serviço na localidade (TX), essa taxa transformada para a amostra seguir uma distribuição normal (TX_QTD), o valor total gasto nas internações hospitalares origindas por esse serviço na localidade (VL_TOTAL), a média (MEDIA) e o desvio padrão (STD) desse serviço no país.

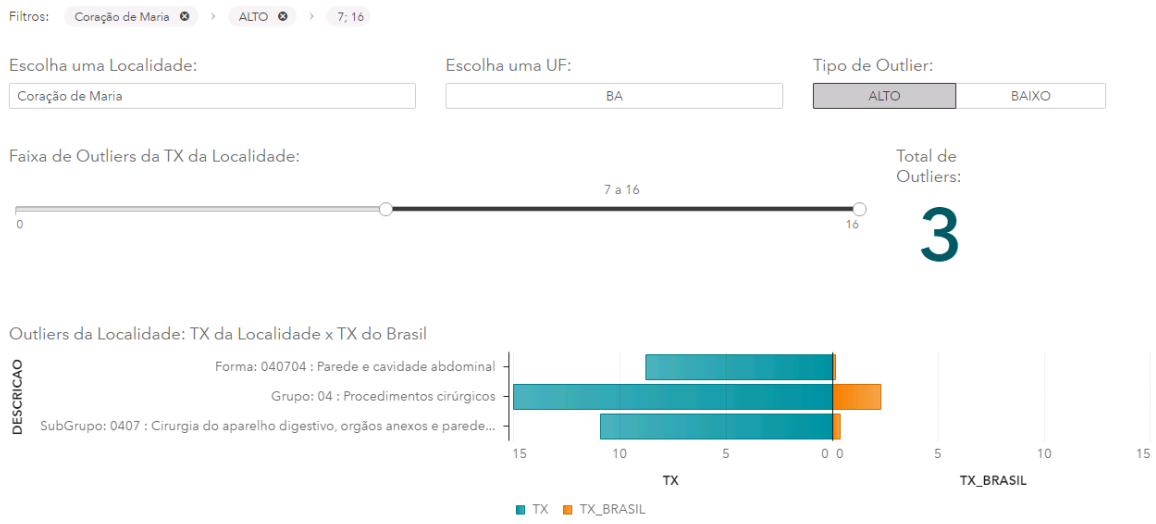
Figura 27: Lista de *Outliers* da Localidade



Fonte: Elaborada pela autora (2020).

Na Figura 28: Comparação da TX da Localidade com a do Brasil, são mostrados os serviços que a localidade selecionada foi considerada *outlier*. No exemplo, o município Coração de Maria foi considerado *outlier* em 16 serviços, entretanto estão sendo exibidos apenas três deles, porque o filtro de Tipo de *Outlier* está selecionado com a opção ALTO (indicando que está exibindo apenas os serviços que a localidade foi considerada um *outlier* alto) e o filtro da Faixa de *Outliers* da TX da Localidade está selecionado com a faixa de 7 a 16.

Figura 28: Comparação da TX da Localidade com a do Brasil



Fonte: Elaborada pela autora (2020).

Nessa tela, pode-se comparar a taxa da localidade, que fez com que ela se tornasse um *outlier* no serviço, com a taxa média do Brasil para o mesmo serviço. Percebe-se que as três taxas desse município que estão sendo exibidas na imagem têm valor muito superior à taxa média do Brasil.

Na Figura 29: *Outliers* por serviço, podem ser vistas as localidades que foram consideradas *outliers* no serviço escolhido (Subgrupo 0407: Cirurgia do aparelho digestivo, órgãos anexos e parede abdominal), totalizando 167 *outliers*, sendo 109 *outliers* altos e 55 *outliers* baixos.

Cada localidade representa uma bolha no mapa. A cor laranja representa as localidades que foram consideradas *outlier* baixos e a cor verde representa as localidades que foram consideradas *outliers* alto. O tamanho da bolha indica o valor da taxa de quantidade do serviço na localidade.

A tela também traz informação sobre os valores da média, do desvio padrão, do tipo de distribuição dos dados, da quantidade de localidades que foram analisadas (frequência) e a taxa média desse serviço no Brasil.

Figura 29: *Outliers* por serviço

Fonte: Elaborada pela autora (2020).

A

Figura 30: População atendida pelo SIH/SUS mostra duas telas do painel que auxilia na análise das respostas das últimas questões levantas no item 1.2 PROBLEMA E JUSTIFICATIVA quanto ao atendimento do princípio da universalidade.

