UNIVERSIDADE MOGI DAS CRUZES
LUIS FERNANDO REIS TAVARES PAIS

ANÁLISE DESCRITIVA DO DESLOCAMENTO DE PACIENTES EM TRATAMENTO DE CÂNCER DE MAMA NO SUS

Mogi das Cruzes, SP
2020

UNIVERSIDADE MOGI DAS CRUZES
LUIS FERNANDO REIS TAVARES PAIS

ANÁLISE DESCRITIVA DO DESLOCAMENTO DE PACIENTES EM TRATAMENTO DE CÂNCER DE MAMA NO SUS

Dissertação apresentada ao Programa de Pós-Graduação da Universidade de Mogi das Cruzes como parte dos requisitos para obtenção do grau de Mestre Profissional em Ciência e Tecnologia em Saúde.

Área de Concentração: Gestão em Saúde

Orientador: Ricardo da Silva Santos

Mogi das Cruzes, SP
2020

FICHA CATALOGRÁFICA
Universidade de Mogi das Cruzes - Biblioteca Central

Pais, Luis Fernando Reis Tavares

Análise descritiva do deslocamento de pacientes em tratamento de câncer de mama no SUS / Luis Fernando Reis Tavares Pais. – 2020.

83 f.

Dissertação (Mestrado Profissional em Ciência e Tecnologia em Saúde) - Universidade de Mogi das Cruzes, 2020
Área de concentração: Gestão em Saúde
Orientador: Prof. Dr. Ricardo da Silva Santos

1. Neoplasias da mama 2. Registros hospitalares de câncer 3. Análise de dados 4. Deslocamento 5. Acesso aos serviços de saúde 6. Mineração de dados I. Santos, Ricardo da Silva

CDD 616.99449

Elaborado por Maisa Martins de Carvalho - CRB-8/7385

UMC
UNIVERSIDADE

Tel.: (11) 4798-7000
0800 19 2001
www.umc.br

ATAS

ATA DA SESSÃO PÚBLICA DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO PROFISSIONAL EM CIÊNCIA E TECNOLOGIA EM SAÚDE DA UNIVERSIDADE DE MOGI DAS CRUZES

Às quinze horas do dia quatorze de agosto de dois mil e vinte, na Universidade de Mogi das Cruzes, realizou-se a defesa de dissertação intitulada: "ANÁLISE DESCRITIVA DO DESLOCAMENTO DE PACIENTES EM TRATAMENTO DE CÂNCER DE MAMA NO SUS" pelo(a) candidato(a) **Luis Fernando Reis Tavares Pais,** para obtenção do grau de Mestre Profissional em Ciência e Tecnologia em Saúde. Tendo sido o número de créditos alcançados pelo(a) mesmo(a) no total de 44 (quarenta e quatro), a saber: 24 unidades de crédito em disciplinas de pós-graduação e 20 unidades de crédito no preparo da dissertação, o(a) aluno(a) perfaz assim os requisitos para obtenção do grau de Mestre Profissional. A Comissão Examinadora estava constituída dos Senhores Professores Ricardo da Silva Santos e Tatiana Ribeiro de Campos Mello, da Universidade de Mogi das Cruzes e José Luiz Barbosa Bevilacqua, do Hospital Sírio Libanês, sob a presidência do primeiro, como orientador da dissertação. A Sessão Pública da defesa foi aberta pelo Senhor Presidenta da Comissão, tendo em seguida o(a) candidato(a) realizado a apresentação da dissertação. Concluída a apresentação, tiveram início as arguições pelos Membros da Comissão Examinadora. A seguir a Comissão, em Sessão Secreta, conforme julgamento discriminado por cada membro, considerou o(a) candidato(a)

Aprovado (aprovado(a) /reprovado(a)) por *unanimidade* (unanimidade/maioria).

Mogi das Cruzes, 14 de agosto de 2020.

Comissão Examinadora	Julgamento
Prof. Dr. José Luiz Barbosa Bevilacqua	*aprovado* (aprovado(a)/reprovado(a))
Prof[a] Dr[a] Tatiana Ribeiro de Campos Mello	*aprovado* (aprovado(a)/reprovado(a))
Prof. Dr. Ricardo da Silva Santos	*aprovado* (aprovado(a)/reprovado(a))

Dedicatória:

Dedico este trabalho à minha esposa Sirlei, pelo carinho e pela paciência nos períodos de ausência. Por ser mais que uma companheira, sempre. A meu filho, Vini, pelos sorrisos singelos nos momentos mais inesperados.”♪Por você, eu faria até um mestrado. ♪”

A meus irmãos, pai e mãe, pelo apoio quando necessário.

Às amigas Gláucia, Ariane e Joelma pelos pitacos informais e por me relembrarem que toda paciente é, antes de tudo, uma pessoa. Aos amigos Alberto e Roberto pelo suporte.

Aos gestores da empresa em que trabalho, pela flexibilidade nos horários.

Ao orientador Professor Dr. Ricardo, à coordenadora Dra. Kátia e demais professores do Mestrado, pela compreensão da complexa realidade dos alunos em conciliar seus trabalhos com os estudos e pela disposição em ensinar e compartilhar seus conhecimentos.

RESUMO

O câncer vem sendo considerado a principal barreira ao aumento da expectativa de vida em todos os países, com taxas de incidência e de mortalidade em crescimento acelerado. No Brasil, estima-se mais de 625 mil novos casos no triênio 2020-2022, sendo o câncer de mama o tipo mais incidente com 15,3% do total. Dentro do sistema de saúde brasileiro, o atendimento oncológico está concentrado em centros especializados, localizados em poucos municípios. Essa centralização dos serviços de saúde dificulta o acesso aos tratamentos de parte da população, fazendo com que tenham de se deslocar por grandes distâncias. Não há, contudo, estudos de amplitude nacional que possibilitem a análise detalhada desses deslocamentos em todas as esferas de saúde relacionadas (estado, macrorregião e região de saúde, e município). Por isso, este trabalho buscou descrever o perfil de deslocamento de pacientes em tratamento de câncer de mama, considerando o período de 2000 a 2016. Foram feitas análises descritivas por métodos estatísticos e por mineração de dados para mais de 490 mil ocorrências do Registro Hospitalar de Câncer. O uso das técnicas estatísticas permitiu a elaboração de análises para quatro grupos de características relacionadas às pacientes, ao tumor, aos tipos de tratamento adotados e aos deslocamentos efetuados. Também foi desenvolvido um painel para navegação e consulta dos dados dentro de cada esfera da saúde, no qual se pode selecionar o período e a região desejada, possibilitando uma ampla gama de diferentes análises. Em paralelo, foram gerados modelos usando as técnicas de agrupamento e associação, da mineração de dados, para descrever padrões encontrados na base de dados. Dentre os resultados, foram encontradas associações entre as variáveis estadiamento grupo e tipo de tratamento com determinadas faixas de deslocamento, bem como uma alta concentração de casos (69,5%) nas distâncias até 60 quilômetros. Tais análises tem o intuito de estimular discussões e reflexões entre os gestores de saúde, responsáveis pelo monitoramento das políticas públicas e pela organização das redes de atenção em saúde.

Palavras-Chave: Neoplasias da Mama. Registros Hospitalares de Câncer. Análise de Dados. Deslocamento. Acesso aos Serviços de Saúde. Mineração de Dados.

ABSTRACT

Cancer has been considered the main barrier to increasing life expectancy in every country of the world, with accelerated incidence and mortality rates. In Brazil, more than 625 thousand new cases are estimated in the 2020-2022 period, with breast cancer being the most prevalent one with 15.3% of the total. The Brazilian health system contains specialized centers for cancer care concentrated in a few municipalities, hindering access to health services for part of the population, making them travel long distances for their treatments. However, there are no studies of national amplitude that allow the detailed analysis of these displacements in all related health spheres (state, health macro-region, health region and municipality). Therefore, this study sought to describe the travel profile of patients undergoing breast cancer treatment, considering the period from 2000 to 2016. Descriptive analyses were performed by statistical methods and data mining for more than 490,000 occurrences of the Hospital Cancer Registry. Statistical techniques allowed analyses for four groups of characteristics related to patients, tumor, adopted treatments and travels performed. A dashboard with these groups was also developed for each health sphere, in which the desired period and region can be selected, enabling a wide range of analyses. Besides that, models using association and clustering techniques, from data mining, were generated to describe patterns found in the database. Among the results, associations were found between the variables group staging and type of treatment with certain travel distance ranges, as well as a high concentration of cases (69.5%) in distances up to 60 kilometers. These analyses aim to stimulate discussions and reflections among health managers, responsible for monitoring public policies and organizing health care networks.

Keywords: Breast neoplasms. Hospital Cancer Registries. Data Analysis. Travel burden. Access to Health Services. Data Mining.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O câncer vem sendo considerado a principal barreira ao aumento da expectativa de vida em todos os países, com taxas de incidência e de mortalidade em crescimento acelerado. As razões para esse aumento são complexas, em parte explicadas pelo envelhecimento populacional, em parte pela mudança nos fatores de risco associados ao câncer, como a urbanização e desenvolvimento socioeconômico (BRAY *et al*., 2018).

A Organização Mundial de Saúde (OMS) estimou que em 2018, no mundo todo, tenham ocorrido mais de 18 milhões de novos casos, com cerca de 9,5 milhões de mortes (BRAY *et al*., 2018). No Brasil, o Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA) estima a ocorrência de mais de 625 mil casos novos no país a cada ano para o triênio de 2020-2022. Com todo esse impacto no sistema de saúde, é essencial aos gestores monitorar e organizar ações de controle e prevenção ao câncer (INCA, 2019).

No intuito de revisar e aprimorar suas políticas nacionais para prevenção e combate a essa doença dentro do Sistema Único de Saúde (SUS), o governo, através da Portaria n° 3.535 do Ministério da Saúde (MS), de 2 de setembro de 1998, estabeleceu os primeiros critérios para cadastramento e habilitação de centros de atendimento em oncologia no SUS (BRASIL, 1998).

Em 19 de dezembro de 2005, a Portaria nº 741 do MS redefiniu as diretrizes para a implantação e credenciamento dos serviços de alta complexidade em oncologia no país. Foram estabelecidos critérios para diferenciar as Unidades de Assistência de Alta Complexidade em Oncologia (UNACON) dos Centros de Assistência de Alta Complexidade em Oncologia (CACON), bem como as condições exigidas para que uma unidade hospitalar possa ser qualificada em um ou mais dos serviços (BRASIL, 2005).

Em 2013, foi instituída no Brasil a “Política Nacional para a Prevenção e Controle do Câncer na Rede de Atenção à Saúde das Pessoas com Doenças Crônicas no âmbito do SUS”, através da Portaria nº 874, reconhecendo o câncer como uma doença crônica prevenível, que demanda cuidado integral. Além disso, enfatizou que deve ser planejada uma “organização de redes de atenção regionalizadas e descentralizadas, com respeito a critérios de acesso, escala e escopo” (BRASIL, 2013).

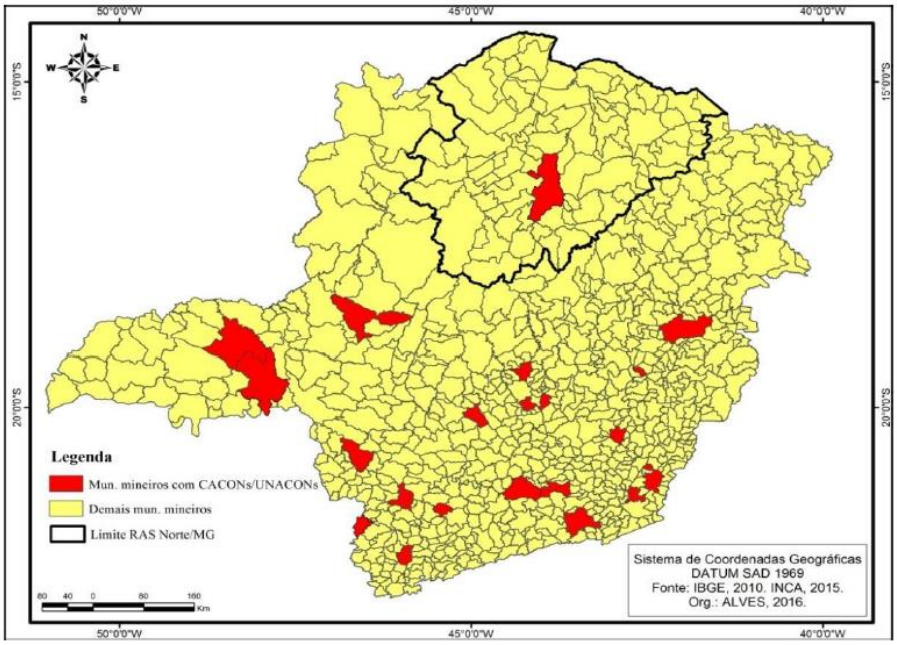
Uma rede de atenção para diagnóstico e tratamento descentralizada pressupõe que o acesso a tais serviços seja feito em unidades próximas ao local de residência dos pacientes. Porém, há alguns fatores no SUS que dificultam a formação de redes regionais homogêneas em relação à prestação dos serviços de saúde: a heterogeneidade do território brasileiro, a

formalização de responsabilidades institucionais em cada um dos níveis de atenção, e a delicada questão político-administrativa da manutenção de autonomia dos governos locais com as regulações federais e estaduais, entre outros (SALDANHA *et al*., 2019).

O Brasil, com seus mais de cinco mil municípios, tinha em 2017 apenas 296 centros de tratamento habilitados para o tratamento oncológico. Desses, quase metade (138) ficavam na região Sudeste, que contém cerca de 42% da população e 30% das cidades, embora detenha apenas 11% da área territorial[1] total do país. Já a região Norte, que tem 8% dos municípios e 9% da população em uma área que corresponde a 45% do país, tinha apenas 30 centros oncológicos, o que demonstra a grande disparidade entre as regiões brasileiras (MELO, 2017).

As diferenças, entretanto, ocorrem até mesmo dentro de um único estado. Os autores Oliveira e Muniz (2017) citam a região de saúde do norte de Minas Gerais como um exemplo em que a distribuição geográfica não está otimizada, tendo somente um centro municipal com serviços oncológicos para atender toda a região, incluindo cidades a mais de 300 quilômetros de distância. A Figura 1 ilustra essa discrepância, destacando em vermelho os municípios que possuem instituições de tratamento para câncer. É possível observar que a maioria dos centros de tratamento estão nas regiões central e sul do estado.

Figura 1 – Espacialização dos centros oncológicos no estado de Minas Gerais



Fonte: Oliveira e Muniz (2017), Mapa 03

[1] Instituto Brasileiro de Geografia e Estatística (IBGE). Acesso em 10 jun 2020. Disponível em: https://cidades.ibge.gov.br/brasil/panorama

A consolidação de centros especializados em oncologia, buscando a otimização dos recursos de saúde, é comum em várias partes do mundo. Contudo, isso faz com que os pacientes muitas vezes tenham de percorrer longas distâncias para receber o diagnóstico e o tratamento adequados. A relação entre o deslocamento dos pacientes e o estágio do câncer no diagnóstico ou a efetividade do tratamento vem sendo alvo de diversos estudos.

Para Payne, Jarrett, Jeffs, (2000), que fizeram uma das primeiras revisões de literatura sobre o impacto do deslocamento dos pacientes durante o tratamento, os resultados encontrados em onze estudos não foram homogêneos. Em uma revisão mais recente Ambroggi *et al*, 2015, os autores concluem que viagens longas estão associadas com: graus avançados da doença ao diagnóstico, tratamentos inapropriados, prognósticos ruins e piora na qualidade de vida.

No cenário brasileiro, há diversos estudos que avaliaram deslocamentos de pacientes em tratamento para câncer de mama sem, entretanto, analisar os impactos associados a diferentes distâncias (OLIVEIRA *et al*., 2011; SALDANHA *et al*., 2019; SILVA *et al*., 2019).

Embora tenham avaliado períodos diferentes, todos indicaram uma alta concentração de atendimentos nas regiões mais povoadas e em poucos centros especializados, com cerca de metade dos tratamentos sendo feito no mesmo município do paciente. Também em comum foi o fato de as análises terem sido feitas considerando o país todo, usando as bases de dados disponibilizadas pelo Departamento de Informática do SUS (DATASUS), que contém dados de internações e atendimentos ambulatoriais.

Em um outro artigo nacional, os autores Renna Junior e Silva (2018) usaram uma base diferente: os Registros Hospitalares de Câncer (RHC) contém dados de atendimentos de todas as unidades hospitalares credenciadas para tratamento de câncer do país, e estão disponíveis publicamente no site do INCA. Os autores avaliaram a tendência temporal para diagnóstico de câncer de mama em estágio avançado no Brasil, para o período de 2000 a 2012. Concluíram que o acesso ao diagnóstico é desigual no país, com mulheres em nível socioeconômico mais baixo tendo uma maior probabilidade de doença mais avançada ao serem diagnosticadas.

O câncer de mama é frequentemente selecionado nas análises da doença por ser o tipo mais comum no Brasil (exceto pelo câncer de pele não melanoma), com mais de 66 mil casos estimados para 2020, correspondendo a cerca de 30% dos casos entre as mulheres (INCA, 2019). Em número de mortes, o câncer de mama é o terceiro no geral e o primeiro entre as mulheres, com 18.442 óbitos estimados (BRAY *et al*., 2018).

Por essa alta incidência e mortalidade do câncer de mama no cenário nacional, e tendo em vista as dificuldades enfrentadas por pacientes que precisam se deslocar para conseguir um diagnóstico em tempo hábil com um tratamento adequado, é essencial que haja mais trabalhos

que forneçam informações úteis sobre os trajetos percorridos e suas variáveis associadas. Esse tipo de estudo pode auxiliar os órgãos de políticas públicas, provendo subsídios essenciais para melhorias na organização de sua rede de atendimento ao tratamento oncológico.

Em virtude da escassez de materiais voltados ao tema, está sendo desenvolvido pelo Laboratório de Computação Aplicada à Medicina (LACAM) da Universidade de Mogi das Cruzes (UMC) um projeto denominado "Técnicas de *Big Data* no Estudo do Impacto de Deslocamento de Pacientes em Tratamento Oncológico". Para que o projeto possa avaliar tais impactos, uma das fases iniciais é a descrição de tais deslocamentos.

Assim, este estudo visa auxiliar no preenchimento dessa lacuna, por meio de análises que descrevam o perfil de deslocamento de pacientes com câncer de mama em tratamento no país. Para que se obtenha uma melhor compreensão desse perfil, são propostas também análises de outras características, associadas ao paciente, ao tumor, ao tipo de tratamento adotado e às especificidades de cada unidade da Federação.

As análises aqui propostas têm o intuito de estimular discussões e reflexões entre os gestores de saúde, responsáveis pelo monitoramento das políticas públicas e pela organização das redes de atenção em saúde. A fim de auxiliá-los em suas análises e considerando a característica regionalizada e descentralizada dos serviços oncológicos no SUS, este estudo fornecerá um painel para navegação e consulta dos dados dentro de cada esfera da saúde.

A base de dados utilizada neste estudo terá como origem o RHC, selecionando registros a partir do ano 2000, de todo o Brasil. O escopo está restrito aos dados públicos disponibilizados pelo INCA e é específico para o câncer de mama, por ser o mais comum no país.

2 OBJETIVOS

O objetivo geral deste estudo é o desenvolvimento de modelos analíticos capazes de descrever os deslocamentos de pacientes em tratamento de câncer de mama, a partir dos Registros Hospitalares de Câncer (RHC) do INCA. Espera-se com essa análise colaborar para auxiliar os órgãos de políticas públicas com subsídios essenciais para a organização de sua rede de atendimento ao tratamento oncológico, buscando futuras melhorias e ampliações.

A fim de atingir o objetivo geral, foram definidos alguns objetivos específicos:

- Atualização do repositório de dados utilizado nas análises, a partir dos dados do RHC;
- Estudos exploratórios da base de dados, agrupando características de deslocamento em cada região, criando perfis por paciente, tipo de tratamento e tumor;
- Montagem de painel de consultas com flexibilidade para diversas análises e filtros por ano/região;
- Disponibilização do painel na *internet*;
- Estudo de técnicas analíticas apropriadas ao conjunto de dados;
- Aplicação de técnicas descritivas de mineração de dados para representação do conhecimento obtido.

Para que as análises sejam amplas e atuais, deve-se alimentar o banco de dados do LACAM com informações consideradas recentes, muito embora os dados públicos costumem ter uma certa demora em sua divulgação.

Para que se possa compreender melhor as características do deslocamento nas regiões em estudo, entende-se que seja importante conhecer também o perfil dos pacientes, do tumor e dos tratamentos envolvidos.

Esse conhecimento será inicialmente feito através de técnicas estatísticas descritivas habituais (como gráficos e tabelas), criando modelos exploratórios que serão também disponibilizados em um painel de consultas interativo na página do LACAM, permitindo diferentes análises por região e período.

Por fim, serão também usadas técnicas de mineração de dados (MD), cujos algoritmos computacionais são otimizados para lidar com grandes volumes de dados, para a elaboração de modelos analíticos descritivos relacionados ao deslocamento.

3 CONCEITUAÇÃO

Serão definidos neste capítulo três tópicos utilizados neste estudo, a fim de tornar mais claros seus conceitos. Foram selecionados: o Registro Hospitalar de Câncer, por ser a principal base de dados usada na elaboração dos modelos descritivos; a divisão da saúde pública vinculada às esferas administrativas estaduais e, por fim, uma sucinta conceituação sobre mineração de dados, enfatizando as duas técnicas descritivas usadas.

3.1 REGISTRO HOSPITALAR DE CÂNCER

A coleta sistemática de informações a respeito de doenças e mortalidade de populações é prática estabelecida há tempos na área da saúde. Classicamente, para o câncer, existem dois tipos de registros: os populacionais e os hospitalares. Os registros de base populacional são direcionados a áreas geográficas específicas, visando conhecer a incidência e mortalidade da doença na população. Já os registros hospitalares são "implantados em hospitais que atendem pacientes com câncer, com o objetivo de conhecer o perfil da população assistida na instituição, os recursos utilizados e a efetividade dos tratamentos oferecidos" (INCA, 2010).

Os Registros Hospitalares de Câncer (RHC) podem ser caracterizados como centros de informação situados em Unidades Hospitalares (UH) de assistência oncológica, cujas atividades incluem: a coleta de dados dos pacientes com diagnóstico confirmado de câncer atendidos nessas unidades; o armazenamento e processamento desses dados; a análise e divulgação dessas informações (PINTO *et al*., 2012).

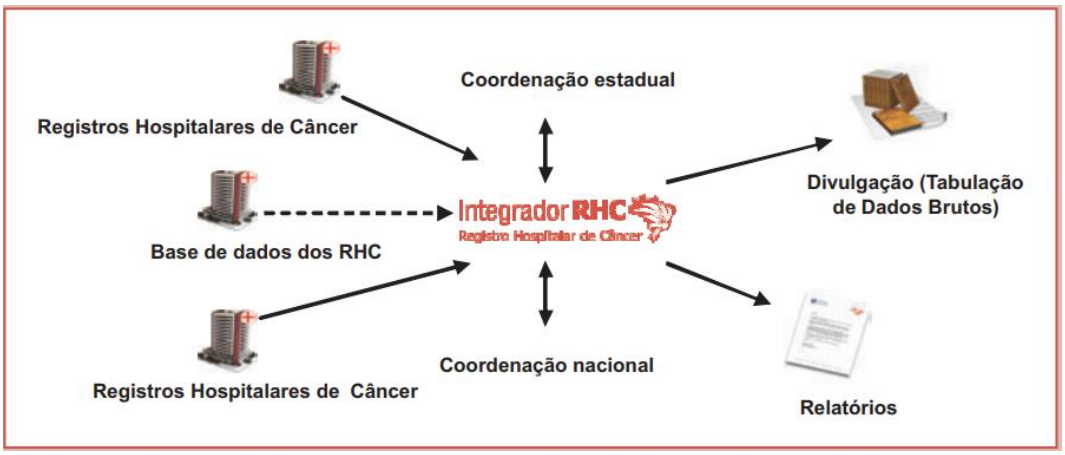
São, portanto, ferramentas importantes para aprimorar a assistência oferecida aos pacientes, uma vez que contém aspectos demográficos, indicam a evolução da doença ao longo do tempo, os tipos de tratamentos utilizados, e permitem avaliar o desempenho da UH na atenção ao paciente oncológico (INCA, 2010).

O INCA é o órgão responsável pela consolidação e publicação em âmbito nacional das informações geradas pelos RHC de todo o Brasil. Para isso, desenvolveu dois sistemas: um deles é o SisRHC, disponibilizado gratuitamente aos centros habilitados em oncologia para auxiliar na padronização de entrada dos dados e no envio de relatórios. Também está disponível desde 2007 o IntegradorRHC, um sistema web que faz a consolidação e divulgação dos dados hospitalares oriundos de todos os RHC (INCA, 2011).

Para manter os sistemas permanentemente atualizados, o INCA realiza oficinas de consenso com especialistas, revisando e aprimorando seus instrumentos de coleta de informações. A cada cinco anos, as fichas de coleta são ajustadas para contemplar as novas rotinas e procedimentos, como novos campos obrigatórios, opcionais ou relevantes para a instituição (INCA, 2010).

As UH enviam os dados através da internet e cabe às coordenações estaduais fazerem uma primeira consolidação estadual. Posteriormente, é feita também uma consolidação nacional, para que só então os dados possam ser divulgados no site do IntegradorRHC (no mínimo uma vez por ano, mas há anos em que são feitas mais atualizações parciais). A principal função dessa consolidação é identificar redundâncias (casos em que o paciente foi registrado em mais de um centro de tratamento) e eliminá-las, evitando assim que o número de ocorrências seja indevidamente superestimado (INCA, 2011). A Figura 2 ilustra este fluxo de informações.

Figura 2 - Fluxo de informações no Integrador RHC



Fonte: INCA, 2011

O primeiro RHC que se tem registro no Brasil foi criado em 1983, no Rio de Janeiro (INCA, 2012). O funcionamento desses registros vem sendo aprimorado ao longo do tempo, amparado por diversas portarias do Ministério da Saúde. A primeira citação foi em 1993, na portaria 171 do Ministério da Saúde, em sua classificação de hospitais de atendimento oncológico do SUS (DOS MINISTÉRIOS, 2007). Em 1998, foram estabelecidos os critérios para cadastramento de centros de atendimento em oncologia, dentre os quais foi previsto que deveriam dispor e manter em funcionamento seus RHC (BRASIL, 1998).

