FLAVIO FERRAZ VIEIRA

O USO DE CIÊNCIA DE DADOS E INTELIGÊNCIA GEOGRÁFICA COMO METODOLOGIAS DE POLÍTICAS PÚBLICAS PARA O DIAGNÓSTICO PRECOCE DE TUMORES

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Gérson Rodrigues dos Santos

Coorientador: Luiz Alexandre Peternelli

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

V658u
2022

Vieira, Flávio Ferraz, 1994-

O uso de ciência de dados e inteligência geográfica como metodologias de políticas públicas para o diagnóstico precoce de tumores: estudo retrospectivo de casos de tumores do Hospital do Câncer de Muriaé / Flávio Ferraz Vieira. – Viçosa, MG, 2022.

1 dissertação eletrônica (60 f.): il. (algumas color.).

Orientador: Gerson Rodrigues dos Santos.

Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Estatística, 2022.

Referências bibliográficas: f. 57-60.

DOI: https://doi.org/10.47328/ufvbbt.2022.205

Modo de acesso: World Wide Web.

1. Estômago - Tumores - Métodos estatísticos. 2. Estômago - Tumores - Muriaé (MG). 3. Análise espacial (Estatística). 4. Aprendizado do computador. I. Santos, Gerson Rodrigues dos, 1974-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.54

Bibliotecário(a) responsável: Euzebio Luiz Pinto CRB 6/3317

FLAVIO FERRAZ VIEIRA

O USO DE CIÊNCIA DE DADOS E INTELIGÊNCIA GEOGRÁFICA COMO METODOLOGIAS DE POLÍTICAS PÚBLICAS PARA O DIAGNÓSTICO PRECOCE DE TUMORES

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 24 de fevereiro de 2022.

Assentimento:

Flavio Ferraz Vieira

Flavio Ferraz Vieira
Autor

Gérson Rodrigues dos Santos
Orientador

A Deus, meus pais Sebastião e Dirlene,
irmãos Diego, Luiza e Paulo,
minha companheira Camila,
e meu filho Cícero.

AGRADECIMENTOS

A Deus, sem ele nada seria possível.

Ao meu orientador Dr. Gérson Rodrigues dos Santos por compartilhar seus conhecimentos e experiências que enriqueceram a minha pesquisa e meu desenvolvimento.

Ao meu coorientador Dr. Luiz Alexandre Peternelli e os membros da banca a Dra. Adriana Maria Rocha Trancoso Santos e a Dra. Lidiane Maria Ferraz Rosa.

A Universidade Federal de Viçosa e o Programa de Pós Graduação em Estatística Aplicada e Biometria por me acolher como estudante.

Aos professores do departamento do PPESTBIO pelo o ensino das disciplinas que cursei.

E aos meus colegas de trabalho e gestora do Departamento de Pesquisa e Desenvolvimento da Fundação Cristiano Varella pelo apoio e tempo disponibilizado para meus estudos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

A dor é inevitável, o sofrimento é opcional. (Tim Hansel)

RESUMO

VIEIRA, Flavio, M.Sc., Universidade Federal de Viçosa, fevereiro de 2022. **O uso de Ciência de Dados e Inteligência Geográfica como Metodologias de Políticas Públicas para o Diagnóstico Precoce de Tumores**. Orientador: Gérson Rodrigues dos Santos. Coorientador: Luiz Alexandre Peternelli.

O Hospital do Câncer de Muriaé é um centro de oncologia que tem total pactuação com 83 municípios da Região Geográfica Intermediária de Juiz de Fora. Os tumores dos órgãos digestivos são os mais frequentes dos pacientes da instituição, sendo 19,16% dos casos entre 2010 e 2020. Os principais fatores de risco são o sobrepeso/obesidade e consumo exagerado de produtos com álcool e tabaco. O estado nutricional desses pacientes é obtido nos prontuários eletrônicos da instituição, porém essa informação possui um viés difícil de ser mensurado, pois tumores nos órgãos digestivos tem como consequência a rápida perda de peso. Com isso, este estudo teve como objetivo estimar o estado nutricional do paciente antes do surgimento do tumor utilizando o algoritmo *Random Forest*, e mapear as classes das variáveis que apresentam maiores proporções dos casos. A estimação pelo algoritmo obteve uma taxa de acurácia de 85,48% sendo considerado satisfatório. O perfil epidemiológico se manteve durante os 11 anos analisados no estudo (2010 - 2020), os perfis mais frequentes foram pacientes masculinos, com idade de 63 anos, casado, com ensino fundamental incompleto, não etilista, tabagista e acima do peso (de acordo com o Índice de Massa Corporal). O teste de Qui-Quadrado de Pearson foi utilizado para verificar a associação entre as variáveis epidemiológicas e letalidade em até 3 anos, o estadiamento e as variáveis de risco, também foi utilizado para verificar o estadiamento e a letalidade entre as variáveis de risco. O teste de associação apresentou significância para quase todas as variáveis, sendo os pacientes com baixa escolaridade com pior taxa. A identificação espacial dos fatores de risco, pacientes com baixa escolaridade e estadiamento avançado apresentou uma associação espacial para a maioria dos municípios do estudo, as regiões imediatas de Ubá e Viçosa apresentaram maior taxa de tabagistas e etilistas do que as outras imediatas, para os pacientes com baixa escolaridade esses foram em proporções maiores em cidades menos desenvolvidas e mais distantes dos municípios de referência das regiões imediatas.

Palavras-chave: ELSA. Entrograma. Aprendizado de Máquina. Hospital do Câncer de Muriaé.

ABSTRACT

VIEIRA, Flavio, M.Sc., Universidade Federal de Viçosa, February, 2021. **Data Science and Geographic Intelligence as methodologies of public policy and early diagnosis of tumor**. Advisor: Gérson Rodrigues dos Santos. Co-adviser: Luiz Alexandre Peternelli.

The Cancer Hospital of Muriaé is a highly complex oncology center that has a total agreement with 83 municipalities in the Intermediate Geographic region of Juiz de Fora. Tumors of digestive organs are the most frequent among patients at the institution, with 19,16% of the cases between 2010 - 2020, the main risk factors being overweight/obesity, excessive consumption of products with alcohol and tobacco. The nutritional status of these patients is obtained from the institution's electronic medical records, but this information has a bias that is difficult to measure, as tumors in digestive organs result in rapid weight loss. This study aimed to estimate the nutritional status of the patient before the tumor using the Machine Learning Random Forest algorithm, and to map the classes of variables that present the highest proportions of cases. The estimation by the algorithm obtained an accuracy rate of 85,48%, being considered satisfactory. The epidemiological profile was maintained during the 11 years analyzed in the study (2010 - 2020), the most frequent profile were male patients, aged 63 years, married, with incomplete elementary education, non-alcoholic, smoker and overweight (according to the Body Mass Index). Pearson's chi-square test was used to verify the existence of an association between epidemiological variables and lethality in up to 3 years, staging and risk variables, it was also used to verify staging and lethality between the variables of risk. The association test showed significance for almost all variables, with patients with low education having the worst rate. The spatial identification of risk factors, patients with low schooling and advanced staging showed a spatial association for most of the municipalities in the study, the immediate regions of Ubá and Viçosa had a higher rate of smokers and alcoholics than the other immediate regions, for patients with low education, these were in greater proportions in less developed cities and more distant from the reference cities in the immediate regions.

Keywords: ELSA. Entrogram. Machine Learning. Câncer Hospital of Muriaé.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Segundo a Organização Mundial da Saúde (OMS), o câncer é a segunda maior causa de morte em todo o mundo, sendo responsável por quase 10 milhões de mortes em 2020. Câncer é um termo genérico que abrange mais de 100 tipos de diferentes doenças malignas que podem começar em quase qualquer órgão ou tecido do corpo e tem em comum o crescimento desordenado de células. Essas células anormais crescem incontrolavelmente, vão além de seus limites habituais para invadir partes adjacentes do corpo e/ou se espalhar para outros órgãos (OMS, 2021).

A OMS destaca que entre 30% e 50% das mortes por câncer poderiam ser evitadas modificando ou evitando os principais fatores de risco e implementando estratégias de prevenção baseadas em evidências existentes. As consequências do câncer também podem ser reduzidas por meio da detecção precoce e de melhores manejos dos pacientes que desenvolvem a doença e o diagnóstico precoce também viabiliza em um menor custo monetário para o tratamento.

A prevenção também oferece a estratégia de longo prazo mais econômica para o controle do câncer. Os fatores de risco mais comuns nos tumores e que devem ser modificados ou evitados são: diminuir/extinguir o uso de tabacos, controle do peso adequado, controle adequado do uso do álcool, reduzir a exposição da radiação ultravioleta, evitar a poluição do ar urbano e a fumaça interna do uso doméstico de combustíveis sólidos. Boas práticas colaboram para a prevenção e o diagnóstico precoce, manter uma dieta balanceada, praticar exercícios físicos e obter cuidados médicos regulares. O conhecimento dos fatores de risco nos pacientes oncológicos é de grande valor para elaboração de estratégias públicas, visando municípios e grupos de pessoas específicos que apresentem alta taxa dos fatores.

O Instituto Nacional do Câncer (INCA) aponta que os tumores de esôfago, estômago, cólon e reto são uns dos mais incidentes na população brasileira, o câncer de cólon e reto é tratável e na maioria dos casos curável quando detectado precocemente; o câncer de esôfago tem seu principal fator de risco o excesso de gordura corporal. O tumor de estômago tem como principal fator de risco excesso de peso e obesidade, consumo de álcool, consumo excessivo de sal, alimentos salgados ou conservados no sal e tabagismo (INCA, 2021c).

O conhecimento das variáveis epidemiológicas dos pacientes, principalmente os que são fatores de risco é útil para encontrar padrões e tendências que auxiliam em estratégias de políticas públicas de combate ao câncer de órgãos digestivos. Algumas informações obtidos nos prontuários eletrônicos do paciente podem possuir viés, diversos tumores alteram a vida e o cotidiano dos pacientes, umas das consequências dos tumores dos órgãos digestivos no paciente é a rápida perda de peso, quando este vai para a consulta médica, está pesando bem menos do que era o habitual. Sa-

ber essas informações antes do surgimento do tumor é difícil de se obter, perguntar diretamente para o paciente pode trazer um viés difícil de se mensurar.

A Ciência de Dados tem se mostrado uma metodologia eficaz que consegue estimar variáveis em diversos tipos de dados, mesmo com banco de estruturas diferentes, por meio dos seus algoritmos de *Machine Learning* (ML), encontram padrões nos dados e a utilizam para estimação de novas entradas. Com isso, essa metodologia é útil no problema citado, podendo estimar o verdadeiro estado nutricional do paciente antes do surgimento do tumor. Para estimar essa informação, foi utilizado o algoritmo de *Random Forest* (RF). Breiman (2001) desenvolveu inicialmente as noções do algoritmo, que utiliza diversos algoritmos de *Decision Tree* (DT) para definir o valor estimado. DT é um algoritmo que pertence a família de algoritmos de ML, é um estimador que utiliza da recursividade para estimar a variável de interesse.

O algoritmo divide o conjunto de dados heterogêneos em subconjuntos homogêneos em relação a variável de interesse para a estimação, isto é, os subconjuntos são parecidos dentro de si e diferentes entre si. A recursão dos dados é finalizada quando uma subdivisão tem todos os mesmos valores da variável de estimação ou quando a divisão não agrega mais valor às estimações. A estimação final da RF é escolhida como a estimação mais frequente de todos os algoritmos de DT, o algoritmo apresenta resultados melhores que o da *Decision Tree* e sempre converge para que não sofra de sobreajuste. Sobreajuste é definido quando algum algoritmo ajusta muito bem ao conjunto de dados utilizado para treinar o modelo, porém o mesmo não aprende os padrões e comportamentos dele, e sim, 'decora' os dados, o que se mostra ineficaz para a estimação de novos dados (BREIMAN, 2001).

Com as variáveis epidemiológicas, encontrar as classes mais incidentes auxilia no planejamento de estratégias de políticas públicas na saúde e a divulgação dessa informação ajuda na conscientização da população. É importante verificar quais cidades possuem as maiores porcentagens dos fatores de riscos, permitindo políticas públicas exclusivas para cada município. HINO (2006) salienta que a distribuição espacial de qualquer doença é um importante instrumento na gestão em saúde, tanto para atividades de vigilância epidemiológica, quanto para o planejamento de ações de prevenção e controle. A Inteligência Geográfica é a metodologia que utiliza de outras metodologias para a avaliação espacial das variáveis. Neste contexto torna-se necessário tanto para a localização geográfica dos casos da doença quanto o perfil epidemiológico dos pacientes. Com os dados georreferenciados, é possível plotar gráficos poligonais de densidade para verificar tendências e padrões espaciais nos dados.

O Indicador Local de Associação Espacial Baseado em Entropia, do inglês *Entropy-based Local Indicator of Spatial Association* (ELSA), é um indicador desenvolvido por Naimi et al. (2019) que quantifica o grau de associação de uma variável observada em um local em relação com as observações dos seus vizinhos. ELSA utiliza o con-

ceito de entropia utilizado na Teoria de Informação, mais precisamente a entropia de Shannon. Naimi et al. (2019) padroniza essa entropia para obter uma estatística que varia entre 0 e 1, valores próximos de 0 indica alta similaridade dos valores de uma região com seus vizinhos, e valores próximos de 1 pouca similaridade. Com este indicador é encontrado padrões de uma região com seus vizinhos próximos ou todo conjunto de locais. É possível utilizá-lo com variáveis quantitativas e qualitativas e dados espacialmente contínuos ou poligonais, tornando uma excelente alternativa que os indicadores antecessores similares.

A partir dessas considerações, este estudo teve como objetivo utilizar a Ciência de Dados e a Inteligência Geográfica como metodologias de políticas públicas e de diagnóstico precoce de tumores, para estimar uma variável de fator de risco, encontrar associações da incidência total, do tumor em estado avançado e do óbito em até 3 anos da data de diagnóstico entre as variáveis epidemiológicas, mapear e identificar as associações espaciais nos municípios com total pactuação com o Hospital do Câncer de Muriaé.

Especificamente, nesse estudo realizou-se:

(i) A identificação dos pacientes com tumores nos órgãos digestivos no Registro Hospitalar de Câncer de 2010 a 2020;

(ii) A estimação da variável fator de risco Estado Nutricional que não é registrado no RHC;

(iii) A associação das características epidemiológicas com a incidência, o estadiamento avançado e o óbito em até 3 anos da data do diagnóstico;

(iv) O Mapeamento das características epidemiológicas e clínicas do câncer;

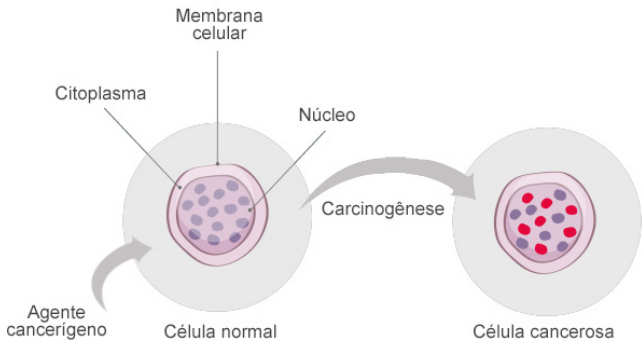
(v) A análise de padrões espaciais das incidência dos tumores e das variáveis epidemiológicas.

2 REFERENCIAL TEÓRICO

2.1 O CÂNCER

Câncer é um termo genérico que abrange mais de 100 tipos de diferentes doenças malignas, e tem em comum o crescimento desordenado de células. Os diversos tipos de células do nosso corpo denominam os tipos de câncer, se começados em tecidos epiteliais são chamados de carcinomas, começados em tecidos conjuntivos de sarcomas. Se as células cancerígenas multiplicam-se rapidamente e são capazes de invadir tecidos e órgãos vizinhos ou distantes, o câncer é conhecido como metástase. Nas Figuras 1 e 2 tem-se a representação do surgimento do câncer.

Figura 1: Representação do surgimento da célula cancerosa.



Fonte: INCA (2021a).

Figura 2: Representação do surgimento do tumor.

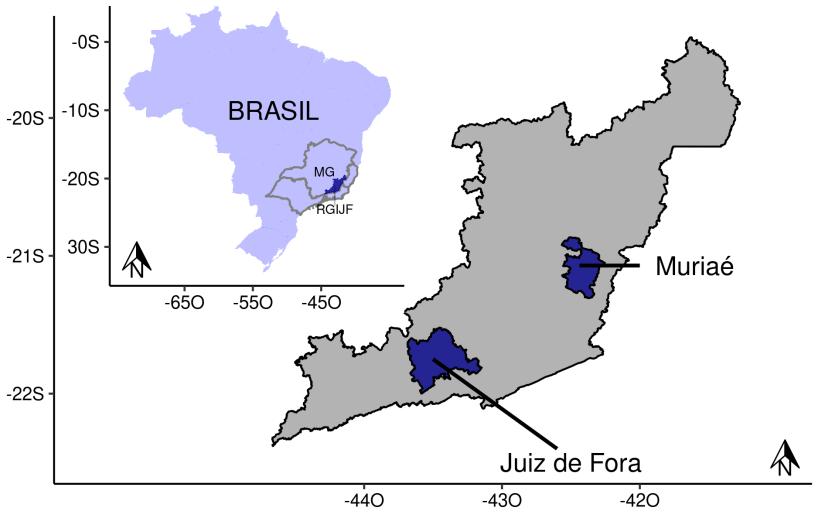
Fonte: INCA (2021a).