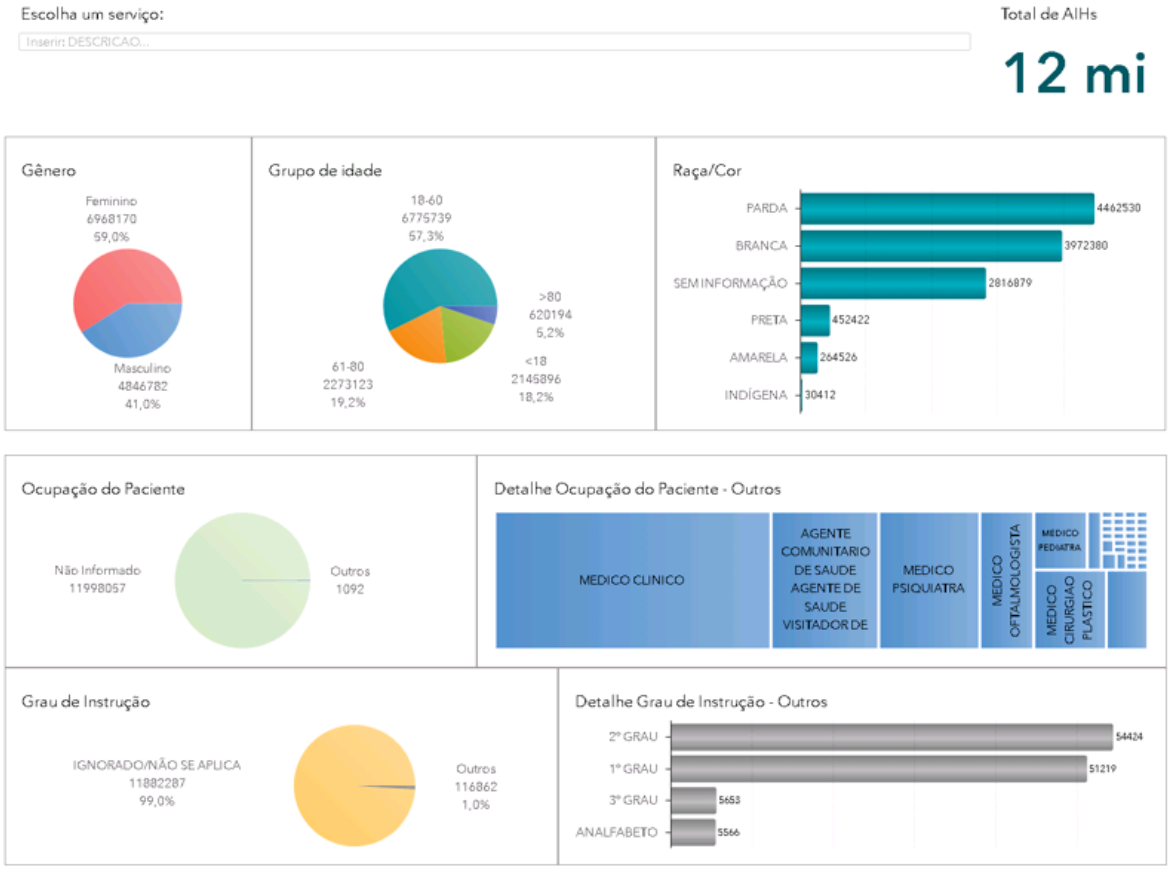
Nela, pode-se verificar a distribuição dos quantitativos de AIHs originadas pelo procedimento selecionado em função do grupo de idade (<18, 18-60, 61-80, >80), grau de instrução (1º grau, 2º grau, 3º grau, analfabeto, ignorado/não se aplica), gênero (feminino ou masculino), raça/cor (parda, branca, preta, amarela, indígena e sem informação) e ocupação do paciente.

Nenhum procedimento foi selecionado nesse exemplo. Dessa forma, tem-se o total de AIHs do ano de 2018: aproximadamente 12 milhões.

Percebe-se que quase toda informação da coluna Ocupação do Paciente está preenchida com valor 'Não Informada', o mesmo ocorre com a coluna Grau de Instrução, que está com 99% da sua informação preenchida com o valor 'Ignorado/Não se aplica', e 23% da coluna Raça/Cor está sem informação, indicando que a qualidade dos dados desses três campos poderia ser melhorada.

Verifica-se que, no Brasil, a maior parte dos atendimentos do SIH/SUS foi para pacientes entre 18 e 60 anos, para mulheres e para pessoas pardas ou brancas. Dentre a pouca informação preenchida das colunas Grau de Instrução e Ocupação do Paciente, a maior parte dos atendimentos foi para pessoas com primeiro e segundo grau, e para médicos e agentes comunitários, respectivamente.

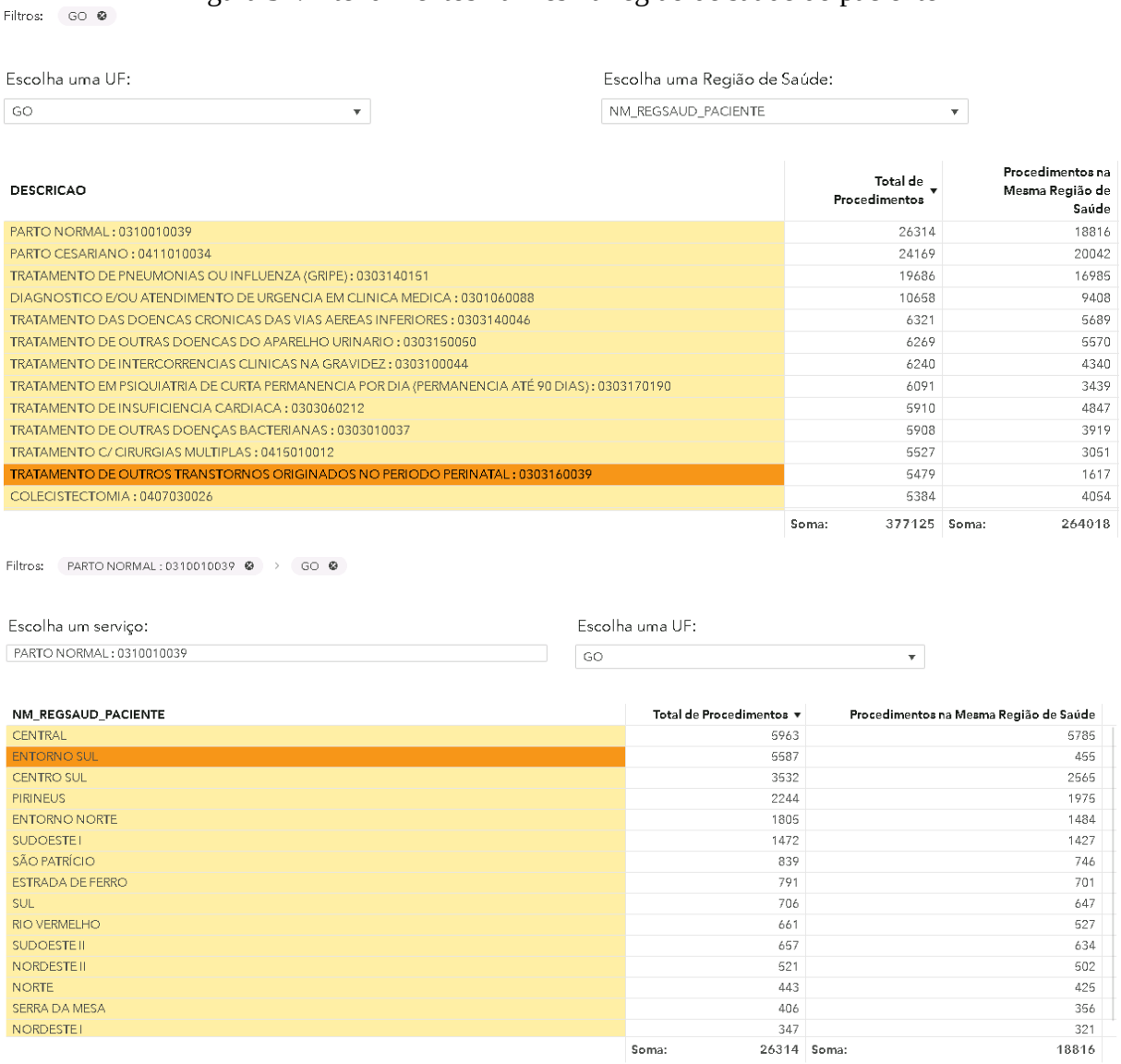
Figura 30: População atendida pelo SIH/SUS