A informatização desses registros foi reafirmada e fortalecida pela Portaria MS Nº741, de dezembro de 2005, que estabeleceu o prazo de até um ano para sua implantação obrigatória.

Com isso, iniciou-se a expansão dos RHC no Brasil, passando de 176, em 2005, para 268, em 2012. A portaria também definiu que o envio das informações para o INCA deveria ter periodicidade anual, no mês de setembro, iniciando a partir de 2007 (BRASIL, 2005; INCA, 2012).

Outro marco importante da portaria MS Nº 741 foi a definição dos critérios que diferenciam uma UNACON de uma CACON. Ambos são serviços credenciados de alta complexidade em oncologia, mas o foco inicial da assistência da UNACON são os cânceres mais prevalentes do Brasil, enquanto na CACON são todos os tipos de câncer. Para ser credenciada como CACON, uma unidade hospitalar deve obrigatoriamente ter quatro serviços: cirurgia oncológica, oncologia clínica, radioterapia e hematologia. Já a UNACON é obrigada a manter somente os dois primeiros (cirurgia e clínica), sendo opcionais os demais, podendo incluir também serviços de oncologia pediátrica (BRASIL, 2005).

Em 2013, na implantação da "Política Nacional para a Prevenção e Controle do Câncer na Rede de Atenção à Saúde das Pessoas com Doenças Crônicas no âmbito do Sistema Único de Saúde (SUS)", foram estabelecidas as responsabilidades para as secretarias estaduais de saúde, reforçando a necessidade de apoiar a implantação dos RHC e fazer o acompanhamento do envio dos dados anualmente junto às unidades credenciadas (BRASIL, 2013).

Recentemente, em 2019, ao redefinir os critérios de habilitação de estabelecimentos de saúde na alta complexidade em oncologia no SUS, a implementação de RHC foi novamente contemplada como uma das competências da unidade hospitalar, bem como o repasse desses dados ao IntegradorRHC do INCA, reforçando assim a importância da coleta e envio das informações (BRASIL, 2019).

3.2 DIVISÃO ADMINISTRATIVA DA SAÚDE PÚBLICA

O Brasil contém um dos maiores e mais abrangentes sistemas de saúde pública do mundo, o SUS. Com ações que vão da prevenção até as emergências, de atendimentos simples até procedimentos de alta complexidade, o SUS garante o acesso integral, gratuito e universal a toda a população do país. A rede que o compõe é ampla e a gestão desses serviços é dividida entre os três entes da Federação: a União, os estados e os municípios[2]. Ressaltando que os municípios são, de fato, os responsáveis imediatos pelo atendimento das demandas de saúde da população e das intervenções saneadoras que se fizerem necessárias.

[2] Ministério da Saúde, portal SUS. Acesso em 06 jun.2020. Disponível em: https://www.saude.gov.br/sistema-unico-de-saude

O SUS foi instituído pela Constituição Federal de 1988 e consolidado pelas leis 8.080/90 e 8.182/93. Por abranger todo o território brasileiro, é norteado por alguns princípios que permitem uma melhor organização de suas entidades, implicando numa hierarquia na operação do sistema. Na esfera federal o Ministério da Saúde atua como órgão planejador das políticas públicas nacionais, em corresponsabilidade com os estados e municípios através de suas respectivas secretarias, que atuam na implementação das políticas e execução das ações de saúde (CARVALHO, 2013).

Visando uma estruturação que busque a otimização dos recursos e o atendimento de todos os cidadãos, numa perspectiva harmônica e integrada, a Norma Operacional da Assistência à Saúde – NOAS 01/2001 – regulamentou a diretriz geral para a organização regionalizada da saúde pública brasileira. Coube a cada estado a responsabilidade de definir seu Plano Diretor de Regionalização da Saúde (PDR), com autonomia para identificar suas macrorregiões e/ou regiões (GUIMARÃES, 2005). Citando trecho da Portaria MS/GM n. 373, que regulamentou a NOAS:

A Portaria MS/GM nº 373, de 2002, regulamentou a NOAS, indicando que o processo de regionalização deveria contemplar um planejamento integrado, respeitando a territorialidade e os limites do município como unidade indivisível, identificando as prioridades para intervenção de forma a garantir o acesso de todos os cidadãos aos serviços necessários para resolver seus problemas de saúde (BRASIL, 2002).

Em 2011, o decreto presidencial nº 7.508 regulamentou as disposições sobre a organização do SUS e do planejamento e assistência em saúde. Em seu artigo 2º, define a Região de Saúde como 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". As regiões são instituídas pelos estados em articulação com os municípios, embora possam ser interestaduais, conforme atos conjuntos dos respectivos entes federativos. Sua finalidade é integrar a organização, o planejamento e a execução de ações e serviços de saúde (BRASIL, 2011).

Portanto, a região de saúde já nasce imbuída de garantir iguais direitos ao cidadão às ações e serviços de saúde próximos de onde vive, sem onerar nenhum ente federativo além de suas possibilidades econômicas, espaciais e demográficas. Para tanto, um planejamento integrado é essencial, no qual os municípios planejam de forma compartilhada e cooperativa, olhando para a região de saúde em que se encontram e considerando as características de cada um. Na prática, isso implica também numa diversificada rede técnico-sanitária, na qual os serviços mais simples ficam mais dispersos e os de maior densidade tecnológica (por exemplo,

unidades hospitalares especializadas) tendem a ser concentrados em menos cidades (SANTOS, 2017).

Conforme mencionado por Duarte *et al*. (2015), a regionalização só faz sentido quando é elaborada junto com a hierarquização, integrando cada uma das esferas administrativas de maneira planejada e organizada. Nesse contexto hierárquico, este estudo considerou que a divisão administrativa da saúde pública brasileira tem, em sua base, o município como entidade principal. É nele que vivem as pessoas que necessitam dos serviços de saúde, e é nele também que se localizam as unidades hospitalares de tratamento oncológico.

Partindo do município, configura-se então a próxima esfera do nível hierárquico: a região de saúde. Em alguns estados, foi definida também a esfera da macrorregião de saúde, que nada mais é que um agrupamento de regiões contíguas. Cada uma das macrorregiões de saúde pertence a um dos 27 entes federativos (estados e distrito federal). Essas foram, portanto, as quatro esferas de saúde utilizadas neste estudo: município, região de saúde, macrorregião de saúde e Unidade da Federação.

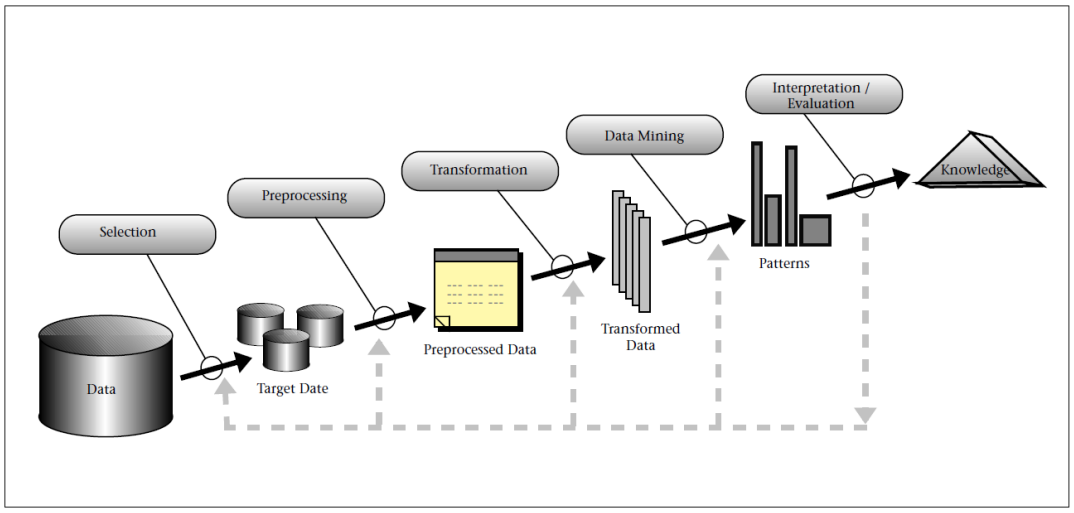
3.3 MINERAÇÃO DE DADOS

3.3.1 Visão Geral

O volume de dados disponível vem aumentando exponencialmente na área da saúde e em muitas outras. Com isso, fica também cada vez mais difícil extrair informações úteis desse gigantesco oceano digital. Para auxiliar na análise desses dados e obtenção do conhecimento inerente a eles, existem diversos métodos e teorias computacionais, que se enquadram num processo chamado de "descoberta de conhecimento em bancos de dados" – ou, em inglês, KDD, acrônimo para *Knowledge Discovery in Databases*, que é a sigla mais comumente usada. (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996).

O processo KDD busca identificar padrões que sejam compreensíveis, úteis, novos e válidos, a partir de grandes e complexos conjuntos de dados. Envolve algumas etapas, que se iniciam com uma seleção de um conjunto de dados, seu pré-processamento e sua transformação, passando pela etapa principal de mineração de dados e culminando com a análise e a avaliação do conhecimento descoberto. É comum que em uma das etapas seja necessário retornar a algum passo anterior e, por isso, o processo é descrito como sendo tanto iterativo quanto interativo. A Figura 3 ilustra uma visão geral de todas essas etapas (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996; MAIMON *et al*., 2009).

Figura 3- Visão geral das etapas do processo KDD



Fonte: Fayyad, Piatetsky-Shapiro, Smyth (1996)

As técnicas descritivas buscam encontrar modelos e padrões que representem o conhecimento obtido em formatos legíveis e compreensíveis a uma pessoa. Já as preditivas visam modelos que permitam prever, baseados num histórico de ocorrências anteriores, os valores dos atributos desejados em situações futuras ou desconhecidas (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996; SANTOS *et al*., 2013).

Existem várias técnicas diferentes, com inúmeros algoritmos para sua implementação. A seguir, serão conceituadas e exemplificadas algumas das técnicas consideradas principais, sendo que neste estudo foram usadas as duas primeiras descritivas.

3.3.2 Agrupamento

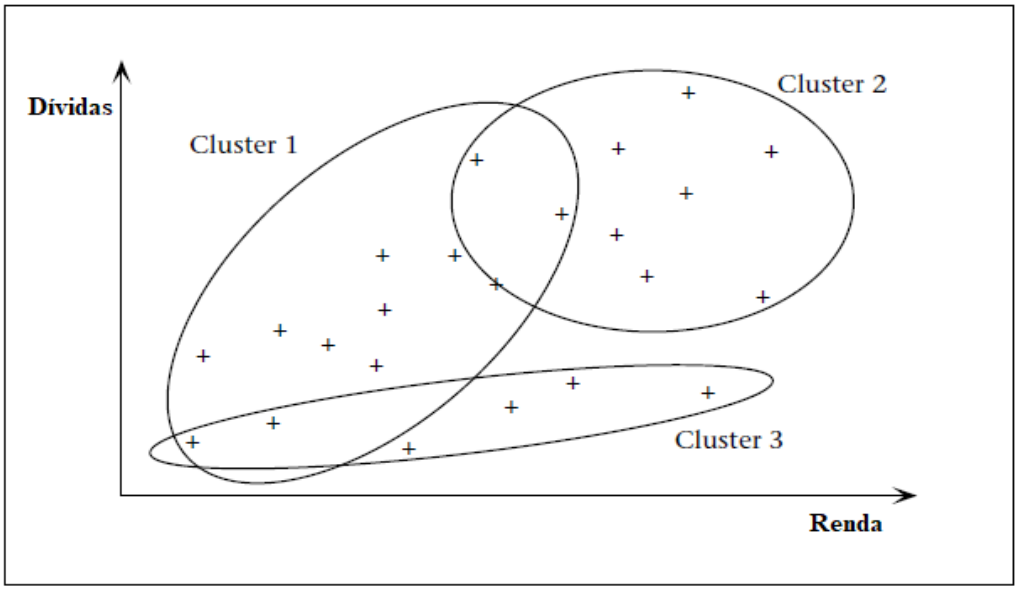
A técnica descritiva chamada de Agrupamento (ou *Clustering*) consiste em identificar grupos que possam ter características ou comportamentos similares, permitindo uma descrição mais sucinta dos dados (GOEBEL, GRUENWALD, 1999). Esses grupos, também chamados de subconjuntos ou categorias, podem ser mutuamente exclusivos e exaustivos, ou ter formatos hierárquicos ou até mesmo elementos sobrepostos (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996).

Há diversas aplicações para uso da técnica de agrupamento. Em Camilo e Silva (2009), os autores citam: reconhecimento de padrões, processamento de imagens, taxonomia de plantas e animais, segmentação de mercado para um nicho de produtos, detecção de comportamentos atípicos (fraudes), pesquisas geográficas, entre outras.

O Agrupamento é, muitas vezes, usado em conjunto com outras técnicas, de modo a se atingir um resultado desejado. Por exemplo: pode ser necessário identificar grupos de clientes que tenham o mesmo comportamento de consumo para, em seguida, usar uma técnica que preveja e sugira novos itens de compra a esses grupos (GOEBEL, GRUENWALD, 1999).

A Figura 4 ilustra um exemplo de agrupamento para um conjunto de variáveis em que o sinal + representa os níveis de renda e endividamento dos clientes de uma empresa. É possível perceber três grupos nesta simulação, com alguns elementos sobrepostos e com quantidades de registros (clientes) diferente em cada.

Figura 4 - Representação simples de uma base de clientes.



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro, Smyth (1996)

Um outro exemplo de *clustering* são as faixas etárias definidas pelo IBGE, que agrupam as idades em intervalos de cinco em cinco anos, para facilitar eventuais análises. A mesma variável (idade) pode também ser agrupada de outra forma, nos chamados grupos etários: jovens (idade de 0 a 19 anos), adultos (20 a 59 anos) e idosos (acima de 60 anos).

No caso deste estudo, a técnica foi usada para identificar as faixas de deslocamento usadas pelas pacientes em tratamento de câncer, uma vez que a distância, de forma isolada, poderia ter uma grande variabilidade (com um intervalo de zero até milhares de quilômetros). Também foi usada para identificar intervalos de tempo relacionados à demora no tratamento – algo que pode variar de poucos dias até anos de espera.

3.3.2 Associação

A outra técnica descritiva usada neste estudo é chamada Associação. Dado um conjunto de itens, ela identifica relacionamentos entre determinados atributos de modo que a presença de um padrão implica na presença de outro (GOEBEL, GRUENWALD, 1999).

Segundo Camilo e Silva (2009), um outro nome pelo qual é bastante conhecida é "Análise de Cesta de Compras", por ser muito usada para identificar produtos que são levados juntos pelos consumidores.

A técnica de Associação demanda que se tenha atributos de entrada (E) e de saída (S) para poder analisar se há alguma ligação entre eles, gerando regras num formato "*Se* atributo E *então* atributo S". Na Análise da Cesta de Compras, o objetivo é identificar o quão provável é que o cliente compre um certo item "S", dado que ele está comprando um item "E". A execução dessa técnica numa base de dados de um supermercado pode gerar diversas regras como, por exemplo, "se compra pão e leite, então compra manteiga" e "se compra vinho, então compra queijo" (CAMILO e SILVA, 2009).

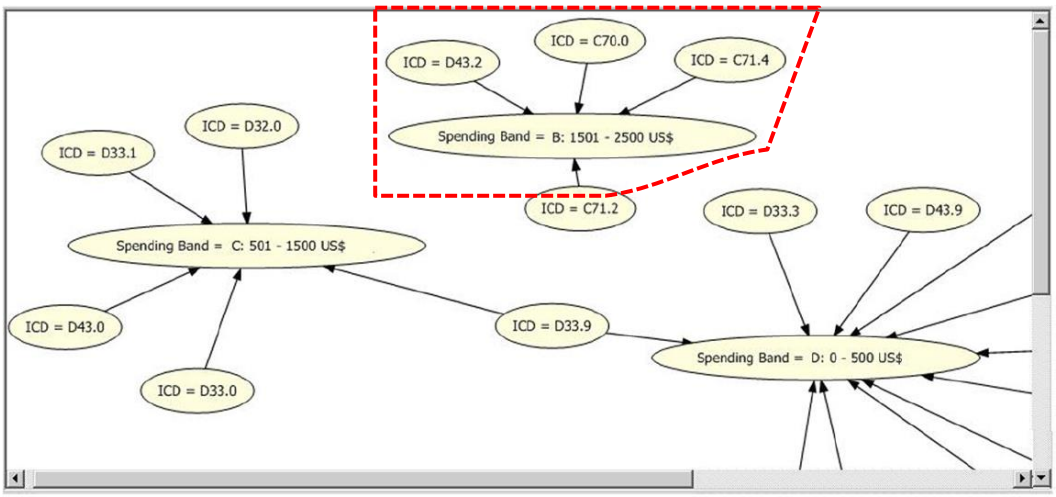
Num volume grande de dados, há inúmeras regras que podem ser geradas e, por isso, alguns conceitos são necessários para selecionar os resultados considerados válidos e significativos. Os principais conceitos usados são: suporte, confiança e importância. Suporte (ou frequência) indica a quantidade ou o percentual de registros na base que se encaixam na regra. Confiança (ou probabilidade) mede o percentual de registros que atendem especificamente a essa regra. Importância (chamada também pelo termo em inglês *lift*) mede o grau de interesse e utilidade da regra e a interdependência entre os atributos (CAMILO e SILVA, 2009).

Desse modo, no exemplo anterior da Cesta de Compras, a primeira regra "se compra pão e leite, então compra manteiga" é gerada com uma probabilidade (confiança) que pode ser maior ou menor que a outra regra "se compra vinho, então compra queijo". Em geral, os softwares utilizados para execução da técnica já calculam os índices de confiança, suporte e importância, possibilitando que sejam avaliados em conjunto para identificar as regras mais significativas conforme o objetivo almejado.

A Figura 5 mostra uma representação gráfica de algumas regras de associação geradas no artigo de Santos *et al*. (2013), relacionando o tipo de tumor cerebral com faixas de gastos hospitalares. O tipo de tumor é o atributo de entrada (indicado pelo Código Internacional de Doenças-CID, ou ICD na sigla em inglês) e a faixa de gastos no hospital é o atributo de saída (elipses com o texto "*Spending Band*" e intervalos de valores em dólares). É possível identificar

associações específicas entre quatro tipos de tumores e gastos de hospitalização maiores, destacadas na área hachurada em vermelho na imagem.

Figura 5 - Representação gráfica de regras de associação entre dois atributos



Fonte: Santos *et al.* (2013)

3.3.2 Técnicas Preditivas

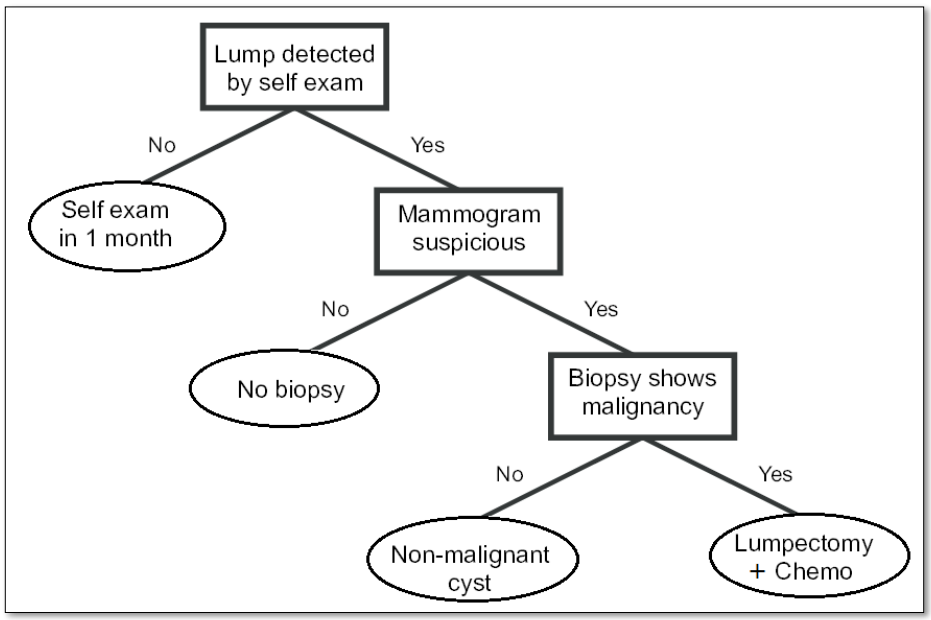
Dentre as técnicas preditivas, uma das mais comuns é a Classificação. Consiste em determinar, baseada num conjunto de categorias (classes) definidas pelo algorítmo, a qual delas um novo item pertence. Por exemplo: baseado no histórico de pacientes e em como eles respondem a diferentes tratamentos, a técnica pode identificar a qual tratamento um novo paciente tem maiores chances de responder bem (GOEBEL, GRUENWALD, 1999).

Há diversas implementações diferentes desta técnica, que podem ser usadas dependendo do tipo de atributo sendo analisado (quantitativo ou qualitativo, isto é, numérico ou não). Os algoritmos de árvores de decisão são bastante usados por sua fácil legibilidade e aplicabilidade, uma vez que os resultados ficam organizados de forma hierárquica e, para identificar a classificação final, basta seguir o fluxo conforme cada nó decisório. Além deles, há muito outros como a classificação Bayesiana, as redes neurais artificiais, SVM (*Support Vector Machines*), os algoritmos genéticos etc. (CAMILO e SILVA, 2009).

A Figura 6 ilustra uma árvore de decisão bastante simplificada (CRUZ, WISHART, 2006), que indica caminhos possíveis para diagnóstico e tratamento de câncer de mama. Cada retângulo (nó) representa uma decisão ou evento, e a partir dele saem os ramos com as opções, levando a uma consequência (folha), representada na imagem pelas elipses. Neste exemplo há

três níveis de decisão e somente duas opções para cada uma, mas em situações reais é possível ter diversos ramos, com percentuais de probabilidade em cada um, levando às decisões seguintes até que se atinja o objetivo desejado.

Figura 6 - Exemplo de árvore de decisão simplificada



Fonte: Adaptado de Cruz, Wishart (2006).

Outra técnica preditiva é a Regressão (sendo por vezes também chamada de predição numérica), que permite a análise de dependência de valores de alguns atributos sobre os valores de outros atributos do mesmo registro e a geração automática de um modelo que preveja esses valores para novos itens na base. Também tem uma gama ampla de métodos, como a regressão linear, a não-linear, a logística, entre outros (GOEBEL, GRUENWALD, 1999; CAMILO e SILVA, 2009).

4 TRABALHOS RELACIONADOS

Foram pesquisados trabalhos sobre deslocamentos de pacientes em tratamento de câncer de mama (CM) no Brasil e no mundo. Para cada trabalho apresentado nesta seção, são apresentados o objetivo, a metodologia e os resultados relevantes, contrapondo com os mesmos tópicos deste estudo.

4.1 TRABALHOS NACIONAIS

Dentre os trabalhos pesquisados restritos ao cenário brasileiro, todos usaram dados públicos. Nenhum artigo utilizou técnicas de MD para identificar padrões relacionados ao deslocamento. Em relação à abrangência geográfica dos deslocamentos, a maioria analisa o Brasil todo e se assemelham a este estudo, enquanto um deles se concentrou num único estado. Também há diferenças sobre o período analisado, pois a maioria restringiu seu escopo a poucos anos, enquanto este estudo analisa um período de 17 anos (de 2000 a 2016).

No artigo de Oliveira, Muniz (2017), os autores tinham por objetivo analisar a dinâmica das condições de acesso das mulheres com CM aos serviços de saúde no norte de Minas Gerais. Para tanto, fizeram um estudo exploratório que usou os dados do RHC de dois hospitais dessa macrorregião de saúde, dos anos de 2004 a 2014, como base para diversos mapeamentos geográficos e epidemiológicos. Parte da metodologia envolveu também pesquisa documental, aplicação de questionários às pacientes e visitas técnicas às unidades de assistência oncológica.

Os autores informam que foi observado aumento na incidência da doença no norte de Minas Gerais, devido a fatores que incluem falta de informação, baixos indicadores sociais e dificuldades para percorrer grandes distâncias para o diagnóstico e tratamento. O aumento no número de casos de CM reportado pelos autores é ratificado pelos números encontrados neste estudo, tanto em relação a todo o estado de Minas Gerais quanto somente da macrorregião de saúde norte.

A partir dos dados do DATASUS, três pesquisas que tiveram abrangência nacional (OLIVEIRA *et al.*, 2011; SALDANHA *et al.*, 2019; SILVA *et al.*, 2019) analisaram os deslocamentos das pacientes em tratamento para câncer de mama, selecionando períodos diferentes: a primeira analisou os anos de 2005 e 2006, a segunda de 2014 a 2016, e a terceira o ano de 2013. Embora haja similaridades com este estudo na amplitude geográfica e na doença

estudada, o principal método de análise usado por essas pesquisas foi outro (redes de grafos). Além disso, em cada uma foi usada uma plataforma diferente para os cálculos das distâncias utilizadas, nem sempre medindo o deslocamento pela malha rodoviária.

No estudo de Oliveira *et al*. (2011), o objetivo era analisar o fluxo de pacientes com CM conforme o tipo de tratamento recebido (cirurgias, radioterapia e quimioterapia). Dentre seus resultados, foi detectado que o padrão de concentração do atendimento cirúrgico é mais acentuado do que o do tratamento ambulatorial (radio e quimio). Outro dado informado é que mais da metade dos atendimentos é no mesmo município: 55,8% das cirurgias, 54,6% das quimioterapias e 48,7% das radioterapias. Os autores também apontaram diferenças de acesso da população aos serviços ambulatoriais, com uma maior concentração da radioterapia em grandes centros, enquanto para quimio a rede de atendimento parece ser mais aberta.

Para Saldanha *et al*. (2019), o objetivo era fazer a análise do fluxo de pacientes de CM que são atendidos fora de seu município residencial, em internações hospitalares e tratamentos de quimio e radioterapia. Os resultados indicaram que 52,2% das pacientes eram atendidas em cidades diferentes de seu domicílio, mas a maioria ainda dentro do próprio estado.