De acordo com o INCA, uma forma de classificar o câncer é o método definido pela União Internacional para o Controle do Câncer (UICC), o sistema TNM de Classificação dos Tumores Malignos. Esse sistema utiliza a extensão anatômica da doença, levando em conta as características do tumor primário (T), as características dos linfonodos das cadeias de drenagem linfática do órgão em que o tumor se localiza (N), e a presença ou ausência de metástases a distância (M). Estes parâmetros recebem graduações, geralmente de T0 a T4, de N0 a N3 e de M0 a M1, respectivamente. Quando as categorias T, N e M são agrupadas em combinações pré-estabelecidas, ficam distribuídas em estadiamento denotados por 0 e I a IV (INCA, 2021b). De forma generalizada, os grupos 0, I e II são considerados estadiamento em estado inicial e os grupos III e IV estadiamento em estado avançado.

Os hospitais oncológicos possuem uma unidade de Registro Hospitalar de Câncer (RHC) que coletam dados epidemiológicos e clínicos dos pacientes sendo esses auditados no mínimo duas vezes, internamente e externamente, tornando um banco de grande valor para melhor compreendimento (causas e desenvolvimento) do câncer, sendo útil para pesquisadores, profissionais de saúde e tomadores de decisão (por exemplo, formuladores de políticas pública) (LEUNG et al., 2020). O Hospital do Câncer de Muriaé (HCM) da Fundação Cristiano Varella (FCV) é um Centro de Alta

Complexidade em Oncologia (CACON) localizado às margens da Rodovia BR-116 em Muriaé, Minas Gerais (MG). O município fica a cerca de 312 km de distância de Belo Horizonte, capital do estado, e a cerca de 159 km de Juiz de Fora, a cidade referência da região geográfica intermediária. A Figura 3 apresenta a localização de Muriaé.

Figura 3: Localização do município de Muriaé-MG



Fonte: Elaborada pelo autor.

O hospital atende regularmente 158 municípios (com abrangência populacional de mais de 2 milhões de habitantes), com uma média de mais de 20.000 atendimentos ao mês, sendo 85% deles pelo Sistema Único de Saúde (SUS). Segundo os dados RHC - HCM (2010 - 2020), a maioria dos pacientes do hospital são pessoas idosas, com idade mediana de 64 anos (sendo 66 anos para homens e 60 anos para mulheres). Os locais mais comuns de tumor foram: órgãos digestivos (19,16%), órgãos genitais masculinos (16,90%), pele (15,54%), mama (12,82%) e órgãos genitais femininos (7,99%). Dos 158 municípios de abrangência do hospital, 89 possuem total pactuação com a instituição pelo SUS, os detalhes estão na Tabela 1.

Os tumores dos órgãos digestivos possuem codificação de C15-C26 pela Classificação Internacional de Doenças para Oncologia (CID-O) e possuem 12 tipos de localizações. Os nomes e a porcentagem entre os tumores dos órgãos digestivos são dados na Tabela 2. Desses tumores se destacam esôfago, estômago, cólon e reto que juntos representam aproximadamente 82% de todas as incidências dos tumores dos órgãos digestivos.

Os pacientes da instituição são avaliados por um nutricionista antes do início do primeiro tratamento no HCM e seu estado nutricional classificado e anotado nos prontuários eletrônicos. As classificações possíveis são desnutrição (valores do Índice de Massa Corporal (IMC) abaixo de 18,5) e esses sub classificados em leve, moderada e severa; eutrófico (IMC entre 18,6 e 24,9: peso adequado), sobrepeso (IMC entre 25,0 e

Tabela 1: Tabela da abrangência populacional do presente estudo incluindo as regiões geográficas imediata, os número de municípios e habitantes.

Região geográfica imediata	**Municípios**	**População**[1]
Além Paraíba	5	57.258
Carangola	9	112.052
Cataguases	10	182.689
Manhuaçu	24	349.553
Muriaé	12	184.701
Ubá	17	282.903
Viçosa	12	169.484
Total	**89**	**1.338.640**

Fonte: Elaborada pelo autor.

Tabela 2: Porcentagem dos pacientes com tumores nos órgãos digestivos do HCM e sua Classificação no CID-O.

CID-0	**Tumores**	**Porcentagem**
C15	Esôfago	22,73%
C16	Estômago	19,29%
C17	Intestino Delgado	01,04%
C18	Cólon	23,19%
C19	Junção Retossigmóide	04,00%
C20	Reto	15,70%
C21	Ânus e Canal Anal	01,60%
C22	Fígado e Vías Biliares Intrahepáticas	03,05%
C23	Vesícula Biliar	01,85%
C24	Outras partes das Vias Biliares	01,54%
C25	Pâncreas	05,87%
C26	Outros Órgãos Digestivos e Localizações Mal Definidas	00,14%
Total	**Órgãos Digestivos**	**100,00%**

Fonte: Elaborada pelo autor.

29,9) e obesidade (IMC acima de 30,0). Dos pacientes que passaram pelo nutricionista na primeira consulta 82,9% estavam em estado eutrófico ou desnutridos, veja a Tabela 3 para mais detalhes. Os tumores dos órgãos digestivos quando não diagnosticado precocemente, tem como um dos principais sintomas a rápida perda de massa corpórea, acredita-se que a média de perda de peso dos pacientes com o tumor avançado que não iniciaram nenhum tratamento é de 10 kg por mês. Este sintoma causa um viés na informação do estado nutricional coletado antes do primeiro tratamento, o que explica a alta taxa de pacientes desnutridos (44,9%) mesmo com sobrepeso e obesidade como fator de risco para tais tumores. Outros dois fatores de risco como já citado, o hábito de consumo de tabacos e bebidas alcoólicas são perguntados diretamente ao paciente nas primeiras consultas.

Tabela 3: Porcentagem do estado nutricional dos pacientes com tumores nos órgãos digestivos do HCM.

Estado Nutricional	Porcentagem
Desnutrição Leve	09,7%
Desnutrição Moderada	13,4%
Desnutrição Severa	21,8%
Eutrófico	38,0%
Sobrepeso	11,7%
Obesidade	05,6%
Total	**100,00%**

Fonte: Elaborada pelo autor.

2.2 CIÊNCIA DE DADOS

Os grandes avanços dos recursos computacionais dos últimos 50 anos trouxeram incontáveis benefícios à sociedade, desde otimizações de tarefas rotineiras, agendas digitais, comunicações remotas até tarefas mais complexas como levar a humanidade a lua. Um desses benefícios foi a digitalização de informações, que foram transferidos do meio físico (papel) para o digital. Os dados em meios digitais facilitaram varias etapas de uma pesquisa, a extração, a manipulação e as análises dos dados nunca foram tão fáceis como são nos dias atuais, resultando em um crescente volume de dados. Segundo Kelleher e Tierney (2018, p.9, tradução do autor) "Estima-se que a quantidade de dados coletados ao longo de cinco milênios desde a invenção da escrita até 2003 tem cerca de 5 exabytes. Desde 2013, os humanos geram e armazenam essa mesma quantidade de dados todos os dias". O grande volume de dados, de diversas estruturas e sem um padrão aparente trouxe um grande desafio à sociedade, as metodologias tradicionais não são eficientes para extrair informações úteis por causa da grande variabilidade de estruturas desses dados. Desse modo, uma metodologia capaz de extrair informações de um banco de dados aparentemente sem valor é fundamental para este contexto.

A chamada Ciência de Dados incorpora elementos variados e se baseia em técnicas e teorias oriundas de muitos campos básicos em engenharia e ciências básicas, sendo assim intimamente ligada com muitas das disciplinas tradicionais bem estabelecidas, porém viabilizando uma nova área altamente interdisciplinar (PORTO; ZIVIANI, 2014). As principais áreas disciplinares que compõem a ciência de dados são a Estatística, Ciência da Computação e a própria metodologia científica. Por meio dos algoritmos de ML, a Ciência de Dados conseguiu ter resultados positivos na extração de informações de dados complexos.

Diversos tipos de áreas podem se beneficiar aplicando a metodologia da Ciência de Dados para buscar *insight* para problemas do cotidiano ou em pesquisas científicas. Com a adoção de prontuários eletrônicos e de sistemas de informação de saúde,

a saúde pública tem lidado rotineiramente com enormes quantidades de dados, e a tendência é de ainda maior expansão no volume de dados num futuro próximo devido ao uso crescente de sensores remotos ou mesmo dispositivos móveis para coleta de dados individualizados em ambientes residenciais ou pré-hospitalares (ESTRIN, 2014). Dalianis et al. (2015) salientam que os dados produzidos no ambiente de saúde são muito valiosos para análises posteriores, desenvolvimento de processos de saúde aprimorados, políticas públicas e produção de mais conhecimento das doenças. As bases extraídos principalmente de prontuários eletrônicos que contém informações epidemiológicas dos pacientes, por exemplo sexo, idade, escolaridade, cidade de residência, estado nutricional, constitui em uma ótima ferramenta para erradicar a doença quando possível, ou elaborar estratégias para amenizar seus efeitos. Os prontuários eletrônicos ainda possuem informações clínicas sobre a doença, estadiamento, o tipo de tumor, lateralidade, óbito, que permite a obtenção do conhecimento da realidade da doença, o quão avançado está no momento da entrada no hospital e o tempo de sobrevida a partir do diagnóstico é um grande material de análise para verificar associações, padrões, tendências a fim de obter decisões para melhorar os indicadores de saúde. Diante da importância dos dados da saúde, Dalianis et al. (2015) mostram-se preocupados com a dificuldade que pesquisadores têm encontrado para obter tais tipos de dados, observando como a área da saúde se beneficiaria se a academia tivesse acesso a seus dados mais facilmente. Os autores propuseram um banco de dados gerais de saúde que respeitam as leis de proteção de dados e éticas (protegendo o anonimato dos pacientes) para facilitar a divulgação dos dados. Consoli, Reforgiato Recupero e Petkovic (2019) observaram que há uma incidência crescente de doenças evitáveis relacionadas ao estilo de vida causadas por fatores de risco como obesidade, tabagismo e consumo de álcool. Encontrar padrões e associações com outras variáveis epidemiológicas podem se mostrar excelentes ferramentas para políticas públicas.

Áreas relacionadas a saúde já se beneficiaram com as metodologias da Ciência de Dados, diagnóstico médico a partir de dados de imagem em medicina, quantificação de dados sobre estilo de vida na indústria de fitness, e isto é evidente que dentro dos *Big Datas* há um conhecimento oculto que pode mudar a vida do paciente ou em grande medida, mudar o mundo em si. Extrair esse conhecimento é o mais rápido, menos caro e mais eficaz caminho para melhorar a saúde das pessoas (CONSOLI; REFORGIATO RECUPERO; PETKOVIC, 2019). Porém, nem sempre os dados disponíveis são enormes e é comum na literatura encontrar autores que afirmam que a premissa para a utilização dos algoritmos de ML é obter uma quantidade enorme de dados, tanto de entradas quanto de variáveis. Kelleher e Tierney (2018) afirmam que tal premissa é um mito, ainda observam que atributos irrelevantes ou redundantes podem ter um efeito negativo no desempenho de muitos dos algoritmos usados para analisar os dados. Concluindo que muitos atributos aumentam a chance do al-

goritmo não encontrar padrões significativos nos dados, o mais importante é ter os dados certos e de qualidade. Leung et al. (2020) testou essa premissa em dois tipos de algoritmos em três conjuntos de dados sobre câncer de mama, concluindo que não foi preciso muitos dados para o conjunto de treinamentos para que a acurácia dos algoritmos obtivesse resultados satisfatórios. Em uma instituição hospitalar os dados são auditados no mínimo duas vezes, uma internamente e uma externamente, garantindo veracidade e qualidade dos dados sendo eles um terreno fértil para estudo.

Por fim, Kelleher e Tierney (2018, p.101, tradução do autor) apontam que "O verdadeiro desafio em usar o ML é encontrar o algoritmo cujo viés de aprendizagem é a melhor correspondência para um determinado conjunto de dados". Quanto mais complexo o algoritmo, mais recursos serão necessários, tornando esse desafio ainda maior, encontrar o melhor algoritmo para os dados que respeitem as limitações de cada pesquisa, i.e., recursos computacionais, financeiros, de dados. Iniciar as análises com algoritmos tradicionais que a maioria das configurações dos computadores atuais suportam os cálculos, criam parâmetros iniciais para buscar modelos mais complexos que melhorem os resultados, caso os próprios algoritmos tradicionais não obtenham resultados satisfatórios. O algoritmo de *Random Forest* é uns dos algoritmos mais tradicionais da Ciência de Dados, a simplicidade da sua teoria e a não exigência de grandes recursos computacionais permitiram o algoritmo chegar no status de tradicional. Breiman (2001) desenvolveu inicialmente as noções do *Random Forest* que a definiu como um classificador que consiste em uma coleção de árvores aleatórias, obtendo maior acurácia e mostrando por meio da Lei Forte dos Grandes Números que eles sempre convergem para que o sobreajuste não seja um problema. O que possibilita o algoritmo ser o primeiro para os testes iniciais na Ciência de Dados e o candidato para a solução do problema proposto.

MACHINE LEARNING

Machine Learning é um subcampo da Ciência de Dados que utiliza algoritmos para explorar um conjunto de dados, encontrar padrões e utilizá-lo para fazer predições, estimações ou classificações com novos dados. O conjunto inicial de exploração é definido como conjunto de treinamento. Este estudo utilizou o ML para estimação de uma variável. Em *Machine Learning* é necessário utilizar medidas de desempenho para avaliar os algoritmos e verificar se são adequados para resolver o problema do pesquisador. Ting (2011) aponta que para um problema de estimação, o mais usual é calcular a taxa de erro denotado por $TE(h)$, onde h é o estimador. Seja z o valor da variável de interesse do conjunto de treinamento, e $\hat{z}$ o valor estimado dado por h. Seja a função indicadora dado a seguir:

$$I(z_i, \hat{z_i}) = \begin{cases} 1, & \text{se } z_i = \hat{z_i} \\ 0, & \text{se } z_i \neq \hat{z_i} \end{cases} ;$$

a fórmula para o cálculo da taxa de erro do estimador h é apresentada na Equação 1.

$$TE(h) = \left\{ 1 - \frac{\sum_{i=1}^{n} I(z_i, \hat{z_i})}{n} \right\} x100\% \quad (1)$$

Consequentemente a taxa de acerto ou a acurácia do algoritmo é calculado por $100\% - TE(h)$, outra forma de medir a qualidade da estimação é analisar a Matriz de Confusão, descrito em (TING, 2011). Essa matriz apresenta os valores da variável do conjunto de treinamento e as estimações do teste, por exemplo, seja uma variável dicotômica que assume valores positivos ou negativos. Após a etapa de estimação com o algoritmo, a matriz de confusão de um exemplo hipotético é dado na Tabela 10. Ela contém o valor positivos e negativos reais da variável, e a estimação positiva (o algoritmo estima valor positivo), estimação negativa (o algoritmo estima valor negativo).

Tabela 4: Exemplo de uma Matriz de Confusão da Estimação de uma Variável Dicotômica.

	Estimação Positiva	Estimação Negativa
Positivo	400	20
Negativo	70	510

Fonte: Elaborada pelo autor.

O valor da matriz que está na linha 1 e coluna 1 indica que 400 entradas dos dados possuem real valor positivo e o algoritmo estimou que o mesmo também seja positivo, o valor da linha 1 e coluna 2 indica que 20 entradas positivas foram estimados erroneamente como negativa, e assim sucessivamente para a outra linha. Por meio da matriz, também é possível o cálculo da taxa de erro e de acerto, a taxa de acerto é a somatória dos valores da diagonal dividido pelo somatório de todos os valores da matriz, e a taxa de erro como 100% menos a taxa de acerto. Neste exemplo, a acurácia foi de 91% e o TE de 9%. A matriz de confusão estende para estimação de variáveis com n classes resultando em uma matriz nxn, apresentada na Tabela 5.

Para encontrar a Matriz de Confusão, o conjunto de treinamento foi dividido em dois subconjuntos, 75% dos dados para o subconjunto de treinamento e 25% para o de teste. O subconjunto de treinamento foi utilizado para treinar o algoritmo de ML, e este estimou os valores do conjunto teste. Com isso, foi possível verificar a acurácia e a matriz de confusão do algoritmo.

Tabela 5: Exemplo de uma Matriz de Confusão da Estimação de uma Variável com n Classes.

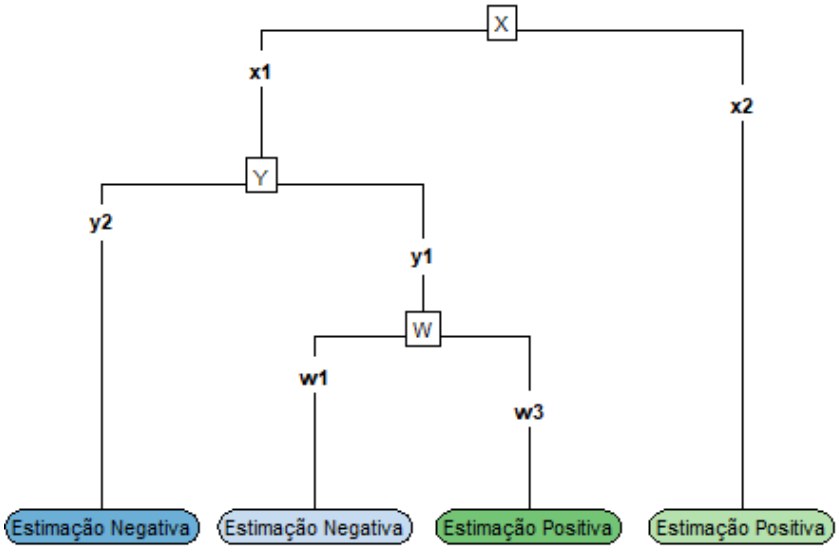
	Estimação Valor 1	Estimação Valor 2	...	Estimação Valor n
Valor 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,n}$
Valor 2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,n}$
...	...	...	...	
Valor n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,n}$

Fonte: Elaborada pelo autor.

