ANÁLISE DOS MODELOS HÍBRIDOS NA DISTRIBUIÇÃO DE VIAGENS. CASO: MEDELLÍN – COLÔMBIA

César Alfonso Parada Sánchez

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Transportes, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Transportes.

Orientador: Marcelino Aurélio Vieira da Silva

Rio de Janeiro

Março de 2021

ANÁLISE DOS MODELOS HÍBRIDOS NA DISTRIBUIÇÃO DE VIAGENS. CASO: MEDELLÍN - COLÔMBIA

César Alfonso Parada Sánchez

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE TRANSPORTES.

Orientador: Marcelino Aurélio Vieira da Silva

Aprovada por: Prof. Marcelino Aurélio Vieira da Silva
Prof. Bruno Morais Lemos
Prof. Claudio Falavigna

RIO DE JANEIRO, RJ - BRASIL
MARÇO DE 2021

Parada Sánchez, César Alfonso

Análise dos modelos híbridos na distribuição de viagens. Caso: Medellín – Colômbia/ César Alfonso Parada Sánchez. – Rio de Janeiro: UFRJ/COPPE, 2021.

XII, 94 p.: il.; 29,7 cm.

Orientador: Marcelino Aurélio Vieira da Silva

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Transportes, 2021.

Referências Bibliográficas: p. 76-82.

1. Modelos híbridos. 2. Distribuição de viagens. 3. Redes Neurais Artificiais. I. Silva, Marcelino Aurélio Vieira da. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Transportes. III. Título.

DEDICATÓRIA

A minha mãe Maria D.,

ao meu pai José N., as minhas irmãs Angélica e Sara e

ao meu sobrinho Juan Sebastian.

“There should be no obstacles to accessing knowledge.”

Alexandra Elbakyan

AGRADECIMENTOS

Gostaria de agradecer a minha família pelo apoio e compreensão recebidos durante toda a minha vida, especialmente quando decidi empreender esta aventura pelo Brasil.

Ao Professor Marcelino Aurélio Vieira da Silva por ter acreditado em mim, pelo apoio e grande disponibilidade neste processo.

Aos Professores Bruno Morais Lemos, Cláudio Falavigna e Licínio Portugal por terem aceitado conformar a minha banca, além de suas contribuições permitirem o melhoramento deste trabalho.

Ao meu colega de mestrado Filipe Nascimento pela ajuda incondicional e desinteressada, esclarecendo dúvidas de programação com sua vasta experiência.

Aos meus amigos Guida, Evelyn, Brandão, Cardoso, Gabriella, Vinicius e Solano por terem estado nos momentos que precisava e por terem me recebido com os braços abertos.

Ao corpo docente e demais funcionários do Programa de Engenharia de Transportes da UFRJ, especialmente à Jane e Dona Helena.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES pelo financiamento desta pesquisa, possibilitando seu desenvolvimento.

A todas as pessoas que de alguma ou outra forma colaboraram para que eu cumprisse este objetivo.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

ANÁLISE DOS MODELOS HÍBRIDOS NA DISTRIBUIÇÃO DE VIAGENS. CASO: MEDELLÍN – COLÔMBIA

César Alfonso Parada Sánchez

Março/2021

Orientador: Marcelino Aurélio Vieira da Silva

Programa: Engenharia de Transporte

No planejamento de transportes, é importante compreender a forma como os deslocamentos são realizados em uma região. Informações que posteriormente permitirão estabelecer uma política de transportes adequada. Para tal, é pertinente conhecer o comportamento dos indivíduos. Esta pesquisa estuda o desempenho de redes neurais a partir da compreensão das origens das viagens, sendo apresentada uma nova abordagem por meio da estruturação de um modelo híbrido que incorpora dados agregados e desagregados. Além das informações próprias do indivíduo (socioeconômicas) e de elementos dissuasores como o tempo de viagem e oportunidades intervenientes. Serão também utilizadas características do entorno (uso do solo) como localização dos pontos de origem, dispersão das atividades (destino) e propósito da viagem. Os resultados mostram que a realização da viagem pode ser explicada por meio da combinação de variáveis agregadas e desagregadas e que a omissão dessas últimas pode levar a estimativas com altos níveis de erro. Por outro lado, observou-se que, para cada modelo estruturado, a atração da zona de destino e as impedâncias possuem relevância nas estimativas das viagens.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

ANALYSIS OF HYBRID MODEL OF THE TRAVEL DISTRIBUTION. CASE: MEDELLIN – COLOMBIA

César Alfonso Parada Sánchez

March/2021

Advisor: Marcelino Aurélio Vieira da Silva

Department: Transportation Engineering

In transportation planning, it is important to understand the way that displacement is carried out in a region. Posteriorly, the information will allow to establish a transport policy. For that, it is relevant to known the individual's behaviors. This research studies the performance of neural networks from the understanding the origins of the journeys, by presented a new approach through the structure of a hybrid model which incorporates aggregate and disaggregate data. In addition, the individual information (socioeconomic) and the dissuasive elements such as time travel and intervening opportunities, will also be used the characteristics around the area (land use) such as location of points of origin, activities dispersal (destinations) and purpose of travel. The results indicates that the travel can be explained by the combination of aggregate and disaggregate variables, and the omission of theses last ones could lead the estimation to high levels of error. On the other hand, it was noticed that each structured model the attraction of destination zone and the impedances are relevant in the travel estimates.