No terceiro artigo nacional (SILVA *et al*., 2019), o objetivo era o mapeamento e análise dos fluxos de rede das pacientes de CM em tratamento quimioterápico para identificar potenciais obstáculos relacionados a assistência farmacêutica. Foi analisado que somente 17 cidades concentraram aproximadamente 50% dos atendimentos, embora haja 156 municípios com possibilidade para este tipo de tratamento. Outro resultado divulgado foi que o percentual de procedimentos realizados fora do município de origem era cerca de 49%. Este valor difere do percentual de 52,2% informado por Saldanha *et al*. (2019) e dos resultados encontrados por Oliveira *et al*. (2011).

4.2 TRABALHOS INTERNACIONAIS

Na literatura internacional também há diversas pesquisas que tratam do tema deslocamento de pacientes, buscando associações com os tratamentos utilizados ou com o diagnóstico para câncer, seja de mama ou outros tipos estudados. Os artigos apresentados têm abrangência regional, em locais sem a mesma diversidade socioeconômica encontrada no Brasil, e quase todos selecionam bases de dados com períodos longos (cinco anos ou mais).

Em três artigos que analisaram diferentes estados norte-americanos, os resultados foram divergentes (HUANG *et al*., 2009; CELAYA *et al*., 2010; SCOGGINS *et al*., 2012). Nesses estudos, o objetivo principal era investigar se a distância de acesso da paciente a centros

de diagnóstico e de mamografia estava de alguma forma associada ao grau de estadiamento no diagnóstico de CM. Em todos, a metodologia cita que foi usado um banco de dados estadual público, com eventuais complementos de instituições privadas.

Em relação à amostragem utilizada, também há diferenças: no estudo de Celaya *et al*. (2010), foram identificadas seis mil mulheres acima de 40 anos, no período de 1998 a 2004, e os fatores associados ao estadiamento foram comparados, com ênfase na distância até o centro de mamografia. No artigo de Huang *et al*. (2009), cerca de doze mil mulheres acima de 40 anos foram identificadas, no período de 1999 a 2003, usando modelos estatísticos para identificar se há relação entre maiores deslocamentos e detecção do CM em estágios mais avançados. Na pesquisa feira por Scoggins *et al*. (2012), a seleção abrangeu pouco mais de quatro mil casos de três tipos de câncer (mama, colorretal e pulmão) entre os anos de 1997 e 2003.

No estado de New Hampshire (CELAYA *et al*., 2010) não foram encontradas relações entre o estadiamento do tumor no diagnóstico e a distância percorrida, enquanto na região de Kentucky (HUANG *et al*., 2009) houve, sim, uma possível associação entre maiores trajetos e uma detecção tardia do tumor. No estado de Washington (SCOGGINS *et al*., 2012), o estudo calculou o que foi chamado de "fardo da viagem", que englobava fatores como tempo e distância percorrida, encontrando uma associação entre fardos maiores e um diagnóstico já em estádios mais avançados para CM.

Neste estudo, uma das técnicas de MD usada permitiu, em certos casos, associar o estadiamento no diagnóstico com as faixas de deslocamento percorridas. Porém, assim como nos três estudos citados anteriormente, é interessante notar que não foi encontrada a mesma regra de associação em todas as regiões pesquisadas, variando conforme o local selecionado.

Os três casos anteriores limitaram suas bases de dados a três estados norte-americanos (Kentucky, New Hampshire e Washington), mas foi possível também avaliar amostras maiores. No artigo de Henry *et al*. (2011), os casos de mais de 160 mil mulheres de dez estados norte-americanos de 2004 a 2006 foram analisados, com o objetivo de investigar se o tempo de viagem da paciente até os centros de diagnóstico ou mamografia mais próximos impactava o estadiamento no diagnóstico. A principal conclusão foi que outros fatores como raça, idade avançada, e situação econômica parecem estar mais associados a estádios mais avançados no diagnóstico do que o tempo de deslocamento da paciente até a instituição.

Neste estudo, não foi possível avaliar a situação econômica das pacientes, pois é uma variável não existente na base de dados do RHC. Contudo, as variáveis raça e idade foram utilizadas em alguns modelos da análise por MD, sem encontrar associações relevantes com o estadiamento ou outras características do tumor.

Na Inglaterra, o estudo de Jones *et al*. (2008), feito para cinco tipos de câncer diferentes (mama, colorretal, pulmão, ovário e próstata), avaliou cerca de 117 mil casos entre 1994 e 2002, buscando associação de alguns fatores como idade, local e tempo de viagem com o estádio no diagnóstico e com a sobrevida. Dentre os resultados encontrados, um diagnóstico tardio para CM e colorretal foi associado com tempos de viagem maiores até o centro de atenção primária, ao passo que o tempo de deslocamento até o hospital não foi um fator significativo para o grau de estádio no diagnóstico. A diferenciação entre os locais de atendimento não foi feita neste estudo, porém uma associação similar entre maiores distâncias e graus mais avançados de estadiamento para CM foi também encontrada em algumas regiões do país.

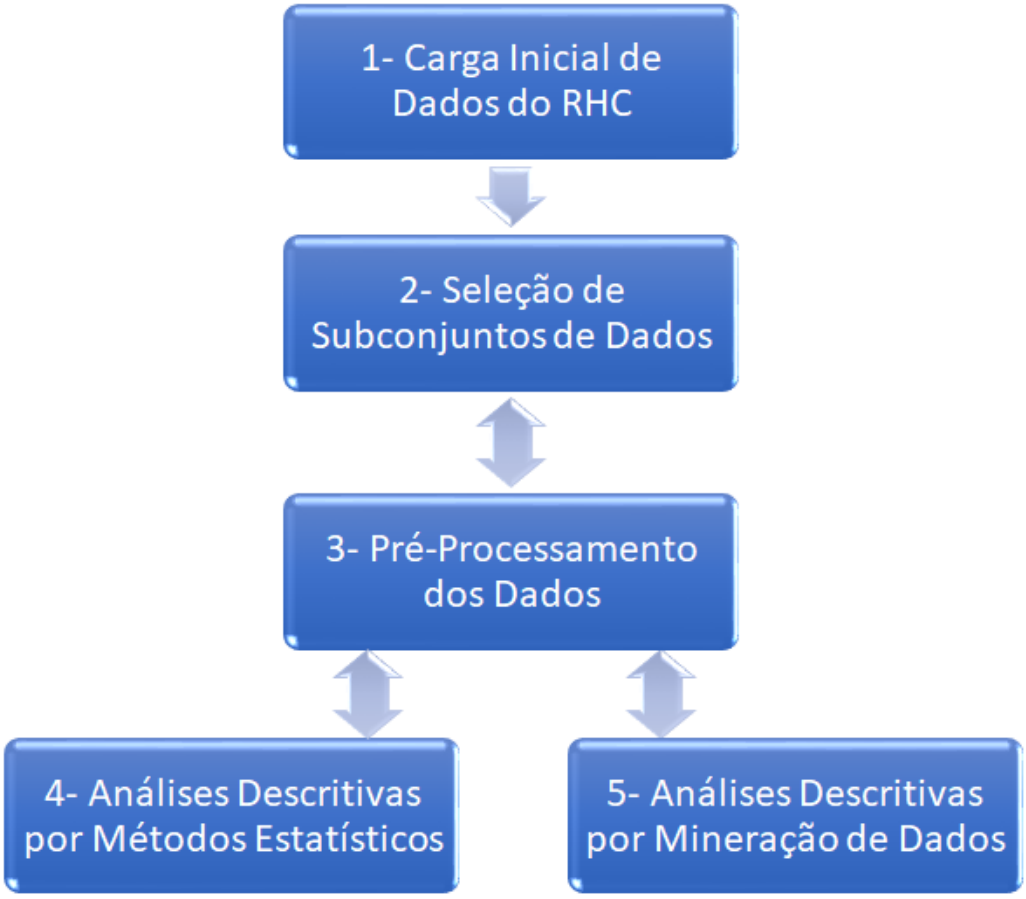
Em outro artigo, Vetterlein *et al*. (2017) apresenta que o volume de dados se assemelhava mais ao deste estudo: foram selecionados quase 776 mil pacientes do banco de dados de câncer nacional norte-americano, no período de 2004 a 2012, diagnosticados com câncer de próstata em estádios diferentes e que passaram por diversos tipos de tratamento. O objetivo era investigar o impacto do deslocamento até o centro de tratamento e sua relação com o risco de mortalidade. Através de análises estatísticas, foi identificada uma associação entre quem viajou longas distâncias e uma sobrevida maior.

Além disso, esse último artigo reforça os resultados de outro já citado (HENRY *et al*., 2011), pois ambos indicam que melhores condições socioeconômicas parecem ter relação com deslocamentos maiores, nos Estados Unidos. O estudo de Vetterlein *et al*. (2017) separou as distâncias em três faixas específicas, definidas por trabalhos anteriores: curta (menor que 12,5 milhas, ou cerca de 20km), média (de 20 a 80km) e longa (de 80 a cerca de 400km – valores acima deste foram excluídos), encontrando respectivamente 54,5%, 33,4% e 12,1% dos pacientes em cada uma delas.

5 METODOLOGIA

A metodologia deste estudo consiste em cinco grandes etapas (Figura 7). Algumas delas retornam a passos anteriores, num modelo iterativo, baseado no processo KDD mencionado na conceituação.

Figura 7 - Fluxograma representando as etapas metodológicas do estudo.



A carga inicial de dados é a primeira etapa, cujo objetivo é alimentar o banco de dados com as informações mais atualizadas do RHC. Nesta etapa, também deve ser feita uma análise prévia dos dados, com o propósito de evitar informações inúteis ou inconsistentes no banco de dados.

A próxima etapa é a seleção dos subconjuntos de dados relevantes ao estudo. Esta fase consiste em filtrar, dentro de todo o repositório de dados, somente aqueles que estão no escopo do estudo e que serão úteis para as análises. É necessário definir os critérios e atributos dessa seleção, revisando-os e complementando-os sempre que as análises exigirem.

O pré-processamento dos dados corresponde à preparação dos dados, visando deixá-los em um formato adequado para as análises. Envolve desde a obtenção de dados complementares (como, por exemplo, a distância entre municípios) até sua formatação e, por isso, pode ser necessário revisitar os subconjuntos selecionados previamente.

As duas últimas etapas correspondem à elaboração dos modelos analíticos propriamente ditos, seja por métodos estatísticos ou por técnicas de mineração de dados. Os primeiros envolvem medidas numéricas (como médias, quantidades, percentuais) exibidas usualmente em forma de gráficos e tabelas. Já os segundos envolvem algoritmos específicos que permitem a elaboração de modelos mais sofisticados. Os dois tipos de análise, porém, ajudam a retroalimentar a própria base, também de forma iterativa.

5.1 CARGA INICIAL DE DADOS

5.1.1 Obtenção dos Dados

Inicialmente, foi necessário atualizar o banco de dados do LACAM com as informações mais recentes do RHC. O laboratório dispõe de um repositório de dados de diversos sistemas do SUS, além de informações demográficas e sociais oriundas do IBGE.

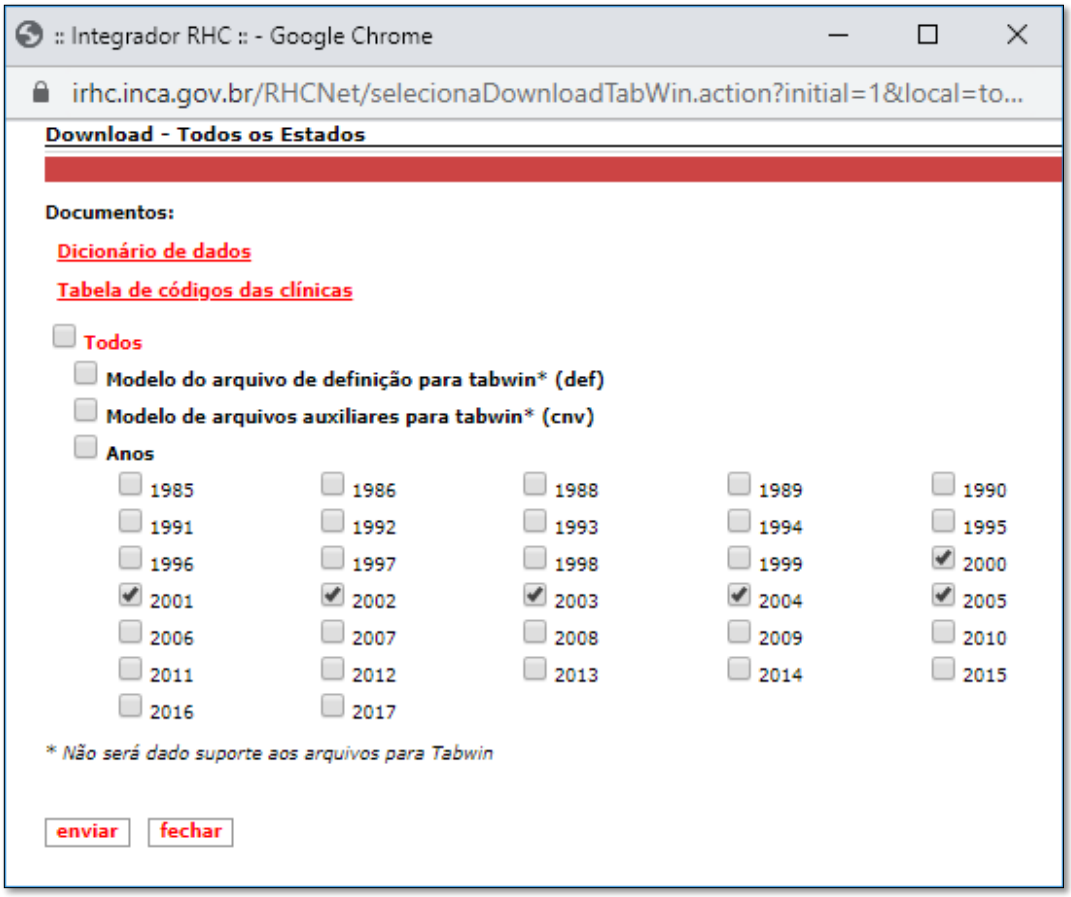
O IntegradorRHC, através de sua opção de *download*, permite que seja selecionada ou uma base estadual, ou a nacional, ou a nacional sem o estado de São Paulo. Após isso, é possível selecionar o(s) ano(s) desejado(s), conforme indicado na Figura 8.

O arquivo é gerado em formato compactado (.zip), para minimizar seu tamanho e facilitar a obtenção desses dados. Após descompactá-lo, o arquivo original apresenta a extensão DBF, que significa "*DataBase File*" (arquivo de banco de dados) e nada mais é que um arquivo texto tabulado, facilmente lido pela maioria dos sistemas gerenciadores de banco de dados (SGBD), além de editores de texto e planilhas.

Os arquivos obtidos foram dos anos 2000 a 2017, que eram os disponíveis na época da última atualização feita para este estudo, ou seja, em agosto de 2019. O ano exibido na tela corresponde ao da primeira consulta, conforme explicado nas Notas Técnicas do RHC[3]: As bases de dados são estruturadas segundo o ano da 1ª consulta no hospital, ou seja, uma base de dados do ano 2000 é composta por todos os casos cadastrados no RHC que tiveram a 1ª consulta relativa ao tumor na respectiva Unidade Hospitalar no ano 2000.

[3] Notas Técnicas do Integrador RHC. Acesso em: 10 mar.2020. Disponível em: https://irhc.inca.gov.br/files/Notas_tecnicas_final.pdf

Figura 8 - Tela de download da base RHC por ano, com os anos de 2000 a 2005 selecionados



Fonte: RHC.INCA[4]

5.1.2 Análise da Carga dos Dados

Para facilitar a análise dos dados obtidos, cada base anual foi convertida em um arquivo do tipo Microsoft Excel, mantendo-se os nomes dos campos. A conversão nesse formato permite uma identificação mais rápida de indicadores gerais (como quantidade total de registros e média de valores), além da possibilidade de filtros por coluna.

Foi avaliada a quantidade total de registros por ano (ou seja, por arquivo), a fim de identificar se haveria algum período muito discrepante dos demais. Foi feita também uma consulta comparativa no *site* do DATASUS, para comprovar se a eventual diferença encontrada em um ano estaria coerente com as demais bases do SUS. Caso fosse confirmada a discrepância, o arquivo deveria ser desconsiderado da carga.

[4] C.f.: https://irhc.inca.gov.br/RHCNet/visualizaTabNetExterno.action.

Para a coluna que informava o ano da consulta, foi usado o filtro para checar se haveria alguma ocorrência incorreta. Por exemplo, na base de 2008, essa coluna só deveria mostrar registros de 2008. Caso fosse encontrada alguma ocorrência de outro ano, deveria ser feita uma segunda análise para avaliar se esse caso seria desconsiderado da base ou apenas remanejado para o arquivo correspondente.

5.1.3 Inclusão dos Dados no Sistema

Após a análise prévia da carga dos dados do RHC, os arquivos em formato Excel foram todos carregados em um banco de dados Microsoft SQL Server, usando a ferramenta de importação MS-SQL Integration Services. A tabela do banco de dados continha a mesma estrutura de campos do RHC e foi sendo alimentada com um arquivo anual por vez.

Muitos campos do RHC vêm em formato de códigos, sem a sua respectiva descrição. As descrições para cada código (por exemplo: código de tipo de tratamento, código do grau de escolaridade, código da raça etc.) ficam armazenadas em tabelas auxiliares, as quais já tinham sido previamente carregadas no repositório de dados do LACAM para estudos anteriores.

Foi necessário fazer a revisão no conteúdo dessas tabelas auxiliares, usando o dicionário de dados atualizado disponibilizado no site do RHC[5]. Em geral, são tabelas com poucos registros e, por isso, caso algum valor estivesse divergente ou faltante, a correção foi feita manualmente e diretamente no SGBD.

5.2 SELEÇÃO DE SUBCONJUNTOS DE DADOS

O banco de dados obtido do RHC contém registros referentes a todos os tipos de câncer. No entanto, o escopo deste estudo é analisar somente os pacientes com câncer de mama. Dessa forma, é necessário selecionar apenas os registros referentes a este tipo de câncer. Tal seleção foi feita a partir do atributo “CID Primário”; foram considerados todos os registros cujos três primeiros dígitos do CID eram “C50”. O código D05 (correspondente a carcinoma *in situ* da mama) não aparece no RHC e, portanto, não foi selecionado.

Os dados selecionados foram armazenados em uma estrutura no SGBD chamada “visão” (ou *view*, em inglês, que é o termo mais comumente utilizado), denominada *V_RHC_C50*. Além da seleção pelo atributo CID Primário, outros critérios de seleção também

[5] Acesso em: 10 mar.2020. Disponível em: https://irhc.inca.gov.br/RHCNet/documentosTabWin.action?doc=dicionario.dados

foram adotados: campo Sexo com o valor igual a “2” (Feminino); e campo UF Residencial diferente de “99” (este código indica uma Unidade da Federação desconhecida ou fora do Brasil, o que impossibilitaria análises relacionadas a deslocamentos).

Com a criação dessa visão, foi possível realizar algumas operações extremamente úteis nesse subconjunto de dados: junções com tabelas auxiliares, criação de novas colunas a partir de atributos pré-existentes, cálculos numéricos e cálculos temporais.

Esse subconjunto principal V_RHC_C50 foi utilizado como base para todas as análises estatísticas descritivas. Para as análises por mineração de dados, porém, foi necessário criar dois subconjuntos diferentes, específicos, devido às características do *software* utilizado.

Foi desenvolvida uma nova visão, chamada *V_RHC_DS1*, usada nas análises por agrupamento e em uma das análises por associação. Os critérios de seleção desse novo subconjunto foram os mesmos da visão principal. Contudo, ela incluiu um atributo sequencial numérico que permite a identificação unívoca de cada registro, algo essencial para execução da rotina de mineração de dados.

Já a terceira visão (*V_RHC_DS2*) foi necessária para que se pudesse fazer o desmembramento do campo “Primeiro Tratamento no Hospital” (usado em uma das análises de associação). Quando há mais de um tipo de tratamento adotado, este campo é concatenado com os diversos valores possíveis. Assim, para se obter uma listagem completa de todos os tipos de tratamento usados em cada ocorrência, usou-se uma função no banco de dados que separa estes códigos, cada um em uma linha, mantendo os demais atributos. Por exemplo, uma ocorrência que originalmente tinha este formato:

UF	**Município**	**Tipo Tratamento**	**Distância**
SP	Suzano	Cir Qui Rad	52 km

Após a execução da função é desmembrada em três linhas:

UF	**Município**	**Tipo Tratamento**	**Distância**
SP	Suzano	Cir	52 km
SP	Suzano	Qui	52 km
SP	Suzano	Rad	52 km

5.3 PRÉ-PROCESSAMENTO DOS DADOS

5.3.1 Definição das Premissas de Cálculo

A base original do RHC contém o município de residência da paciente e o município da unidade hospitalar que registrou a ocorrência, mas não tem um detalhamento maior, tal como o CEP. Dentro deste contexto, o cálculo da distância de deslocamento foi feito somente entre as cidades e, para torná-lo o mais fiel possível à realidade, foram assumidas algumas premissas.

A primeira delas foi que a distância a ser considerada seria zero (0 km) quando os municípios de residência e da unidade hospitalar fossem os mesmos; independente do porte da cidade, isto foi necessário porque não estavam disponíveis os endereços completos das pacientes.

Uma vez que, não há como identificar para cada ocorrência qual foi o transporte utilizado, adotou-se como segunda premissa considerar o meio de transporte rodoviário como principal forma de deslocamento. Aqui podem ser incluídos carros, ônibus e similares, excluindo-se, portanto, aviões, trens, barcos e outros.

A última premissa foi que o ano do tratamento não foi considerado como restrição para o cálculo. Se, por acaso, no ano 2000 não houvesse uma estrada pavimentada entre dois municípios e, naquela época, o deslocamento pudesse ter sido maior, isto não seria considerado nesta análise. O cálculo da distância de deslocamento foi feito em outubro de 2019 e refletiu, portanto, o caminho disponível nessa ocasião.

Caso as condições anteriores não pudessem ser atendidas, a distância seria calculada usando a diferença geodésica entre dois pontos, através de suas latitude e longitude. O cálculo geodésico não leva em consideração o relevo, as estradas e as barreiras naturais entre municípios e, por isso, não é de todo correto, mas foi considerado neste estudo como uma aproximação válida. Tal cálculo poderia ser útil, por exemplo, em municípios localizados em ilhas (como Fernando de Noronha) ou cujo acesso é somente através de barcos.

5.3.2 Cálculo da Distância de Deslocamento

Para calcular a distância de deslocamento de cada ocorrência da base, foi feita inicialmente uma consolidação dos dados, a fim de identificar somente os pares origem-destino unívocos. Através de uma planilha, os dados da base principal foram agrupados e foram

selecionadas apenas as quatro colunas necessárias para calcular a distância de deslocamento: a UF residencial, o município residencial, a UF da UH, e o município da UH.

A forma utilizada para calcular a distância em metros foi através de uma interface (*Application Programming Interface*, ou API) disponibilizada pelo Google em sua plataforma relacionada a mapas[6]. Em linhas gerais, deve ser montada uma *string* com os locais de origem e destino em uma URL que é passada via HTTP ao servidor do Google, o qual retorna também uma *string* com o resultado em formato XML ou JSON.

Foi criada uma rotina no Excel na linguagem *Visual Basic for Applications* (VBA) que lia cada linha da planilha com as quatro colunas, fazia a chamada da interface e retornava a menor distância percorrida na rota mais eficiente (segundo os critérios usados pelo Google). A rotina também trazia a latitude e longitude das cidades, que seriam usadas posteriormente.

A fim de atualizar o banco de dados com as distâncias obtidas, foi criada uma tabela auxiliar que continha, além das colunas de origem, destino e distância, mais duas do tipo geográfico – que armazenam informações espaciais baseadas na latitude e longitude de um local. Assim, a planilha Excel pôde ser totalmente importada para o SGBD.

Para os pares origem-destino não contemplados no método citado, a distância foi obtida diretamente no SQL Server, graças a uma função específica para cálculos geográficos: a *STDistance*[7]. Foi feita uma rotina que seleciona as linhas sem valor de deslocamento e, usando essa função, calculava a distância geodésica em metros.

Após a tabela auxiliar ter sido concluída, com todos os pares de distâncias calculados, a visão V_RHC_C50 foi reajustada, incluindo a distância de deslocamento da paciente do seu município de residência até o município do hospital.

5.3.3 Inclusão das Esferas Administrativas

Para uma melhor exibição dos deslocamentos entre municípios, foi necessário também incluir na visão V_RHC_C50 as esferas administrativas existentes entre a UF e o município, de acordo com a hierarquia administrativa da saúde pública brasileira. Isso possibilitou que os dados fossem organizados em diferentes níveis hierárquicos, desde o município da ocorrência, passando pelas regiões e macrorregiões de saúde, pelo estado, até a região geográfica.

---

[6] Plataforma do Google Maps Documentação. Disponível em: https://developers.google.com/maps/documentation?hl=pt-br. Acesso em: 04 abr. 2020.

[7] Método do SQL Server em instâncias geográficas. Disponível em: https://docs.microsoft.com/pt-br/sql/t-sql/spatial-geography/stdistance-geography-data-type?view=sql-server-ver15. Acesso em: 04 abr. 2020

As informações geográficas vieram de tabelas do IBGE, as quais já tinham sido previamente carregadas no repositório de dados do LACAM para estudos anteriores. As tabelas contêm os relacionamentos entre cada esfera, indicando o nível imediatamente superior ao qual um local pertence. Ou seja, um município pertence a uma certa região de saúde, a qual pertence a uma macrorregião, que por sua vez está dentro de um estado. Desse modo, a partir de um município, é possível identificar hierarquicamente todas as esferas às quais ele pertence.

Essa inclusão foi feita tanto para o município residencial quanto para o município da unidade hospitalar. Com esses dados, foi possível também criar atributos para identificar rapidamente se um determinado tratamento ocorreu dentro ou fora da esfera em análise. Por exemplo, uma ocorrência poderia estar dentro do estado e da macrorregião de saúde, mas o paciente pode ter se deslocado até um município de outra região de saúde.