DECISION TREE

Decision Trees (DT) pertencem a família de algoritmos de ML que utilizam da recursividade para estimar a variável de interesse. Dado um conjunto de treinamento, o algoritmo divide o conjunto de origem, definido como o nó raiz da árvore, em subconjuntos, os nós folhas. Este processo é repetido em cada nó filho, de uma maneira recursiva chamada particionamento recursivo . A recursão é finalizada quando um nó filho tem todos os mesmos valores da variável de estimação ou quando a divisão não agrega mais valor às estimações. Para o *Decision Tree* estimar novos dados, percorre-se a árvore a partir do nó raiz e desvia-se em cada nó folha até chegar a o último nó folha disponível, atribuindo à classe correspondente desse nó a os dados novos. A Figura 4 apresenta um exemplo de estrutura de uma *Decision Tree*.

Figura 4: Exemplo de uma Estrutura de *Decision Tree*.



Fonte: Elaborada pelo autor.

A escolha das variáveis para a divisão dos nós da árvore não é arbitrária, é necessário uma medida utilizada que mostre qual a variável é a melhor para a divisão, as medidas mais populares são a Impureza de Gini que foi utilizado no presente estudo

e a estatística Qui-Quadrado (χ^2). A medida de impureza de Gini decide a divisão ideal de um nó raiz e as divisões subsequentes, é o índice mais popular e fácil de dividir um *Decision Tree*, funciona apenas com alvos categóricos pois faz apenas divisões binárias. Seja a variável X de um conjunto de treinamentos, a variável de interesse para a estimação. A impureza de Gini (IG) de uma divisão D utilizando uma variável Y_j é calculada usando a fórmula da equação 2,

$$Gini_D(Y_j) = \frac{\#D_1}{\#D} Gini(D_1) + \frac{\#D_2}{\#D} Gini(D_2). \tag{2}$$

Onde $\#D_w$ é o número de elementos após a divisão da variável Y_j, $\#D$ o número de elementos do nó antes da divisão e o $Gini(D_w)$ é obtido da equação 3

$$Gini(D_w) = 1 - \sum_{i=1}^{N_C} (p_i)^2. \tag{3}$$

O N_C é a quantidade de classes da variável X e p_i é a proporção das classes, e é estimado pelo número de valores pertencentes à classe C_i dividido por número total de elementos da variável X no nó, ou seja, $\frac{\#C_i}{\#X}$. Como a divisão por meio do Gini faz apenas divisões binárias, as variáveis contínuas são dividas por meio de um ponto de corte a_j, ou seja, valores menores que a_j e valores maiores que a_j. É calculado o IG para todos os pontos de corte possíveis da variável. A medida de Impureza de Gini pode assumir valores de 0 até 1, valores próximos de zero indica homogeneidade dos dados (dados mais puros) e próximos de um, heterogeneidade.

Etapas para o cálculo da Impureza de Gini e a seleção da melhor variável para a divisão, está descrito a seguir:

- Para todas variáveis e suas divisões, calcule individualmente a Impureza de Gini de cada nó filho usando a equação 3;
- Calcule a Impureza de Gini de cada divisão como a média ponderada da equação 2 das Impurezas de Gini dos nós filhos;
- Selecione a divisão com o menor valor de Impureza Gini.

O algoritmo de DT foi introduzido por Breiman et al. (1984), o próprio autor apontou anos após a publicação que uma árvore pode crescer muito, ajustando-se muito bem ao conjunto de treinamento, "aprendendo" muito bem especificamente para o conjunto de treinamento, o que o torna ineficaz para prever novos resultados, pois o algoritmo também aprende os erros do conjunto de treinamento (BREIMAN, 2001). Esse evento é conhecido como sobreajuste (*overfitting*) e pode ser resolvido com um processo denominado de poda. Uma poda de um *Decision Tree* consiste em retirar da estrutura da DT nós que não contribuem para a estimação, outra forma mais eficaz

que a poda é utilizar outro algoritmo que utiliza AD, mas que não ocorre o problema de sobreajuste. Esse algoritmo é o *Random Forest*.

RANDOM FOREST

Neste estudo foi utilizado o algoritmo de *Machine Learning* (ML), *Random Forest* (RF), suas noções foram inicialmente desenvolvidas por Breiman (2001). Seu artigo se tornou texto base para qualquer pesquisador/analista que pretende utilizar o algoritmo em suas análises. O algoritmo *Random Forest* é um método que utiliza diversos algoritmos de *Decision Tree*. Breiman (2001) afirma que o RF obtém uma maior acurácia que DT e mostrou por meio da Lei Forte dos Grandes Números que eles sempre convergem para que o sobreajuste não seja um problema, a Definição 1, apresenta a definição formal do algoritmo proposto por Breiman.

Definição 1 *(BREIMAN, 2001): Um Random Forest é um estimador que consiste em uma coleção de estimadores de Decision Tree* $\{h(\boldsymbol{x}, \Theta_k), k = 1, ...\}$ *onde os* Θ_k *são vetores aleatórios, independentes e identicamente distribuídos e cada árvore lança um voto para uma classe* $\boldsymbol{x}$*, à classe mais popular é a classificação final.*

O algoritmo utiliza um conjunto de *Decision Tree* sem poda, aliado a 2 métodos de aleatorização, *Bagging* (BREIMAN, 1996) e a seleção aleatória das variáveis (BREIMAN, 2001). Com esses dois métodos e um conjunto elevado de DTs, o estimador resulta em melhores resultados de acurácia que a DT e torna o modelo resistente ao sobreajuste (BREIMAN, 2001). Cada *Decision Tree* da floresta é treinada com um subconjunto da base de dados, amostrado aleatoriamente com repetição e em cada nó, um subconjunto das variáveis da base de dados é amostrado, para que a função de divisão avalie apenas aqueles atributos. O algoritmo possui dois parâmetros principais: *ntree* é o número de árvores que serão utilizados e *mtry* a quantidade de variáveis amostradas em cada nó das árvores. A estimação de novos dados por uma RF é dado por um sistema de votação envolvendo cada DT que compõe a floresta, onde a estimação mais frequente das árvores é a estimação dado pela RF. O valor padrão dos do parâmetros *mtry* do algoritmo é $\sqrt{n}$, sendo n a quantidade de variáveis do conjunto. O parâmetro *ntree* é comum utilizar o valor 500.

2.3 INTELIGÊNCIA GEOGRÁFICA

A distribuição espacial de dados procedentes de fenômenos ocorridos no espaço constitui um desafio para o esclarecimento de questões relacionadas às diversas áreas do conhecimento, seja em saúde, ambiente, geologia, agronomia, dentre outros (DRUCK et al., 2004). A preocupação com a variabilidade espacial é antiga, principalmente no início do século XX quando Smith (1910) estudou a disposição de parcelas no campo

em experimentos de rendimento da variabilidade do milho, Montgomery (1913) preocupado com o efeito do nitrogênio no rendimento do trigo, fez um experimento com 224 parcelas nas quais mediu o rendimento do milho. Porém o início da formalização teórica da variabilidade espacial começou quando Krige (1951) concluiu que não haveria sentido nas variâncias das amostras se não considerasse as distâncias entre elas. Matheron (1963,1971), baseado nessa conclusão desenvolveu a teoria a qual chamou de Teoria das Variáveis Regionalizadas e que contém os fundamentos da Geoestatística (VIEIRA, 2000).

A partir do surgimento dos fundamentos da Geoestatística, até o final do século XX vários pesquisadores utilizaram a teoria predominantemente na área de solos (VIEIRA, 2000), porém pesquisadores de outras áreas começaram a notar que a Geoestatística seria uma ferramenta útil para seus estudos. Na área da saúde, Cameron e Jones (1983) em seu artigo escreveram sobre o estudo do médico inglês John Snow, considerado o pai da epidemiologia, que entre 1840 e 1860 procurou entender o surto de cólera de Londres utilizando de outras teorias além da teoria miasmática o qual na época era mais aceita para surtos de cólera e peste negra. Na ocasião, apesar de um pouco mais de 100 anos antes do surgimento dos fundamentos das variáveis regionalizadas, Snow encontrou um padrão espacial nos casos de cólera e identificou que estes estavam sendo causados por uma bomba de água específica, a qual estava poluída. Fechando o acesso a bomba houve um decréscimo rápido dos casos, resultando no fim do surto de cólera. Por causa de Snow, pesquisadores começaram a utilizar de outras teorias e ferramentas na epidemiologia. No início do século XXI estudos utilizando estatística espacial ou artigos salientando a importância dessa ferramenta na área da saúde começaram a ser publicados. Ricketts (2003) aborda a importância dos Sistemas de Informação Geográfica (SIG) para compreender a distribuição de doenças e problemas da saúde pública. Os SIGs georreferenciam as variáveis dos dados possibilitando a visualização espacial dos mesmos, por meio de mapas. Waller e Gotway (2004) publicou um livro aplicando a estatística espacial em dados de saúde pública, apresentando ferramentas úteis, fonte de dados e mapas para a melhor compreensão dos dados. Em trabalhos mais recentes, Pinto (2013) utilizou em sua tese os fundamentos da Geoestatística para identificar áreas de risco de hipertensão e diabetes e encontrou dependência espacial nos riscos, tornando assim a metodologia uma ótima ferramenta para políticas públicas.

A metodologia clássica da Geoestatística proposta por Matheron possui uma certa condição dos dados, eles precisam ser contínuos no espaço. Essa condição inviabiliza diversos tipos de estudos pois fazer experimentos com dados contínuos no espaço demanda muitos recursos financeiros e pessoais, além disso, dados públicos dificilmente são georreferenciados em dados contínuos, eles possuem dados georreferenciados em locais determinados geopoliticamente, bairros, cidades, estados, país. Um indicador

de associação espacial que possa ser usado em dados contínuos ou categóricos é importante.

Naimi et al. (2019) propôs em seu artigo um indicador de associação espacial baseado em entropia local (*entropy-based local indicator os spatial association - ELSA*) que pode ser usado para dados espaciais contínuos e categóricos. O ELSA foi comparado com estatísticas existentes e os estudos mostraram que ele é robusto e consistente, portanto, adequado para diversas áreas, suprindo assim, uma necessidade que a Geoestatística tinha para dados espaciais categóricos. Alguns pesquisadores já aplicaram a nova metodologia em seus estudos.

Dobarco et al. (2021) calcularam a entropia do nível de pedogênese na fazenda Nowley na Austrália, pedogênese é o processo de formação dos solos, produzidos a partir da degradação ou decomposição das rochas, além da junção de fatores químicos, físicos e biológicos. Seus resultados indicaram um nível de ELSA homogêneo em toda a área de estudo, com um valor máximo de 0,33. Houve alguma variação no grau de associação espacial, maiores ELSA, nos extremos leste e oeste da fazenda. Dobarco conclui seus estudos afirmando que os resultados da entropia dos pedogêneses, é útil para determinar o tamanho da amostra para possíveis estudos futuros na região, não necessitando de amostras grandes, já que houve uma homogeneidade do ELSA na área. Nuijten et al. (2021) utilizaram o novo indicador de entropia local para o monitoramento da restauração ecológica, os resultados mostraram associação espacial em alguns pontos da área observada, o autor conclui que o arranjo espacial das estruturas mapeadas pode melhorar a compreensão dos padrões da vegetação, o que pode ajudar a descrever as variáveis ambientais motrizes e os processos de sucessão ecológica na paisagem.

Apesar do surgimento da Geoestatística em mineralogia e ciência dos solos, essa pode ser útil em diversas áreas, na saúde com trabalhos e artigos abordados no texto, essa metodologia é imprescindível para a compreensão das doenças e problemas de saúde pública.

ELSA: *ENTROPY-BASED LOCAL INDICATOR OF SPATIAL ASSOCIATION*

O Indicador local de Associação Espacial Baseado em Entropia, do inglês ELSA, é um indicador que pode ser usado tanto para variáveis categóricas como contínuas. Desenvolvido por Naimi et al. (2019), ELSA quantifica o grau de associação espacial de uma variável em cada local em relação a mesma variável nos seus vizinhos. A análise espacial explora e identifica associações geográficas. Essas associações quantificam o grau em que um valor de uma variável medida em um local é dependente dos valores da mesma variável medida a uma distância geográfica específica deste local. ELSA utiliza o conceito da entropia utilizado na Teoria de Informação, a medida de entropia de Shannon denotada por H e é apresentada na Equação 4.

$$H = -\sum_{k=1}^{m} p_k \log_2 p_k. \tag{4}$$

Onde H mensura a entropia de um sistema com finito números m de possíveis eventos e p_k a probabilidade de um evento k. Naimi et al. (2019) propôs uma padronização para H dividindo o por $\log_2 m$, fornecendo uma medida de entropia que varia entre 0 e 1. A estatística ELSA, utiliza da entropia padronizada para fornecer uma medida de associação espacial de um local i com seus vizinhos de uma determinada distância. Suponha que $\mathbf{x} = (x_1, x_2, ..., x_n)$ são n observações de uma variável nos locais $\mathbf{u} = (u_1, u_2, ..., u_n)$, a ELSA (estatística E) é definido na Equação 5.

$$E_i = E_{ai} \times E_{ci}; \tag{5}$$

$$E_{ai} = \frac{\sum_j \omega_{ij} d_{ij}}{max(d) \sum_j \omega_{ij}}, j \neq i; \tag{6}$$

$$E_{ci} = -\frac{\sum_{k=1}^{m_\omega} p_k \log_2(p_k)}{\log_2 m_i}, j \neq i; \tag{7}$$

$$m_i = \begin{cases} m, & \text{se } \sum_j \omega_{ij} > m \\ \sum_j \omega_{ij}, & \text{caso contrário.} \end{cases}; \tag{8}$$

$$d_{ij} = |x_i - x_j|. \tag{9}$$

Onde ω_{ij} especifica se o local j está dentro de uma distância especificada do local i, d_{ij} descreve a diferença absoluta entre x_i e x_j e $max(d)$ é a máxima dissimilaridade possível entre qualquer par de observações no conjunto de dados. Existem m categorias em todo o conjunto, p_k é a probabilidade da $k-$ésima categoria de m_ω categorias dentro da distância determinada do local i e m_i é o número máximo possível de categorias dentro da distância do local i. Isso significa que se o número de observações dentro da distância determinada do sítio i, incluindo ele, é maior que o número de categorias do de todo conjunto de dados ($\sum_j \omega_{ij} > m$), então m_i é o número de categorias, caso contrário m_i será $\sum_j \omega_{ij}$.

O primeiro termo (E_{ai}) para calcular ELSA na Equação 5 é calculado usando a Equação 6. Este coeficiente varia entre 0 e 1 e resume a dissimilaridade do local e seus vizinhos. Baixo valores de E_{ai} indicam alta similaridade entre o local i e seus vizinhos e altos valores indicam pouca similaridade. O cálculo de E_{ci} é apresentado na Equação 7, a entropia de Shannon da Equação 4 padronizada, esse coeficiente também varia entre 0 e 1 com a mesma interpretação.

Uma etapa fundamental para calcular ELSA para dados contínuos é que a variável

deve ser primeiro categorizada (categorizado ou discretizado) em várias classes; um procedimento que pode causar perda de informações. Morrison (1972) propôs uma estimativa do número ótimo de categorias que minimiza a perda de informações em uma categorização. Este número ótimo é o número mínimo de categorias que é capaz de reproduzir os dados espaciais estatisticamente (ou seja, que minimiza a perda de informações por meio da categorização). Para encontrar o número ideal de classes, o procedimento usa a classificação de Spearman do coeficiente de correlação, ρ, como uma medida de informação entre a variável contínua e a variável categorizada. Se a quantidade de informações não é afetada pela categorização, o valor observado da correlação deve ser igual a um. Qualquer perda de informação resultaria na correlação observada ser menor que um. Portanto, a magnitude da diferença fornece uma medida de perda de informações (QUESTER; DION, 1997). Naimi et al. (2019) apresenta o procedimento de seleção do número ideal que envolve as seguintes etapas:

1. O procedimento de categorização começa com um número mínimo de categorias (m=2);

2. O procedimento atribui um número de classificação (entre 1 e m, onde m é o número total de classes) para cada categoria;

3. O coeficiente ρ entre os valores contínuos e as classificações atribuídas é calculado para cada iteração;

4. As etapas 1 a 3 são repetidas, sempre considerando mais uma categoria (ou seja, aumentando m), até que um limite de convergência (por exemplo, 0,005) seja alcançado. A convergência é definida como o diferença entre os coeficientes ρ das iterações atual e anterior;

5. A regra de um erro padrão (JAMES et al., 2013) é aplicada para selecionar o número ideal de categorias. Primeiro, o erro padrão dos coeficientes é calculado e, em seguida, o ótimo número de categorias seria o menor número para o qual o coeficiente ρ está dentro de um erro padrão do coeficiente mais alto.

Supõe-se que a perda de informação devido à categorização não é substancial quando o número ótimo é usado.

Naimi et al. (2019) apresenta uma forma de verificar a associação espacial global por meio da análise do entrograma. Similar ao variograma, o entrograma apresenta resumidamente as estatísticas ELSA de todos os locais para todas as distâncias até um limite pré determinado. A fórmula do entrograma é apresentado na Equação 10.

$$E(h) = \frac{\sum_{i=1}^{n_i} E_i(h)}{n_h}. \tag{10}$$

O n_h é o número total de locais dentro da distância h utilizado para calcular a ELSA.