Fonte: Elaborada pela autora (2020).

A Figura 31: Atendimentos na mesma região de saúde do paciente apresenta, na primeira lista, os procedimentos realizados pela população de Goiás e, na segunda lista, como foi o atendimento do procedimento Parto Normal para as regiões de saúde desse estado. Para cada procedimento, é apresentado o quantitativo total que a população utilizou e quantos deles foram realizados na mesma região de saúde do paciente. Todas as linhas que possuem o percentual menor do que 50% estão destacadas com a cor laranja. As demais estão com a cor amarela.

Figura 31: Atendimentos na mesma região de saúde do paciente



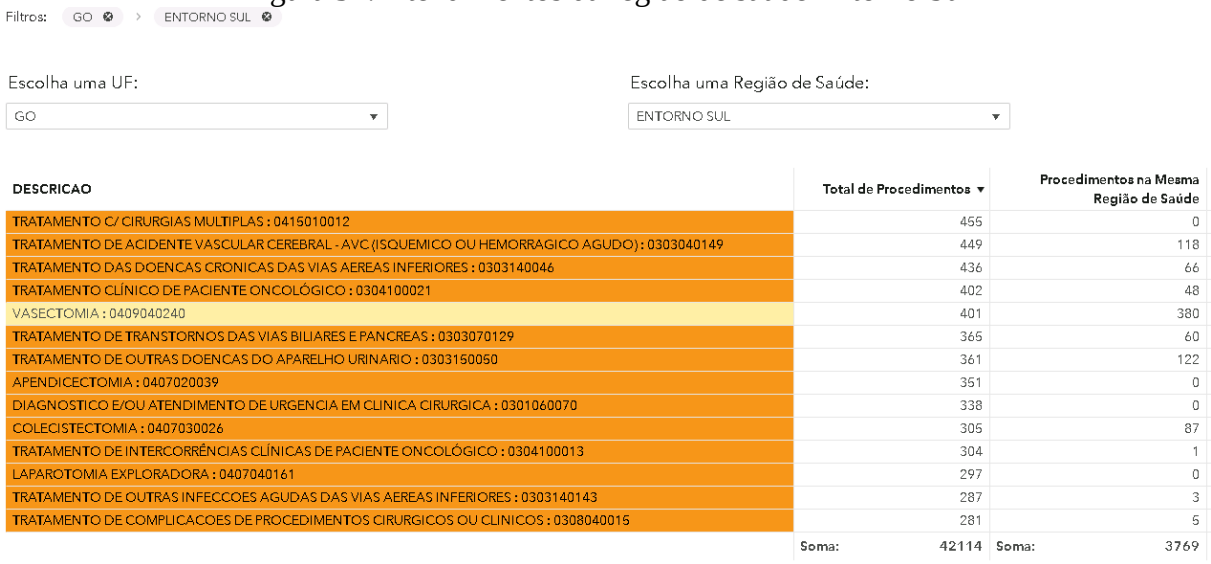
DESCRICAO	Total de Procedimentos	Procedimentos na Mesma Região de Saúde
PARTO NORMAL : 0310010039	26314	18816
PARTO CESARIANO : 0411010034	24169	20042
TRATAMENTO DE PNEUMONIAS OU INFLUENZA (GRIPE) : 0303140151	19686	16985
DIAGNOSTICO E/OU ATENDIMENTO DE URGENCIA EM CLINICA MEDICA : 0301060088	10658	9408
TRATAMENTO DAS DOENCAS CRONICAS DAS VIAS AEREAS INFERIORES : 0303140046	6321	5689
TRATAMENTO DE OUTRAS DOENCAS DO APARELHO URINARIO : 0303150050	6269	5570
TRATAMENTO DE INTERCORRENCIAS CLINICAS NA GRAVIDEZ : 0303100044	6240	4340
TRATAMENTO EM PSIQUIATRIA DE CURTA PERMANENCIA POR DIA (PERMANENCIA ATÉ 90 DIAS) : 0303170190	6091	3439
TRATAMENTO DE INSUFICIENCIA CARDIACA : 0303060212	5910	4847
TRATAMENTO DE OUTRAS DOENÇAS BACTERIANAS : 0303010037	5908	3919
TRATAMENTO C/ CIRURGIAS MULTIPLAS : 0415010012	5527	3051
TRATAMENTO DE OUTROS TRANSTORNOS ORIGINADOS NO PERIODO PERINATAL : 0303160039	5479	1617
COLECISTECTOMIA : 0407030026	5384	4054
	Soma: 377125	Soma: 264018

NM_REGSAUD_PACIENTE	Total de Procedimentos	Procedimentos na Mesma Região de Saúde
CENTRAL	5963	5785
ENTORNO SUL	5587	455
CENTRO SUL	3532	2565
PIRINEUS	2244	1975
ENTORNO NORTE	1805	1484
SUDOESTE I	1472	1427
SÃO PATRÍCIO	839	746
ESTRADA DE FERRO	791	701
SUL	706	647
RIO VERMELHO	661	527
SUDOESTE II	657	634
NORDESTE II	521	502
NORTE	443	425
SERRA DA MESA	406	356
NORDESTE I	347	321
	Soma: 26314	Soma: 18816

Fonte: Elaborada pela autora (2020).