5.4 ANÁLISES DESCRITIVAS POR MÉTODOS ESTATÍSTICOS

As análises descritivas por métodos estatísticos foram separadas em quatro grupos distintos: Tipo de Tratamento, Paciente, Tumor e Deslocamento, os quais foram em seguida divididos em subgrupos. Esses agrupamentos foram feitos para segregar as análises com características semelhantes, permitindo um estudo exploratório das características por região e possibilitando assim uma melhor compreensão do deslocamento nas regiões envolvidas.

Para cada grupo e subgrupo foram criadas páginas *web* com botões e atalhos para auxiliar na navegação. As análises dos três primeiros grupos foram desenvolvidas usando o Excel, através dos recursos de tabelas dinâmicas e gráficos. Em cada subgrupo, é possível filtrar os dados de um determinado estado e compará-lo com os resultados do Brasil.

As análises do quarto grupo exigiram cálculos mais específicos e foram elaboradas no software Microsoft Power BI. Este aplicativo é uma ferramenta de *business intelligence* que possibilita a construção de painéis interativos, faz cálculos dinâmicos complexos, tem recursos de navegação *web* e opções de segmentação dos dados. Permite também que seja feita a comunicação direta com o banco de dados e facilita o agrupamento de dados hierárquicos.

Os detalhes de cada grupo serão apresentados no capítulo Resultados. De modo geral, nas telas onde há gráficos é também possível acessar a tabela com os dados consolidados do subgrupo. Os gráficos só mostram dados dos registros em que o atributo está preenchido com um valor válido (ou seja, diferente de “vazio” ou “sem informação”). Os tipos de tratamento Imunoterapia e Transplante de Medula Óssea (TMO) foram adicionados à coluna “Outros”, por serem quantidades muito pequenas em relação aos demais.

5.5 ANÁLISES DESCRITIVAS POR MINERAÇÃO DE DADOS

As análises por mineração de dados são recomendadas quando há grandes volumes de dados a serem estudados, possibilitando a descoberta de padrões e regras difíceis de serem encontradas pelos meios tradicionais (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996).

Como um dos objetivos deste estudo é desenvolver modelos analíticos descritivos para o deslocamento de pacientes em tratamento de câncer de mama, foram selecionadas duas técnicas descritivas de *data mining* para auxiliar na obtenção desses modelos.

Para a elaboração das análises descritivas por mineração de dados foi usado o *software* MS-SQL Analysis Service. As características do *software* que foram consideradas relevantes para este estudo são: implementações eficazes dos algoritmos exigidos para executar as técnicas; conexão direta com o SGBD; flexibilidade de parâmetros dos algorítmos; ampla documentação de uso e diversos recursos gráficos.

A primeira técnica usada foi o Agrupamento, que possibilita agrupar dados com características similares. Foi utilizada na discretização dos atributos numéricos da base que possuem um intervalo de valores muito amplo e variado. A separação em grupos torna a análise mais simples e eficaz, sendo possível inclusive reutilizá-los na própria base, em outros tipos de análise. Havia apenas três atributos numéricos na base e todos se encaixaram no critério de ampla variabilidade: “distância de deslocamento”, “dias até o tratamento”, e “idade”.

O algoritmo usado na elaboração dos modelos descritivos foi o *Microsoft Clustering*[8], que possibilita a definição de filtros, a seleção do atributo desejado e o ajuste de parâmetros para sua execução. Os parâmetros relevantes para criação dos modelos foram:

- *Cluster_Count*: indica a quantidade de grupos a ser gerada. O valor padrão do *software* é de dez *clusters*, mas o objetivo é que se tenha menos grupos para facilitar as análises. O ideal é ter a formação de no máximo cinco grupos, pois análises com mais *clusters* que esse número se torna complexo e de difícil visualização em gráficos. Durante as execuções dos modelos, este parâmetro foi sendo ajustado para obter grupos que tivessem menor desvio padrão;

[8] Referência técnica do algoritmo Microsoft Clustering. Disponível em: https://docs.microsoft.com/pt-br/analysis-services/data-mining/microsoft-clustering-algorithm-technical-reference?view=asallproducts-allversions. Acesso em 05 abr. 2020

- *Clustering_Method*: indica qual método do algoritmo será usado. O padrão é o de Maximização de Expectativa evolutivo (EMe), mas em algumas execuções também foi avaliado o uso do outro método disponível, o K-Means evolutivo (KMe).

Todas as execuções dos modelos, os parâmetros efetivamente usados e os grupos encontrados através da técnica de Agrupamento estão descritos no capítulo de Resultados e Discussão.

A segunda técnica usada foi a Associação, que identifica relações entre itens do conjunto, calculando também os percentuais de confiança e de importância das associações encontradas. Para poder montar essas relações, a técnica exige que sejam informados os atributos de entrada e de saída. Contudo, nem sempre são encontradas regras significativas, o que faz com que o processo possa ter muitas iterações.

O algoritmo utilizado para a elaboração dos diversos modelos foi o *Microsoft Association Rules*[9], o qual permite definir filtros, indicar os atributos de entrada/saída, e ajustar parâmetros de execução. Os parâmetros relevantes para criação dos modelos descritivos foram:

- *Maximum_Itemset_Size:* o tamanho máximo do conjunto de itens que pode ser encontrado pelo algoritmo. Foi aplicado o valor 0 (zero), que indica que não há tamanho máximo: podem ser formadas regras com qualquer número de elementos.

- *Minimum_Probability*: especifica a probabilidade mínima (também chamada confiança) a partir da qual as regras geradas são consideradas. Ou seja, com o valor padrão (0,4), não são geradas regras cuja probabilidade seja menor que 40%. Esse parâmetro pode ser ajustado (diminuindo ou aumentando) até que sejam encontradas regras relevantes para a análise sendo feita.

Para o atributo de “Entrada”, foram selecionados dois campos: Estadiamento Grupo, que é extremamente importante para identificar o quão avançado pode estar o câncer; e Tipo de Tratamento, que indica quais os tratamentos mais comuns em uso no país. Ambos são de preenchimento obrigatório no RHC (não tendo, portanto, ocorrências com valores vazios) e têm poucas opções de valores, o que os tornam de fácil leitura ao descrever as regras. Já o atributo

[9] Referência técnica do algoritmo de associação da Microsoft. Disponível em: https://docs.microsoft.com/pt-br/analysis-services/data-mining/microsoft-association-algorithm-technical-reference?view=asallproducts-allversions. Acesso em 05 abr. 2020

de “Saída” foi o mesmo para os dois modelos: Faixa de Deslocamento, que é o escopo deste estudo.

Definidos os atributos, os modelos finais foram então elaborados, divididos da seguinte forma (sem restrição de ano em nenhum deles):

- Para o Brasil, contemplando os dados de forma geral, sem restrição de local;

- Para cada uma das cinco regiões geográficas: Centro-Oeste, Nordeste, Norte, Sudeste e Sul, a fim de avaliar se há diferenças regionais significativas;

- Para os dois estados mais populosos (São Paulo e Minas Gerais), que são também os que mais tem casos de câncer de mama no país.

Além das duas combinações selecionadas (estadiamento x deslocamento e tratamento x deslocamento), todas as outras combinações de atributos usadas em alguma simulação estão também descritas no capítulo Resultados e Discussão, mesmo as que não geraram resultados significativos.

6 RESULTADOS E DISCUSSÃO

Os resultados obtidos a partir das análises do banco de dados do RHC e suas respectivas discussões estão separados em três seções. A primeira mostra a preparação dos dados antes de sua utilização no desenvolvimento das análises. A segunda apresenta as análises descritivas estatísticas, detalhando a montagem do painel de navegação inicial e os diferentes grupos de análises criados, mostrando exemplos de cada um. A terceira seção apresenta as análises descritivas efetuadas por mineração de dados, separadas pela técnica utilizada.

6.1 PREPARAÇÃO DOS DADOS

As três etapas iniciais da metodologia envolveram análises preliminares, carga de dados, criação de subconjuntos dos dados e formatação dos atributos, de modo que fosse possível realizar as demais etapas de análise. Os resultados dessas atividades estão descritos a seguir.

6.1.1 Carga Inicial de Dados

Após o download das bases do RHC, foi necessário analisar as informações obtidas. Para isso, o principal critério foi avaliar a quantidade total de registros por ano. As quantidades de ocorrências identificadas em cada arquivo estão listadas na Tabela 1, separadas por ano de primeira consulta.

Foi verificado que o ano de 2017 continha uma quantidade de registros muito inferior aos anos anteriores: 88.636 registros, que correspondem a apenas 40% dos registros de 2016. Considerando, a partir de uma consulta no site do DATASUS, que não houve essa redução drástica nos dados de câncer em 2017[10], o arquivo desse ano foi desconsiderado da carga final.

Para cada arquivo anual, também foi avaliado se havia alguma ocorrência que não pertencia àquele ano. Caso ocorresse tal situação, seria necessária uma segunda análise para determinar se esses casos seriam remanejados ou excluídos da base. Porém, isso não ocorreu e cada arquivo anual tinha somente registros correspondentes ao ano da consulta selecionado. A base total que foi carregada no repositório de dados do LACAM ficou ao final com 3.176.138

[10] DATASUS (tabnet), seção de informações Epidemiológicas e Morbidades, opção “Tempo até o início do tratamento oncológico – Painel Oncologia”. Acesso em: 16 abr. 2020. Disponível em: http://tabnet.datasus.gov.br/cgi/dhdat.exe?PAINEL_ONCO/PAINEL_ONCOLOGIABR.def

registros obtidos do RHC, considerando os anos 2000 a 2016.

Tabela 1 - Quantidade de registros obtidos do RHC por ano

Ano	Quantidade	Ano	Quantidade
2000	70.824	2009	220.018
2001	91.880	2010	238.936
2002	104.989	2011	256.433
2003	112.557	2012	259.365
2004	126.334	2013	277.319
2005	148.978	2014	273.473
2006	155.694	2015	245.165
2007	176.282	2016	218.925
2008	198.966	2017	88.636
		Total	**3.264.774**

6.1.2 Seleção de Subconjuntos de Dados

Para selecionar somente os dados relacionados a câncer de mama dentre os mais de três milhões de registros, foram criadas três visões diferentes, conforme critérios já informados na metodologia. Ao executarmos a visão principal (V_RHC_C50), foram encontrados 492.023 registros na base.

Dentre as operações possibilitadas com a criação de visões, a primeira delas foi fazer junções nas tabelas auxiliares disponíveis no repositório de dados, substituindo os códigos por suas respectivas descrições. Exemplo: ao mostrar a lateralidade do tumor é exibido “Direita” no lugar do código “1”.

A *view* também permite a criação de novas colunas a partir de atributos pré-existentes. Usando esse recurso, a coluna “Classificação Etária” foi criada. Dessa forma, a idade pôde ser agrupada em cinco faixas específicas, definidas por um especialista em saúde pública, facilitando as análises: “00-19”, “20-39”, “40-49”, “50-59”, e “60 +”. Uma lógica semelhante foi aplicada posteriormente para o campo “Faixa de Deslocamento”.

Outra, importante coluna criada foi “Dias até o Tratamento”, que mostra a quantidade de dias decorridos entre o diagnóstico e o início do tratamento. A lista final de todos os campos da visão principal, sua descrição e exemplo de valor pode ser encontrada no Quadro 1.

(continua)

Quadro 1- Lista de campos da visão V_RHC_C50

Campo	Exemplo de Valor	Descrição
TPCASO	*1*	Tipo de caso: analítico (1) ou não (2)
IDADE	*56*	Idade da paciente
CLAS_ETARIA	*50-59*	Classificação etária da paciente
RAÇA	*Parda*	Raça/cor
INSTRUÇÃO	*Nível Médio*	Escolaridade
HIST_FAMILIAR	*Não*	Histórico familiar de câncer
ALCOOL	*Nunca*	Histórico de consumo de bebida alcoólica
TABACO	*Sim*	Histórico de consumo de tabaco
EST_CONJUGAL	*Casado*	Estado conjugal atual
UFNAS	*MG*	UF de nascimento
REGIAO_RES	*Sudeste*	Região residencial
UF_RES	*SP*	UF residencial
MUNIC_RES	*Mogi das Cruzes*	Município de procedência (residência)
REGS_RES	*Alto do Tietê*	Região de saúde residencial
MRS_RES	*RRAS 02*	Macrorregião de saúde residencial
ANOPRIDI	*2013*	Ano do primeiro diagnóstico
ANOTRI	*2013*	Ano da triagem
ANOPRICON	*2013*	Ano da primeira consulta
ANOINITRAT	*2013*	Ano do início do 1º tratamento específico para o tumor, no hospital
DTTRIAGE	*13/09/2013*	Data da triagem
DATAPRICON	*13/09/2013*	Data da primeira consulta
DTDIAGNO	*16/07/2013*	Data do primeiro diagnóstico
DATAINITRAT	*24/10/2013*	Data do início do 1º tratamento específico para o tumor, no hospital
DATAOBITO	*04/09/2015*	Data do óbito
DIAS_ATE_TRAT	*100*	Quantidade em dias entre a Data de Início do tratamento e a Data de Diagnóstico.
LOC_TUM_PRIM	*Porção central da mama*	Localização detalhada do tumor primário (subcategoria da topografia do CID-O)
TIPO_HIST	*8500/3*	Tipo histológico do tumor primário (Codificação pela CID-O)
HISTOLOGIA	*Carcinoma ductal infiltrante*	Descrição do tipo histológico do tumor primário
LATERALIDADE	*Direita*	Lateralidade do tumor
MAISUMTU	*Não*	Ocorrência de mais um tumor primário
TNM	*410*	Codificação do estádio clínico segundo classificação TNM
ESTADIAM	*3B*	Estadiamento clínico do tumor (TNM)
ESTADIAG	*3*	Estadiamento clínico do tumor (TNM) – Grupo
OUTROEST	*88*	Codificação do grupamento do estádio clínico segundo outras classificações que não a TNM

CLIATEN	*Mastologia*	Clínica do primeiro atendimento - entrada da paciente
CLITRAT	*Radioterapia*	Clínica de início do tratamento
DIAGANT	*Com Diagnostico Sem Tratamento*	Diagnóstico e tratamento anteriores
RZNTR	*Óbito*	Principal razão para a não realização do tratamento antineoplásico no hospital
PRI_TRAT_H	*423*	Código do Primeiro Tratamento no Hospital
PTH_DESC	*Qui Cir Rad*	Descrição Abreviada do Primeiro Tratamento no Hospital
T1_NENHUM	*0*	Fez (1) ou não (0) o tratamento: nenhum
T2_CIRURGIA	*1*	Fez (1) ou não (0) o tratamento: cirurgia
T3_RADIO	*1*	Fez (1) ou não (0) o tratamento: radio
T4_QUIMIO	*1*	Fez (1) ou não (0) o tratamento: quimio
T5_HORMONIO	*0*	Fez (1) ou não (0) o tratamento: hormônio
T6_TMO	*0*	Fez (1) ou não (0) o tratamento: TMO
T7_IMUNO	*0*	Fez (1) ou não (0) o tratamento: imuno
T8_OUTROS	*0*	Fez (1) ou não (0) o tratamento: outros
T9_SEMINFO	*0*	Fez (1) ou não (0) o tipo de tratamento: sem informação
ESTDFIMT	*Doença estável*	Estado da doença ao final do primeiro tratamento no hospital
CNES	*2726998*	Número do CNES do hospital
UH_NOME	*Fundação Pio XII*	Nome da unidade hospitalar
TP_PREST	*Particular*	Tipo de prestador (Público/Particular/Outros)
UH_REGIAO	*Sudeste*	Região da unidade hospitalar
UH_UF	*SP*	UF da unidade hospitalar
UH_MUNIC	*Barretos*	Município da unidade hospitalar
UH_REGS	*Norte-Barretos*	Região de saúde da unidade hospitalar
UH_MRS	*RRAS 13*	Macrorregião de saúde da UH
ORIENC	*SUS*	Origem do encaminhamento
DISTÂNCIA	*483,07*	Distância calculada (em quilômetros) entre o município residencial e o da UH
MM_MUNIC	*F*	Indica se o tratamento foi Dentro (D) ou Fora (F) do município residencial
MM_REGS	*F*	Indica se o tratamento foi Dentro (D) ou Fora (F) da região de saúde residencial
MM_MRS	*F*	Indica se o tratamento foi Dentro(D) ou Fora(F) da macrorregião de saúde residencial
MM_UF	*D*	Indica se o tratamento foi Dentro (D) ou Fora (F) da UF residencial
FX_DESLOC	*4-Acima de 300 km*	Faixa de deslocamento
OCORRENCIA	*1*	Valor fixo, indica uma (1) ocorrência.

Ao executar a visão auxiliar V_RHC_DS1, foi encontrada a mesma quantidade que na visão principal (492.023). Isso era esperado, uma vez que os critérios de seleção são os mesmos

e a única diferença é o atributo chave a mais. A outra visão auxiliar (V_RHC_DS2) apresentou uma quantidade de registros maior (988.917), pois neste caso um registro pode ter sido desmembrado em vários.

6.1.3 Pré-Processamento dos Dados

Para realizar o cálculo da distância de deslocamento, os quase 500 mil registros da visão principal foram inicialmente exportados para uma planilha Excel. Em seguida, foi feita a consolidação desses dados, para agrupar os pares origem-destino iguais. Com essa agregação, obteve-se um valor bem menor de registros para o cálculo: 15.939. Isso evitou cálculos redundantes para a mesma rota (por exemplo: dos 724 casos de Mogi das Cruzes, 458 trataram-se na capital São Paulo, e a distância entre as duas cidades só precisou ser calculada uma vez).

O principal método de cálculo foi a interface disponibilizada pelo Google Maps (API). Com ela, foi possível calcular 99,1% dos quase dezesseis mil registros do arquivo. Os restantes 0,9% (143 registros) não puderam ser obtidos pela API por serem locais em que não havia transporte rodoviário disponível. Na imensa maioria, cidades da região Norte, como Afuá no Pará e Parintins no Amazonas – cidades acessíveis somente por via fluvial ou aérea – ou, ainda, a ilha de Fernando de Noronha em Pernambuco. Para esses 143 casos não contemplados no método citado, foi obtida a distância geodésica, através de uma função diretamente no SQL Server.

Após a inclusão na visão principal dos campos relacionados a deslocamento, ela foi novamente executada. A base completa manteve-se com 492.023 registros e foi então exportada para Excel. Nessa etapa, foram realizados alguns ajustes manuais e em seguida foram gerados arquivos separados. Esses ajustes tinham por objetivo facilitar o agrupamento e filtro de informações, além de diminuir o tamanho de cada arquivo, melhorando sua performance para abertura e leitura. Por exemplo:

- Substituição dos valores com o texto “NULL” por vazio;
- Troca de datas inválidas (8888, 1899) pelo padrão 9999;
- Substituição de valores "/ /" e datas inválidas (99/99/9999) por vazio;
- Correção pontual de algumas datas (3015 p/ 2015, por exemplo);
- Substituição do texto “Sem Informação” por “Sem Info”;
- Remoção de colunas não relevantes para o grupo de análise sendo feita;

- Conversão das colunas de Tipo de Tratamento para tipo lógico (verdadeiro/falso).

6.2 ANÁLISES DESCRITIVAS POR MÉTODOS ESTATÍSTICOS

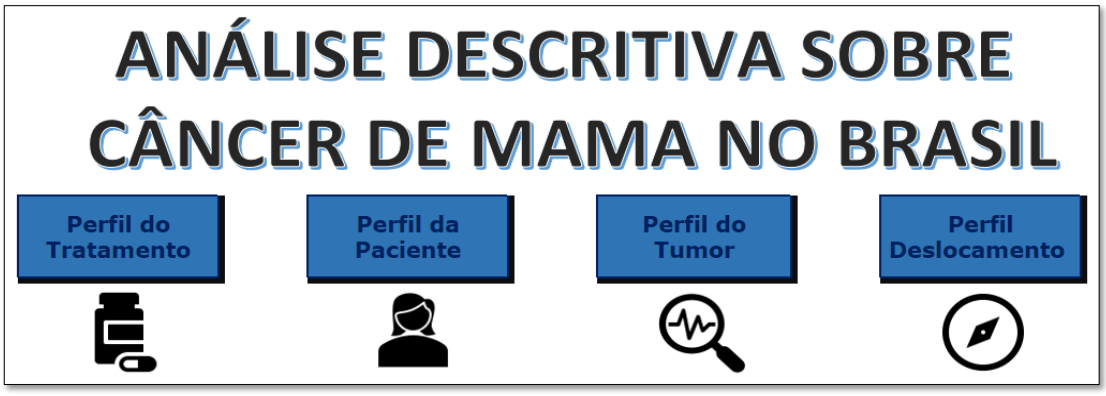
As análises estatísticas podem ser acessadas através de um painel de navegação inicial, cuja montagem e organização são informadas na primeira seção. A seguir, cada um dos quatro grupos de análises será descrito com uma explicação geral de montagem dos subgrupos e pelo menos um exemplo de gráfico para cada grupo.

Os gráficos apresentados foram obtidos por captura de tela da página *web* correspondente. Em cada um, foram simulados cenários e períodos diferentes, com o intuito de demonstrar os recursos e possibilidades do sistema. Entretanto, não serão discutidas todas as variações de filtros e colunas de cada subgrupo, pois seriam milhares de combinações possíveis.

6.2.1 Tela de Navegação Inicial

Para organizar as análises descritivas estatísticas por grupos e subgrupos, foi criada uma tela de navegação inicial (Figura 9). Cada botão desse painel inicial abre uma página específica, que leva a diferentes análises.

Figura 9 - Tela de navegação inicial com os quatro grupos de análise



De modo geral, buscou-se uma padronização nos gráficos, cores e botões de navegação, para que o manuseio fosse de fácil assimilação. Quando se percebeu necessário, foram colocadas instruções na tela para auxiliar o usuário. O link para acesso ao painel estará disponível na página do LACAM.

6.2.2 Análises por Tipo de Tratamento

As análises descritivas estatísticas relacionadas aos tipos de tratamento usados foram separadas em três páginas:

- Por Tipo de Tratamento
- Por Estado
- Evolução Temporal

Todas elas mostram de forma gráfica os dados percentuais anuais do Brasil e das unidades da federação. Os percentuais exibidos referem-se à quantidade de casos em que houve um determinado tratamento em relação à quantidade total de ocorrências naquele mesmo período. Por exemplo: em 2009, houve no país 17.944 tratamentos de quimioterapia (Q), de um total de 36.173 casos (T). Portanto, o valor percentual para este ano, deste tipo de tratamento, e especificamente para esta unidade territorial (Brasil), foi de quase 50% (Q / T = 0,4961).

Além dos gráficos, em todas as páginas também é possível consultar as tabelas de origem dos valores, usando o botão “Dados da Amostra”. Nas duas primeiras páginas, foram feitos filtros para mostrar nos gráficos somente os dados de um ano específico (entre 2000 e 2016) ou a somatória do período todo. Além disso, os percentuais individuais somados podem ser maiores do que 100%, pois cada ocorrência pode ter tido mais de um tratamento.

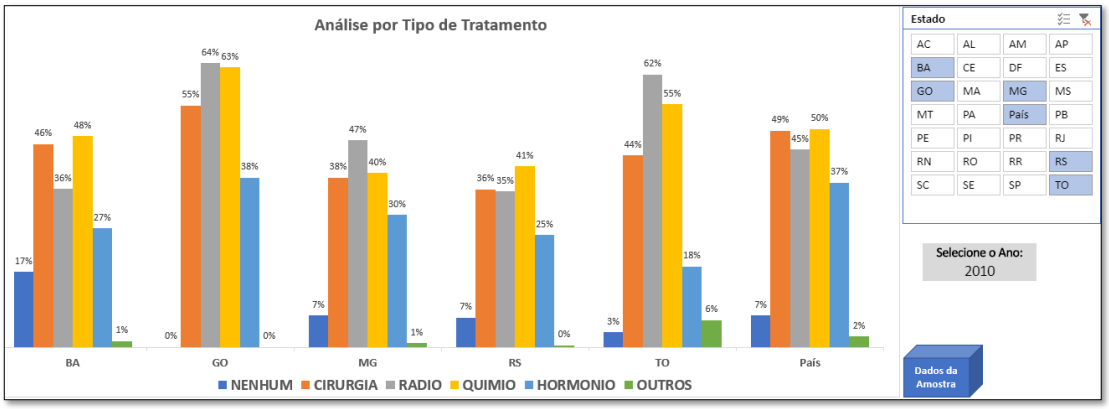
A primeira página de consulta (Por Tipo de Tratamento) permite que se comparem quais são os tipos de tratamento mais usados em um estado. O gráfico mostra os seis Tipos de Tratamento agrupados para cada região selecionada. Cada um deles é exibido em uma cor diferente e na mesma sequência em cada estado.

Na Figura 10, em que foram selecionados um estado de cada região brasileira e o país, é possível ver algumas diferenças e semelhanças regionais. Nesse exemplo, foi usado o filtro por ano para selecionar apenas 2010.

É possível identificar que a Bahia apresenta 17% dos pacientes que não fizeram nenhum tratamento para o ano em consulta. É o maior percentual dentre os estados selecionados e acima da média brasileira de 7%. O estado que mais fez “outros” tratamentos foi Tocantins (6%), bem acima de todos os demais, que tem 1% ou 0%. O percentual de radioterapia é bem alto em Goiás (64%) e no Tocantins (62%), enquanto na Bahia esse tratamento foi usado em 36% dos casos. Os valores encontrados para cirurgia também têm algumas variações, com o

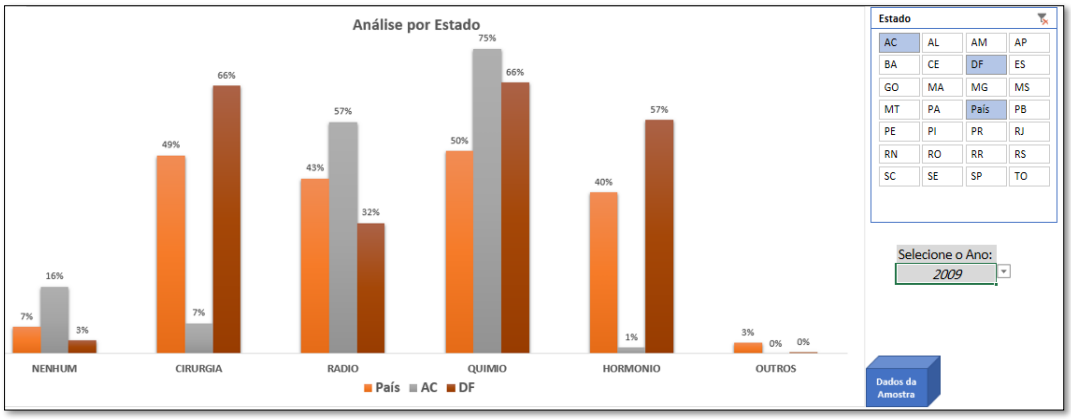
menor índice no Rio Grande do Sul (36%) e o maior em Goiás (55%), enquanto a média nacional foi de 49% em 2010.