2.4 TESTES ESTATÍSTICOS

TESTE QUI-QUADRADO DE PEARSON (χ^2)

O teste de Qui-Quadrado de Pearson é um teste não paramétrico utilizado para verificar associação entre duas variáveis categóricas. O teste foi primeiramente investigado por Pearson (1900), a estatística denotada por χ^2 eleva ao quadrado a subtração da frequência observada dos valores bi variados com o seu valor esperado (sob a hipótese nula de que não há associação entre as variáveis) e o divide pelo valor esperado. Seja a Tabela 6 a notação de uma tabela de contingência para duas variáveis X e Y qualitativas, uma classificada em r categorias e outra em k categorias, respectivamente.

Tabela 6: Notação para tabelas de contingência.

X / Y	c_1	c_2	...	c_k	Total
l_1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
l_2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
...	...	...	...	...	...
l_r	n_{r1}	n_{r2}	...	n_{rk}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	$n_{..}$

Fonte: Elaborada pelo autor.

Na tabela, n_{ij} é o número de elementos pertencentes à i-ésima categoria de X e a j-ésima categoria de Y, $n_{i.}$ os elementos pertencentes a $i-$ésima categoria de X, $n_{.j}$ os elementos pertencentes a $j-$ésima categoria de Y e $n_{..}$ o número total de elementos. A estatística χ^2 cuja fórmula está descrita na Equação 11 possui uma distribuição qui-quadrado com grau de liberdades igual a $(r-1)\times(k-1)$.

$$\sum_{j=1}^{k}\sum_{i=1}^{r}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)\times(k-1)}. \tag{11}$$

O p-valor é obtido por $P(\chi^2_{(r-1)\times(k-1)} \geq \chi^2_{observado})$. Este teste foi utilizado para verificar associação estatística das variáveis óbito em até 3 anos, estadiamento e de fatores de risco com as variáveis epidemiológicas e a associação do óbito em até 3 anos e estadiamento com as de fatores de risco.

TESTE DE MANN-WHITNEY

O teste de Mann-Whitney é um teste não paramétrico que testa sob a hipótese nula, se as distribuições de duas variáveis quantitativas são as mesmas. O teste foi

formalizado por Mann e Whitney (1947), para encontrar a estatística U utilizada no teste, considere n_1 o número de elementos da primeira variável, n_2 o número de elementos da segunda variável e $n = n_1 + n_2$, é unido em um único conjunto os elementos de ambas as variáveis e denotados postos (ordem) para cada elementos, do menor para o maior de 1 até n e depois o conjunto dos postos separado pelas duas variáveis. A estatística U é o menor valor de U_1 e U_2 da Equação 12.

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \text{ e } U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2. \tag{12}$$

Os valores R_1 e R_2 são a soma dos postos para a primeira e a segunda variáveis. A estatística U é padronizada para se aproximar da distribuição normal padrão. A Equação 13 apresenta a padronização.

$$Z = \frac{U - \mu_R}{\sigma_R} \sim N(0,1). \tag{13}$$

Sendo $\mu_R = \frac{n_1 n_2}{2}$ e $\sigma_R = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$. O p-valor é obtido pelo cálculo da $P(Z \geq |z_{observado}|)$. Este teste foi utilizado para verificar se as distribuições das idades podem ser consideradas iguais para as variáveis epidemiológicas com duas classes e para o teste *post-hoc* do teste de Kruskall Wallis.

TESTE DE KRUSKALL WALLIS

O teste de Kruskall Wallis estende o teste de Mann-Whitney para comparação de mais de duas variáveis quantitativas, proposto por Kruskal e Wallis (1952), o teste utiliza da estatística H descrita na Equação 14, o cálculo para encontrar essa estatística se assemelha ao do teste de Mann-Whitney, primeiro é juntado em um único grupo todos os dados de todas as variáveis e denotado postos (ordem) de 1 até N do menor até o maior, sendo N o número total de elementos desse conjunto. É atribuído a quaisquer valores repetidos a média dos postos que eles teriam recebido se não fossem repetidos.

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i.} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}. \tag{14}$$

O valor de n_i é o número de observações do grupo i, r_{ij} é o posto da observação j do grupo i, $\bar{r}_{i.} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ e $\bar{r} = \frac{1}{2}(N+1)$. O pvalor é aproximado por $P(\chi^2_{g-1} \geq H)$. Este teste foi utilizado para verificar se as distribuições das idades são consideradas iguais para as variáveis epidemiológicas acima de duas classes.

TESTE PARA ASSOCIAÇÃO ESPACIAL LOCAL

Naimi et al. (2019) propôs um teste de aleatorização *bootstrap* não paramétrica para testar a associação espacial local contra uma distribuição nula. A abordagem é baseada em amostragem repetida de uma distribuição $\hat{F}_0$, que satisfaz a hipótese nula relevante descrita por Davison e Hinkley (1997). Suponha que $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_m)$ são possíveis eventos de um processo espacial com n observações, $\mathbf{x} = (x_1, x_2, ..., x_n)$ nas localizações $\mathbf{u} = (\mathbf{u_1}, \mathbf{u_2}, ..., \mathbf{u_n})$. A superfície nula pode ser obtida reorganizando ou misturando os locais (ANSELIN, 1995). Uma vez que a superfície nula $\hat{F}_0$ é construído, um simulação de Monte Carlo com R execuções é usada para desenhar uma amostra com tamanho da distribuição n por meio de um procedimento de amostragem *bootstrap* para cada execução. A estatística ELSA observada no local i (E_i) pode ser comparada a R valores independentes da estatística obtida a partir de as amostras correspondentes simuladas independentemente sob a hipótese nula (ou seja, sem autocorrelação). Se esses valores simulados no local forem denotados por $E^*_{1i}, ..., E^*_{Ri}$, então a probabilidade de aceitar a hipótese nula no local i pode ser aproximado por:

$$P_i = \frac{1 + \#\{E_i \geq E^*_{ir}\}}{R + 1}; \tag{15}$$

onde $\#\{E_i \geq E^*_{ir}\}$ indica o número de vezes que o ELSA observado no local i é maior ou igual o ELSA calculado da amostras de *bootstrap* retiradas da distribuição nula.

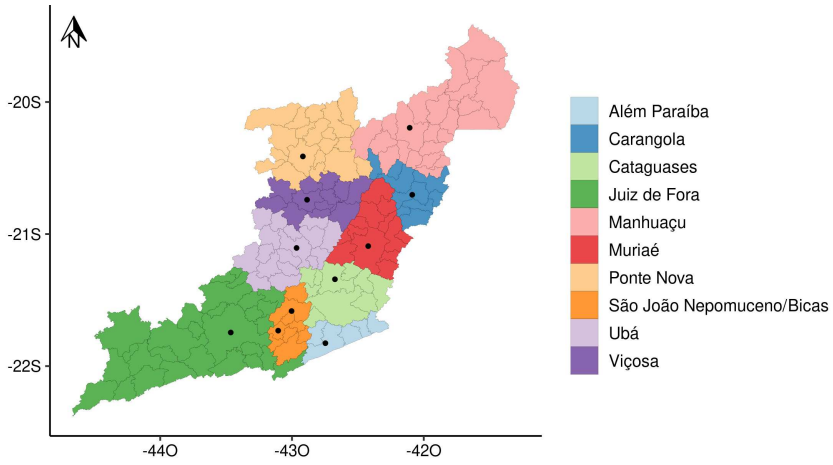
3 MATERIAL E MÉTODOS

3.1 LOCALIZAÇÃO E DADOS DO ESTUDO

Este estudo utilizou dados epidemiológicos e clínicos dos pacientes oncológicos dos municípios da região geográfica intermediária de Juiz de Fora com total pactuação dos tratamentos custeados pelo SUS com o HCM e que fizeram algum tratamento no hospital. A pactuação dos municípios com a instituição, dar-se á quando todos os pacientes oncológicos que necessitem de algum tratamento que será custeado pelo SUS são encaminhados para o hospital, a pactuação pode ser apenas para algum tratamento específico ou para todos os tratamentos oncológicos disponíveis (cirurgia oncológica e/ou reestruturação plástica, quimioterapia, radioterapia, hormonoterapia, iodoterapia, paliativo e outros). Caso a pactuação seja para todos os tratamentos, o município tem total pactuação com a instituição para tratamentos oncológicos.

Regiões geográficas são divisões regionais definidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE) que está em vigor desde 2017, substituindo as antigas divisões meso e micro regiões. A intermediária de Juiz de Fora está localizado no estado de Minas Gerais, é dividida em 10 regiões geográficas imediatas, as imediatas de Além Paraíba, Carangola, Cataguases, Juiz de Fora, Manhuaçu, Muriaé, Ponte Nova, São João Nepomuceno/Bicas, Ubá e Viçosa. A Figura 5 contém detalhes dos limites geopolíticos das regiões imediatas e os municípios pertencentes. As regiões imediatas recebem o nome da cidade referência dos municípios de cada uma, a localização desses municípios e é indicado na Figura 5 com o símbolo de ponto (·).

Figura 5: Regiões geográficas imediatas da região geográfica de Juiz de Fora.



Fonte: Elaborada pelo autor.

De acordo com HCM, as regiões geográficas imediatas de Além Paraíba, Carangola, Cataguases, Manhuaçu, Muriaé, Ubá e Viçosa são as regiões imediatas com

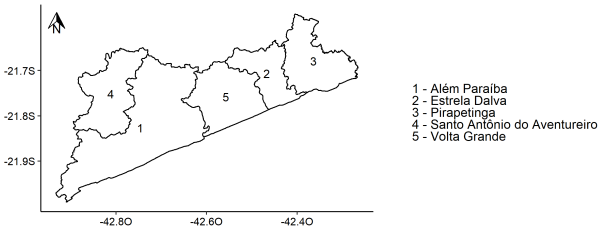
municípios com total pactuação com o hospital. As de Juiz de Fora e São João Nepomuceno/Bicas não possuem municípios pactuados, a de Ponte Nova possui pactuação apenas para tratamento radioterapêuticos. Os municípios de Guarani, Mercês, Piraúba, Rio Pomba e Tabuleiro da imediata de Ubá não são pactuados. No total, são 89 municípios da intermediária de Juiz de Fora com total pactuação, suas respectivas regiões imediatas e as divisões geopolíticas estão representadas na Figura 21.

As bases de dados do Registro Hospitalar do Câncer do HCM e dos prontuários eletrônicos dos pacientes foram os objetos de pesquisa do presente estudo, estes bancos contém informações dos pacientes que estão ou já fizeram algum tratamento no hospital. Os detalhes das variáveis disponíveis no banco de dados do RHC podem ser obtidos em (INCA, 2010). Para este estudo foi utilizado as variáveis sexo, idade, escolaridade, raça, ocupação, estado civil, etilismo, tabagismo, histórico familiar, tipo de câncer (classificação do CID-0), estadiamento e diagnóstico anterior (variável com informação se o paciente fez algum tratamento oncológico prévio em outra instituição) coletados do RHC, o estado nutricional antes do primeiro tratamento oncológico foi coletado dos prontuários eletrônicos. Outro banco de dados utilizado foi o Índice de Desenvolvimento Humano Municipal (IDH-M) que foi obtido no site do IPEAGEO (2021). Os bancos foram unidos em um único banco, os dados do RHC e prontuários eletrônicos foram unidos pelo Cadastro de Pessoa Única (CPF) do paciente, o IDH-M com o banco da junção anterior com o nome da cidade, sendo este o banco final utilizado no estudo. O critério de inclusão do estudo foi, pacientes residentes dos municípios com total pactuação com o HCM da região geográfica intermediária de Juiz de Fora, que fizeram tratamento de 2010 a 2020 no hospital, a exclusão foram pacientes que não possuíam CPF registrado na base do RHC.

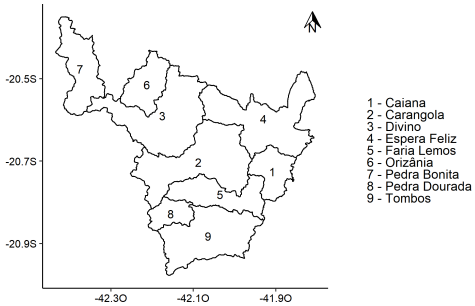
Algumas manipulações dos dados foram realizados, a variável estado nutricional possui 6 classes, a partir dessa variável originou outra denominada de Acima do Peso, que indica se o paciente é ou não acima do peso de acordo com o IMC, à classe "Sim" agrupou as classes "Obesidade" e "Sobrepeso" e a "Não" agrupou as outras categorias. As variáveis Tabagismo e Etilismo possuem 4 classes com informações: "Consumidor", "Ex-Consumidor", "Nunca Consumiu" e "Não se Aplica", quando essas duas variáveis assumem esta última classe, indica que o paciente é menor de 14 anos, para esse caso foi substituído para Nunca Consumiu. Duas novas variáveis dicotômicas foram criadas a partir dessas, "Tabagista" e "Etilista" com as classes "Sim" e "Não" agrupando Consumidor/Ex-Consumidor e Nunca Consumiu respectivamente. Os pacientes objeto do estudo, foram os que entraram no hospital no período de 2010 a 2020, com tumores nos órgãos digestivos procedentes das cidades com total pactuação com a instituição e que tinham registro do CPF nos prontuários eletrônicos.

Figura 6: Regiões geográficas imediatas da região intermediária de Juiz de Fora e os municípios pertencentes.

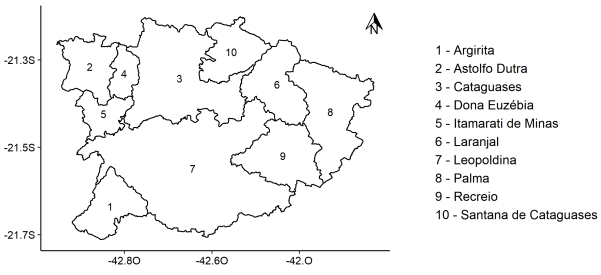
(a) Imediata de Além Paraíba.



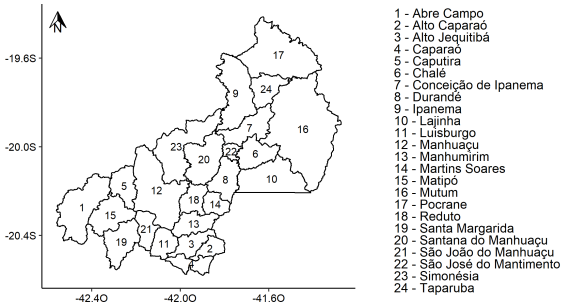
(b) Imediata de Carangola.



(c) Imediata de Cataguases.

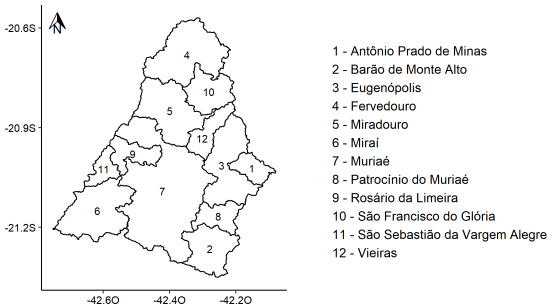


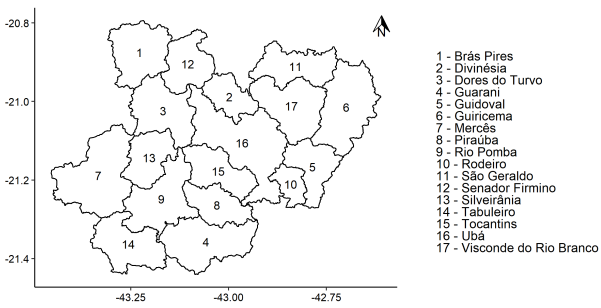
(d) Imediata de Manhuaçu.



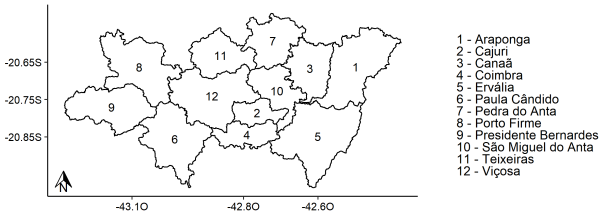
(e) Imediata de Muriaé.

(f) Imediata de Ubá.





(g) Imediata de Viçosa.



Fonte: Elaborada pelo autor.

3.2 MÉTODOS

Primeiramente no presente estudo, foi realizado uma análise exploratória dos dados. Analisando os novos casos por ano, a distribuição da frequência dos tumores dos órgãos digestivos, o perfil epidemiológico e a quantidade de entradas sem informações nas variáveis estudadas. A segunda parte, foi utilizado a Ciência de Dados. Na base final do estudo (junção da base do RHC, Prontuários Eletrônicos e IDH-M) existem algumas entradas e variáveis com valores sem informações (não disponíveis). Essas entradas foram imputadas antes da estimação do estado nutricional dos pacientes, a imputação foi utilizando o algoritmo de RF. Após essa etapa, foi testado diversos conjuntos de treinamento para treinar o algoritmo visando encontrar o conjunto que resultasse em uma acurácia satisfatória. Para definir os valores dos parâmetros do RF, foi testado diversos pares e comparado a acurácia dos algoritmos. O par escolhido foi o que resultou na melhor taxa de acerto e os valores dos parâmetros da RF foram *ntree* igual a 1000 e *mtry* igual 2.