Verifica-se que em Goiás a maior parte do procedimento Parto Normal foi realizada na mesma região de saúde do paciente, mas isso não aconteceu na região de saúde Entorno Sul, inclusive a maioria dos procedimentos que a população dessa região de saúde utilizou ocorreu em um estabelecimento pertencente a outra região de saúde, conforme figura abaixo.

Figura 32: Atendimentos da região de saúde Entorno Sul



DESCRICAO	Total de Procedimentos	Procedimentos na Mesma Região de Saúde
TRATAMENTO C/ CIRURGIAS MULTIPLAS : 0416010012	455	0
TRATAMENTO DE ACIDENTE VASCULAR CEREBRAL - AVC (ISQUEMICO OU HEMORRAGICO AGUDO) : 0303040149	449	118
TRATAMENTO DAS DOENCAS CRONICAS DAS VIAS AEREAS INFERIORES : 0303140046	436	66
TRATAMENTO CLÍNICO DE PACIENTE ONCOLÓGICO : 0304100021	402	48
VASECTOMIA : 0409040240	401	380
TRATAMENTO DE TRANSTORNOS DAS VIAS BILIARES E PANCREAS : 0303070129	365	60
TRATAMENTO DE OUTRAS DOENCAS DO APARELHO URINARIO : 0303150050	361	122
APENDICECTOMIA : 0407020039	351	0
DIAGNOSTICO E/OU ATENDIMENTO DE URGENCIA EM CLINICA CIRURGICA : 0301060070	338	0
COLECISTECTOMIA : 0407030026	305	87
TRATAMENTO DE INTERCORRÊNCIAS CLÍNICAS DE PACIENTE ONCOLÓGICO : 0304100013	304	1
LAPAROTOMIA EXPLORADORA : 0407040161	297	0
TRATAMENTO DE OUTRAS INFECCOES AGUDAS DAS VIAS AEREAS INFERIORES : 0303140143	287	3
TRATAMENTO DE COMPLICACOES DE PROCEDIMENTOS CIRURGICOS OU CLINICOS : 0308040015	281	5
	Soma: 42114	Soma: 3769

Fonte: Elaborada pela autora (2020).

8.3. PRÓXIMOS PASSOS

Nesta fase, também foram definidos os próximos passos a serem realizados, que compreendem o aumento do escopo das análises a fim de incrementar o resultado do trabalho, que são:

- Ampliar a análise para todos os anos disponíveis nos dados do SIH (no trabalho, a análise foi feita apenas com os dados de 2018);
- Ampliar a análise para outras variáveis, como valor total da internação (no trabalho, a análise de detecção de *outlier* foi realizada baseada na coluna quantidade de serviços realizados);
- Inserir na análise dados de outros SIS, como o Sistema de Informação Ambulatorial (SIA) do SUS;
- Para os municípios de grande porte (como São Paulo), incluir a análise para um nível de localidade menor do que o município: distritos de saúde;
- Realizar analise de série temporal comparando um município com ele mesmo no decorrer dos anos;
- Estudar e aplicar outros métodos de detecção de anomalias;
- Levar em conta as características (sexo, idade, raça, cor etc.) da população brasileira para ver como cada parcela está realmente sendo atendida pelos serviços do SUS;
- Organizar e comentar o código para disponibilizá-lo para a comunidade.

8 IMPLANTAÇÃO

A última fase do ciclo consiste na implantação, no ambiente de produção, dos modelos produzidos, validados e testados nas fases anteriores.

Esta fase ainda está em execução. A estratégia para a implantação deste trabalho consiste em apresentar, para a SecexSaúde, os resultados de todas as rodadas de análise executadas na fase de modelagem e os painéis correspondentes criados na fase de avaliação.

A SecexSaúde deverá escolher quais tipos de análises podem ser melhor aproveitados na unidade: no nível de município ou de região de saúde; a partir do nível da forma ou a partir do nível de procedimento; e contabilizando apenas as localidades que utilizaram o serviço ou contabilizando todas as localidades, independente de terem utilizado o serviço (nesse caso, ficando com a quantidade desse serviço igual a zero).

Após definição da SecexSaúde, serão realizados três passos:

1. Colocar no ambiente de produção o painel de visualização dos dados para ser acessado por todos os servidores do TCU, em especial, pelos servidores da SecexSaúde;
2. Agendar o script de carga para continuar atualizando o LabContas com os dados disponíveis no FTP do DATASUS;
3. Agendar o script de análise para ser executado mensalmente de forma automática e atualizar o painel de visualização com seu resultado.

9 CONSIDERAÇÕES FINAIS

Em todas as rodadas de análise realizadas na fase de Modelagem, foram encontradas localidades (municípios ou regiões de saúde) que foram consideradas anômalas quanto à utilização de um dado serviço (procedimento, forma, subgrupo ou grupo) que deu origem à internação hospitalar financiada pela SUS, por terem taxa muito acima ou muito abaixo da média do Brasil. Existiram casos em que a taxa de atendimento de um serviço na localidade era superior a 30 desvios-padrão da média nacional, como, por exemplo, o procedimento 'Revascularização Miocárdica c/ Uso de Extracorpórea (c/ 2 ou Mais Enxertos)' no município de Campina Grande do Sul (PR) e os procedimentos da forma 'Cirurgia do aparelho digestivo, órgãos anexos e parede abdominal' no município de Coração de Maria (BA).

Nas análises em que foram acrescentadas as localidades cuja população não utilizou determinado serviço, ou seja, que o quantitativo do serviço para essa localidade foi preenchido com valor zero, verificou-se que uma boa parte das localidades não tiveram acesso a maioria dos serviços. Na segunda rodada, por exemplo, 85,2% dos dados totais analisados era de municípios que não utilizaram determinado serviço.