Figura 10 - Análise por Tipo de Tratamento com alguns estados selecionados



A segunda página de consulta (por estado) apresenta os mesmos dados que a anterior, mas em posições diferentes no gráfico. Para cada tratamento, é possível identificar quais os estados que mais o utilizam. Conforme ilustrado pela Figura 11, os tipos de tratamento são apresentados no eixo horizontal, enquanto as UF's aparecem como colunas, em cores diferentes.

Figura 11 - Análise por Estado do ano de 2009 para Acre, DF e Brasil



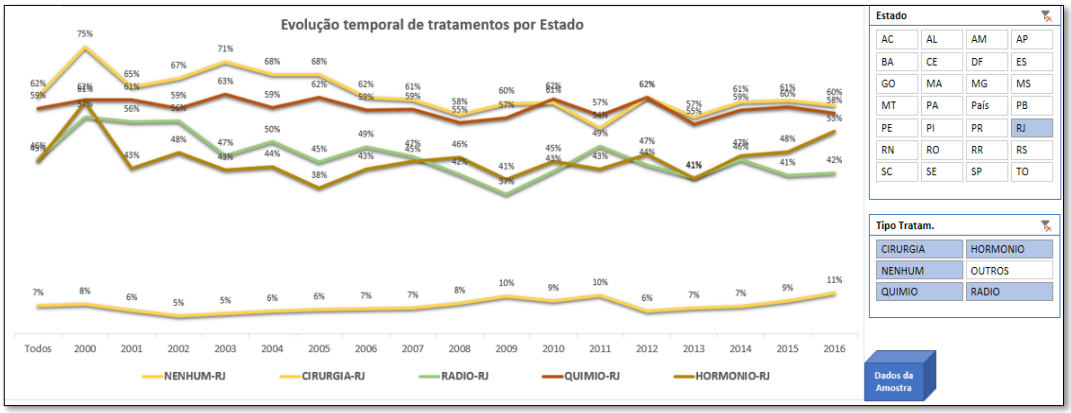
Pelo gráfico disponibilizado, é possível detectar que, para quimioterapia, os percentuais no Acre (75%) e no Distrito Federal (66%) foram bem acima da média nacional (50%) em 2009. Já a quantidade relativa tanto de hormonioterapia (1%) quanto de cirurgias (7%) foram significativamente menores que a do Brasil (40% e 49%, respectivamente). Em

Brasília, o percentual de radioterapia no período (32%) foi menor que nos outros dois lugares (Brasil com 43% e Acre com 57%).

A terceira página de análises (Evolução Temporal) mostra um gráfico de linhas, com duas opções de filtro: o tipo de tratamento e o estado. Seu objetivo é mostrar os dados numa sequência temporal e, por isso, não permite filtrar o ano. Pode-se analisar a evolução temporal de um determinado tipo de tratamento comparando vários estados simultaneamente ou, então, diversos tratamentos ao longo do tempo restringindo a uma única UF.

O gráfico exibe no eixo X os anos (de 2000 a 2016) e no eixo Y o percentual de uso para cada tratamento em cada estado, separando-os em cores diferentes. Como exemplo, a Figura 12 mostra a evolução temporal de cinco tipos de tratamentos para o estado do Rio de Janeiro.

Figura 12 - Evolução temporal dos tipos de tratamento no Rio de Janeiro



É possível identificar que a cirurgia era o tratamento mais recorrente até 2010, quando então a quimioterapia a alcançou e ambas vem dividindo o "topo" do ranking nos últimos anos. Visualmente se observa também que a radioterapia vem perdendo espaço desde o início desta série histórica, enquanto que "nenhum" tratamento vem crescendo de 2012 até o final do período – apesar de, felizmente, ainda ser menos usado em comparação com os demais.

6.2.3 Análises por Paciente

As páginas de análises descritivas relacionadas a pacientes foram divididas em oito subgrupos:

- Classificação Etária
- Grau de Instrução

- Tabagismo
- Alcoolismo
- Histórico Familiar de Câncer
- Raça
- Mais de um Tumor
- Origem do Encaminhamento

Em todas, padronizou-se o formato com dois gráficos: à esquerda da tela, exibem-se os dados gerais do Brasil para o subgrupo em questão e, na parte central, são exibidos os dados dos estados selecionados. O filtro para escolher uma ou mais UF's foi colocado na área direita superior da tela, e o campo para seleção do ano na área direita central. É possível restringir o período a apenas um ano, ou exibir a somatória total (de 2000 a 2016).

A única página que tem um filtro adicional a esses é a primeira (Classificação Etária), que permite também que se filtrem as faixas requeridas. Para consultar as tabelas de origem dos dados basta clicar no botão "Dados da Amostra", que leva a uma página com as quantidades detalhadas da base e os valores previamente filtrados.

Assim como as análises descritas na seção anterior, os dados exibidos nos gráficos referem-se apenas a valores informados ao sistema (ou seja: campos vazios, com o valor "não se aplica", "sem informação", ou "não avaliado", não são usados nos cálculos).

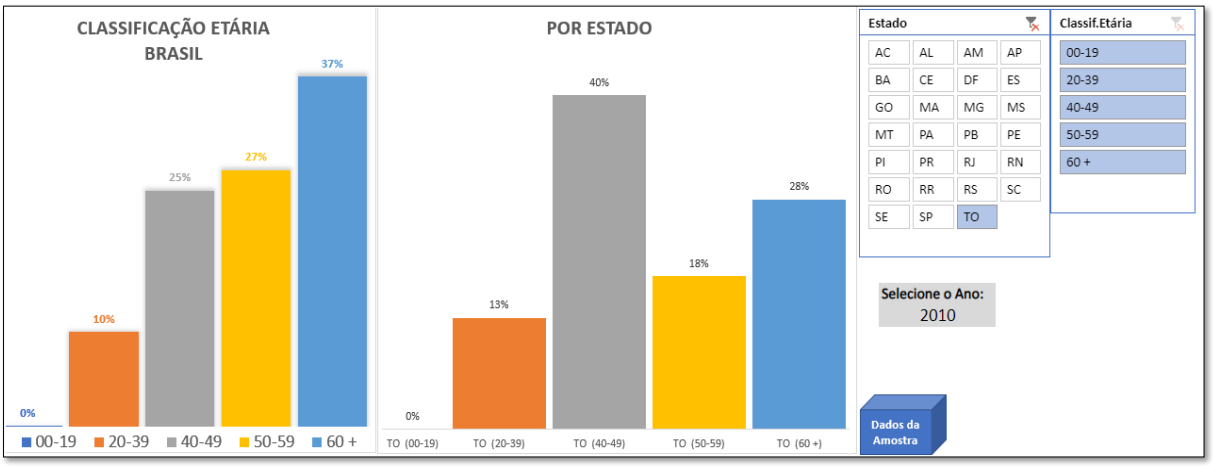
O cálculo dos percentuais foi realizado considerando os valores existentes para cada opção disponível no subgrupo, restringindo ao lugar e ano previamente escolhidos. Por exemplo: para a categoria "Histórico Familiar de Câncer", as opções são "Sim" ou "Não". No ano 2010, no Acre, houve um total de 29 casos (T), sendo 10 registrados sem histórico (S) de casos na família e os outros 19 com histórico (C). O sistema calcula, portanto, que 34% não tem histórico familiar e os restantes 66%, sim (S/T =0,344 e C/T=0,655).

Na página de análise por Classificação Etária, é possível ver os percentuais dos grupos etários correspondentes ao Brasil como um todo no gráfico de colunas à esquerda. Conforme ilustrado na Figura 13, as colunas da parte central mostram os percentuais do estado selecionado (nesse caso, Tocantins). A caixa de opções no canto direito permite a restrição complementar de uma ou mais faixas etárias. O ano selecionado nesse exemplo foi 2010.

O gráfico nacional apresenta números condizentes com os dados do INCA, uma vez que a incidência de câncer em idosos costuma ser maior que na população mais jovem (INCA, 2019). O curioso nessa análise é que, para esse ano específico, o percentual de mulheres em

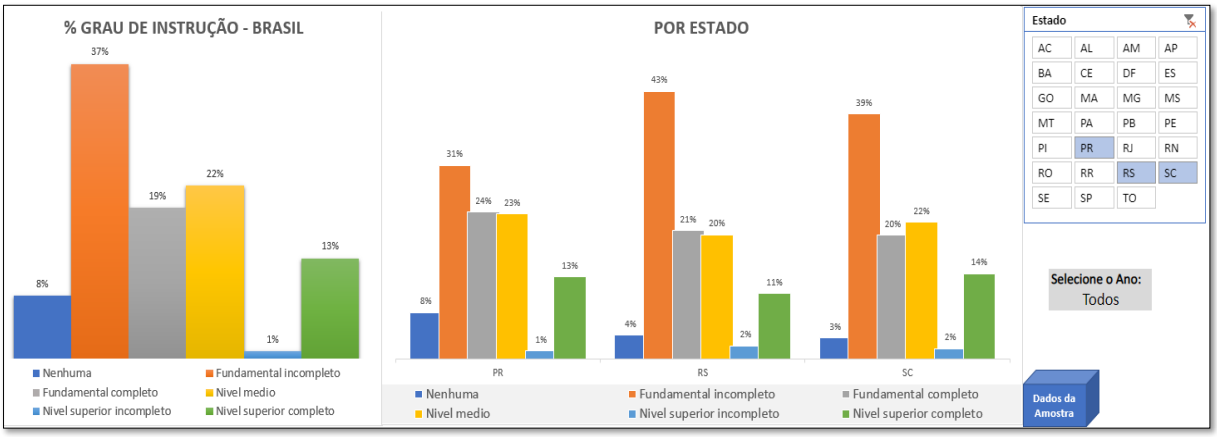
tratamento na faixa de 40 a 49 anos (40%) está bem acima da média brasileira (25%), ocupando a maior parcela dentre todas as faixas desse estado. O mais provável é que tenha sido um comportamento pontual desse ano pois, quando é selecionado o período acumulado, o número desse intervalo muda para 27%, ficando abaixo da faixa de maiores de 60 anos (com 31%).

Figura 13 - Classificação Etária do Tocantins e do Brasil em 2010



A análise do Grau de Instrução das pacientes mostra, também de forma comparativa, os dados nacionais e dos estados selecionados. A Figura 14 mostra os três estados da região Sul selecionados, no período total, indicando perfis parecidos com o restante do Brasil.

Figura 14 - Graus de instrução no Brasil e nos estados da região Sul



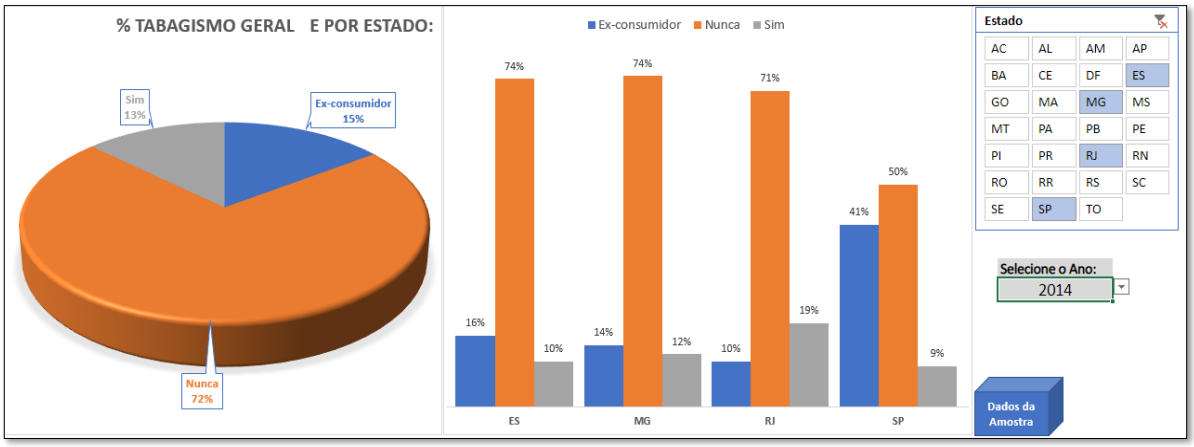
A quantidade de pacientes em tratamento de câncer de mama com ensino superior (completo ou incompleto) é um pouco maior em Santa Catarina (total de 16%), bem como o percentual de pessoas sem nenhuma instrução é o menor da região. Já o percentual de pessoas

com fundamental incompleto é o que atinge o maior valor em todas as esferas, com um alto índice no Rio Grande do Sul (43%), acima da média nacional (37%).

Alguns dos subgrupos (Tabagismo, Alcoolismo, Mais de um Tumor, Histórico Familiar, e Origem do Encaminhamento) mostram os valores referentes ao Brasil em um gráfico de setores (pizza), enquanto para os estados a apresentação continua sendo no formato gráfico de barras (colunas). A Figura 15 exemplifica esta situação para o Tabagismo, selecionando os estados da região Sudeste no ano de 2014.

No ano selecionado, as residentes em São Paulo apresentam um perfil bem diferente do resto da região sudeste e do Brasil, com 41% de ex-fumantes. O índice é bem mais alto que dos outros três estados. O Rio de Janeiro mostra o maior percentual de fumantes desta amostra, com 19% do total de pacientes.

Figura 15 - Perfil de uso de tabaco na região Sudeste e no Brasil, em 2014



Os demais subgrupos de consulta seguem o mesmo conceito apresentado até aqui. As diversas combinações possíveis de análise atingem o objetivo de estudos exploratórios relacionados às características de pacientes disponíveis na base de dados.

6.2.4 Análises por Tumor

Para os subgrupos relacionados ao tumor, foram criadas três páginas para análises descritivas:

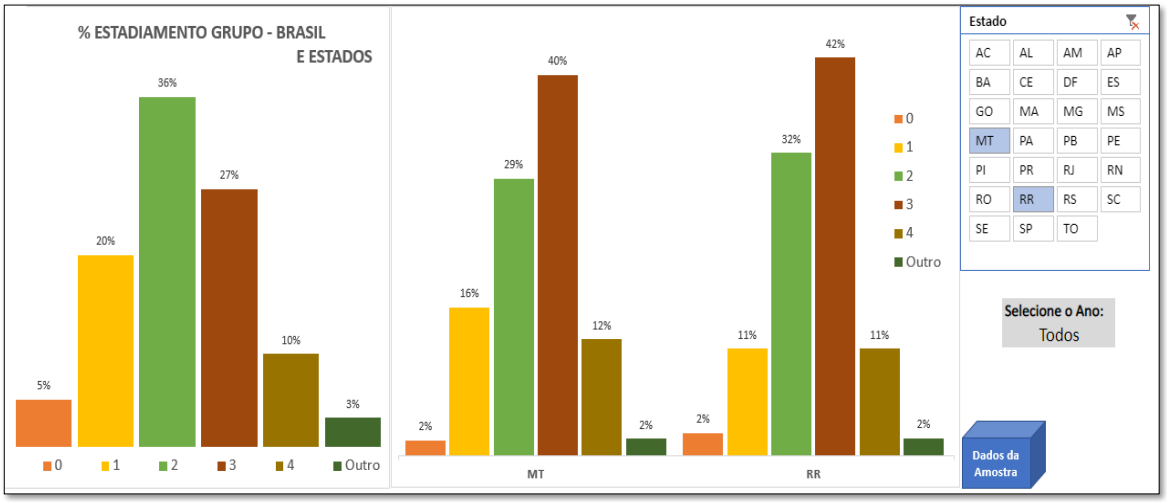
- Por Estadiamento Grupo
- Por Lateralidade
- Por Localização Primária

A fim de facilitar o uso da ferramenta, em todas elas houve a mesma padronização no formato dos gráficos e filtros já descrita na seção anterior: um gráfico à esquerda exibe os dados percentuais do Brasil e outro, à direita, dos estados selecionados.

Os gráficos podem ser filtrados por ano e foi disponibilizada a consulta das tabelas de origem dos dados. O cálculo dos percentuais segue o mesmo método adotado na seção anterior, em que valores vazios ou com preenchimento incompleto não foram considerados.

A análise considerando o primeiro subgrupo (Estadiamento Grupo) está representada na Figura 16, na qual foram selecionados os dois estados que destoaram do perfil brasileiro, quando considerado o período todo (2000 a 2016). Esta seleção foi feita manualmente, clicando uma a uma as siglas estaduais a fim de identificar situações atípicas. Em todas as outras Unidades da Federação o estádio 2 é o que tem os maiores percentuais, enquanto em Mato Grosso e em Roraima o estádio 3 é que fica em primeiro (com, respectivamente, 40% e 42%).

Figura 16 - Grau de estadiamento do câncer de mama no Brasil

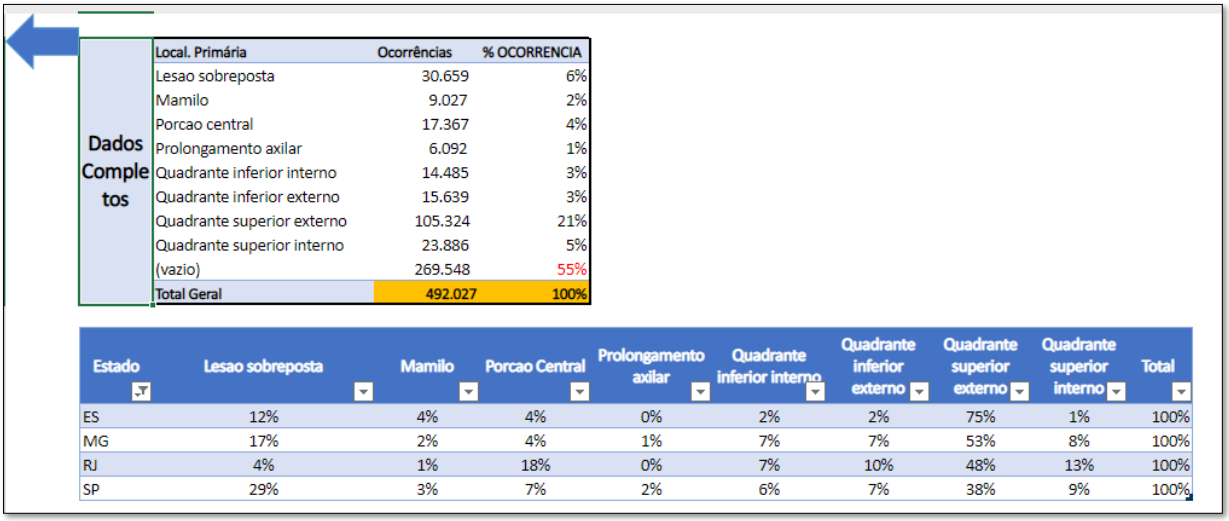


Em todas as páginas de análises descritivas citadas até aqui (nesta seção e nas duas anteriores), está disponível o botão “Dados da Amostra”, que leva a uma outra página com os dados exibidos em formato tabular. A Figura 17 mostra um exemplo de parte dessa tela, referente ao subgrupo Localização Primária, com os dados percentuais consolidados para os quatro estados da região Sudeste.

Na tabela de “Dados Completos” é possível identificar todas as opções disponíveis do subgrupo em questão, sendo que as não utilizadas aparecem com o percentual na cor vermelha. Os valores selecionados na tela anterior para estados e ano podem ser alterados manualmente

também nessa página. A tabela permite filtros por outros valores do atributo e seu conteúdo pode ser copiado para uma planilha.

Figura 17 - Tela de Dados da Amostra para Localização Primária da região Sudeste



Dados Completos

Local. Primária	Ocorrências	% OCORRENCIA
Lesao sobreposta	30.659	6%
Mamilo	9.027	2%
Porcao central	17.367	4%
Prolongamento axilar	6.092	1%
Quadrante inferior interno	14.485	3%
Quadrante inferior externo	15.639	3%
Quadrante superior externo	105.324	21%
Quadrante superior interno	23.886	5%
(vazio)	269.548	55%
Total Geral	492.027	100%

Estado	Lesao sobreposta	Mamilo	Porcao Central	Prolongamento axilar	Quadrante inferior interno	Quadrante inferior externo	Quadrante superior externo	Quadrante superior interno	Total
ES	12%	4%	4%	0%	2%	2%	75%	1%	100%
MG	17%	2%	4%	1%	7%	7%	53%	8%	100%
RJ	4%	1%	18%	0%	7%	10%	48%	13%	100%
SP	29%	3%	7%	2%	6%	7%	38%	9%	100%

6.2.5 Análises de Deslocamento

A página das análises relacionadas ao deslocamento contém um menu de navegação dividido em dois grandes blocos, conforme exibido na Figura 18. A parte da esquerda contém os atalhos para navegar pelos dois principais subgrupos: Distâncias e Esferas de Tratamento. Já o bloco da direita subdivide-se em cinco opções interativas:

- a quantidade total de registros selecionados aparece no cabeçalho;
- o período de análise pode ser alterado através da opção “Ano da Consulta”;
- um grupo de campos permite o filtro por região e estado residencial do paciente;
- um outro grupo de campos permite o filtro por região ou estado da UH;
- o ícone “Filtros Diversos” (em formato de funil) leva a outra página com mais campos para seleção.

Uma funcionalidade bastante útil da ferramenta é que, caso algum filtro seja utilizado em uma análise, ele se mantém ativo pelas demais telas navegadas. Por exemplo, se na página inicial o ano da consulta for restrito ao decênio de 2000 a 2009, esta seleção é levada para as

demais consultas e a quantidade de registros selecionados é atualizada automaticamente em todas elas.

Figura 18 - Tela inicial da página Deslocamento

Os dois subgrupos (Distância e Esferas de Tratamento) tem opções de análises bem diferentes e por isso serão apresentados separadamente a seguir.

6.2.5.1 Visão: Distâncias

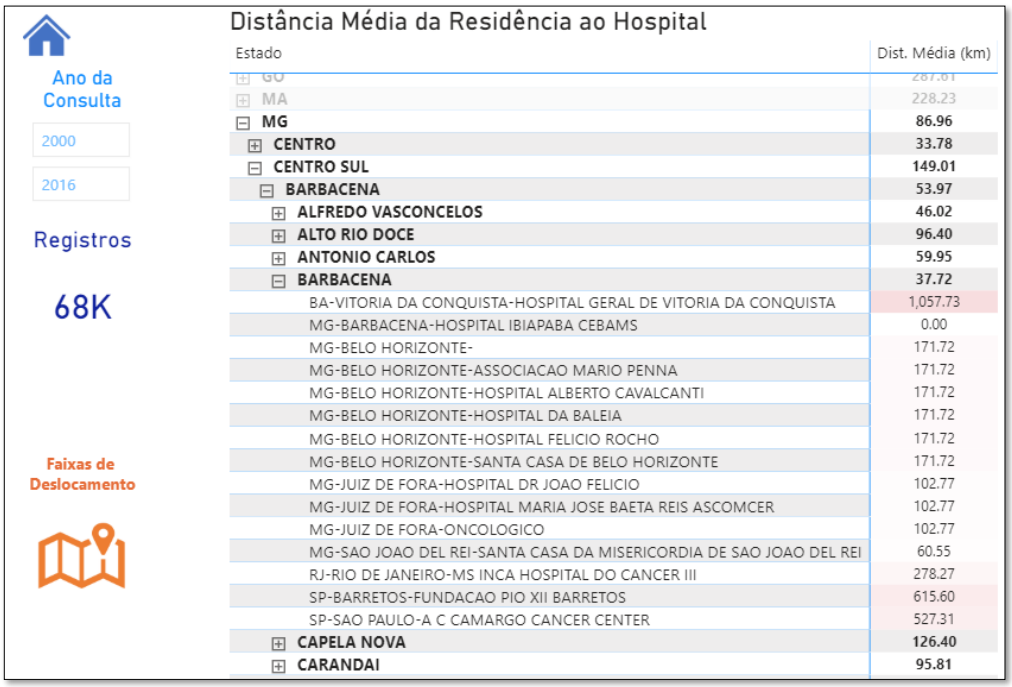
A visão Distâncias exibe na tela principal uma tabela com duas colunas. A primeira coluna mostra a esfera selecionada (inicialmente agrupada por estados) e a segunda mostra o valor da distância média calculada para cada linha, em quilômetros. Os valores são exibidos com uma coloração graduada do branco ao vermelho indicando, respectivamente, as menores para as maiores distâncias (quanto mais forte o tom de vermelho, maior a distância média do grupo).

Ao clicar no ícone “mais” (+) ao lado do estado, pode-se abrir a hierarquia seguinte (macrorregião de saúde); da mesma forma, uma ação semelhante abre o próximo nível (região de saúde) e, em seguida, a hierarquia mais baixa (município). Ao abrir o nível municipal, o que aparecem são as unidades hospitalares às quais as pacientes daquela cidade foram se tratar – para facilitar a identificação, aparecem também a UF e a cidade de destino.

A Figura 19 mostra um exemplo desses recursos. O estado de Minas Gerais foi destacado, selecionando em seguida a macrorregião de saúde Centro-Sul, depois a região de saúde Barbacena, e por fim o município de Barbacena. Mesmo com alguns deslocamentos distantes para Bahia, São Paulo e Rio de Janeiro, a distância média é de 37,72 km. Subindo um nível, é possível ver que a média da região de saúde é de 53,97 km, e esse número aumenta ao se considerar a macrorregião de saúde Centro-Sul, para 149,01 km. Contudo, a média de distância em Minas Gerais é mais baixa, de 86,96 km, pois engloba todas as ocorrências de residentes no estado e há outras unidades hospitalares espalhadas em regiões específicas.