Na terceira parte, foram realizados testes estatísticos. Para as inferências realizadas foram utilizados testes não paramétricos. Para comparação de duas variáveis qualitativas o teste de Qui-Quadrado de Pearson (χ^2), para comparação de duas variáveis uma quantitativa e uma qualitativa com duas categorias o teste de Mann-Whitney, para mais de duas categorias o teste de Kruskal-Wallis (em caso de significância estatística, um teste *post-hoc* foi realizado utilizando o teste Mann-Whitney).

Por fim, a Inteligência Geográfica foi utilizada para a análise espacial. A estatística E foi calculada nos municípios do estudo e calculado a associação espacial em até 100 quilômetros de distância considerando a fronteira das cidades. Para testar a associação espacial das variáveis, foi utilizado o teste de associação espacial local *bootstrap* descrito por Naimi et al. (2019).

O software e os pacotes utilizados no presente estudo foram, R: *Language and Environment for Statistical Computing* (R CORE TEAM, 2021), versão 4.0.1, *Tidyvese* (2019) para manipulação dos dados e plotagem dos gráficos, *Ranger* (2017) para a utilização da *Random Forest* para predição, *missRanger* (2021) para a imputação dos valores sem informação utilizando *Random Forest*, *ELSA* (2019) para o cálculo da entropia dos dados, IBGE (2021) para obter a malha dos municípios e *RGDAL* (2021) para a leitura e manipulação da malha. O nível de significância adotado para os testes estatísticos foi de 5%. Este estudo foi aprovado pelo Comitê de Ética em Pesquisa e está registrado com o CAAE: 52012521.3.0000.5105.

4 RESULTADOS E DISCUSSÃO

4.1 ANÁLISE EXPLORATÓRIA

De 2010 a 2020, 31.021 pacientes oncológicos fizeram algum tratamento no Hospital do Câncer de Muriaé sendo 88,66% de municípios da Região Geográfica Intermediária de Juiz de Fora (RGIJF); 19,2% são pacientes com tumores nos órgãos digestivos e 17,21% com as duas condições anteriores. Dos 5.339 pacientes com tumores nos órgãos digestivos da RGIJF, 46 não tinham registro do CPF no RHC e 176 residiam em municípios que não eram 100% pactuados com o HCM. Detalhes da frequência e a porcentagem da distribuição dos tumores dos órgãos digestivos dos 5.133 pacientes objeto do estudo estão descritos na Tabela 7. Cólon, esôfago, estômago e reto são 80,70% dos casos.

Tabela 7: Frequência e porcentagem dos casos de tumores nos órgãos digestivos dos pacientes objeto do estudo.

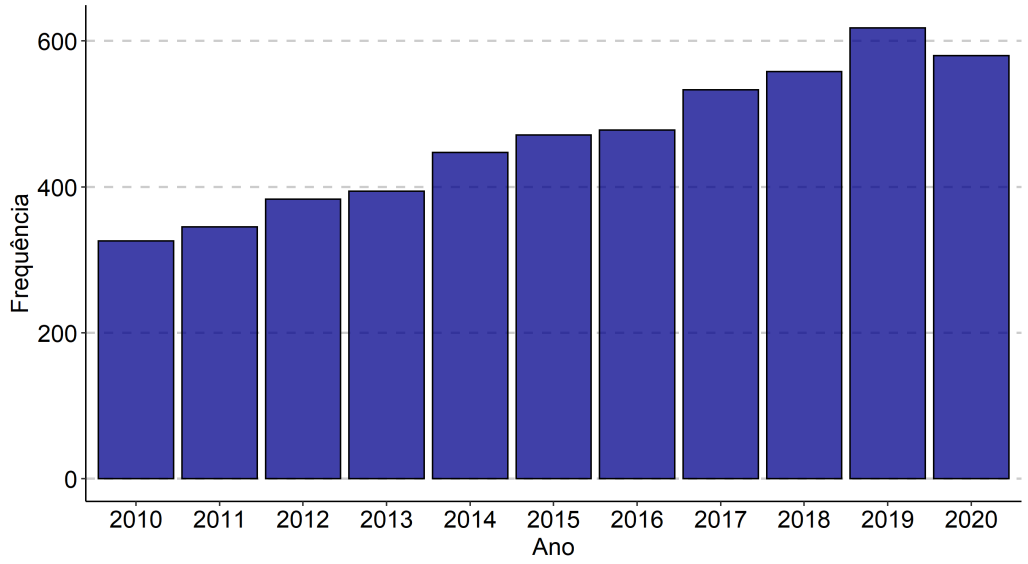
Localização do Tumor	Frequência	Porcentagem
Cólon	1.185	23,09%
Esôfago	1.173	22,85%
Estômago	986	19,21%
Reto	798	15,55%
Pâncreas	309	06,02%
Junção retossigmóide	207	04,03%
Vesícula biliar	172	03,35%
Fígado e vias biliares	161	03,13%
Ânus e canal anal	80	01,56%
Intestino delgado	55	01,07%
Localizações mal definidas	7	00,14%
Total	**5.133**	**100,00%**

Fonte: Elaborada pelo autor.

Os dados do RHC apontam que a média de novos casos anual foi de 466 pacientes, detalhes na Figura 7. O número de novos casos aumentou de 2010 a 2019, sendo nos últimos 4 anos a entrada média foi de 572 pacientes.

O perfil médio epidemiológico dos pacientes do estudo foram homens com 63 anos, brancos, casados, com o ensino fundamental incompleto, não etilistas; tabagistas e não eram acima do peso. O perfil se manteve durante os anos de 2010 a 2020, a Figura 8 apresenta as evoluções anuais das porcentagem dessas variáveis. Homens foram os pacientes mais frequentes da instituição, no período de 2010 a 2018 corresponderam cerca de 60%. A idade mediana sempre ficou entre 63 e 64 anos, pelo menos metade dos pacientes são idosos. A raça branca foi a mais observada, exceto em 2013, porém o resultado observado neste ano foi atípico. Do total, brancos corres-

Figura 7: Frequência anual de novos casos de pacientes com tumores nos órgãos digestivos do HCM.



Fonte: Elaborada pelo autor.

ponderam cerca de 45%, os pardos 40% e pretos 18%. As raças indígenas e amarela foram pouco observados (1 e 2) nos 11 anos. A escolaridade foi outra variável que manteve um padrão, pacientes com o fundamental incompleto foram aproximadamente 60% dos casos, durante todos os anos 74,35% dos pacientes da instituição ou eram apenas alfabetizados ou analfabetos. Pacientes com familiares que já foram diagnosticados com algum tumor são mais frequentes, representando quase 70% nos últimos 5 anos.

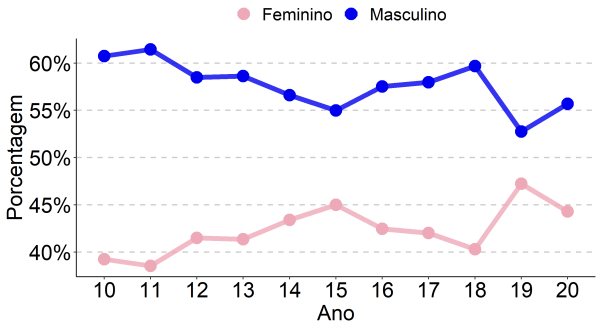
Em relação as ocupações dos pacientes, as frequências e as porcentagens das 10 profissões mais observadas estão descritas na Tabela 8. Do lar e funcionário público foi a mais observada, sendo que a ocupação Do lar classifica as pessoas que trabalham exclusivamente para a própria família, não exercendo atividade remunerada. Acredita se que seja a maioria da categoria. As ocupações que utilizam trabalhos braçais somados juntos são 2.049 (39,91%) pacientes.

A evolução anual dos fatores de risco estão descritos na Figura 9. Até 2016 não houve predominância entre os pacientes etilistas e os não etilistas, a partir desse ano esse último foi o mais frequente. Os tabagistas configuraram a maioria dos pacientes, em 2013 sendo 59% dos casos, essa taxa diminuiu nos anos posteriores chegando a 53% no ano de 2020.

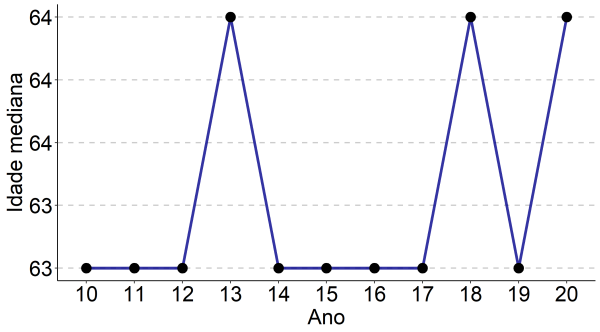
A evolução anual do estadiamento e o óbito em até 3 anos dos pacientes é exibido na Figura 10. O estadiamento avançado é um problema desde 2010, em todos os anos 70% dos pacientes entraram no hospital com estadiamento III ou IV. A letalidade em até 3 anos após o diagnóstico até 2017 sempre foi acima de 30% dos pacientes, as

Figura 8: Evolução anual do perfil dos pacientes objeto do estudo.

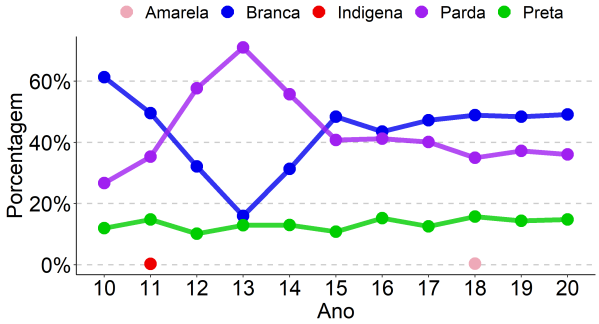
(a) Sexo.



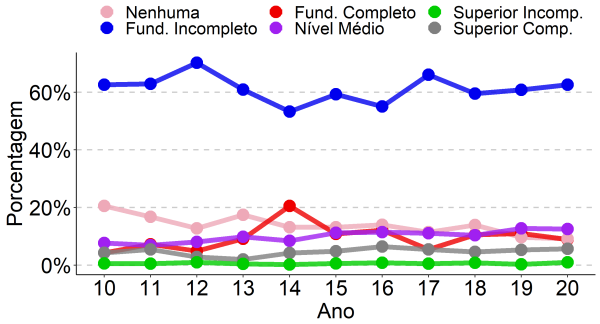
(b) Idade mediana.



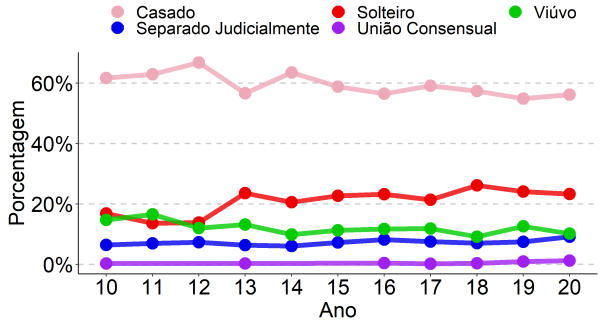
(c) Raça.



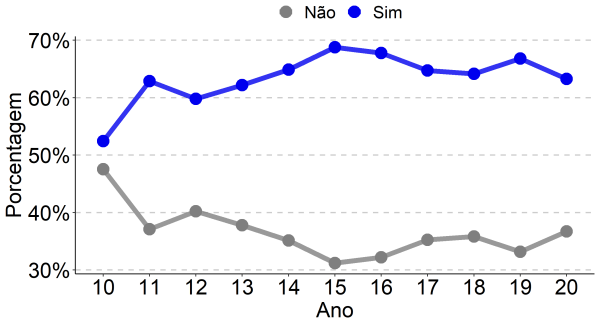
(d) Escolaridade.



(e) Estado civil.



(f) Histórico familiar de câncer.



Fonte: Elaborada pelo autor.

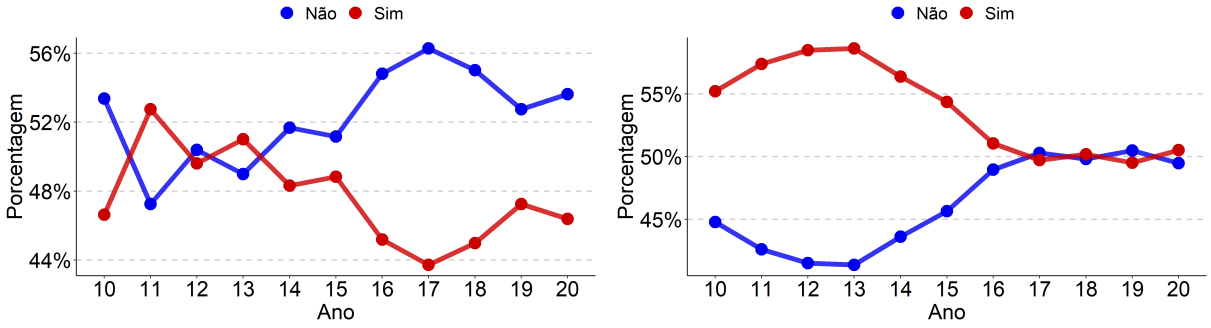
taxas dos anos de 2018 a 2020 podem aumentar pois ainda não completaram 3 anos de seguimento.

Dos dados de todos os pacientes oncológicos da instituição procedentes da RGIJF; 0,4% da raça; 4,26% da escolaridade; 0,05% do estado civil e 24,20% do histórico familiar de câncer não tinham informações registradas no RHC, das variáveis de fatores de risco; 12,25%; 10,26% e 35,95% não existiam informações para o etilismo, tabagismo e estado nutricional, respectivamente.

Figura 9: Evolução anual dos fatores de risco dos pacientes objeto do estudo.

(a) Etilismo dos pacientes.

(b) Tabagismo dos pacientes.

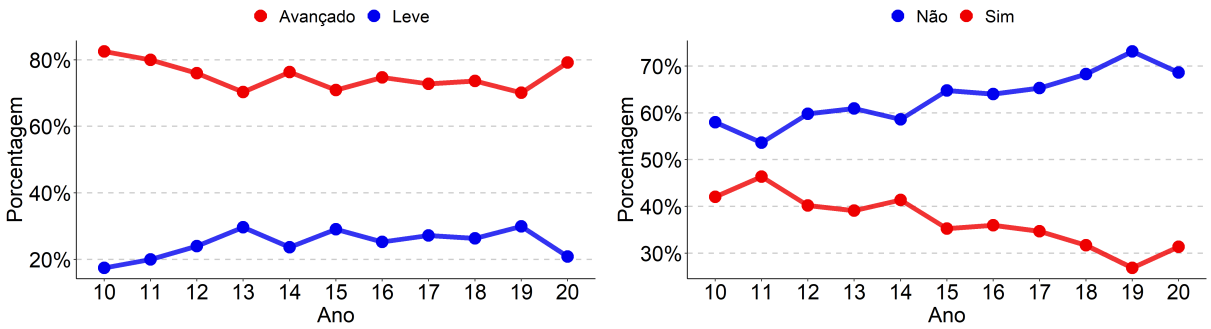


Fonte: Elaborada pelo autor.

Figura 10: Evolução anual do perfil clínico dos pacientes objeto do estudo.

(a) Estadiamento dos tumores.

(b) Óbito em até 3 anos do diagnóstico.



Fonte: Elaborada pelo autor.

Tabela 8: Porcentagem das 10 ocupações mais frequentes dos pacientes objeto do estudo.

Ocupação	**N**	**%**
Do lar e funcionário público	1487	28,97%
Agropecuarista	1432	27,90%
Serviços gerais	196	03,82%
Pedreiro	193	03,76%
Motoristas	184	03,58%
Comerciantes	168	03,27%
Professores	133	02,59%
Trabalhadores de serventia	118	02,30%
Trabalhadores agrícolas	110	02,14%
Costureiros	93	01,81%

Fonte: Elaborada pelo autor.

4.2 CIÊNCIA DE DADOS

Os dados sem informações para as variáveis epidemiológicas foram imputados pelo RF exceto para a variável Acima do Peso. A taxa de acerto da imputação foi calculado, do conjunto de dados inicial foram excluídos todas as entradas que possuíam valores sem informação e desse conjunto foram retirados de forma aleatória as informações em algumas entradas para todas as variáveis dos dados e esses imputados pelo algoritmo. O *Random Forest* para a estimação de valores sem informação obteve uma acurácia de 46,12%. Com os dados sem informações imputados, foi realizada uma busca de todo o conjunto de dados dos 31.021 pacientes de todos os tumores de 2010 a 2020 para encontrar um conjunto de treinamento que apresentasse um resultado com uma acurácia satisfatória. Para diminuir o viés das consequências do tumor no estado nutricional dos pacientes, foram excluídos do conjunto de dados os pacientes que entraram na instituição com o estadiamento classificado como II, III e IV, pacientes que fizeram tratamentos oncológicos previamente em outras instituições e pacientes com tumores com fatores de risco o sobrepeso/obesidade, o tabagismo, o etilismo ou as variáveis epidemiológicas utilizadas neste estudo. O conjunto originado por essas remoções, foi separado pelos tumores dos pacientes e para cada subconjunto desses dados foi utilizado o algoritmo de RF e selecionado o conjunto que apresentasse a melhor acurácia. Os dados dos pacientes com câncer de pele foi o subconjunto que apresentou melhor acurácia para a estimação da variável Acima do Peso; 85,48% (detalhes da matriz de confusão para o teste da estimação utilizando os dados dos pacientes com tumores na pele na Tabela 9). Com o algoritmo treinado, foi estimado os valores da variável para os pacientes do estudo, a Tabela 10 descreve os valores do estado nutricional dos 3.276 pacientes que continham essa informação nos prontuários eletrônicos e as estimações dado pelo algoritmo. Desses pacientes; 779 o

algoritmo estimou que este estava acima do peso que antes da consulta não pertenciam a essa categoria, apresentando um panorama diferente do que é observado na instituição; 23,78% dos pacientes dos que continham informação do estado nutricional e 15,17% de todos os pacientes perderam peso em decorrência do tumor. Para a estimação final da variável Acima do Peso, foi utilizado a informação inicial dessa variável; para os pacientes que eram classificados como "Sim" no dia da consulta permaneceram com a mesma classificação, dificilmente ganham peso e mudam de classe de estado nutricional após o diagnóstico, para os outros pacientes a estimação final é dado pela RF.