Dessa forma, analisando os resultados descritos nos dois parágrafos anteriores, pode-se observar que não há um atendimento igualitário à população de todos os municípios brasileiros. Isso pode ocorrer por alguns motivos, dentre eles: o serviço analisado ser muito específico e raro, e realmente ocorrer poucas vezes no país durante o ano; falha no preenchimento dos dados no SIH/SUS; dificuldade e/ou falta de acesso a serviços do SUS para a população de alguns municípios.

No painel desenvolvido na fase de Avaliação, foi verificado que quase toda informação da coluna Ocupação do Paciente está preenchida com valor 'Não Informada', 99% da coluna Grau de Instrução está preenchida com valor 'Ignorado/Não se aplica' e 23% da coluna Raça/Cor está sem informação, demonstrando deficiência no preenchimento desses dados.

Foi verificado também que a maior parte dos atendimentos de internação hospitalar financiada pelo SUS foi para pacientes entre 18 e 60 anos, para mulheres e para pessoas pardas ou brancas. Apenas 6,2% dos atendimentos foram para pessoas da raça/cor preta, amarela ou indígena.

Para se ter uma avaliação mais fidedigna com a realidade, é importante levar em consideração qual é o percentual de cada característica analisada na população brasileira, para

ver as parcelas da população que realmente estão sendo atendidas e as que estão com déficit de atendimento.

Ainda há muito a ser feito, muitas análises novas a serem implementadas. Entretanto, o presente trabalho já consegue indicar alguns indícios de que:

- Há falha no preenchimento dos dados do sistema de informação hospitalar;
- Nem todos os municípios estão provendo acesso adequado aos serviços do SUS para sua população, que seria um dos resultados esperados do princípio da descentralização; e
- Nem toda a população do Brasil está tendo acesso de forma igualitária aos serviços do SUS, desrespeitando, assim, o princípio da universalidade.

BIBLIOGRAFIA

AGUIAR, F. P. et al. Confiabilidade da informação sobre município de residência no Sistema de Informações Hospitalares - Sistema Único de Saúde para análise do fluxo de pacientes no atendimento do câncer de mama e do colo do útero. **Cadernos Saúde Coletiva**, Rio de Janeiro, 21, jun. 2013.

AHAD, N. A. et al. Sensitivity of Normality Tests to Non-normal Data. **Sains Malaysiana**, jun. 2011. 637–641. Disponivel em: <http://journalarticle.ukm.my/2511/1/15_NorAishah.pdf>.

ALAUDEEN, M. N.; ENGLAND, A.; CHOPRA, R. **Data Science with Python**. Birmingham: Packt Publishing, 2019.

ANS. Questionamento sobre Região de Saúde. **Site da Agência Nacional de Saúde Suplementar**, 03 jan. 2020. Disponivel em: <http://www.ans.gov.br/aans/index.php?option=com_centraldeatendimento&view=pergunta&resposta=963&historico=21974463>.

API Sklearn. **Site Scikit-learn**. Disponivel em: <https://scikit-learn.org/stable/modules/classes.html>.

ASSUNÇÃO, R. M. et al. DETECÇÃO DE ANOMALIAS NOS PAGAMENTOS. **XV Congresso Brasileiro de Informática em Saúde**, Goiania, 2016. Disponivel em: <http://docs.bvsalud.org/biblioref/2018/07/906376/anais_cbis_2016_artigos_completos-459-468.pdf>.

BOXCOX. **Site SciPy**. Disponivel em: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.boxcox.html>.

BRASIL. Lei Nº 8.080. **Site da Presidência da República**, 19 set. 1990. Disponivel em: <http://www.planalto.gov.br/ccivil_03/leis/l8080.htm>. Acesso em: 15 fev. 2020.

CARMONA, G. Conheça tudo sobre o Boxplot e entenda suas funções. **Gradus**. Disponivel em: <https://www.gradusct.com.br/voce-realmente-conhece-o-boxplot-2/>. Acesso em: 26 jan. 2020.

COELHO, F. C. Projeto PySUS. **Site PyPI**, 201-? Disponivel em: <https://pypi.org/project/PySUS/>. Acesso em: 01 nov. 2019.

DESFORGES; JACOB; COOPER, 1998 APUD FREITAS; IGOR. **Um estudo comparativo de técnicas de detecção de outliers no contexto de classificação de dados**. Mossoró: [s.n.], 2019. Disponivel em: <http://repositorio.ufersa.edu.br/bitstream/prefix/1093/1/IgorWSF_DISSERT.pdf>.

DESHPANDE, B.; KOTU, V. **Data Science**. 2ª. ed. Cambridge: Elsevier, 2018.

DISTRIBUIÇÃO Normal. **Portal Action**. Disponivel em: <http://www.portalaction.com.br/probabilidades/62-distribuicao-normal>. Acesso em: 15 fev. 2020.

ESKIN, 2000 APUD FREITAS; IGOR. **Um estudo comparativo de técnicas de detecção de outliers no contexto de classificação de dados**. Mossoró: [s.n.], 2019. Disponivel em: <http://repositorio.ufersa.edu.br/bitstream/prefix/1093/1/IgorWSF_DISSERT.pdf>.

ESTATCAMP. Testes de Normalidade. **Portal Action**, 20–. Disponivel em: <http://www.portalaction.com.br/inferencia/testes-de-normalidade>. Acesso em: 16 jan. 2020.

EXTENSÃO do arquivo CNV. **Site sobre extensões de arquivos**, 2014. Disponivel em: <https://www.file-extension.info/pt/format/cnv>.

FIOCRUZ. **Site da Fundação Oswaldo Cruz**, 20–. Disponivel em: <https://pensesus.fiocruz.br/>. Acesso em: 28 fev. 2020.

FUENTES, A. CRISP-DM and other approaches. In: ___ **Hands-On Predictive Analytics with Python**. Birmingham: Packt, 2018.

GARNER, H. The log-normal distribution. In: __ **Clojure for Data Science**. Birmingham: Packt Publishing, 2015.