Na mesma figura, é possível ver que, ao selecionar somente o estado de Minas Gerais, a quantidade de registros diminuiu para 68 mil, ao invés dos cerca de 492 mil da base completa. Isto é feito automaticamente pelo sistema ao dar ênfase em alguma linha específica. De modo análogo, é possível pesquisar as distâncias e ocorrências de outros lugares do país, seja filtrando a região desejada ou navegando pela lista de opções dessa tabela.

Figura 19 - Exemplo da tabela de Distância Média da Residência ao Hospital

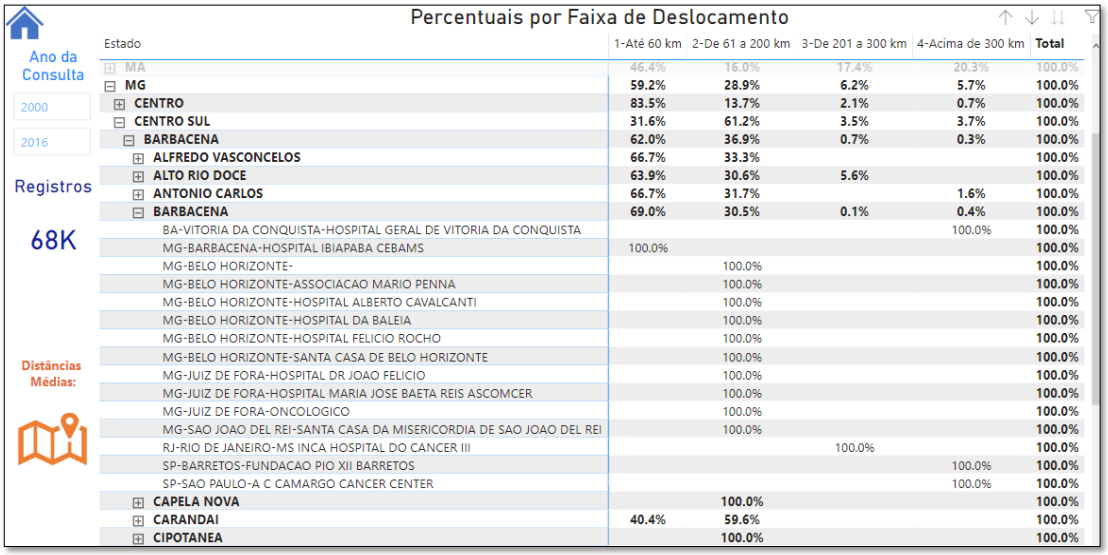


Ao clicar no ícone de “Faixas de Deslocamento”, no canto inferior esquerdo da tela, é exibida uma nova página, também com os dados em forma de tabela. Na coluna da esquerda, aparece a mesma estrutura hierárquica geográfica anterior (do estado até o município). A parte direita exibe outras colunas: mostra os percentuais de ocorrência para cada faixa de deslocamento. Da mesma forma que na página anterior, nesta também é possível navegar pelos

níveis inferiores da hierarquia geográfica até o município e, deste, identificar os destinos e UH's nos quais os tratamentos foram feitos.

A Figura 20 exibe um exemplo desta tela, para a mesma região de saúde e município indicados na figura anterior. É possível verificar que houve alguns deslocamentos (cerca de 30%) na faixa entre 61 e 200 quilômetros para a capital mineira e outros municípios relativamente próximos a Barbacena. A ampla maioria das ocorrências (69%) foram atendidas na unidade hospitalar da própria cidade. Os deslocamentos nos trajetos mais extensos (para outros estados como Bahia, Rio de Janeiro e São Paulo) encontram-se nas duas faixas finais, com cerca de 0,5% do total de casos no período selecionado de 2000 a 2016.

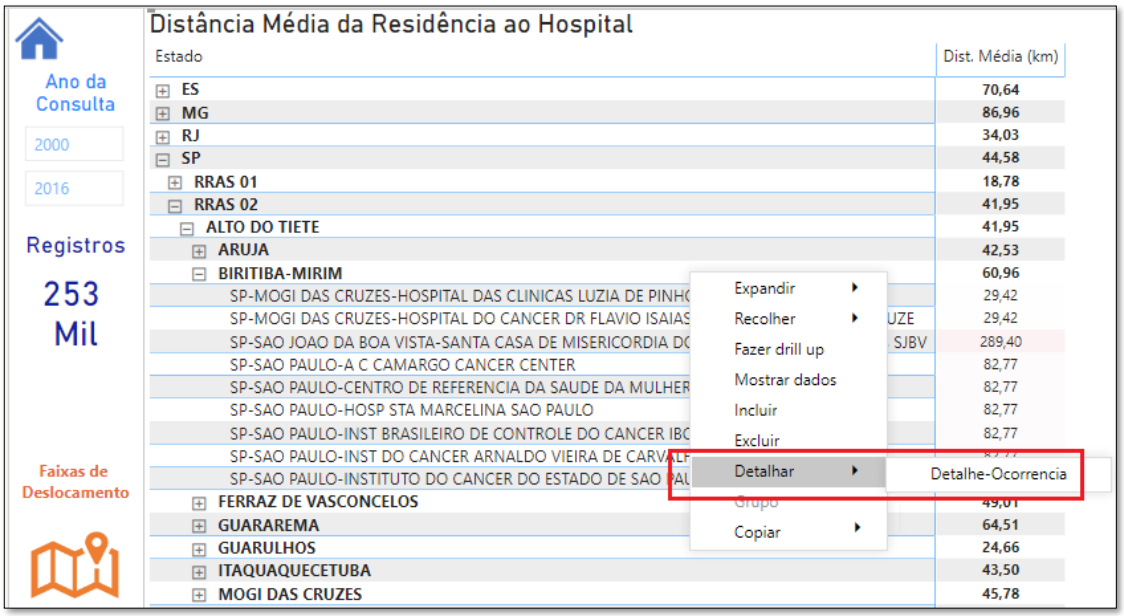
Figura 20 - Exemplo da tela de Faixas de Deslocamento para parte do estado de Minas Gerais



Estado	1-Até 60 km	2-De 61 a 200 km	3-De 201 a 300 km	4-Acima de 300 km	Total
MA	46.4%	16.0%	17.4%	20.3%	100.0%
MG	59.2%	28.9%	6.2%	5.7%	100.0%
CENTRO	83.5%	13.7%	2.1%	0.7%	100.0%
CENTRO SUL	31.6%	61.2%	3.5%	3.7%	100.0%
BARBACENA	62.0%	36.9%	0.7%	0.3%	100.0%
ALFREDO VASCONCELOS	66.7%	33.3%			100.0%
ALTO RIO DOCE	63.9%	30.6%	5.6%		100.0%
ANTONIO CARLOS	66.7%	31.7%		1.6%	100.0%
BARBACENA	69.0%	30.5%	0.1%	0.4%	100.0%
BA-VITORIA DA CONQUISTA-HOSPITAL GERAL DE VITORIA DA CONQUISTA				100.0%	100.0%
MG-BARBACENA-HOSPITAL IBIAPABA CEBAMS	100.0%				100.0%
MG-BELO HORIZONTE-		100.0%			100.0%
MG-BELO HORIZONTE-ASSOCIACAO MARIO PENNA		100.0%			100.0%
MG-BELO HORIZONTE-HOSPITAL ALBERTO CAVALCANTI		100.0%			100.0%
MG-BELO HORIZONTE-HOSPITAL DA BALEIA		100.0%			100.0%
MG-BELO HORIZONTE-HOSPITAL FELICIO ROCHO		100.0%			100.0%
MG-BELO HORIZONTE-SANTA CASA DE BELO HORIZONTE		100.0%			100.0%
MG-JUIZ DE FORA-HOSPITAL DR JOAO FELICIO		100.0%			100.0%
MG-JUIZ DE FORA-HOSPITAL MARIA JOSE BAETA REIS ASCOMCER		100.0%			100.0%
MG-JUIZ DE FORA-ONCOLOGICO		100.0%			100.0%
MG-SAO JOAO DEL REI-SANTA CASA DA MISERICORDIA DE SAO JOAO DEL REI		100.0%			100.0%
RJ-RIO DE JANEIRO-MS INCA HOSPITAL DO CANCER III			100.0%		100.0%
SP-BARRETOS-FUNDACAO PIO XII BARRETOS				100.0%	100.0%
SP-SAO PAULO-A C CAMARGO CANCER CENTER				100.0%	100.0%
CAPELA NOVA		100.0%			100.0%
CARANDAI	40.4%	59.6%			100.0%
CIPOTANEA		100.0%			100.0%

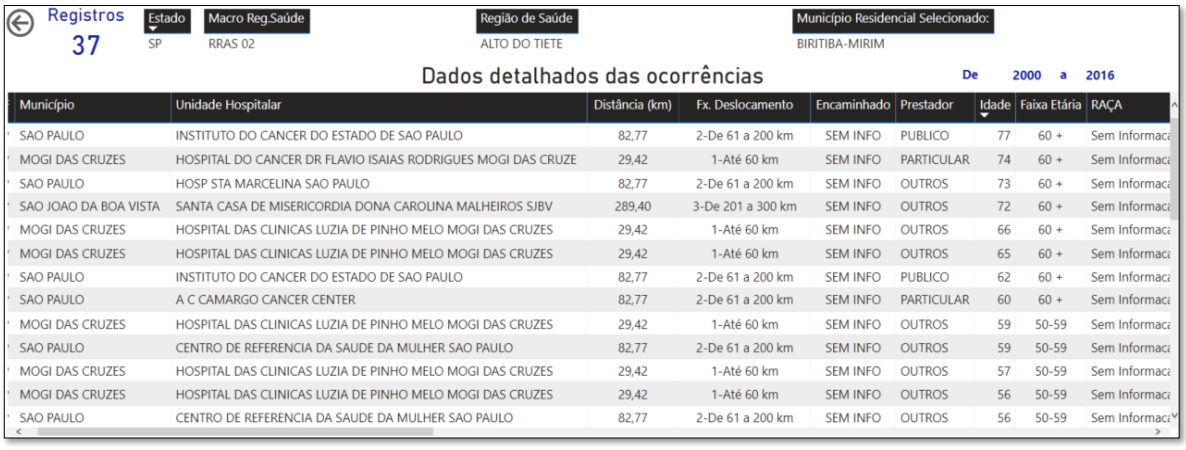
Uma outra funcionalidade disponível em todas as análises de deslocamento, ao se chegar ao nível municipal, é a possibilidade de exibir a lista completa das ocorrências, num formato tabular. Para isso, deve-se usar o botão direito do *mouse* para escolher a opção "Detalhar-Detalhe Ocorrência" (em destaque na Figura 21), que irá então navegar para a página com os atributos disponíveis na base, permitindo que se identifiquem eventuais características relevantes para a análise sendo feita.

Figura 21 - Exemplo de ações disponíveis na página "Visão Distâncias"



A Figura 22 ilustra o resultado do detalhamento indicado na Figura 21. São exibidas as ocorrências da cidade de Biritiba-Mirim (pertencente à região de saúde do Alto do Tietê e macrorregião de saúde RRAS 02, do estado de São Paulo), no período de 2000 a 2016.

Figura 22 - Exemplo da tela com os dados detalhados das Ocorrências para um município



Município	Unidade Hospitalar	Distância (km)	Fx. Deslocamento	Encaminhado	Prestador	Idade	Faixa Etária	RAÇA
SAO PAULO	INSTITUTO DO CANCER DO ESTADO DE SAO PAULO	82,77	2-De 61 a 200 km	SEM INFO	PUBLICO	77	60 +	Sem Informaca
MOGI DAS CRUZES	HOSPITAL DO CANCER DR FLAVIO ISAIAS RODRIGUES MOGI DAS CRUZE	29,42	1-Até 60 km	SEM INFO	PARTICULAR	74	60 +	Sem Informaca
SAO PAULO	HOSP STA MARCELINA SAO PAULO	82,77	2-De 61 a 200 km	SEM INFO	OUTROS	73	60 +	Sem Informaca
SAO JOAO DA BOA VISTA	SANTA CASA DE MISERICORDIA DONA CAROLINA MALHEIROS SJBV	289,40	3-De 201 a 300 km	SEM INFO	OUTROS	72	60 +	Sem Informaca
MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	29,42	1-Até 60 km	SEM INFO	OUTROS	66	60 +	Sem Informaca
MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	29,42	1-Até 60 km	SEM INFO	OUTROS	65	60 +	Sem Informaca
SAO PAULO	INSTITUTO DO CANCER DO ESTADO DE SAO PAULO	82,77	2-De 61 a 200 km	SEM INFO	PUBLICO	62	60 +	Sem Informaca
SAO PAULO	A C CAMARGO CANCER CENTER	82,77	2-De 61 a 200 km	SEM INFO	PARTICULAR	60	60 +	Sem Informaca
MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	29,42	1-Até 60 km	SEM INFO	OUTROS	59	50-59	Sem Informaca
SAO PAULO	CENTRO DE REFERENCIA DA SAUDE DA MULHER SAO PAULO	82,77	2-De 61 a 200 km	SEM INFO	OUTROS	59	50-59	Sem Informaca
MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	29,42	1-Até 60 km	SEM INFO	OUTROS	57	50-59	Sem Informaca
MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	29,42	1-Até 60 km	SEM INFO	OUTROS	56	50-59	Sem Informaca
SAO PAULO	CENTRO DE REFERENCIA DA SAUDE DA MULHER SAO PAULO	82,77	2-De 61 a 200 km	SEM INFO	OUTROS	56	50-59	Sem Informaca

A distância média é calculada através da média simples dos deslocamentos nas ocorrências pertencentes àquela esfera. As faixas de deslocamento foram definidas durante a etapa de mineração de dados, usando a técnica Agrupamento (descrita na seção correspondente). Os valores são autoexplicativos: "0-Até 60 km" indica que as distâncias nesta

faixa vão de zero até sessenta quilômetros, e assim sucessivamente. O percentual indicado é calculado dividindo a quantidade de registros pelo número total de casos daquele nível.

Um dos resultados encontrados ao se fazer a análise da base completa, sem filtro algum, é que a distância média de deslocamento das pacientes em tratamento de câncer de mama no Brasil foi de 88,4 km. Filtrando os dados somente do primeiro ano e em seguida do último, foi verificado que o deslocamento médio nacional aumentou, passando de 79,0 km no ano 2000 para 92,8 km em 2016.

Ao ser analisada a base histórica completa, considerando a visão das faixas de deslocamento, foi encontrado o percentual de 69,5% na faixa inicial (que abrange deslocamentos até 60 km). No ano 2000, esse percentual era de 72,6%, tendo diminuído para 65,8% em 2016. Essa diminuição refletiu-se no acréscimo das demais faixas, com destaque para a segunda (entre 61 e 200 km), que cresceu de 16,9% no primeiro ano para 21,4% no último, coincidindo com a distância média crescente relatada no parágrafo anterior.

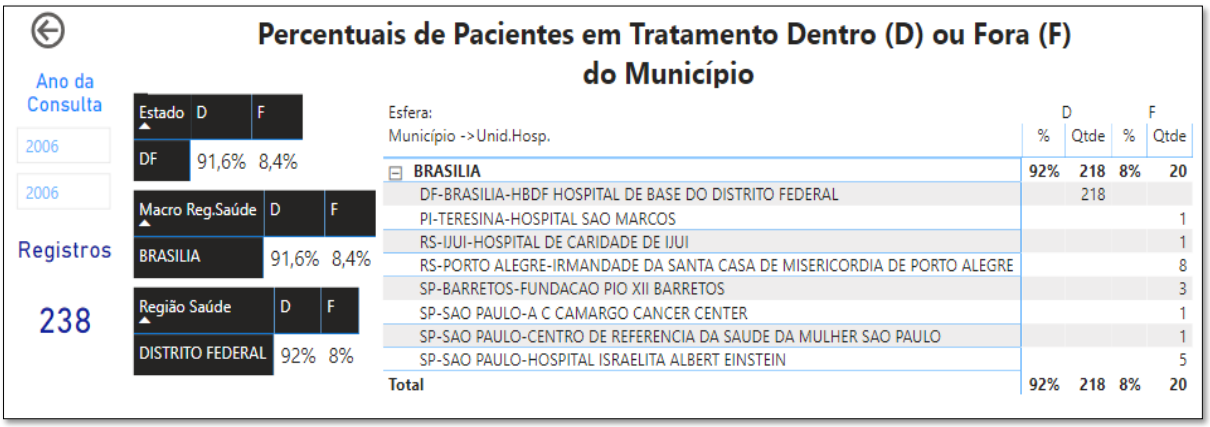
6.2.5.2 Visão: Esferas de Tratamento

Esta análise foi construída com a ideia de que se possa navegar por todas as divisões administrativas da saúde pública, numa sequência lógica, hierárquica, do nível mais amplo para o mais refinado. Em cada uma das páginas, é exibido o mesmo tipo de informação: os percentuais de tratamentos feitos dentro (D) ou fora (F) daquela esfera de tratamento.

O primeiro nível começa com os estados e os percentuais D/ F de cada um, em formato tabular. Pode-se clicar no estado desejado e navegar (usando a opção “Detalhar”) para a esfera seguinte. Aparecem então todas as macrorregiões de saúde existentes para o estado pré-selecionado, com seus percentuais de tratamento D/F. Usando o mesmo conceito, pode-se navegar para o detalhamento seguinte (região de saúde) e, mais uma vez, até a lista de municípios que compõem aquela divisão hierárquica. Ao chegar a esse último nível (esfera municipal), pode-se clicar no ícone “mais” (+) ao lado do município para exibir as UH’s (com suas cidades e estados) em que as pacientes foram se tratar.

A Figura 23 ilustra essa análise com a página aberta no último nível, ou seja, a esfera municipal. Nos blocos em fundo preto são exibidos os níveis superiores (estado, macrorregião e região de saúde). No lado esquerdo aparece a quantidade de registros atualizada e, à direita, a tabela com os municípios dessa região – nesse caso, somente Brasília – exibindo os locais de tratamento para seus 238 casos.

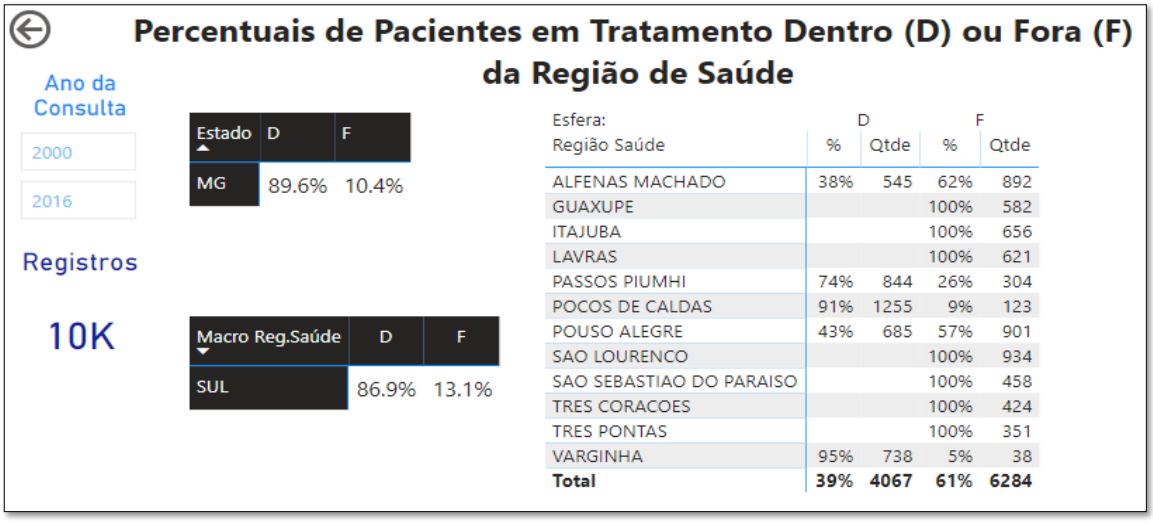
Figura 23 - Exemplo da página "Visão Esferas de Tratamento" no nível municipal



A Figura 24 mostra outro exemplo de análise, com a tela aberta na esfera da região de saúde (e, portanto, já com o estado e a macrorregião de saúde selecionados e exibidos na tela). É possível verificar que o contador à esquerda apresenta o número arredondado de 10 mil registros, que é a soma dos totais de ocorrências exibidos na tabela (dentro: 4.067, fora: 6.284).

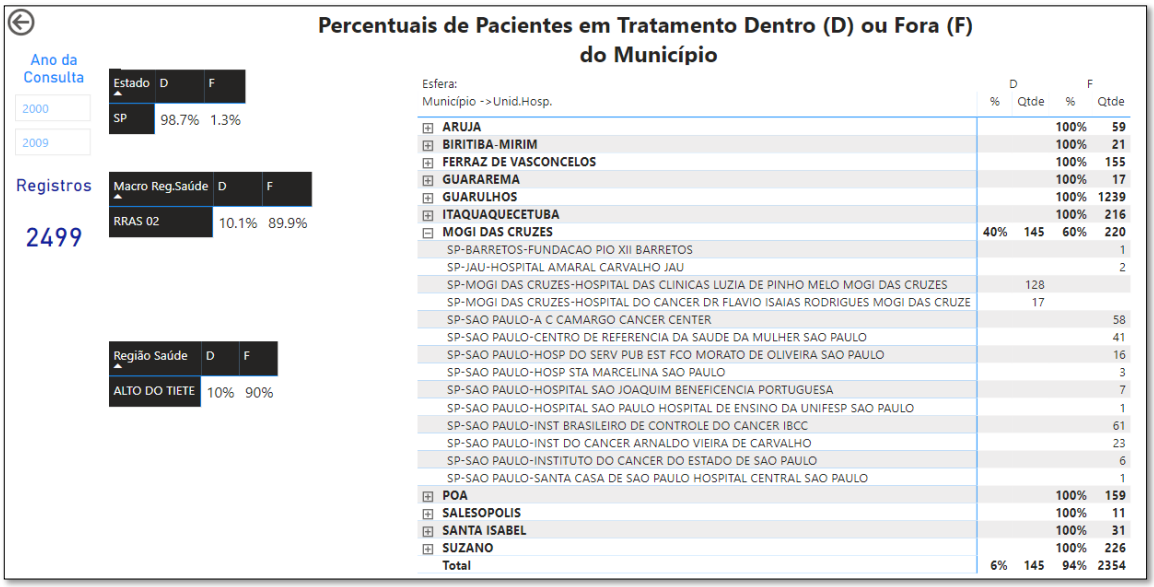
Pela imagem, pode-se ver que apenas cinco das doze regiões de saúde exibidas devem ter unidades hospitalares de tratamento oncológico ("Alfenas Machado", "Passos Piumhi", "Poços de Caldas", "Pouso Alegre" e "Varginha"), pois nas outras o percentual de tratamento externo é de 100% dos casos. É visível também que as regiões com maior predominância de atendimento interno são as de Varginha e Poços de Caldas, com respectivamente 95% e 91% das ocorrências nessa situação.

Figura 24 - Tela da visão Esferas de Tratamento, nível da região de saúde, para a macroregião de saúde Sul de Minas Gerais



Foi feita também uma análise para a região de saúde do Alto Tietê, apresentada na Figura 25, restrita à primeira década deste século. Nesta esfera, em 10 dos 11 municípios as pacientes deslocaram-se para outras cidades. Porém, mesmo em Mogi das Cruzes, a maioria das pacientes (60%) foi se tratar na capital, a despeito de haver duas instituições localizadas na cidade. Os demais municípios não contam com UH's oncológicas disponíveis e, por isso, todos apresentam o índice de 100% de tratamento externo.

Figura 25 - Percentual de pacientes residentes em Mogi das Cruzes em tratamento dentro ou fora do município entre os anos de 2000 e 2009



Ao se atingir o nível municipal de qualquer análise de deslocamento, a funcionalidade que permite ver os dados detalhados das ocorrências é habilitada, exibindo uma nova página com uma lista de diversos atributos, em formato de tabela. A Figura 26 exemplifica esta situação, exibindo uma parte dos registros da cidade de Mogi das Cruzes (a partir dos filtros da análise da Figura 25).

No cabeçalho da tela, aparecem os dados das esferas pré-selecionadas durante a navegação, permitindo que se identifique mais rapidamente a que região essa listagem pertence. O contador também é atualizado, mostrando somente os 365 itens (145 dentro e 220 fora do município).

Figura 26 - Parte da tela com o detalhe das ocorrências para a cidade de Mogi das Cruzes

Registros 365 | Estado: SP | Macro Reg.Saúde: RRAS 02 | Região de Saúde: ALTO DO TIETE | Município Residencial Selecionado: MOGI DAS CRUZES

Dados detalhados das ocorrências — De 2000 a 2009

UF	Município	Unidade Hospitalar	Distância (km)	Fx. Deslocamento	Encaminhado	Prestador	Idade	Faixa Etária	RAÇA	INSTRUÇ
SP	BARRETOS	FUNDACAO PIO XII BARRETOS	483.07	4-Acima de 300 km	SEM INFO	PARTICULAR	24	20-39	Sem Informacao	Sem info
SP	JAU	HOSPITAL AMARAL CARVALHO JAU	358.25	4-Acima de 300 km	SEM INFO	OUTROS	63	60 +	Sem Informacao	Fundame
SP	JAU	HOSPITAL AMARAL CARVALHO JAU	358.25	4-Acima de 300 km	SEM INFO	OUTROS	63	60 +	Sem Informacao	Fundame
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	25	20-39	Sem Informacao	Fundame
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	31	20-39	Sem Informacao	Nivel me
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	32	20-39	Sem Informacao	Fundame
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	33	20-39	Sem Informacao	Nivel me
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	34	20-39	Sem Informacao	Fundame
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	34	20-39	Sem Informacao	Nivel me
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	34	20-39	Sem Informacao	Nivel me
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	34	20-39	Sem Informacao	Nivel sup
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	36	20-39	Sem Informacao	Fundame
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	36	20-39	Sem Informacao	Nivel me
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	36	20-39	Sem Informacao	Nivel me
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	37	20-39	Sem Informacao	Fundame
SP	MOGI DAS CRUZES	HOSPITAL DAS CLINICAS LUZIA DE PINHO MELO MOGI DAS CRUZES	0.00	1-Até 60 km	SEM INFO	OUTROS	38	20-39	Sem Informacao	Fundame

Para calcular o percentual Dentro/Fora de cada esfera, o sistema considera a quantidade total de ocorrências e o valor encontrado de D/F em cada respectiva coluna “mm_uf”, “mm_mrs”, “mm_regs” ou “mm_munic”. Pelo exemplo da figura anterior, da região de saúde do Alto Tietê, houve 2499 ocorrências (T), sendo que 145 se trataram dentro (D) e 2354, fora (F). Com isso, o sistema identifica para essa esfera os valores arredondados respectivos de 6% e 94% (F/T = 0,058 e D/T = 0,942).