Sumário

Lista de Figuras

Lista de Tabelas

1. INTRODUÇÃO

1.1. Contextualização

Conhecer o padrão de deslocamento populacional é primordial no planejamento estratégico urbano e de transportes em uma determinada região. Para isso, deve-se compreender os fatores que influenciam as escolhas dos indivíduos.

O ser humano baseia a tomada de decisão na experiência diária, ponderando as alternativas que possui e comparando-as com as condições próprias da cotidianidade. Um claro exemplo disto é a maneira de realizar viagens na cidade.

A realização de uma viagem, como primeira medida, depende do objeto ou razão que a satisfaz. Segundo Ortuzar & Willumsen (2011), as viagens são realizadas com o objetivo de satisfazer uma necessidade (trabalho, lazer, educação, saúde, entre outras) e, em consequência, a demanda por transporte deve ser abordada mediante a compreensão da distribuição espacial dessas atividades.

Além desse aspecto, a viagem é influenciada por condições operativas do sistema (alternativa modal existente no lugar de origem), temporais (tempo que está disposto a gastar com deslocamentos, faixa horária do dia e sazonalidade), geoespaciais, socioeconômicas e demográficas (ORTUZAR & WILLUMSEN, 2011).

Nesse sentido, o estudo das viagens poderia ser considerado a base do planejamento de transportes e urbano, sendo importante sua compreensão no cenário atual quanto à predição destas em um horizonte de tempo dado.

Para Vuchic (2005), existem quatro fatores que auxiliam no entendimento para se estimar as viagens indiferentemente do modo de transporte, são eles:

- O grau de precisão da estimativa depende da validez e da confiabilidade dos dados utilizados como crescimento ou decrescimento populacional, desenvolvimentos econômicos, hábitos e atitudes da população entre outras informações;

- Essas predições serão válidas enquanto as condições dos cálculos dos modelos permanecerem constantes, pois o procedimento de planejamento baseia-se em modelos que representam as atuais condições e relacionamentos;
- O volume de viagens em um sistema não apenas é influenciado por fatores externos, também depende da concepção ou do tipo de sistema de transporte existente na cidade;
- O sistema de transporte possui a propriedade de alterar as condições do uso do solo e dessa maneira alterar os padrões de viagens, portanto, a concepção desse tipo de modelo deve ser considerada como um elemento interdependente do planejamento urbano nas cidades.

Ao longo dos anos, foram desenvolvidas técnicas de planejamento de transportes que procuravam prever o comportamento de um grupo de viajantes na rede de transporte. A maioria delas estruturam as escolhas feitas pelos viajantes mediante a sequenciação básica de escolhas discretas, como é o caso do modelo quatro etapas (THOMPSON, 2019).

Da mesma forma, foram criadas técnicas de mineração de dados que aliadas com o avanço computacional têm permitido o processamento e a análise de grandes volumes de dados, uma delas são as redes neurais artificiais.

Neste trabalho, é apresentada uma nova abordagem por meio da estruturação de um modelo híbrido que incorpora dados agregados e desagregados. Por um lado, serão utilizadas informações próprias do indivíduo (socioeconômicas), elementos dissuasores como o tempo de viagem e oportunidades intervenientes e, por outro, características do entorno (uso do solo) como localização dos pontos de origem e destino e dispersão das atividades econômicas e finalmente o propósito da viagem.

Diante do exposto, este trabalho pretende responder a dois questionamentos: Primeiro, por meio do uso de informação mista (combinação de variáveis agregada e desagregada), pode-se explicar a distribuição de viagens existente em uma determinada região? Segundo, em que nível essas informações influenciam com maior grau os padrões de viagens?

A partir de testes dos modelos nos níveis macro (desempenho global do modelo), meso (análise de classes) e micro (ranqueamento de classes), pretende-se obter a resposta desses questionamentos.

1.2. Objetivos geral e específicos

O desenvolvimento desta pesquisa busca gerar um procedimento para a criação e aplicação de um modelo híbrido baseado em redes neurais na distribuição de viagens, estritamente para viagens com motivo trabalho e realizadas por meio do sistema de transporte público coletivo. Para isso, serão desenvolvidos os seguintes objetivos específicos:

- Estudar a influência das características socioeconômicas do indivíduo e a impedância na escolha do destino da viagem;
- Verificar a importância das variáveis agregadas em relação às desagregadas.

1.3. Justificativa

O modelo clássico referente ao planejamento de transportes é o modelo quatro etapas que, por meio da execução de passos sequenciais, visa a estimar a demanda futura de um sistema de transporte assim como o seu desempenho (MCNALLY, 2016).

Cada uma dessas etapas gera informação de saída utilizada como insumo da fase posterior, porém, essa falta de integração entre etapas foi considerada com um dos principais problemas do modelo.