GHOSH, T.; BALI, R.; SARKAR, D. CRISP-DM. In: ___ **Hands-On Transfer Learning with Python**. Birmingham: Packt, 2018.

GRIFFITHS, D. Using the Normal Distribution: Being Normal. In: __ **Head First Statistics**. Cambridge: O'REILLY, 2008.

HALDER, S.; OZDEMIR, S. Anomaly detection. In: __ **Hands-On Machine Learning for Cybersecurity**. Birmingham: Packt Publishing, 2018.

HERSCU, R. K. Desenvolvimento e implantação de um modelo de detecção de fraude em cheques, São Paulo, 2017. Disponivel em: <http://pro.poli.usp.br/wp-content/uploads/2017/11/TF_RodrigoHerscu_v12.pdf>.

KAMBER, M.; PEI, J.; HAN, J. **Data Mining:** Concepts and Techniques. 3. ed. Waltham: Morgan Kaufmann, 2011.

LAURAE. Large amount of observations: Statistical Test not so Statistical, 13 nov. 2016. Disponivel em: <https://medium.com/data-design/large-amount-of-observations-statistical-test-not-so-statistical-3d8ed0e94be>. Acesso em: 20 fev. 2020.

MACHADO, J. P.; MARTINS, M.; LEITE, I. D. C. Qualidade das bases de dados hospitalares. **REV BRAS EPIDEMIOL**, 2016. Disponivel em: <https://www.scielosp.org/article/ssm/content/raw/?resource_ssm_path=/media/assets/rbepid/v19n3/1980-5497-rbepid-19-03-00567.pdf>. Acesso em: 28 fev. 2020.

MCNEESE, D. B. Box-cox Transformation. **SPC for Excel**, 2016. Disponivel em: <https://www.spcforexcel.com/knowledge/basic-statistics/box-cox-transformation>.

MERELES, C. SAÚDE MUNICIPAL: O QUE PODE E DEVE SER FEITO NESSA ESFERA? **Politize**, 23 set. 2016. Disponivel em: <https://www.politize.com.br/saude-municipio-qual-a-responsabilidade/>.

MICHAEL KLEEHAMMER. Projeto pyodbc. **Site PyPI**, 201-? Disponivel em: <https://pypi.org/project/pyodbc/>. Acesso em: 01 jul. 2018.

MICROSOFT. BULK INSERT - SQL SERVER. **Site da Microsoft**, 2020. Disponivel em: <https://docs.microsoft.com/pt-br/sql/t-sql/statements/bulk-insert-transact-sql?view=sql-server-ver15>.

MINISTÉRIO DA SAÚDE. **O SUS no seu município:** garantindo saúde para todos. 2. ed. Brasília: EDITORA MS, 2009. Disponivel em: <http://bvsms.saude.gov.br/bvs/publicacoes/sus_municipio_garantindo_saude.pdf>. Acesso em: 01 mar. 2020.

MINISTÉRIO DA SAÚDE. **SIH – Sistema de Informação Hospitalar do SUS:** Manual Técnico Operacional do Sistema. Brasília: [s.n.], 2017. Disponivel em: <http://www.saude.sp.gov.br/resources/ses/perfil/gestor/homepage/auditoria/manuais/manual_sih_janeiro_2017.pdf>.

MINISTÉRIO DA SAÚDE. Sistema Nacional de Saúde, 24 abr. 2017. Disponivel em: <https://www.saude.gov.br/component/content/article/681-institucional/40029-sistema-nacional-de-saude>.

MOLIN, S. Comparing Models. In: __ **Hands-On Data Analysis with Pandas**. Birmingham: Packt Publishing, 2019.

NETO, M. V. D. G. **O processo CRISP-DM aplicado na construção de uma solução para Análise de Risco de Crédito**. Recife: Universidade Federal de Pernambuco, 2018. Disponivel em: <https://www.cin.ufpe.br/~tg/2018-2/TG_CC/tg_mvgn.pdf>.

PANDAS documentation. **Site sobre a biblioteca Pandas do Python**, 2008. Disponivel em: <https://pandas.pydata.org/pandas-docs/stable/index.html>. Acesso em: 20 jun. 2018.

PRATES, W. R.; HOPPEN, J. Outliers, o que são e como tratá-los em uma análise de dados?, 25 set. 2017. Disponivel em: <https://www.aquare.la/o-que-sao-outliers-e-como-trata-los-em-uma-analise-de-dados/>.

PROJECT JUPYTER. Project Jupyter. **Site sobre a ferramenta Jupyter**, c2020. Disponivel em: <https://jupyter.org/>. Acesso em: 10 jun. 2018.

PYTHON. **Site sobre a liguagem Python**, 2001. Disponivel em: <https://www.python.org/>. Acesso em: 25 fev. 2020.

PYTHON SOFTWARE FOUNDATION. ftplib - FTP protocol client. **Site sobre o Python**, c2001-2020. Disponivel em: <https://docs.python.org/3/library/ftplib.html>. Acesso em: 01 out. 2019.

READ DBF Files with Python. **Site sobre dbfread**, 20–. Disponivel em: <https://dbfread.readthedocs.io/en/latest/>. Acesso em: 01 out. 2019.

SANTOS, A. C. D. **Sistema de Informações Hospitalares do Sistema Único de Saúde: documentação do sistema para auxiliar o uso das suas informações**. Rio de Janeiro. 2009.

SAS VISUAL ANALYTICS. **Site do SAS**, 2019. Disponivel em: <https://www.sas.com/pt_br/software/visual-analytics.html>. Acesso em: 07 jan. 2020.

SCIPY DEVELOPERS. SciPy library. **Site SciPy**, c2020. Disponivel em: <https://scipy.org/scipylib/index.html>. Acesso em: 15 jan. 2020.

TABWIN. **Site do DATASUS**, 2008. Disponivel em: <http://www2.datasus.gov.br/DATASUS/index.php?area=060805&item=3>.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introdução ao data mining:** Mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

THE SCIPY COMMUNITY. Notes - Shapiro. **Site SciPy**, 19 dez. 2019. Disponivel em: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html>. Acesso em: 15 jan. 2020.