6.3 ANÁLISES DESCRITIVAS POR MINERAÇÃO DE DADOS

As análises por mineração de dados estão divididas em duas seções, conforme a técnica usada:

1. **Agrupamento** – detalha a execução dos modelos selecionados, com os parâmetros utilizados em cada rodada e as tabelas finais com as quantidades de ocorrências em cada grupo.

2. **Associação** – apresenta os parâmetros utilizados e os modelos gerados pelo sistema, comparando os resultados encontrados entre as regiões ou estados. Também é exibido um quadro resumindo as características de cada modelo, para cada par de atributos analisados.

6.3.1 Análises por Agrupamento

Foram selecionados inicialmente três atributos da base para serem usados no Agrupamento: distância de deslocamento, dias até o tratamento, e idade. Os dois primeiros geraram grupos que puderam ser reutilizados em outras etapas e serão descritos nas seções a seguir.

As faixas de distância de deslocamento foram convertidas em um novo atributo durante a seleção dos subconjuntos de dados e, assim, possibilitaram uma opção adicional para a análise descritiva estatística do deslocamento. Do mesmo modo, as faixas de tempo até o tratamento também foram inseridas na base, com o intuito de usá-las na etapa seguinte.

Para o terceiro atributo (“idade”), foram elaborados cerca de dez modelos diferentes, de forma análoga aos anteriores. Entretanto, os grupos identificados não se mostraram relevantes para as análises e, por isso, foram descartados. Para as faixas etárias, optou-se por utilizar a classificação indicada por especialista da saúde pública.

6.3.1.1 Modelos para Distância de Deslocamento

Foram elaborados sete modelos diferentes para o atributo distância, executados em sequência, de modo a identificar os melhores intervalos possíveis. Em cada execução, o programa exibia uma representação gráfica dos *clusters* gerados, além de outras informações estatísticas como desvio padrão, quantidade de registros e média dos valores agrupados. Os parâmetros e filtros utilizados estão resumidos no Quadro 2.

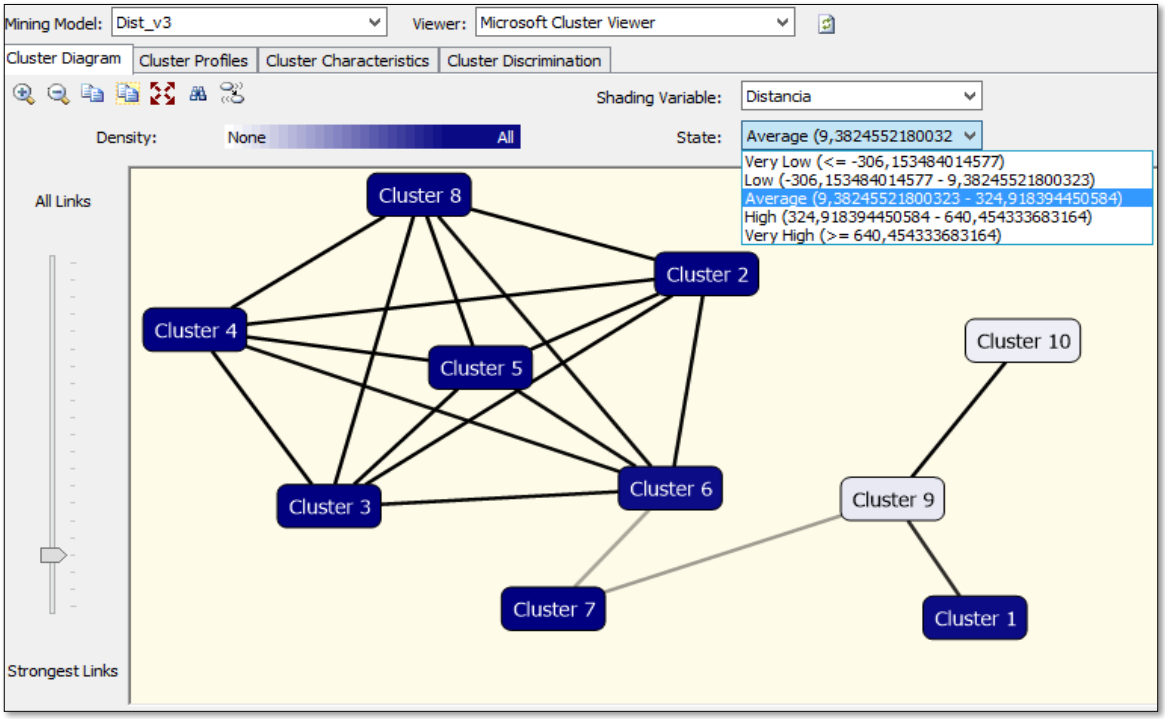
Quadro 2 - Comparativo entre os modelos de Agrupamento para Distância

Modelo	Parâmetro: *Cluster_Count*	Parâmetro: *Cluster_Method*	Filtro
Modelo 1	10	EMe	-
Modelo 2	10	KMe	-
Modelo 3	10	EMe	> 0
Modelo 4	10	KMe	> 0
Modelo 5	10	EMe	> 0 e < 1280
Modelo 6	4	EMe	> 0 e < 643
Modelo 7	5	EMe	> 0 e < 643

Os cinco primeiros modelos, executados com o padrão de dez grupos, trouxeram vários *clusters* intermediários com valores muito próximos entre si, indicando que talvez pudessem ser reagrupados. Percebeu-se que a quantidade grande de ocorrências na base com distância zero (pouco menos de 50%) acabava deturpando os grupos, por isso foi feito no terceiro modelo um filtro com valores maiores que zero.

O *software* sempre colocava os grupos gerados em cinco condições (muito baixo, baixo, médio, alto e muito alto), cujos valores mudavam conforme os parâmetros e filtros selecionados. A cor de cada grupo variava de intensidade dependendo de qual condição estava selecionada. No exemplo da Figura 27, referente ao modelo 3, pode-se ver que a condição Média (*Average*, no termo em inglês) está selecionada e oito dos dez clusters estão com a cor mais intensa, indicando que todos eles se encaixam dentro dessa ampla faixa que vai de 9,38 até 324,91.

Figura 27 - Tela do software de mineração de dados exibindo o modelo 3 do Agrupamento para Distância



Além disso, nas duas vezes em que o método KMe foi usado, identificou-se que o algoritmo gerava a maioria dos grupos com populações muito pequenas (em torno de 20 itens), compostas quase sempre de valores únicos, e a imensa parcela restante das ocorrências era

inserida num *cluster* genérico. Por isso ele foi descartado e o método EMe acabou sendo selecionado para as demais rodadas.

A partir da quinta execução, a lógica para obter os grupos mais bem separados foi pegar o valor médio do cluster com mais *outliers* da melhor execução anterior e usá-lo também como filtro superior (uma vez que o filtro inferior já estava definido como sendo valores maiores que zero).

Foi feita uma sexta execução com a mesma lógica anterior (eliminar os *outliers*), que diminuiu o valor do filtro superior para 643 km. Usando a sugestão do *software* para classificação populacional da base (Baixa distância, Média, Alta e Muito Alta), o parâmetro de grupos foi reduzido para quatro. Neste modelo, os grupos tinham o menor desvio padrão dentre todas as iterações e por isso foi considerado o vencedor.

A sétima e última execução foi feita apenas alterando o número de *clusters* criados pelo sistema para cinco, a fim de verificar se haveria melhoria significativa nos agrupamentos. No entanto, dois deles acabaram ficando dentro da mesma condição e, por isso, a sugestão de quatro grupos foi mantida.

As faixas obtidas no modelo vencedor estão indicadas na Figura 28. Os valores "muito baixos" (*Very Low*), exibidos na primeira condição, são negativos, e, como na prática isso não ocorre, este primeiro intervalo foi desconsiderado do resultado. As outras quatro faixas foram arredondadas, definidas numa sequência de 1 a 4 e nomeadas de forma autoexplicativa. Suas descrições e distribuição podem ser vistas na tabela 2.

Figura 28 - Tela do modelo 6 do Agrupamento para Distância com os valores das faixas obtidas

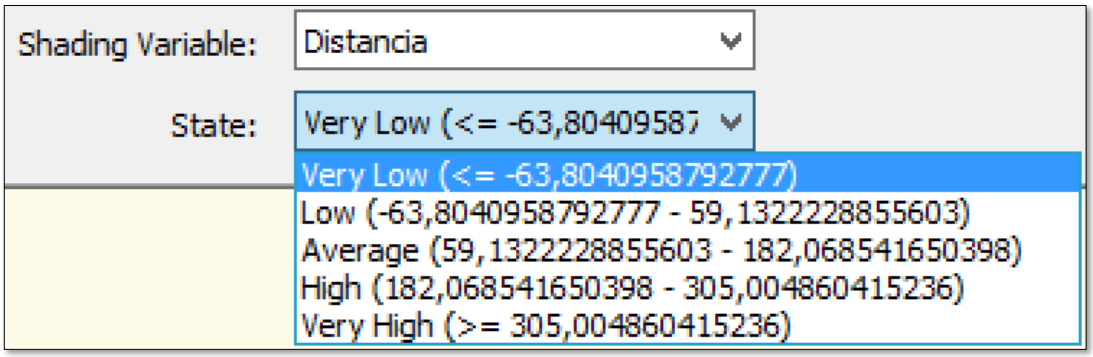


Tabela 2 - Distribuição da base de dados por Faixa de Deslocamento

Faixas de Deslocamento	Quantidade	% do Total
1 - Até 60 km	342.064	69,5%
2 - De 61 a 200 km	91.741	18,6%
3 - De 201 a 300 km	24.271	4,9%
4 - Acima de 300 km	33.947	6,9%
Total:	**492.023**	**100%**

A primeira faixa contempla, portanto, todos os registros até 60 quilômetros, incluindo aqueles que, por serem do mesmo município, não puderam ser calculados. Esse deslocamento pode ser considerado um valor abrangente também nesses casos, uma vez que seriam poucos casos em que os pacientes de uma mesma cidade teriam de se deslocar mais do que 60 quilômetros.

Observa-se nesta análise que, a despeito das dimensões continentais do Brasil, o tratamento de câncer de mama é extremamente concentrado em distâncias curtas (até 60 quilômetros) e médias (entre 61 e 200 quilômetros), chegando a cerca de 88% das ocorrências. Resultado este bem próximo ao do estudo de Vetterlein *et al*. (2017), que englobou todo o território norte-americano, e indicou alta concentração nas duas faixas iniciais. Os autores definiram distâncias curtas como aquelas menores que 20km, e distância média entre 20 e 80km, encontrando 87,9%, dos casos nessas faixas.

Outro dado encontrado foi que 47,1% das ocorrências correspondem a pacientes que tratam no seu município, ou seja, tem distância zero (231.778 casos), considerando o período todo da base (2000 a 2016). Este resultado é semelhante aos encontrados por Saldanha *et al*. (2019), que fez análises das internações e tratamentos de químio e radioterapia entre 2014 e 2016, e ao encontrado na análise de Oliveira *et al*. (2011) para radioterapia entre os anos de 2005 e 2006. É, porém, ligeiramente menor do que o estudo de Silva *et al*. (2019), que está restrito somente a quimioterapia e ao ano de 2013.

As faixas definidas nessa etapa foram usadas nos dois métodos de análise do estudo:

- análise descritiva do deslocamento por métodos estatísticos (vide seção prévia 6.2.5)

- análise descritiva por mineração de dados usando a técnica de Associação (vide seção posterior 6.3.2)

6.3.1.2 Modelos para Dias Até o Tratamento

Foram elaborados cinco modelos diferentes para o atributo “dias até o tratamento”, executados em sequência, de modo a identificar os melhores intervalos possíveis. A metodologia foi bastante semelhante à da seção anterior, mas sem a necessidade de várias tentativas iniciais para se descobrir o melhor método. Foi feita uma única execução com o algoritmo KMe, que teve o mesmo comportamento inadequado já descrito previamente e, portanto, foi descartado. Os filtros e ajustes em parâmetros estão resumidos no Quadro 3 abaixo.

O valor do atributo “dias até o tratamento” mostra a quantidade de dias decorridos entre a data do diagnóstico e a data do início do tratamento. Na rodada inicial, foram identificados muitos valores negativos, que indicam que o tratamento teria começado antes do diagnóstico. Isso pode ser sinal de mau preenchimento da base ou, de fato, o tratamento pôde ser iniciado antes do diagnóstico conclusivo. De qualquer forma, tais valores foram filtrados e agrupados numa faixa específica, para isolá-los das demais análises.

Quadro 3 - Comparativo entre os modelos de Agrupamento para Dias até o Tratamento

Modelo	**Parâmetro: *Cluster_Count***	**Parâmetro: *Cluster_Method***	**Filtro**
Modelo 1	10	EMe	-
Modelo 2	10	KMe	-
Modelo 3	10	EMe	>= 0 e < 1223
Modelo 4	5	EMe	>= 0 e < 670
Modelo 5	5	EMe	> 0 e < 670

No terceiro modelo, foram colocados dois filtros: um limite inferior maior ou igual a zero e um limite superior (1.223 dias) obtido do valor médio do grupo com mais *outliers* da primeira execução. Contudo, os dez *clusters* gerados ainda mostravam uma grande redundância, com possíveis reagrupamentos.

Na quarta rodada foi mantida a lógica de eliminar os *outliers* e diminuiu-se o valor do filtro superior para 670 dias. Para poder usar novamente a sugestão do *software* de classificação populacional da base (Muito Baixa, Baixa, Média, Alta e Muito Alta), o parâmetro de grupos foi reduzido para cinco. Com estes ajustes, o modelo já apresentava valores satisfatórios.

Como, porém, havia uma quantidade grande de ocorrências com zero dias, foi feita mais uma modelagem, alterando o filtro inferior para valores maiores que zero. Nesta quinta

execução os grupos encontrados tinham o menor desvio padrão dentre todas as execuções e, por isso, foi considerado o modelo vencedor.

A sugestão do *software* foi avaliada e, a partir desta recomendação, alguns números foram arredondados para facilitar a identificação em meses. Por exemplo: o valor de 260 dias foi ajustado para 270, a fim de permitir a conversão para nove meses. A descrição final de cada faixa e sua distribuição dentro da base de dados está exibida na Tabela 3.

Assim como o agrupamento anterior, as faixas encontradas foram usadas nas análises por Associação. É possível notar uma boa divisão populacional entre os grupos, sem predominância de um em específico. Entretanto, um cenário com maiores percentuais nas faixas de tempo mais curtas indicaria maior agilidade no tratamento, o que não parece ser o caso.

Tabela 3 - Distribuição da base de dados por Faixa de Tempo para Tratamento

Faixas de Tempo para Tratamento[(a)]	Quantidade	% do Total
0 - Sem Info (< 0 ou nulo)	55.481	11,3%
1 - Abaixo de 1 mês (>= 0 e <= 30)	108.914	22,1%
2 - De 1 a 2 meses (> 30 e <= 60)	89.560	18,2%
3 - De 2 a 5 meses (> 60 e <= 150)	134.022	27,2%
4 - De 5 a 9 meses (> 150 e <= 270)	54.046	11,0%
5 - Acima de 9 meses (> 270)	50.000	10,2%
Total:	**492.023**	**100 %**

(a) Os números entre parênteses correspondem a dias.

As duas faixas iniciais indicam tratamentos em até 60 dias e englobam 40,3% das pacientes. As pacientes que levaram mais de 60 dias para iniciar o tratamento correspondem a cerca de 48% dos casos, o que é bem acima do resultado de 36% encontrado por Renna Junior, Silva (2018) – o que pode ser justificado pelos diversos critérios de exclusão utilizados por esses autores em seu artigo, diferentemente deste estudo.

6.3.2 Análises por Associação

Foram executados no decorrer do estudo dezenas de modelos diferentes, buscando alguma associação interessante nos atributos da base. Ao final, somente a Faixa de deslocamento mostrou-se como opção relevante de atributo de saída, quando associada isoladamente a dois atributos de entrada: Estadiamento Grupo e Tipo de Tratamento, cujos resultados serão descritos detalhadamente nas seções seguintes.

As combinações de atributos testados e descartados foram:

- Dias até o tratamento: foi usada a variável numérica como atributo de saída, deixando o próprio algoritmo de Associação fazer uma categorização dinâmica. Porém, os grupos gerados não fizeram sentido (por exemplo, um intervalo específico de 197 a 1574 dias);

- Distância: assim como no item anterior, o agrupamento dinâmico desta variável numérica não gerou resultados com algum significado prático;

- Faixas de tempo para tratamento: mesmo usando os grupos criados previamente pela técnica de *Clustering*, não foram geradas regras relevantes. Foram feitas tentativas de usá-lo como variável de entrada e de saída, sem sucesso;

- Campos relacionados ao paciente: foram tentadas associações com os atributos de raça, faixa etária, grau de instrução, estado conjugal, uso de tabaco e de álcool (como entrada) e a distância ou tempo de tratamento (como saída). Alguns dos campos não são de preenchimento obrigatório no RHC, por isso há uma quantidade de registros em branco muito grande, o que impediu uma análise eficaz;

- Campos relacionados ao tumor: foram executadas associações também com histologia, lateralidade, ocorrência de mais de um tumor, histórico familiar, localização primária, estadiamento (como entrada) e distância ou tempo de tratamento (como saída), todos juntos. Verificou-se que regras com muitas variáveis não se mostravam efetivas também pelo mau preenchimento da base (muitos registros com o valor “Sem Informação”);

- Campos de tipo de tratamento: os campos lógicos (verdadeiro ou falso) referentes aos tipos de tratamento possíveis também foram usados, sem sucesso. Foi então revisada a forma de processamento e criado um subconjunto de dados desmembrando os tipos de tratamento linha a linha (conforme explicado na metodologia).

6.3.2.1 Associações entre Estadiamento e Faixa de Deslocamento

Foram gerados oito modelos para a combinação Estadiamento e Faixas de Deslocamento. Em todos eles, dois tipos de filtros foram configurados: um geral, para o atributo de entrada Estadiamento Grupo, removendo os registros que estivessem preenchidos com os valores “sem informação” ou “não se aplica”; e um outro individual, referente à localização (envolvendo as regiões ou os estados residenciais).

Durante as primeiras execuções dos modelos, o parâmetro *minimum_probability* estava com o valor padrão do *software* (0,4). Entretanto, as regras geradas não trouxeram associações interessantes e o valor foi reduzido para 0,3, com o objetivo de encontrar mais regras. Ainda assim, poucas regras surgiram, o que levou a uma nova tentativa de redução do parâmetro. Com o novo valor de 0,2, foram então encontradas regras de associação relevantes e, deste modo, o valor do parâmetro mínimo da probabilidade foi definido como 0,2 para todos os modelos.

O quadro 4 apresenta um resumo das principais características de cada modelo para o estadiamento. Em cada coluna, foram exibidas somente duas regras. A regra com maior probabilidade (RMP) é aquela cujo percentual encontrado para esse parâmetro foi o mais alto dentre todas as regras. A regra de maior importância (RMI) é aquela com o maior percentual no parâmetro importância. Nos casos em que essas duas regras são diferentes, foram informados os índices para comparação. Também foi exibida a descrição da principal regra gerada para cada um dos modelos, no formato “entrada => saída”.

É possível ver que, em quase todos os modelos, a RMP encontrada acabou sendo também a RMI, com exceção das regiões Nordeste e Sul. Outro fato evidenciado durante as análises é que a RMP encontrada para o Brasil reflete a regra da região Sudeste, que por sua vez reflete a do estado de São Paulo. Isso ocorre pelo grande volume populacional do estado e da região, que acaba influenciando no cálculo da probabilidade.

Além disso, somente as regiões Centro-Oeste e Norte apresentaram uma RMP cuja faixa de deslocamento associada não era a menor (até 60 km). Ambas associaram o estadiamento “outro” com as maiores distâncias de locomoção (a faixa 4, acima de 300 km). Outras similaridades encontradas nessas regiões foram que a segunda regra mais importante associa o estadiamento 0 (inicial) com a faixa 4 (dos trajetos acima de 300 km), assim como a terceira regra associa o estádio 4 (mais avançado) também com essa mesma faixa de distâncias mais longas. Ou seja, tanto em estádios iniciais quanto em estádios desconhecidos e avançados

há uma associação com a faixa de deslocamentos acima de 300 km. Por serem as regiões mais extensas do país, é um resultado esperado quando a análise é feita isoladamente.

Quadro 4 - Resumo dos modelos gerados para a associação Estadiamento e Deslocamento

Modelo:		*BR*	*CO*	*N*	*NE*	*S*	*SE*	*SP*	*MG*
Filtro local		-	R. Centro-Oeste	Região Norte	Região Nordeste	Região Sul	Região Sudeste	Estado de SP	Estado de MG
Filtro Estad.		≠ 'Sem Informação' e ≠ 'Não se Aplica'							
RMP	**Prob.:**	76,9 %	81,6%	87,5%	65,7%	74,4%	80,8%	85,6%	62,5%
	Imp.:	4,4 %	41,1%	50,0%	4,6%	3,3%	3,4%	2,9%	4,5%
	Regra	Outro =>Fx1(1)	Outro =>Fx4(2)	Outro =>Fx4	Est 0 =>Fx1	Est 0 =>Fx1	Outro => Fx1	Outro => Fx1	Est 1 => Fx1
RMI	**Prob.:**	mesma regra acima	mesma regra acima	mesma regra acima	24,1%	26,4%	mesma regra acima	mesma regra acima	mesma regra acima
	Imp.:				20,6%	5,2%			
	Regra				Outro => Fx4	Est3=> Fx2 (3)			

(1) Faixa 1 = Distâncias até 60 km

(2) Faixa 4 = Distâncias maiores que 300 km

(3) Faixa 2 = Distâncias de 61 a 200 km

Ao examinar modelo a modelo no *software*, é possível fazer uma análise gráfica de forma bastante dinâmica. A tela *Dependency Network* (ou rede de dependências) exibe todos os vínculos encontrados entre os valores dos atributos e permite que se selecione o nível de importância dessas ligações. À medida que se aumenta a força desejada entre os elos, somente as setas indicando as ligações mais robustas permanecem.

Para entender melhor este comportamento, foram feitas algumas capturas de tela de dois modelos que apresentaram resultados interessantes e semelhantes entre si: a região Sul e o estado de Minas Gerais. As imagens podem ser vistas na Figura 29 (que mostra a lista descritiva total de regras, ordenadas por probabilidade) e na Figura 30 (que mostra uma representação gráfica da rede de dependências correspondente, com as quatro regras mais fortes selecionadas).

Figura 29 - Comparativo entre as regras de associação geradas para o estado de MG e a região Sul.