Tabela 9: Matriz de confusão da estimação teste para os pacientes com tumores de pele.

		Estimado	
Acima do peso		Não	Sim
Real	Não	29	6
	Sim	3	24

Fonte: Elaborada pelo autor.

Tabela 10: Matriz de confusão da estimação para os pacientes objeto do estudo.

		Estimado	
Acima do peso		Não	Sim
Real	Não	2.030	779
	Sim	368	220

Fonte: Elaborada pelo autor.

4.3 TESTES ESTATÍSTICOS

A Tabela 11 contém detalhes do perfil epidemiológico dos pacientes nos 11 anos, a idade mediana e a porcentagem de óbitos em até 3 anos após o diagnóstico para cada categoria das variáveis. As variáveis Sexo, Raça, Escolaridade e Histórico familiar de câncer apresentaram associação estatística com a letalidade em até 3 anos, homens morrem 9,1% pontos percentuais a mais que as mulheres; para a raça pacientes pretos tiveram a pior letalidade (a raça amarela foi desconsiderada por possui uma amostra pequena), porém a raça não apresenta uma predisposição conhecida para tal evento, outros fatores como socio econômicos e baixa escolaridade influenciaram no resultado. Dos pretos 86,04% possuem até o ensino fundamental incompleto, pardos 75,94% e brancos 69,16%. Mais de 70% dos pacientes não concluíram o ensino fundamental, esse resultado mostra os cuidados necessários na interação do corpo clínico (médicos, enfermeiros, técnicos de enfermagem e equipe multidisciplinar) com os pacientes. A baixa instrução pode causar dificuldade na compreensão de termos técnicos ou difíceis, nos questionários da anamnese e nas informações dos cuidados necessários durante o tratamento. A baixa escolaridade (até fundamental incompleto) resulta em uma pior letalidade, mais de 36% foram a óbito em menos de 3 anos. Para nenhuma escolaridade a taxa sobe para 45,24%, cerca de 10% em média de pontos percentuais a mais que a dos outros pacientes. Os testes para a idade, apresentou significância para a raça, escolaridade e estado civil. Dos brancos, 50% entraram no hospital acima de 65 anos, pacientes com nenhuma escolaridade acima de 71 anos e viúvos acima de 75. O teste *post-hoc* mostrou diferença entre todas as classes para a raça e o estado civil, para a escolaridade a distribuição das idades dos pacientes com superior completo e fundamental completo são iguais, entre os outros pares o teste apresentou diferença.

Dos 5.115 pacientes; 47,37% consumiam bebidas alcoólicas e 53,16% eram tabagistas. Essas variáveis apresentaram associação com as variáveis epidemiológicas descritas na Tabela 12. Homens etilistas e tabagistas são em proporção mais que o dobro que as mulheres, para a raça os pretos também apresentaram maior proporção relativa dos fatores de risco, com quase 20% de diferença se comparado com os pardos, a segundà classe com mais proporção de tabagistas e etilistas. Da escolaridade, pacientes com até o fundamental completo são em proporção mais etilistas e tabagistas do que os pacientes com grau de instrução maior. Para o estado civil, as porcentagens foram maiores para os solteiros e os em união consensual do que para as outras classes.

O estadiamento e o sobrepeso/obesidade apresentaram associação com as variáveis epidemiológicas descritas na Tabela 13. Os acima do peso configuram 35,14% dos pacientes e 74,73% com estadiamento avançado. Os homens dão entrada no hospital

Tabela 11: Perfil epidemiológico e letalidade em até 3 anos dos pacientes objeto do estudo.

Variável	N	%	Idade M.	Óbito < 3 Anos	
				Não	Sim
Sexo				**	
Feminino	2.190	42,67%	64,00	69,41%	30,59%
Masculino	2.943	57,33%	63,00	60,31%	39,69%
Raça			*	**	
Amarela	2	0,04%	60,00	50,00%	50,00%
Branca	2.242	43,68%	65,00	65,83%	34,17%
Indígena	1	00,02%	31,00	100,00%	00,00%
Parda	2.196	42,78%	63,00	64,25%	35,75%
Preta	692	13,48%	61,00	58,67%	41,33%
Escolaridade			*	**	
Nenhuma	683	13,31%	71,00	54,76%	45,24%
Fund. incompleto	3.133	61,04%	64,00	63,13%	36,87%
Fund. completo	503	9,80%	59,00	68,19%	31,81%
Nível médio	534	10,40%	57,00	72,47%	27,53%
Superior incompleto	34	00,66%	55,50	76,47%	23,53%
Superior completo	246	04,79%	59,00	76,02%	23,98%
Estado Civil			*		
Solteiro	1.102	21,47%	58,00	62,07%	37,93%
União Consensual	20	00,39%	54,00	60,00%	40,00%
Casado	3.028	58,99%	64,00	65,32%	34,68%
Separado Judicialmente	376	7,33%	61,00	66,22%	33,78%
Viúvo	607	11,83%	75,00	61,29%	38,71%
Hist. Fam. de Câncer				**	
Não	1.850	36,04%	64,00	60,65%	39,35%
Sim	3.283	63,96%	63,00	66,19%	33,81%
Total	**5.133**	**100%**	**63**	**64,19%**	**35,81%**

Fund.: Fundamental; Hist. Fam. de Câncer: Histórico Familiar de Câncer; M.: Mediana;
**: p-valor teste de Kruskall Wallis < 0,05; **: p-valor teste de Qui-Quadrado de Pearson < 0,05.*

Fonte: Elaborada pelo autor.

com o estadiamento avançado em 78,80% dos casos, a alta porcentagem da taxa de risco e a demora para procurar atendimento médico (alta idade) justificam esse resultado. Em contrapartida as mulheres possuem mais sobrepeso/obesidade. A raça preta também apresentou uma alta taxa de estadiamento e os brancos de obesidade. Na variável escolaridade, a porcentagem de estadiamento avançado diminui e pacientes acima do peso aumentam conforme aumenta o grau de instrução. Para o estado civil, os casados e os viúvos apresentaram a menor porcentagem de tumor avançado entre as outras categorias.

Comparando o estadiamento e o óbito com as variáveis de risco (detalhes na Tabela 14), estes apresentaram associação para todas as comparações. Os pacientes eti-

Tabela 12: Etilismo e tabagismo entre as variáveis epidemiológicas dos pacientes objeto do estudo.

		Etilista		**Tabagista**	
		Não	Sim	Não	Sim
Sexo		**		**	
	Feminino	82,01%	17,99%	69,77%	30,23%
	Masculino	30,75%	69,25%	29,77%	70,23%
Raça		**		**	
	Amarela	100,00%	00,00%	50,00%	50,00%
	Branca	59,95%	40,05%	53,39%	46,61%
	Indígena	100,00%	00,00%	100,00%	00,00%
	Parda	50,14%	49,86%	45,54%	54,46%
	Preta	36,56%	63,44%	29,62%	70,38%
Escolaridade		**		**	
	Nenhuma	54,76%	45,24%	42,02%	57,98%
	Fund. incompleto	51,42%	48,58%	45,55%	54,45%
	Fund. completo	48,91%	51,09%	45,73%	54,27%
	Nível médio	55,24%	44,76%	53,37%	46,63%
	Superior incompleto	64,71%	35,29%	61,76%	38,24%
	Superior completo	62,20%	37,80%	62,60%	37,40%
Estado Civil		**		**	
	Solteiro	43,10%	56,90%	39,47%	60,53%
	União Consensual	45,00%	55,00%	65,00%	35,00%
	Casado	53,20%	46,80%	47,29%	52,71%
	Separado Judicialmente	44,41%	55,59%	40,43%	59,57%
Total		**52,63%**	**47,37%**	**46,84%**	**53,16%**

*Fund.: Fundamental; **: p-valor teste de Qui-Quadrado de Pearson < 0,05.*

Fonte: Elaborada pelo autor.

listas, tabagistas apresentaram um estadiamento mais avançando do que os que não estavam, para os acima do peso, apesar de ser um dos principais fatores de riscos, os pacientes com sobrepeso/obeso entraram no hospital com o estadiamento menos avançado dos que não eram. O mesmo ocorre para a letalidade em até 3 anos, os tabagistas e etilistas morreram mais que os que não eram e os acima do peso morreram menos do que não eram acima do peso.

Tabela 13: Estadiamento e Peso entre as variáveis epidemiológicas dos pacientes objeto do estudo.

		Estadiamento		**Acima do Peso**	
		Inicial	Avançado	Não	Sim
Sexo			**		**
	Feminino	30,73%	69,27%	57,49%	42,51%
	Masculino	21,20%	78,80%	70,34%	29,66%
Raça			**		**
	Amarela	50,00%	50,00%	100,00%	00,00%
	Branca	26,09%	73,91%	47,90%	52,10%
	Indigena	100,00%	00,00%	00,00%	100,00%
	Parda	26,46%	73,54%	77,87%	22,13%
	Preta	18,64%	81,36%	78,47%	21,53%
Escolaridade			**		**
	Nenhuma	22,11%	77,89%	80,09%	19,91%
	Fund. incompleto	24,13%	75,87%	67,35%	32,65%
	Fund. completo	26,44%	73,56%	58,45%	41,55%
	Nível médio	28,84%	71,16%	46,63%	53,37%
	Nível superior incompleto	35,29%	64,71%	47,06%	52,94%
	Nível superior completo	36,99%	63,01%	45,93%	54,07%
Estado Civil			**		**
	Solteiro	22,41%	77,59%	69,87%	30,13%
	União Consensual	10,00%	90,00%	40,00%	60,00%
	Casado	26,02%	73,98%	62,88%	37,12%
	Separado Judicialmente	24,73%	75,27%	52,93%	47,07%
	Viúvo	27,51%	72,49%	73,81%	26,19%
Total		**25,27%**	**74,73%**	**64,86%**	**35,14%**

*Fund.: Fundamental; **: p-valor teste de Qui-Quadrado de Pearson < 0,05.*

Fonte: Elaborada pelo autor.

Tabela 14: Estadiamento e letalidade entre os fatores de risco dos pacientes objeto do estudo.

		Estadiamento		Óbito < 3 anos	
		Inicial	Avançado	Não	Sim
Etilista			**		**
	Não	30,47%	69,53%	72,71%	27,29%
	Sim	19,49%	80,51%	69,20%	30,80%
Tabagista			**		**
	Não	31,24%	68,76%	72,46%	27,54%
	Sim	20,01%	79,99%	69,81%	30,19%
Acima do Peso			**		**
	Não	22,11%	77,89%	67,35%	32,65%
	Sim	31,10%	68,90%	77,88%	22,12%

***: p-valor teste de Qui-Quadrado de Pearson < 0,05.*

Fonte: Elaborada pelo autor.

4.4 INTELIGÊNCIA GEOGRÁFICA

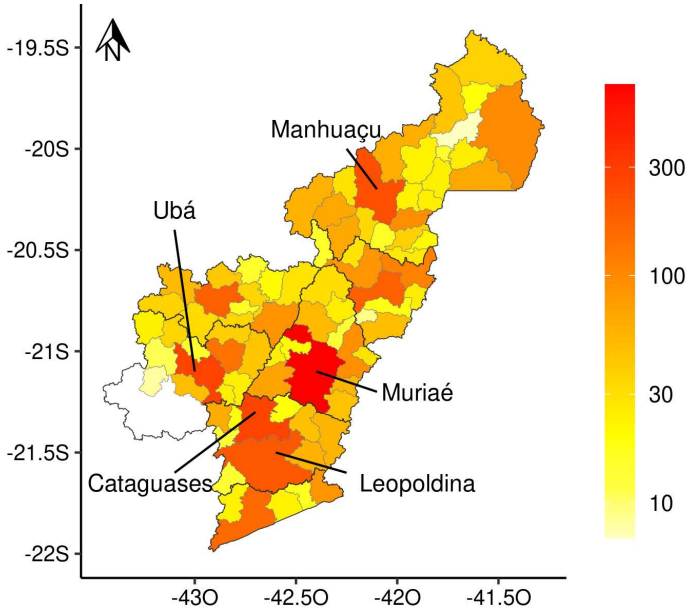
Esta subseção apresenta os resultados das estatísticas espaciais para as frequências de casos e as variáveis Acima do Peso, Tabagismo, Etilismo, Escolaridade e Estadiamento. As Figuras 12, 13, 14, 15, 16 apresentam os gráficos de pontos, box plot, densidade, entrograma das proporções de pacientes com sobrepeso/obesos, tabagistas, etilistas, baixa escolaridade (até o ensino fundamental incompleto) e o estadiamento avançado por município, respectivamente.

As cidades mais populosas e referências das regiões imediatas do estudo foram as com mais casos observados, sendo Muriaé correspondendo 13,46% do total (691), Ubá 05,35% (275), Cataguases 05,26% (270), Manhuaçu 04,51% (232) e Leopoldina 03,93% (202), as frequências para outras cidades é apresentada na Figura 11.

Os pacientes acima do peso representam 57% ou mais em 1/10 dos municípios. As cidades Antônio Prado de Minas, Alto Caparaó e Conceição de Ipanema foram as com mais pacientes com sobrepeso/obesidade; 82,80%, 71,40% e 71,40%; respectivamente. A maioria dos municípios tiveram entre 20% e 40% de casos desses pacientes. A característica do estado nutricional apresentou uma associação espacial mensurada pela estatística E em até 100 km.

Os tabagistas foram observados com mais de 60,19% dos casos em 1/4 dos municípios, nas cidade de Cajuri, Caputira, São Sebastião da Vargem Alegre, Silveirânia, Pedra do Anta e Miradouro foram observados mais de 70% dos casos. Como abordado, o tabagismo apresentou associação tanto com o estadiamento avançado como a letalidade em até 3 anos, esse resultado mostra um panorama preocupante na re-

Figura 11: Número de casos por procedência dos pacientes objeto do estudo.



Fonte: Elaborada pelo autor.

gião do estudo já que 75% dos municípios, mais da metade dos seus pacientes são tabagistas. Esse perfil de pacientes apresentou uma alta associação espacial em uma distância de até 150 km para cada local.

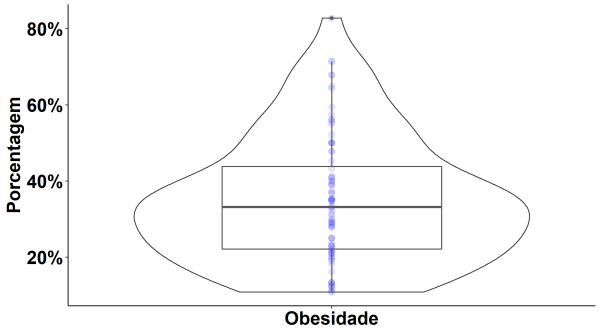
Da proporção dos etilistas, 1/20 das procedência apresentaram acima de 64,58% dos casos. As cidades que apresentaram mais de 70% dos casos de pacientes etílicos foram Argirita e São Sebastião da Vargem Alegre. A maioria dos municípios observaram entre 40% e 60% de pacientes etílicos. Essa proporção apresentou associação espacial em até 200 km.

Os pacientes com baixa escolaridade foram mais de 80% dos casos em 55% das cidades. Araponga, Conceição de Ipanema, Divinésia, Pedra do Anta, Caputira e Dores do Turvo apresentaram mais de 90% dos casos. A grande maioria das cidades apresentou mais de 70% de casos com pacientes poucos instruídos, o entrograma apontou que essa característica teve associação espacial em até 150 km para cada local.

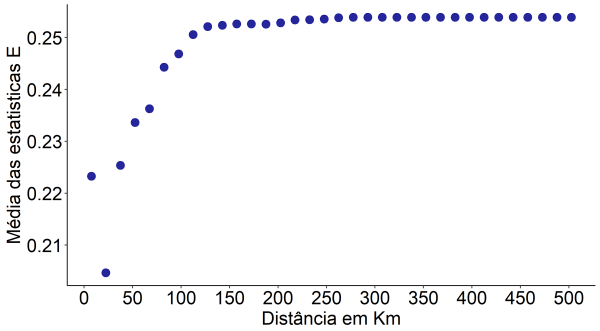
Por fim, o estadiamento em estado avançado foi observado acima de 80,65% em 1/3 das cidades. Os municípios de Silveirânia, Canaã, Cajuri, Pedra Bonita, Santana do Manhuaçu, Santo Antônio do Aventureiro e Volta Grande foram os que apresentaram um cenário crítico com mais de 90% dos casos. Para 3/4 das cidades, os pacientes entraram com estadiamento III e IV em 70% dos casos. Esse perfil apresentou associação em até 150 km dos locais.

Figura 12: Gráficos de boxplot, densidade, pontos e entrograma da variável Acima do Peso dos pacientes objeto do estudo.

(a) Boxplot, densidade e pontos.