TOCCI, A.; JAIME. **Técnicas de Detecção de Anomalias**, 31 out. 2018. Disponivel em: <https://blog.dp6.com.br/t%C3%A9cnicas-de-detec%C3%A7%C3%A3o-de-anomalias-3d9e216bf82e>. Acesso em: 20 fev. 2020.

TOCCI, A.; JAIME. **Técnicas de Detecção de Anomalias Parte II**, 05 dez. 2018. Disponivel em: <https://blog.dp6.com.br/t%C3%A9cnicas-de-detec%C3%A7%C3%A3o-de-anomalias-parte-ii-99ecb90f93b>. Acesso em: 01 mar. 2020.

TRANSFERÊNCIA de Arquivos DATASUS. **Site do Datasus**, 2020. Disponivel em: <http://datasus.saude.gov.br/transferencia-de-arquivos/>. Acesso em: 23 fev. 2020.

W3SCHOOLS. Python open() Function. **Site w3schools**, c1999-2020. Disponivel em: <https://www.w3schools.com/python/ref_func_open.asp>.

WHAT IS FILE EXTENSION TEAM. O que é DBF File Extension? Como abrir DBF? **Site sobre extensão de arquivo**, 20–. Disponivel em: <https://www.whatisfileextension.com/pt/dbf/>. Acesso em: 12 dez. 2019.

YU-WEI, C.; BHATIA, A. **Machine Learning with R Cookbook - Second Edition**. Birmingham: Packt Publishing, 2017.

Z-SCORES review. **Khan Academy**. Disponivel em: <https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/z-scores/a/z-scores-review>.

APÊNDICE A – DEFINIÇÃO DO TESTE DE NORMALIDADE

Em Ahad et al. (2011), foi feita uma comparação do desempenho de quatro testes estatísticos amplamente usados para verificar a normalidade dos dados: Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling e Cramer-von Mises.

Para a comparação, foram utilizadas várias amostras de diferentes tamanhos e tipos de distribuição dos dados. Como resultado, o teste Shapiro_Wilk foi apresentado como melhor teste de normalidade, porque ele consegue rejeitar a hipótese de normalidade com amostras de tamanhos menores do que os demais testes comparados conseguem. O teste Anderson-Darling foi apontado como o segundo teste de normalidade mais eficaz.

O teste Shapiro-Wilk consegue rejeitar a hipótese de que os dados são normais a partir de amostras com tamanho maior ou igual a 40. O teste Anderson-Darling consegue o mesmo feito com amostras maiores ou iguais a 48.

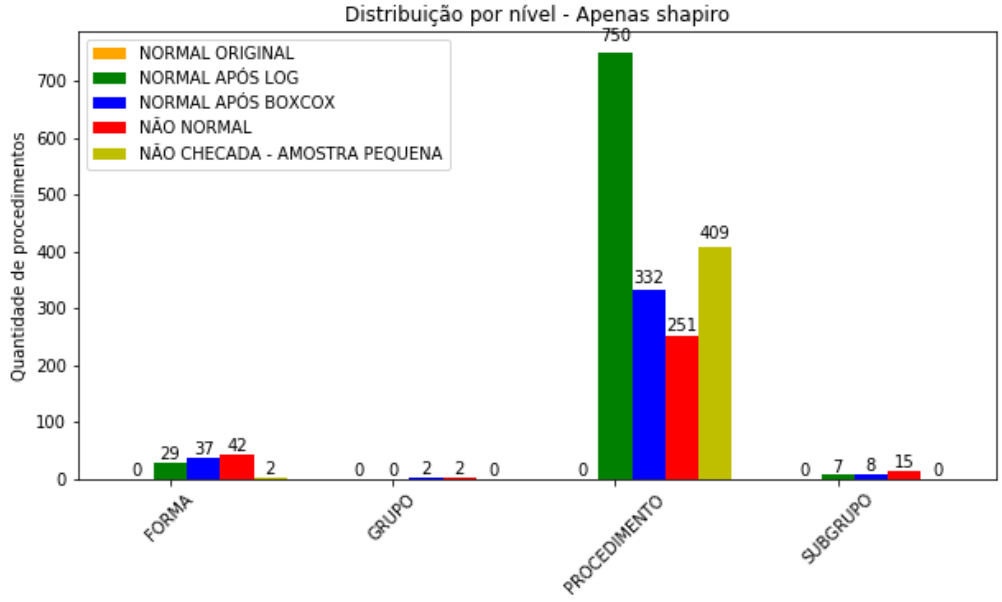
Dessa forma, foi adotado no trabalho um posicionamento mais conservador, a fim de minimizar falsos positivos, que consiste em usar os dois métodos mais eficazes e usar o número 48 como tamanho mínimo da amostra a ser testada, já que os dois métodos escolhidos são eficazes para esse valor.

Para se definir a forma de como seria feito o teste de normalidade o trabalho, foram analisados os 1886 serviços do SIGTAP que foram utilizados em 2018 pela população brasileira, sendo quatro grupos, 30 subgrupos, 110 formas e 1742 procedimentos.

Primeiramente, verificou-se como os dados seriam classificados se fossem usados apenas o teste Shapiro-Wilk, apontado como o melhor teste de normalidade (AHAD, YIN, *et al.*, 2011). Foi usado o fluxograma descrito na Figura 13: Fluxograma das Transformações e Testes de Normalidade. O resultado pode ser visto na Figura 33: Distribuição por nível de serviço usando apenas o teste Shapiro-Wilk.

Nessa análise, 310 serviços foram classificados como ‘Não Normal’, ou seja, os dados não seguem uma distribuição normal, mesmo após uma transformação logarítmica ou aplicando a função boxcox.