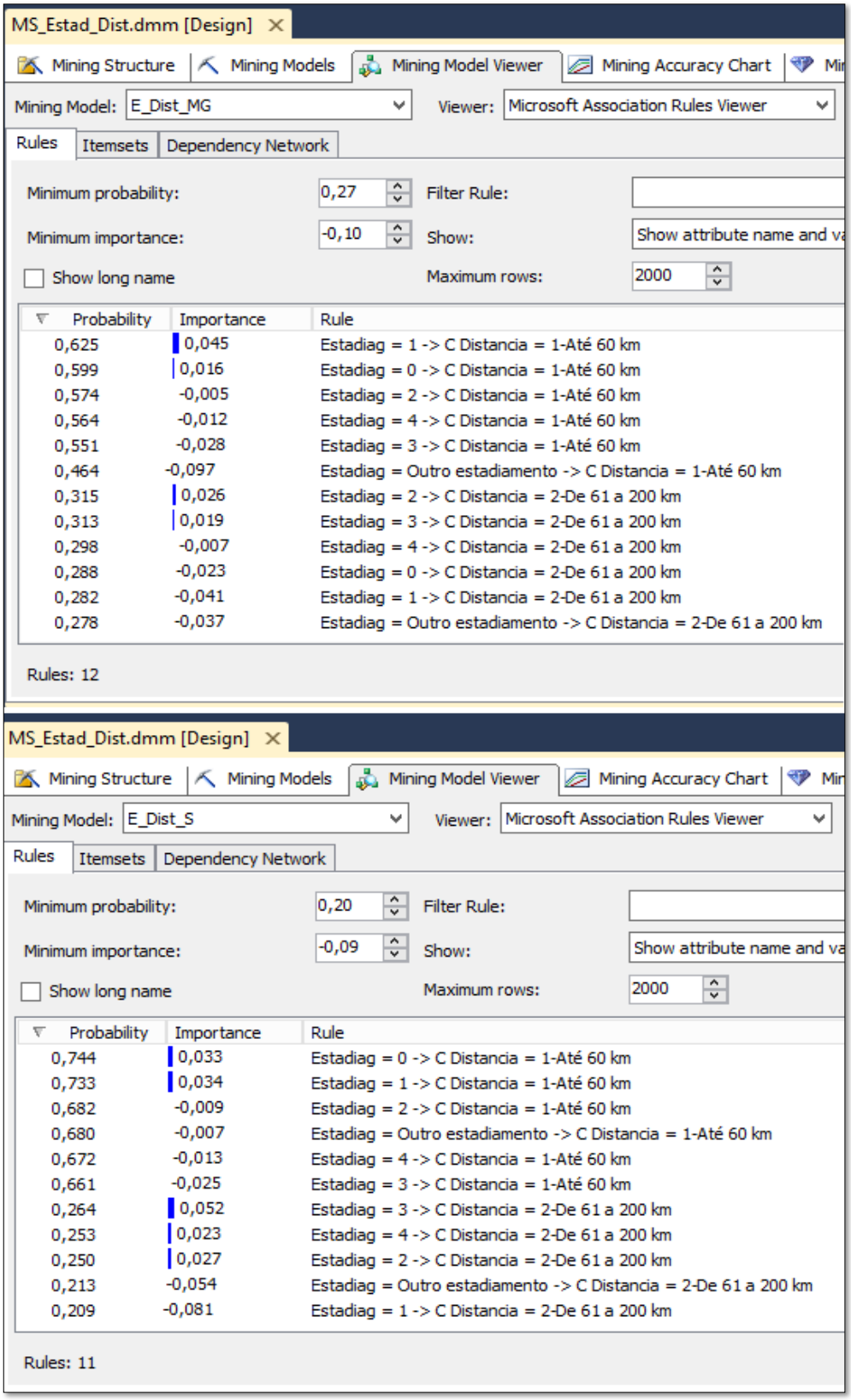
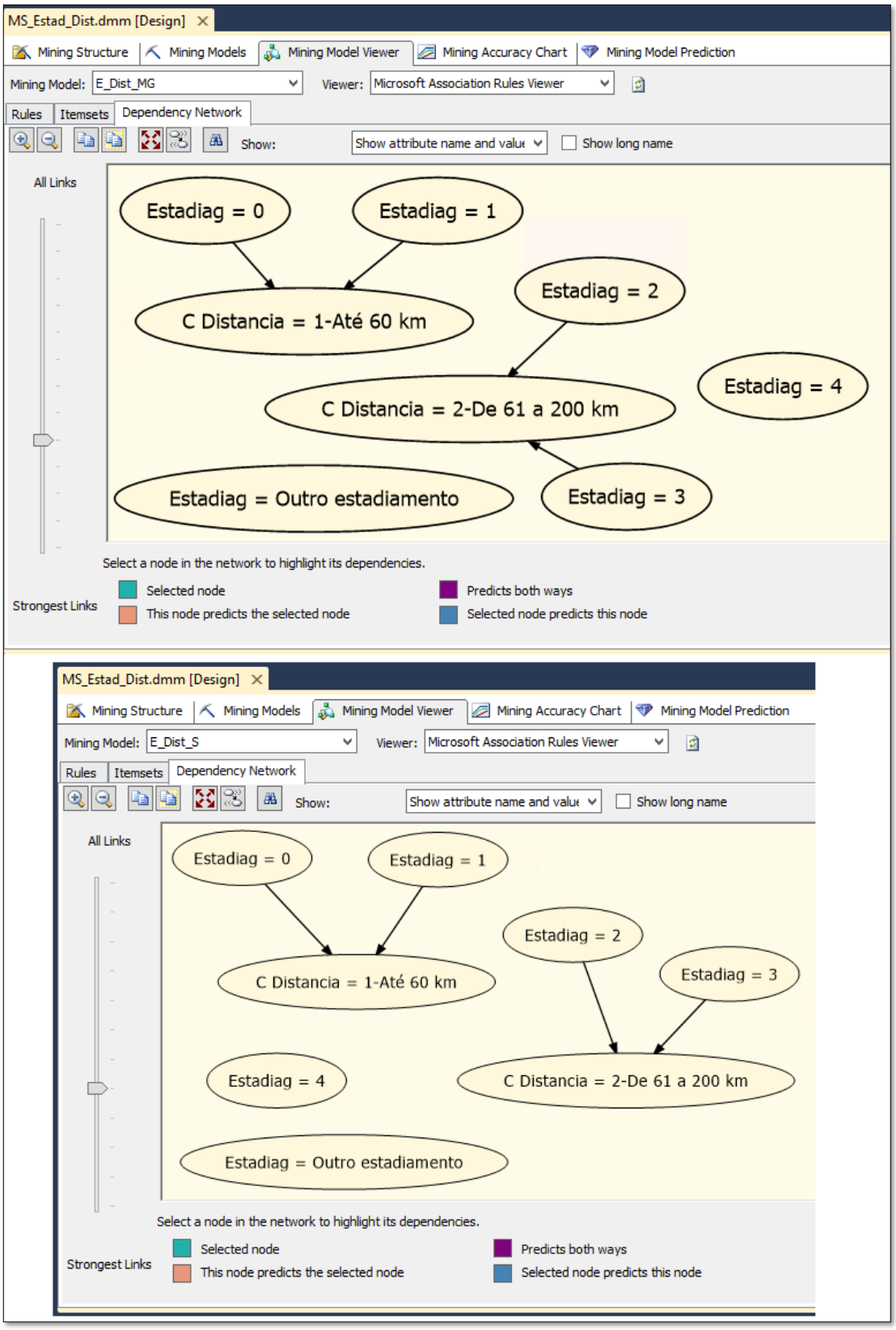


Figura 30- Comparativo entre as redes de dependência do estado de MG e da região Sul



Além de associar estadiamentos baixos (0 e 1) com deslocamentos na faixa inicial, os dois modelos indicaram que para estadiamentos mais altos (02 e 03) as distâncias percorridas foram um pouco maiores, tanto na região Sul quanto no estado de Minas Gerais.

Ao fazer uma pesquisa no site do DATASUS, foi verificado que, em dezembro de 2016, o Brasil dispunha de 414 estabelecimentos habilitados relacionados a oncologia[11]. Desse total, a região Sul tinha 95 habilitações, enquanto no estado de Minas Gerais a quantidade encontrada foi 50. Nos dois locais, pouco mais de 20% das habilitações estavam concentradas nas respectivas capitais. Não foi possível identificar, porém, uma relação entre a localização do restante dos centros de atendimento e as regras descritas no parágrafo anterior.

6.3.2.2 Associações entre Tipo de Tratamento e Faixa de Deslocamento

De forma análoga às análises da seção prévia, foram gerados oito modelos para a combinação Tipo de Tratamento e Faixas de Deslocamento. Após algumas iterações, foi definido o mesmo valor da seção anterior no parâmetro *minimum_probability* (0,2) e os mesmos filtros individuais referentes à região e estado residencial. A única diferença foi no filtro geral para o atributo de entrada Tipo de Tratamento, pois neste caso os registros removidos eram os que estivessem preenchidos com os valores "sem informação" ou "outro".

Assim como na análise da seção anterior, foi feito um resumo das principais características identificadas nos modelos do Tipo de Tratamento, que pode ser visto no Quadro 5. As duas principais regras (de maior probabilidade e de maior importância) foram exibidas, seguindo a mesma lógica e formato previamente mencionados. Também foi exibida a descrição da principal regra gerada para cada um dos modelos, no formato "entrada => saída".

Novamente, foi detectado que a RMP gerada para o estado de São Paulo acabava sendo refletida na região Sudeste e no Brasil: nos três modelos, o tipo de tratamento hormonioterapia foi associado à faixa inicial de deslocamento (até 60 km). Essa mesma regra foi identificada nas regiões Centro-Oeste, Norte e Sul, mas não para a região Nordeste e o estado de Minas Gerais. Para o Nordeste, a RMP gerada foi o tratamento por imunoterapia associado à faixa inicial (até 60 km), enquanto em Minas Gerais foi a cirurgia vinculada a essa mesma faixa inicial.

[11] DATASUS (tabnet), seção CNES-Estabelecimentos da Rede Assistencial, opção "Habilitação", filtro pelas habilitações iniciadas por 17. Acesso em: 10 maio 2020. Disponível em: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/habbr.def

Quadro 5 - Resumo dos modelos gerados para a associação Tipo de Tratamento e Deslocamento

Modelo:		***BR***	***CO***	***N***	***NE***	***S***	***SE***	***SP***	***MG***
Filtro local		-	R. Centro-Oeste	Região Norte	Região Nordeste	Região Sul	Região Sudeste	Estado de SP	Estado de MG
Filtro Tipo Tratamento		≠ 'Sem Informação' e ≠ 'Outro'							
RMP	**Prob.:**	73,0%	58,3%	63%	71,9%	72,2%	76,8%	80,6%	60,0%
	Imp.:	2,4%	2,5%	3,7%	7,3%	3,2%	1,1%	0,7%	2,1%
	Regra	Hor (*) =>Fx1(1)	Hor => Fx1	Hor => Fx1	Imu => Fx1	Hor => Fx1	Hor => Fx1	Hor => Fx1	Cir =>Fx1
RMI	**Prob.:**	mesma regra acima	42,3%	32,3%	mesma regra acima	mesma regra acima	21,0%	mesma regra acima	31,8%
	Imp.:		19,8%	13,0%			7,7%		3,3%
	Regra		Nen => Fx4 (2)	Nen =>Fx4			Nen =>Fx2(3)		Rad =>Fx2

(*) Tipos de tratamento indicados pelas três letras iniciais: Hormônio, Imunoterapia, Cirurgia, Radioterapia, Nenhum.

(1) Faixa 1 = Distâncias até 60 km

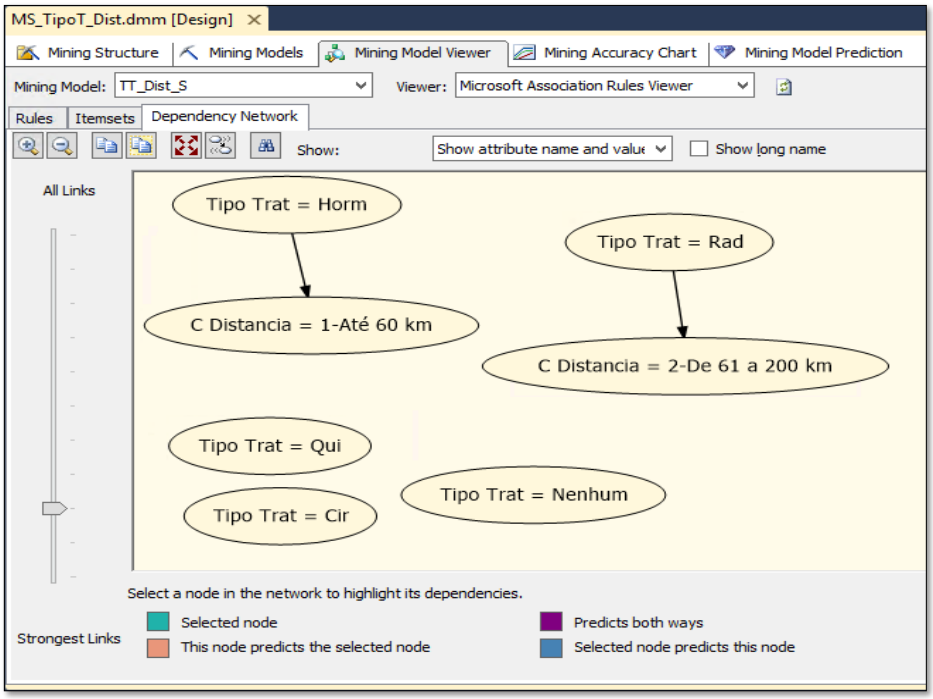
(2) Faixa 4 = Distâncias maiores que 300 km

(3) Faixa 2 = Distâncias de 61 a 200 km

Em quatro dos oito modelos a RMP foi igual a RMI. Nas regiões Norte e Centro-Oeste, a associação mais importante foi outra: o tipo de tratamento "nenhum" indicava um deslocamento na última faixa. Em outras palavras: nessas regiões há uma correlação entre as pacientes que não fizeram nenhum tratamento e deslocamentos maiores de 300 quilômetros. Na região Sudeste, a RMI associou a ausência de tratamento com a segunda faixa (de 60 a 200 km), enquanto em Minas Gerais essa mesma faixa foi associada à radioterapia.

Na região Sul, ao se aprofundar na análise somente das regras mais fortes, é possível ver que a radioterapia estava associada a deslocamentos da segunda faixa, enquanto a hormonioterapia indicava distâncias mais curtas. Isto pode ser visto na Figura 31, que mostra a rede de dependências desse modelo.

Figura 31 - Tela da rede de dependências na análise por associação da região Sul

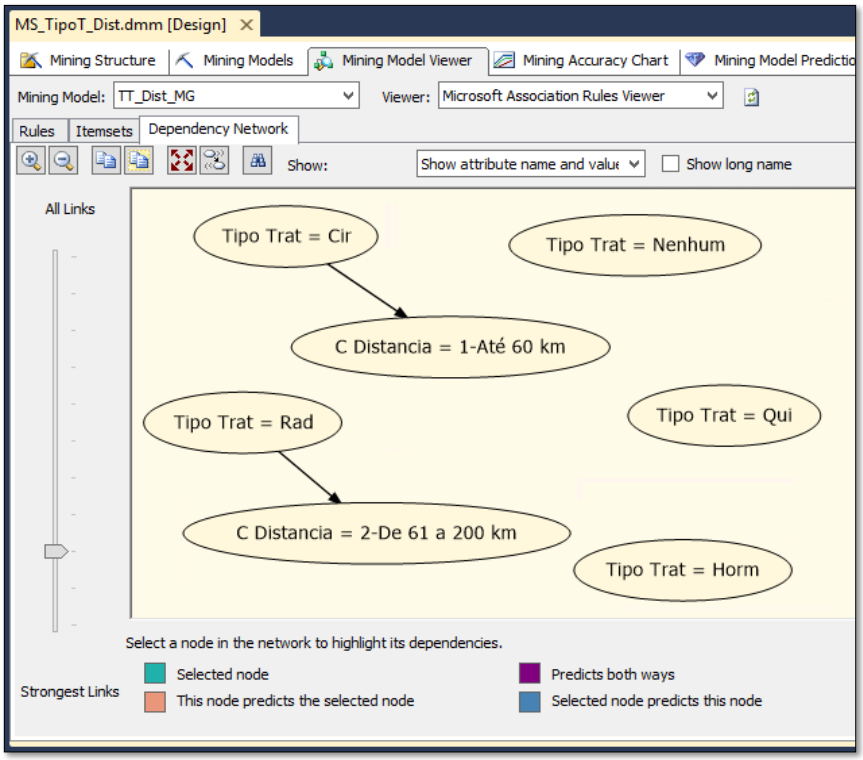


O estado de Minas Gerais novamente trouxe resultados diferentes dos demais modelos. Neste local, cirurgia associada a trajetos curtos é a RMP, enquanto nas outras regiões essa mesma regra aparece em segunda, terceira ou até quarta posição. A associação entre radioterapia e a segunda faixa de deslocamento foi a RMI neste modelo, mas só foi encontrada (na segunda posição) na região Sul.

As duas principais regras desse estado podem ser vistas na Figura 32, que mostra a rede de dependências encontrada. Aqui novamente a análise de um especialista em saúde pública poderia ser útil, ao avaliar as diversas variáveis envolvidas (políticas públicas adotadas localmente, condições específicas do estado, concentração de centros especialistas em regiões determinadas por outros fatores que não a quantidade de casos etc.).

Tanto para a região Sul quanto para Minas, mesmo com o baixo parâmetro de probabilidade (0,2) não houve nenhuma regra produzida para a quarta faixa de deslocamento (acima de 300 km). Esta ausência pode indicar que os registros nessa situação estão numa quantidade inferior à que o sistema precisaria para achar associações adequadas.

Figura 32 - Tela da rede de dependências na análise por associação de Minas Gerais



Uma outra característica comum a todos os dezesseis modelos de associação (os oito desta seção e os oito da anterior) foi que a terceira faixa de deslocamento (que vai de 200 a 300 km) não apareceu em nenhuma regra. Isso pode ser resultado da baixa quantidade de ocorrências nessa faixa (cerca de 5% do total) em comparação com os outros três intervalos.

7 CONCLUSÃO

A partir de um subconjunto de dados do RHC, este estudo desenvolveu modelos analíticos que descrevem os deslocamentos de pacientes em tratamento de câncer de mama, para o período de 2000 a 2016. Os modelos foram criados usando técnicas estatísticas e técnicas de mineração de dados descritivas.

As análises estatísticas foram agrupadas conforme características relacionadas ao tumor, tipos de tratamento, pacientes e seus deslocamentos. Esses quatro grupos de características foram então organizados num painel de navegação interativo, disponibilizado publicamente, que possibilita uma ampla gama de análises a especialistas e gestores em saúde pública.

Assim, a primeira grande contribuição deste estudo foi traçar o perfil de deslocamento das pacientes com câncer de mama no país, possibilitando também análises separadas para todas as esferas de saúde do SUS. O painel de análises estatísticas descritivas desenvolvido para o deslocamento permitiu que fossem avaliadas características como distância média, percentuais por faixa de deslocamento e percentuais de tratamentos dentro/fora não apenas para o país como um todo, mas também para cada uma das unidades da Federação, suas macrorregiões, regiões de saúde, até o município. Dentre os resultados encontrados pode-se destacar a identificação de uma alta concentração (69,5%) de tratamentos na faixa de distâncias curtas (até 60 km) e um percentual próximo à metade dos tratamentos (47,1%) ocorrendo no mesmo município.

Outra significativa contribuição deste estudo foi analisar a correlação entre certas características do deslocamento de pacientes, através de técnicas de mineração de dados como Agrupamento e Associação.

Foram avaliadas as correlações entre as faixas de deslocamento das pacientes com o estadiamento grupo ao diagnóstico e com os tipos de tratamento adotados. Graças a esses métodos, identificou-se quais são as associações de maior ou menor probabilidade entre essas variáveis, em todos os locais selecionados, auxiliando na avaliação do impacto da distância percorrida pelas pacientes.

Identificou-se que o estadiamento grupo do tumor está associado, nas regiões Centro-Oeste e Norte do país, a deslocamentos mais longos em metade (03) dos seis estádios avaliados. Já para a região Sul e o estado de Minas Gerais, encontrou-se o mesmo resultado: as duas principais associações foram entre estadiamentos iniciais (0 e 1) com distâncias curtas (até 60km) e entre graus mais avançados de estádio (2 e 3) com distâncias médias (de 61 a 200km).

Em relação ao tipo de tratamento adotado, houve uma correlação nas regiões Norte e Centro-Oeste entre não fazer nenhum tratamento e faixas de deslocamento maiores que 300km. Para Minas Gerais, foi identificada uma associação entre cirurgias oncológicas e deslocamentos mais curtos (até 60km), enquanto a radioterapia foi associada a deslocamentos entre 61 e 200km. Na região Sul, o resultado indicou uma correlação entre hormonioterapia e deslocamentos até 60 km, enquanto radioterapia foi novamente associada a distâncias na faixa de 61 a 200 km.

Reconhecendo a magnitude do problema do câncer e, especificamente, do tipo mais incidente – câncer de mama – em todo o ciclo do diagnóstico ao tratamento, tais resultados podem ser analisados por secretarias municipais de saúde e demais órgãos públicos, para efeitos de acompanhamento e organização de suas políticas, vislumbrando a possibilidade de atendimento mais próximo para tratamento de câncer de mama e a redefinição de unidades hospitalares na rede de atenção.

Sugere-se como trabalho futuro a análise descritiva do deslocamento de pacientes para outros tipos de câncer, a partir da base do RHC, usando tanto técnicas estatísticas quanto de mineração de dados. Também é proposto que seja feita a análise do deslocamento de pacientes em tratamento de câncer usando os dados dos sistemas de informações ambulatoriais do DATASUS, o que permitiria uma ótica diferente da gerada pelos registros hospitalares.

REFERÊNCIAS

AMBROGGI, M., *et al*. Distance as a barrier to cancer diagnosis and treatment: review of the literature. **The oncologist**, v. 20, n. 12, p. 1378, 2015.

BRASIL. Ministério da Saúde. Portaria nº 3535, de 02 de setembro de 1998. Estabelece critérios para cadastramento de centros de atendimento em oncologia. **Diário oficial da União**, 1998

BRASIL. Ministério da Saúde. Regionalização da assistência à saúde: aprofundando a descentralização com equidade no acesso: **Norma Operacional da Assistência à Saúde**: NOAS-SUS 01/02 e Portaria MS/GM nº 373, de 27 de fevereiro de 2002 e regulamentação complementar. 2. ed. Brasília, DF, 2002.

BRASIL. Ministério da Saúde. Portaria nº 741, de 19 de dezembro de 2005. Define as unidades de assistência de alta complexidade em oncologia, os centros de assistência de alta complexidade em oncologia (CACON) e os centros de referência de alta complexidade em oncologia e suas aptidões e qualidades. **Diário Oficial da União**, 2005.

BRASIL. Ministério da Saúde. Portaria nº 874, de 16 de maio de 2013. Institui a Política Nacional para a prevenção e controle do câncer na rede de atenção à saúde das pessoas com doenças crônicas no âmbito do Sistema Único de Saúde (SUS). **Diário oficial da União**, 2013.

BRASIL. Ministério da Saúde. Portaria nº 1399, de 17 de dezembro de 2019. Redefine os critérios e parâmetros referenciais para a habilitação de estabelecimentos de saúde na alta complexidade em oncologia no âmbito do SUS. **Diário oficial da União**, 2019.

BRAY, F. *et al*. **Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries**. CA: a cancer journal for clinicians, Hoboken, v. 68, n. 6, p. 394-424, Nov. 2018.

CAMILO, C. O.; SILVA, J. C. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas.** Universidade Federal de Goiás (UFG), p. 1-29, 2009.

CARVALHO, G. A saúde pública no Brasil. **Estudos avançados**, v. 27, n. 78, p. 7-26, 2013.

CELAYA, M. O. *et al*. Breast cancer stage at diagnosis and geographic access to mammography screening (New Hampshire, 1998–2004). **Rural and remote health**, v. 10, n. 2, p. 1361, 2010.

CRUZ, J. A.; WISHART, D. S. Applications of machine learning in cancer prediction and prognosis. **Cancer informatics**, v. 2, p. 117693510600200030, 2006.

DOS MINISTÉRIOS, Esplanada. Integração de informações dos registros de câncer brasileiros. **Rev Saude Publica**, v. 41, n. 5, p. 865-68, 2007.

DUARTE, L. S. *et al*. Regionalização da saúde no Brasil: uma perspectiva de análise. **Saúde e Sociedade**, v. 24, p. 472-485, 2015.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-37, 1996.

GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. **ACM SIGKDD explorations newsletter**, v. 1, n. 1, p. 20-33, 1999.

GUIMARÃES, R. B. Regiões de saúde e escalas geográficas. **Cadernos de Saúde Pública**, v. 21, n. 4, p. 1017-1025, 2005.

HENRY, K. A. *et al*. Breast cancer stage at diagnosis: is travel time important? **Journal of community health**, v. 36, n. 6, p. 933, 2011.

HUANG, B. *et al*. Does distance matter? Distance to mammography facilities and stage at diagnosis of breast cancer in Kentucky. **The Journal of Rural Health**, v. 25, n. 4, p. 366-371, 2009.

Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Registros hospitalares de câncer: planejamento e gestão Instituto Nacional de Câncer. 2 ed. – Rio de Janeiro: INCA, 2010.

Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). IntegradorRHC: **Ferramenta para a Vigilância Hospitalar de Câncer no Brasil.** Rio de Janeiro: INCA, 2011.

Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). **Informação dos registros hospitalares de câncer como estratégia de transformação:** perfil do Instituto Nacional de Câncer José Alencar Gomes da Silva em 25 anos. Rio de Janeiro: INCA, 2012.

Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Estimativa 2020: **incidência de câncer no Brasil**. – Rio de Janeiro: INCA, 2019.

JONES, A. P. *et al*. Travel times to health care and survival from cancers in Northern England. **European journal of cancer**, v. 44, n. 2, p. 269-274, 2008.

MAIMON, O; ROKACH, L. Introduction to knowledge discovery and data mining. In: **Data mining and knowledge discovery handbook**. Springer, Boston, MA, 2009. p. 1-15.

MELO, N. A assistência em oncologia no SUS: onde tratar? **Observatório de Oncologia**, 1 set. 2017. Acessado em 31 de maio de 2020. Disponível em: https://observatoriodeoncologia.com.br/a-assistencia-em-oncologia-no-sus-onde-tratar/

OLIVEIRA A., M.; MUNIZ M., S. C. A GEOGRAFIA DO CÂNCER DE MAMA NO NORTE DE MINAS GERAIS: DO DIREITO AO ACESSO À SAÚDE. **Hygeia - Revista Brasileira de Geografia Médica e da Saúde**, v. 13, n. 26, p. 13 - 32, 7 dez. 2017.

OLIVEIRA, E.; *et. al*. Acesso à assistência oncológica: mapeamento dos fluxos origem-destino das internações e dos atendimentos ambulatoriais. O caso do câncer de mama. **Cad Saúde Pública** 2011; 27:317-26.

PAYNE, S.; JARRETT, N.; JEFFS, D. The impact of travel on cancer patients' experiences of treatment: a literature review. **European journal of cancer care**, v. 9, n. 4, p. 197-203, 2000.

PINTO, I. V. *et al*. Completude e consistência dos dados dos registros hospitalares de câncer no Brasil. **Cad Saúde Colet (Rio J.)**, v. 20, p. 113-20, 2012.

RENNA JUNIOR, N. L.; SILVA, G. A. Diagnóstico de câncer de mama em estado avançado no Brasil: análise de dados dos registros hospitalares de câncer (2000-2012). **Revista Brasileira de Ginecologia e Obstetrícia**, v. 40, n. 3, p. 127-136, 2018.

SALDANHA, RF *et al*. Estudo de análise de rede do fluxo de pacientes de câncer de mama no Brasil entre 2014 e 2016. **Cad. Saúde Públi**ca, Rio de Janeiro, v. 35, n. 7, e00090918, 2019. Epub July 22, 2019. https://doi.org/10.1590/0102-311x00090918.

SANTOS, R. S. *et al*. A data mining system for providing analytical information on brain tumors to public health decision makers. **Computer methods and programs in biomedicine**, v. 109, n. 3, p. 269-282, 2013.

SANTOS, L. Região de saúde e suas redes de atenção: modelo organizativo-sistêmico do SUS. **Ciência & Saúde Coletiva**, v. 22, n. 4, p. 1281-1289, 2017.

SCOGGINS, J.F. *et al*. Is distance to provider a barrier to care for medicaid patients with breast, colorectal, or lung cancer? **The Journal of Rural Health**, v. 28, n. 1, p. 54-62, 2012.

SILVA, M.; MELO, E.; OSORIO-DE-CASTRO, C. Fluxos origem-destino para quimioterapia para o câncer de mama no Brasil: implicações para a assistência farmacêutica. **Ciênc. saúde coletiva**, Rio de Janeiro, v. 24, n. 3, p. 1153-1164, Mar. 2019.

VETTERLEIN, M. W. *et al*. Impact of travel distance to the treatment facility on overall mortality in US patients with prostate cancer. **Cancer**, v. 123, n. 17, p. 3241-3252, 2017.