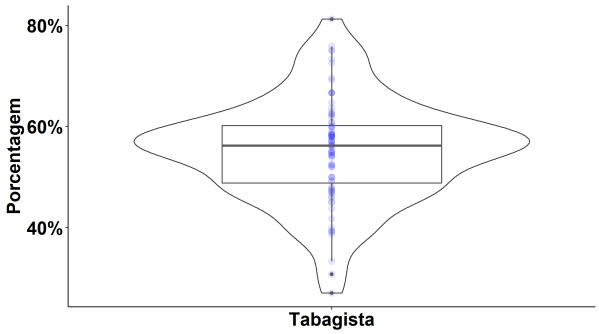
(b) Entrograma.



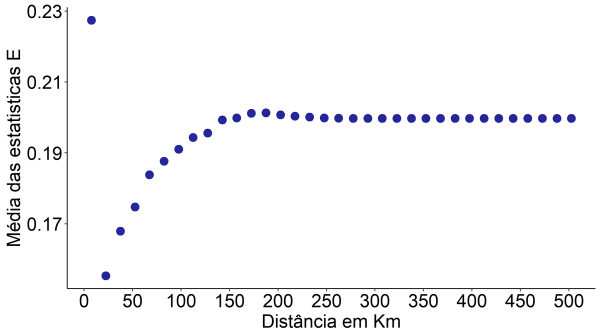
Fonte: Elaborada pelo autor.

Figura 13: Gráficos de boxplot, densidade, pontos e entrograma da variável Tabagista dos pacientes objeto do estudo.

(a) Porcentagem.



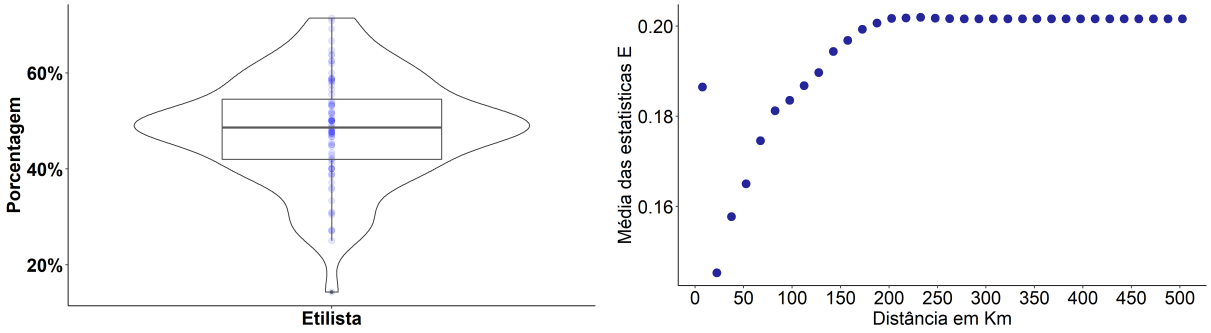
(b) Entrograma.



Fonte: Elaborada pelo autor.

Figura 14: Gráficos de boxplot, densidade, pontos e entrograma da variável Etilista dos pacientes objeto do estudo.

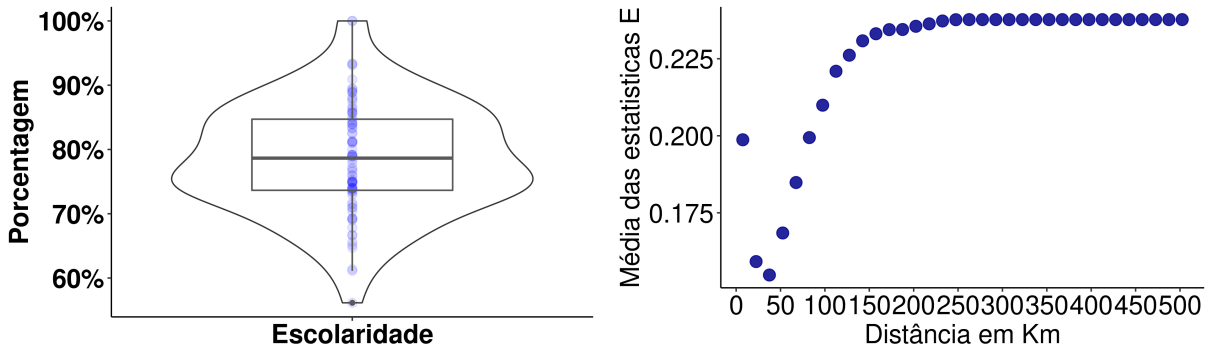
(a) Porcentagem. (b) Entrograma.



Fonte: Elaborada pelo autor.

Figura 15: Gráficos de boxplot, densidade, pontos e entrograma dos pacientes com baixa escolaridade do objeto do estudo.

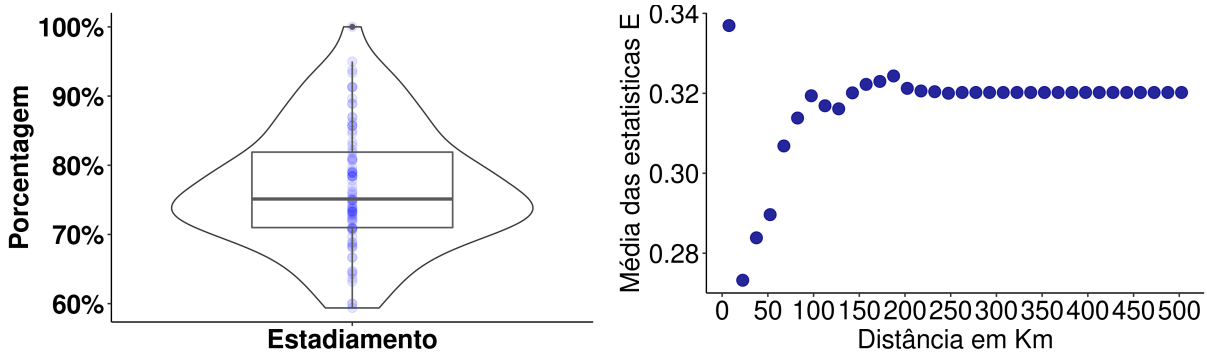
(a) Porcentagem. (b) Entrograma.



Fonte: Elaborada pelo autor.

Figura 16: Gráficos de boxplot, densidade, pontos e entrograma dos pacientes com estadiamento avançado do objeto do estudo.

(a) Porcentagem. (b) Entrograma.



Fonte: Elaborada pelo autor.

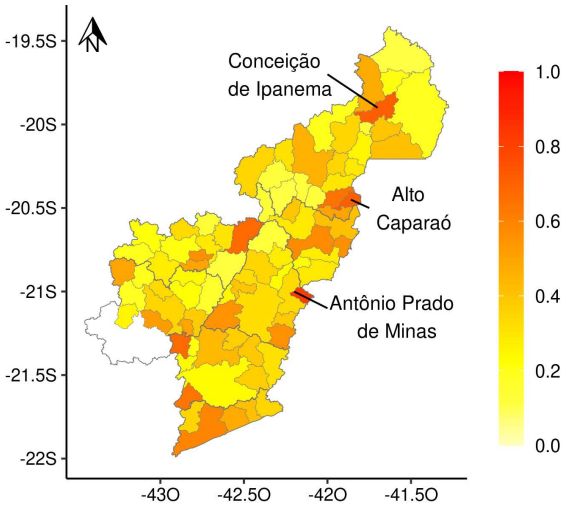
As Figuras 17, 18, 19, 20, 21 apresentam as proporções, a estatística E e o p-valor obtido pelo teste de associação local para cada município dos pacientes acima do peso, tabagistas, etilistas, baixa escolaridade e estadiamento III e IV respectivamente.

As cidades mais desenvolvidas observaram as proporções mais baixa de pacientes acima do peso, na maioria dos casos, os municípios mais afastados das cidades referências das regiões imediatas tiveram as mais alta proporção desses pacientes. A ELSA mostrou uma associação espacial significativa para diversos municípios, o teste apontou que 57,14% tiveram p-valor menor que 0,05. Para a proporção de pacientes tabagistas, as regiões imediadas de Ubá e Viçosa foram as que apresentaram maior proporção que as demais, novamente as cidades mais distantes das referências das regiões foram as que obtiveram o pior resultado. A estatística ELSA apresentou alta associação espacial, observando um valor abaixo de 0,1 para quase todas as cidades, a imediata de Manhuaçu foi a que apresentou menor associação. As imediatas de Além Paraíba e Cataguases resultaram em um associação espacial para todas as cidades. No total o teste de associação encontrou significância estatística para 70,23% das cidades. As imediatas de Ubá e Viçosa foram as regiões que apresentaram mais municípios com a proporção de pacientes etílicos e a de Manhuaçu com a menor. A associação espacial apresentou significância em 83,33% das cidades com total pactuação, as imediatas de Além Paraíba e Carangola apresentaram para todos os municípios. As imediatas de Manhuaçu e Ubá foram as que apresentaram maiores porcentagens de pacientes com baixa escolaridade dos municípios e as cidades de Muriaé, Carangola, Cataguases e Leopoldina foram as cidades que apresentaram as menores proporções. O teste de associação apresentou significância em em 85,71% dos municípios. Para a proporção de casos com o estadiamento classificado como III e IV; 67,85% das ci-

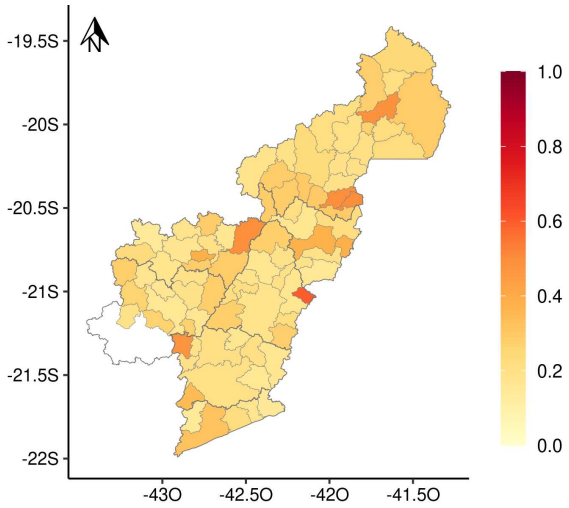
dades apresentaram associação espacial para o estadiamento avançado. Em 3/4 das cidades, os pacientes com estadiamento avançado foram observados em mais de 70% dos casos.

Figura 17: Proporção, ELSA e P-valor do ELSA dos pacientes objeto do estudo acima do peso.

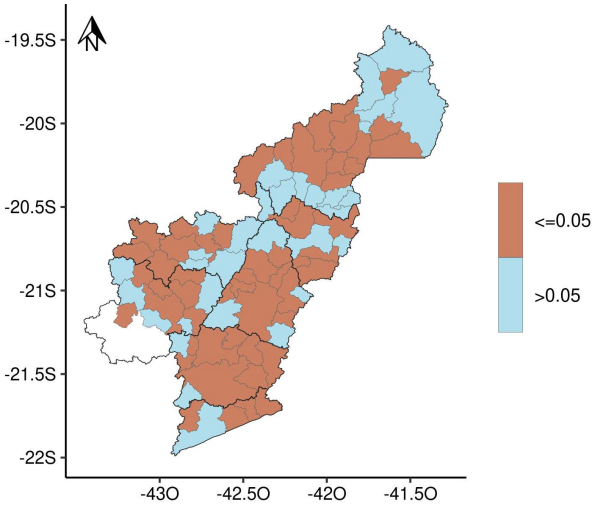
(a) Proporção.



(b) ELSA.



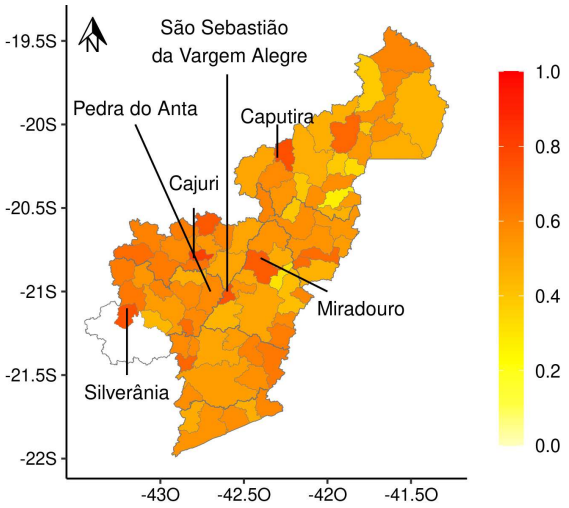
(c) Pvalor.



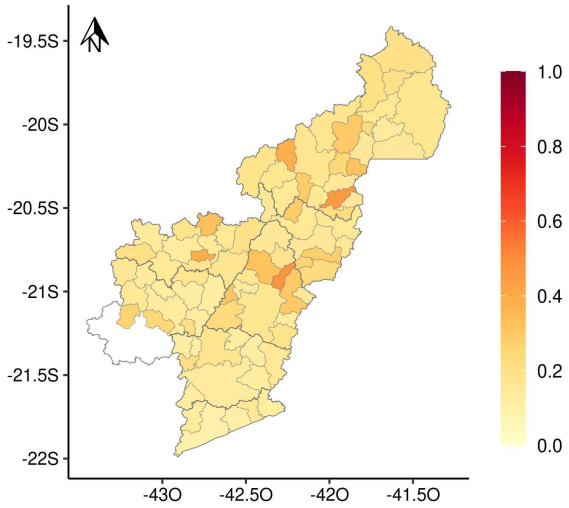
Fonte: Elaborada pelo autor.

Figura 18: Proporção, ELSA e P-valor do ELSA dos pacientes objeto do estudo tabagistas.

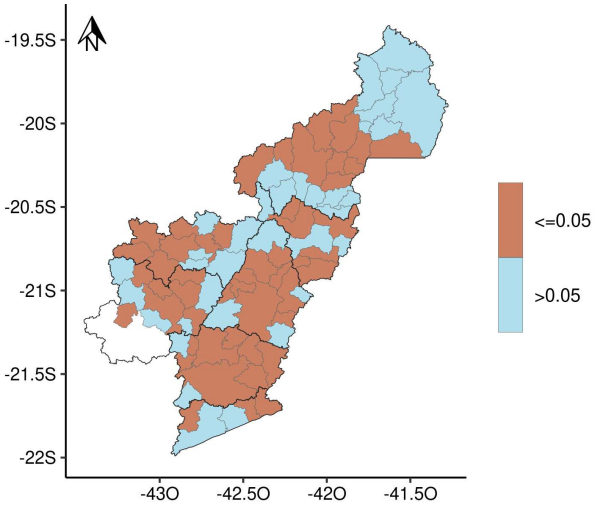
(a) Proporção.



(b) ELSA.



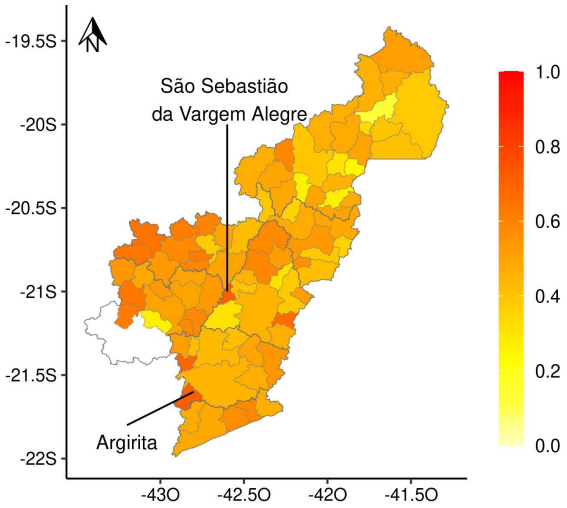
(c) P-valor.



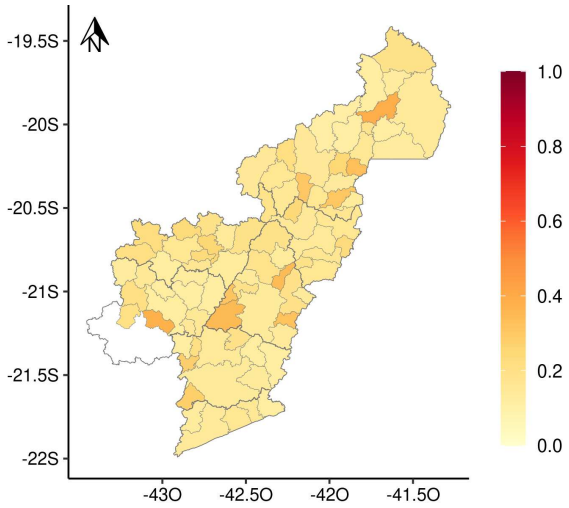
Fonte: Elaborada pelo autor.

Figura 19: Proporção, ELSA e P-valor do ELSA dos pacientes objeto do estudo etilistas.

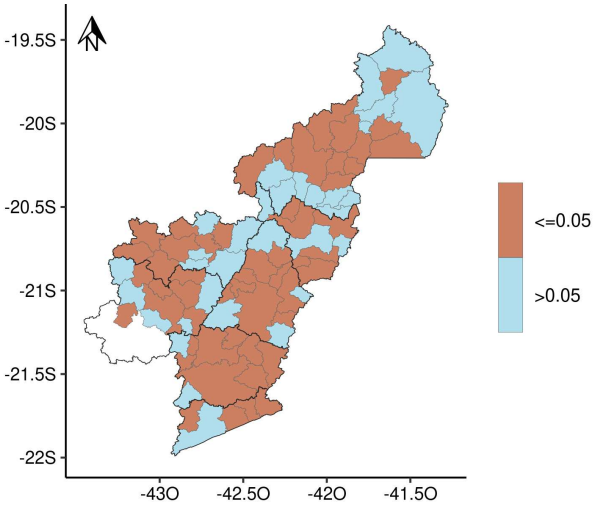
(a) Proporção.