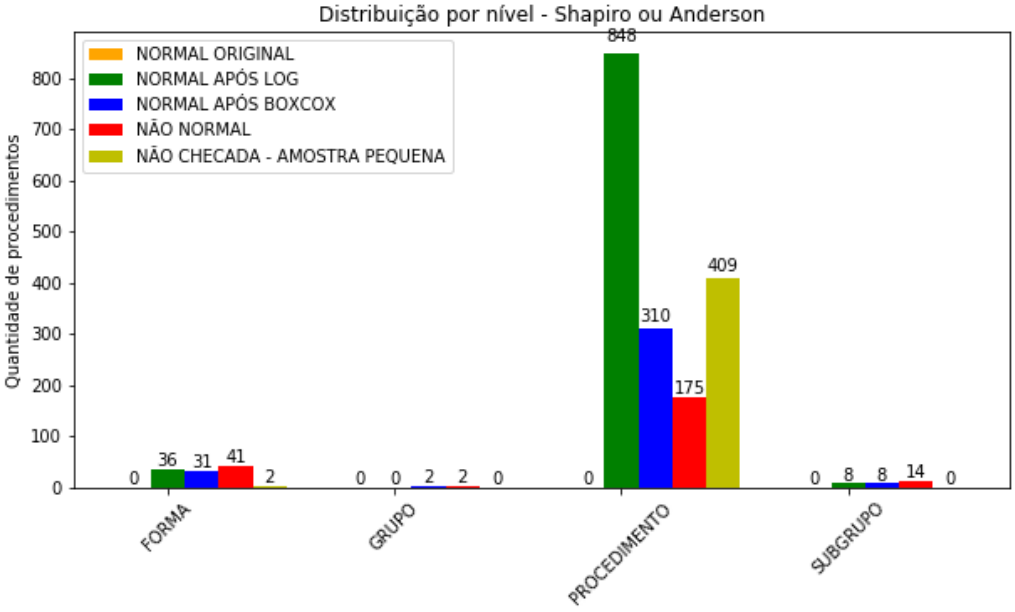
Figura 33: Distribuição por nível de serviço usando apenas o teste Shapiro-Wilk



Fonte: Elaborada pela autora (2020).

Depois, verificou-se como seria a classificação se os dados fossem considerados seguindo uma distribuição normal caso passassem no teste de Shapiro-Wilk ou no teste de Anderson-Darling, ou seja, para a amostra dos dados ser considerada seguindo uma distribuição normal, bastava ela ter passado em um dos dois testes. O resultado pode ser visto na figura abaixo.

Figura 34: Distribuição por nível de serviço usando Shapiro-Wilk ou Anderson-Darling

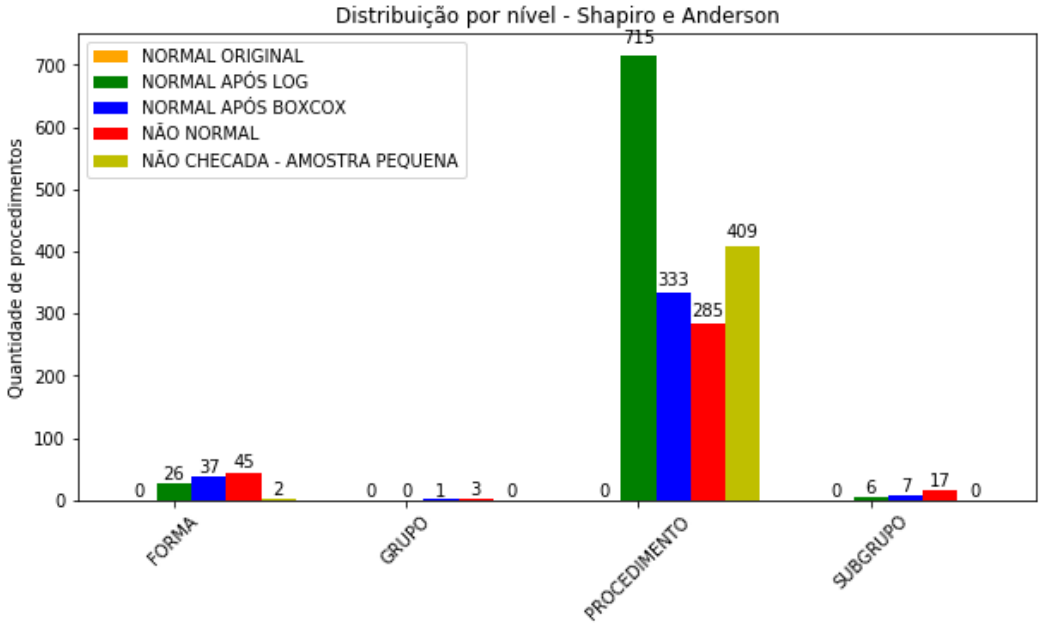


Fonte: Elaborada pela autora (2020).

Nessa análise, 232 serviços foram classificados como ‘Não Normal’.

Por fim, verificou-se como seria a classificação se os dados fossem considerados seguindo uma distribuição normal caso passassem nos testes de Shapiro-Wilk e de Anderson-Darling simultaneamente, ou seja, para a amostra ser considerada seguindo uma distribuição normal, ela teria que passar nos dois testes. O resultado pode ser visto na figura abaixo.

Figura 35: Distribuição por nível de serviço usando Shapiro-Wilk e Anderson-Darling



Fonte: Elaborada pela autora (2020).

Nessa análise, 350 serviços foram classificados como ‘Não Normal’.

É importante destacar que as três análises foram realizadas nos serviços cujas amostras de dados eram maiores que 48 e menores que 5000. As amostras menores do que 48 não foram testadas e foram classificadas como ‘NÃO CHECADA – AMOSTRA PEQUENA’. Para as amostras com tamanho maior do que 5000, foi usado apenas o teste de Anderson-Darling, devido a limitação do teste de Shapiro-Wilk para esse tipo de amostra (THE SCIPY COMMUNITY, 2019).

Para o trabalho, continuou-se optando por uma abordagem mais rígida, a fim de minimizar os falsos positivos de que uma amostra de dados seguia uma distribuição normal quando na verdade não seguia. Dessa forma, foi escolhida a terceira análise, em que uma amostra de dados só passava no teste de normalidade se passasse tanto nos testes de Shapiro-Wilk quanto de Anderson-Darling.