(b) ELSA.



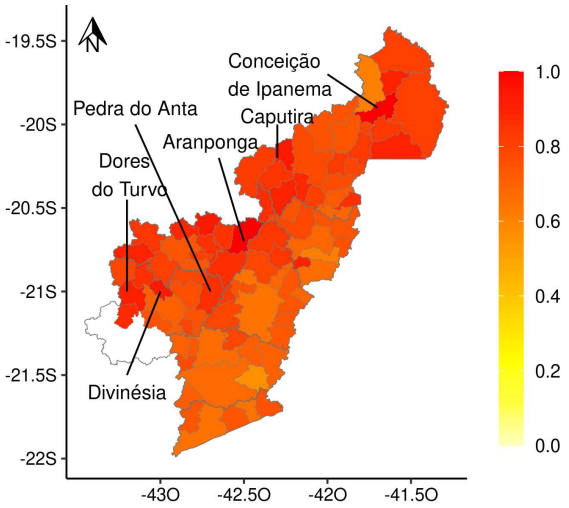
(c) P-valor.



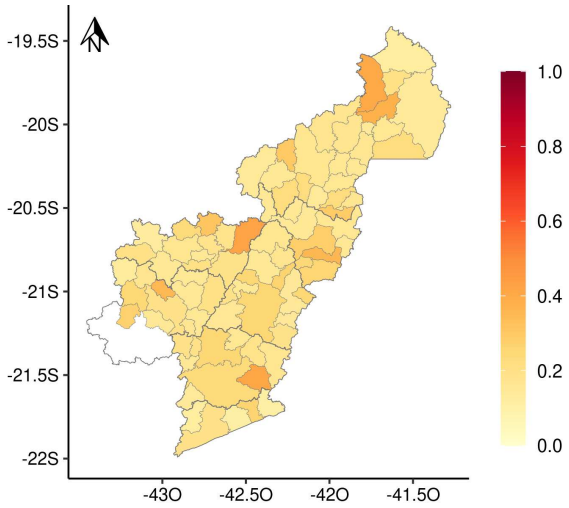
Fonte: Elaborada pelo autor.

Figura 20: Proporção, ELSA e P-valor do ELSA dos pacientes objeto do estudo com baixa escolaridade.

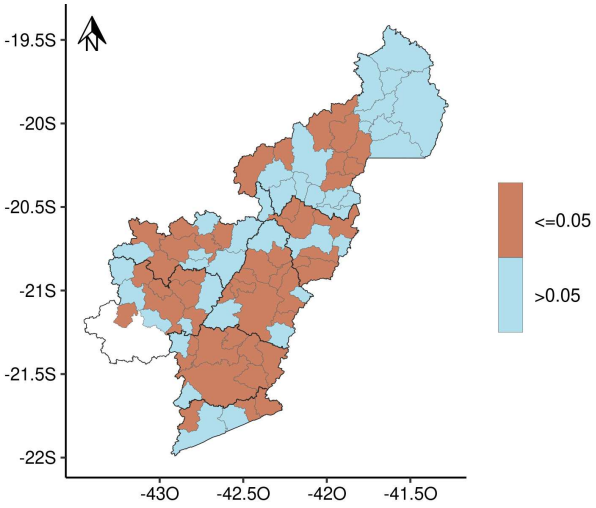
(a) Proporção.



(b) ELSA.



(c) P-valor.

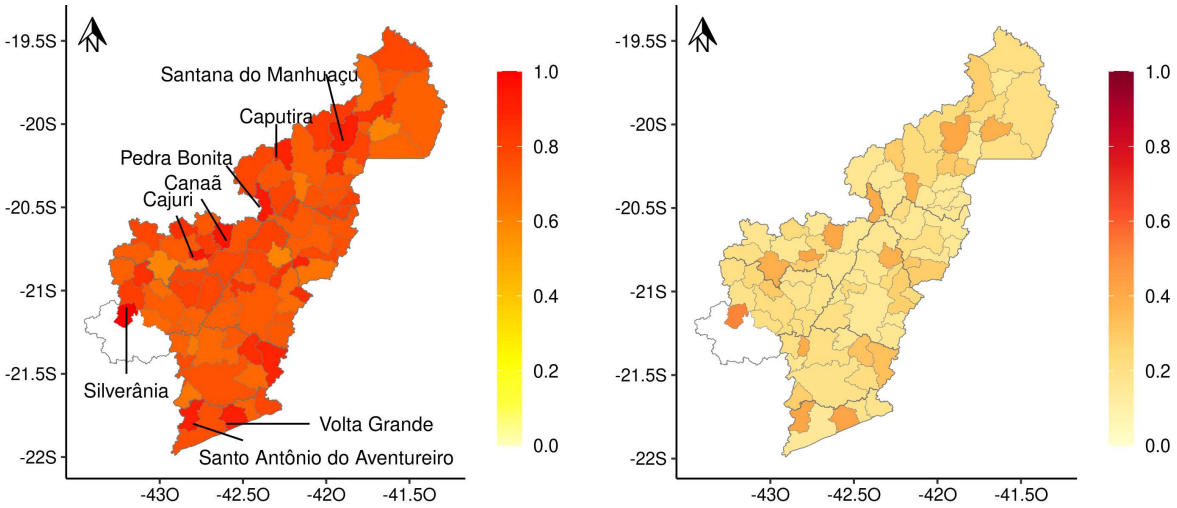


Fonte: Elaborada pelo autor.

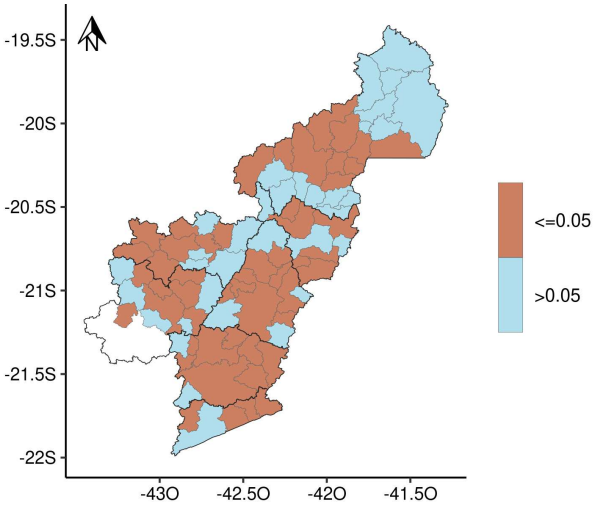
Figura 21: Proporção, ELSA e P-valor do ELSA dos pacientes objeto do estudo com estadiamento avançado.

(a) Proporção.

(b) ELSA.



(c) P-valor.



Fonte: Elaborada pelo autor.

5 CONCLUSÕES

A estimação do sobrepeso/obesidade dos pacientes com tumores nos órgãos digestivos procedentes das cidades com total pactuação com o Hospital do Câncer de Muriaé pelo algoritmo de *Machine Learning Random Forest* obteve um resultado de 85,48% de acurácia, sendo considerado satisfatório para o problema proposto. Das variáveis epidemiológicas utilizadas no estudo Sexo, Raça, Escolaridade e Histórico Familiar de Câncer apresentaram significância estatística com a letalidade em até 3 anos, a baixa escolaridade (até o ensino fundamental incompleto) dos pacientes apresentou a pior taxa sendo a característica de maior risco. Dos fatores de risco, comparando com as categorias das mesma variável, homens, pretos, baixa escolaridade e solteiros são mais em proporção etílicos e tabagistas, do estadiamento em estado avançado homens, pretos, baixa escolaridade e solteiros resultaram nas piores taxas. Mulheres, brancos, alto grau de instrução e pessoas separadas judicialmente apresentaram as piores taxas de acima do peso. O estadiamento e a letalidade entre as variáveis de risco, etílicos e tabagistas apresentaram os piores resultados para o estadiamento avançado e o óbito.

Da distribuição das proporções dos pacientes acima do peso, tabagistas, etilistas, com baixa escolaridade e estadiamento pelos municípios, todas essas variáveis apresentaram alta associação espacial, e a maioria dos municípios resultou em uma significância estatística. As cidades menos desenvolvidas tiveram as maiores proporções de tabagismo, etilismo, baixa escolaridade e estadiamento. As regiões imediatas de Ubá e Viçosa foram as que apresentaram maior proporção de tabagistas e etilistas do que as outras imediatas. A estatística ELSA e o teste de aleatorização *bootstrap* não paramétrica para testar a associação espacial foram adequadas para identificar as regiões com as classes de risco e as mais críticas para o estadiamento avançado e óbito dos pacientes com tumores de órgãos digestivos.

Como recomendações para pesquisas futuras, a utilização do *Random Forest* para estimar outras variáveis que são fatores de risco para tumores nos órgãos digestivos, utilizar outros algoritmos de *Machine Learning* para encontrar melhores acurácias que a encontrado nesse estudo. Utilizar desses métodos para outros tipos de tumores e/ou em outras instituições. E utilizar a análise de sobrevida para as variáveis estimadas.

Referências

ANSELIN, Luc. Local Indicators of Spatial Association—LISA. **Geographical Analysis**, v. 27, n. 2, p. 93–115, 1995. DOI: https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

BIVAND, Roger; KEITT, Tim; ROWLINGSON, Barry. **rgdal: Bindings for the 'Geospatial' Data Abstraction Library**. [S.l.], 2021. R package version 1.5-27. Disponível em: <https://CRAN.R-project.org/package=rgdal>.

BREIMAN, L. Bagging predictors. **Machine Learning**, 1996. DOI: 10.1007/BF00058655.

_____. Random Forests. **Machine Learning**, v. 45, p. 5–32, out. 2001. DOI: 10.1023/A:1010950718922.

BREIMAN, Leo et al. **Classification And Regression Trees**. [S.l.: s.n.], out. 1984. p. 1–358. ISBN 9781315139470. DOI: 10.1201/9781315139470.

CAMERON, Donald; JONES, Ian G. John Snow, the Broad Street Pump and Modern Epidemiology. **International Journal of Epidemiology**, v. 12, n. 4, p. 393–396, jan. 1983. ISSN 0300-5771. DOI: 10.1093/ije/12.4.393. Disponível em: <https://doi.org/10.1093/ije/12.4.393>.

CONSOLI, Sergio; REFORGIATO RECUPERO, Diego; PETKOVIC, Milan. **Data Science for Healthcare: Methodologies and Applications**. [S.l.: s.n.], jan. 2019. ISBN 978-3-030-05248-5. DOI: 10.1007/978-3-030-05249-2.

DALIANIS, Hercules et al. HEALTH BANK - A workbench for data science applications in healthcare. **CEUR Workshop Proceedings**, v. 1381, p. 1–18, jan. 2015.

DAVISON, Anthony; HINKLEY, D. Bootstrap Methods and Their Application. **Journal of the American Statistical Association**, v. 94, jan. 1997. DOI: 10.2307/1271471.

DOBARCO, Mercedes Román et al. A modelling framework for pedogenon mapping. **Geoderma**, v. 393, p. 115012, jul. 2021. DOI: 10.1016/j.geoderma.2021.115012.

DRUCK, S. et al. **Análise Espacial de Dados Geográficos**. [S.l.]: Planaltina: Empresa Brasileira de Pesquisa Agropecuária, 2004. p. 208. ISBN 85-7883-260-6.

ESTRIN, D. Small data, where n = me. **Communication of the ACM**, 2014. DOI: https://doi.org/10.1145/2580944.

HINO, P. et al. Geoprocessamento aplicado à área da saúde. **Revista Latino-Americana de Enfermagem**, São Paulo (SP), 2006.

IBGE. **IBGE - Malha Municipal**. Dezembro 2021. Disponível em: <https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html?=&t=downloads>. Acesso em: 16 nov. 2021.

INCA. **Como surge o câncer?** Nov. 2021. Disponível em: <https://www.inca.gov.br/como-surge-o-cancer>. Acesso em: 16 nov. 2021.

_____. **O que é estadiamento?** Dezembro 2021. Disponível em: <https://www.inca.gov.br/estadiamento>. Acesso em: 16 nov. 2021.

_____. **Registros hospitalares de câncer: planejamento e gestão**. Rio de Janeiro (RJ): [s.n.], 2010. p. 1–538. ISBN 9781315139470.

_____. **Tipos de Câncer: Câncer de Estômago**. Nov. 2021. Disponível em: <https://www.inca.gov.br/tipos-de-cancer/cancer-de-estomago>. Acesso em: 16 nov. 2021.

IPEA. **Índice de desenvolvimento humano.** Dezembro 2021. Disponível em: <https://www.ipea.gov.br/ipeageo/bases.html>. Acesso em: 16 nov. 2021.

JAMES, Gareth et al. **An Introduction to Statistical Learning with Applications in R**. New York, NY: [s.n.], 2013. ISBN 978-1-4614-7138-7. DOI: https://doi.org/10.1007/978-1-4614-7138-7.

KELLEHER, John D.; TIERNEY, Brendan. **Data Science**. Cambridge, MA: MIT Press, 2018. (MIT Press Essential Knowledge Series). ISBN 978-0-262-53543-4.

KRIGE, D.G. A statistical approach to some basic mine valuation problems on the Witwatersrand. **Journal of the Southern African Institute of Mining and Metallurgy**, v. 52, n. 6, p. 119–139, 1951. DOI: 10.10520/AJA0038223X_4792.

KRUSKAL, William H.; WALLIS, W. Allen. Use of Ranks in One-Criterion Variance Analysis. **Journal of the American Statistical Association**, Taylor Francis, v. 47, n. 260, p. 583–621, 1952. DOI: 10.1080/01621459.1952.10483441.

LEUNG, Carson K. et al. Data Science for Healthcare Predictive Analytics. In: (IDEAS '20). ISBN 9781450375030. DOI: 10.1145/3410566.3410598. Disponível em: <https://doi.org/10.1145/3410566.3410598>.

MANN, H. B.; WHITNEY, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 18, n. 1, p. 50–60, 1947. DOI: 10.1214/aoms/1177730491.

MATHERON, G. Principles of geostatistics. **Economic Geology**, v. 58, n. 8, p. 1246–1266, dez. 1963. DOI: 10.2113/gsecongeo.58.8.1246. Disponível em: <https://doi.org/10.2113/gsecongeo.58.8.1246>.

_____. The Theory of Regionalized Variables and Its Applications. **T. Les Cahiers du Centre de Morphologie Mathematique in Fontainebleu**, Paris, France, 1971.

MAYER, Michael. **missRanger: Fast Imputation of Missing Values**. [S.l.], 2021. R package version 2.1.3. Disponível em: <https://CRAN.R-project.org/package=missRanger>.

MONTGOMERY, E. G. Experiments in wheat breeding: experimental error in thenursery and variation in nitrogen and yield. Dept. Agric., Washington, U.S., v. 1381, p. 61, 1913.

MORRISON, Donald G. Regressions with Discrete Dependent Variables: The Effect on R2. **Journal of Marketing Research**, v. 9, n. 3, p. 338–340, 1972. DOI: 10.1177/002224377200900318.

NAIMI, Babak et al. ELSA: An Entropy-based Local indicator of Spatial Association. **Spatial Statistics**, v. 29, p. 66–88, 2019. DOI: 10.1016/j.spasta.2018.10.001.

NUIJTEN, Rik et al. Monitoring the Structure of Regenerating Vegetation Using Drone-Based Digital Aerial Photogrammetry. **Remote Sensing**, v. 13, p. 1942, mai. 2021. DOI: 10.3390/rs13101942.

OMS. **Cancer**. Nov. 2021. Disponível em: <https://www.who.int/health-topics/cancer>. Acesso em: 16 nov. 2021.

PEARSON, Karl. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, Taylor Francis, v. 50, n. 302, p. 157–175, 1900. DOI: 10.1080/14786440009463897.

PINTO, E. S. O. **Estatística Espacial Aplicada na Caracterização de Áreas de Riscopara Hipertensão e Diabetes no Estado de Minas Gerais**. 2013. Diss. (Mestrado) – Universidade Federal de Viçosa, Viçosa - MG.

PORTO, Fabio; ZIVIANI, Artur. Ciência de Dados. **Laboratório Nacional de Computação Científica (LNCC)**, Petrópolis (RJ), 2014.

QUESTER, Pascale; DION, Emanuel. Scaling Numerical Variables and Information Loss: An Appraisal of Morrison's Work. **MARKETING BULLETIN-DEPARTMENT OF MARKETING MASSEY UNIVERSITY**, Massey University, v. 8, p. 59–65, 1997.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>.

RICKETTS, Thomas C. Geographic Information Systems and Public Health. **Annual Review of Public Health**, v. 24, n. 1, p. 1–6, 2003. DOI: 10.1146/annurev.publhealth.24.100901.140924.

SMITH, L. H. Plot arrangement for variety experiment with corn. **Proc. Am. Soc.Agron.**, v. 1381, 1:84–89, 1910.

TING, Kai Ming. **Encyclopedia of machine learning**. [S.l.: s.n.], 2011. ISBN 978–0–387–30164–8.

VIEIRA, S. R. **Tópicos em Ciência do Solo**. Viçosa (MG): [s.n.], 2000.

WALLER, L. A.; GOTWAY, C. A. **Applied Spatial Statistics for Public Health Data**. Canada: John Wiley Sons, Inc, 2004.

WICKHAM, Hadley et al. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019. DOI: 10.21105/joss.01686.

WRIGHT, Marvin N.; ZIEGLER, Andreas. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. **Journal of Statistical Software**, v. 77, n. 1, p. 1–17, 2017. DOI: 10.18637/jss.v077.i01.