Para Bazzan & Klügl (2005), essa situação impede uma modelagem consistente devido à informação chave ficar bloqueada no passo de uma etapa a outra. Exemplificando a situação anterior, na fase de geração de viagem conta-se com informação relevante concernente a características socioeconômicas da população, uso do solo, motivo da viagem, entre outras. No entanto, na seguinte etapa (distribuição de viagens), apenas é usada informação referente ao ponto origem *i* e ponto final *j* da viagem.

Ao longo dos anos, as técnicas associadas aos modelos de quatro etapas melhoraram, porém, permanece uma limitação fundamental, pois o processo de tomada de decisão associado ao comportamento de viagem individual não está bem representado na abordagem estruturada (ITE, 2016).

De acordo com Thompson (2019), embora o processo seja mostrado como sendo de natureza serial, sabe-se que os viajantes não podem tomar cada uma dessas decisões, uma de cada vez, nesta ordem particular, ou independente de outras viagens.

Com relação à etapa de distribuição de viagens, foco desta pesquisa, Bruton (1979) classificou os modelos em dois grupos. O primeiro deles é denominado como métodos análogos, nos quais o uso de fatores de crescimento é fundamental na estimativa de viagens. O segundo grupo chamado de modelos sintéticos buscam, mediante a compatibilização de leis do comportamento físico e o estudo das causas que originam os deslocamentos, prever os padrões de viagens de uma região (BRUTON, 1979).

Esses últimos possuem grande popularidade devido a seu uso recorrente e à base matemática e teórica que trazem consigo. A esse grupo pertencem o Modelo Gravitacional (MG), Modelo de Oportunidades Intervenientes (MOI) e o Modelo de Destino Limitados, também conhecido como Modelo Gravitacional de Oportunidades, que combina as melhores características dos modelos supracitados.

Cabe destacar que Ortuzar & Willumsen (2011) mantêm a mesma categorização para os modelos sintéticos. Já o nome do grupo de modelos análogos foi mudado para métodos de fator de crescimento e, finalmente, foi criado um conjunto adicional no qual se aborda a maximização da entropia que tem sido usada na geração de uma ampla gama de modelos, incluindo o modelo gravitacional, modelos de compras e modelos de localização.

Richards (1974) indica que esses tipos de modelos possuem um problema que envolve o nível de agregação do modelo e a relação entre variáveis. Esse autor acrescenta que “se a relação entre a variável dependente e as variáveis independentes for não linear, a forma funcional e os coeficientes provavelmente variarão com diferentes sistemas de

agregação. Assim, um modelo agregado que não é linear nessas variáveis só é válido para o sistema de agregação para o qual foi estimado".

Adicionalmente o modelo gravitacional assume uma distribuição espacial homogênea das oportunidades e consequentemente superestima o número de viagens de longa distância (VEENSTRA *et al.*, 2010), ademais apresenta uma alta taxa de erro em comparação com os valores observados, levando à superestimação ou subestimação da demanda (AGUIAR JÚNIOR, 2004).

Outra característica a ser considerada nesses tipos de modelos é que se destacam por ter uma abordagem agregada. A informação agregada com a qual trabalham os modelos estatísticos tradicionais limita o poder preditivo dos mesmos porque homogeneíza as características próprias de cada unidade zonal (zona de tráfego) da área em estudo, facilitando dessa maneira a realização dos cálculos em detrimento da precisão do modelo.

De acordo com Richards (1974), o uso de dados agregados implica em uma média sobre as unidades comportamentais individuais dentro, por exemplo, da zona de tráfego, e tal procedimento resulta em uma grande perda de variabilidade, a menos que todos os indivíduos dentro de uma zona sejam perfeitamente homogêneos nas características relevantes para a demanda de viagens.

Além disso, o nível de agregação influi na perda de informação, podendo afetar nas demais etapas do modelo sequencial. Segundo Manout & Bonnel (2018), um alto nível de agregação produz a omissão das viagens intrazonais que, consequentemente, não são consideradas na etapa de alocação de fluxo, gerando resultados subestimados no processo de planejamento do transporte.

Para Pitombo *et al.* (2017), os modelos de escolha de destino ou de distribuição de viagens de abordagem agregada usualmente ignoram sua origem individual. Espera-se que com a utilização de modelos híbridos, como o proposto neste trabalho, apresentar uma solução a esse problema.

Em alternativa, os modelos com abordagem desagregada se centram na interpretação de comportamentos e atitudes individuais, além de reduzir a perda de informação em relação aos modelos tradicionais. Segundo Nazem (2014), a abordagem desagregada possibilita melhor consideração dos modelos dinâmicos e modelos que levam em conta as características sociodemográficas.

Para Ghasri *et al.* (2017), os modelos desagregados, além de capturar o comportamento real dos indivíduos, possuem uma maior sensibilidade a mudanças no entorno, pois o impacto sobre as decisões do indivíduo podem ser estimadas.

Em contrapartida, esse tipo de modelo requer maior número de dados e, consequentemente, a coleta dessas informações acarretaria maiores investimentos.

Com o passar do tempo, foram desenvolvidos diversos modelos em prol da análise da dinâmica de deslocamentos da população. Modelos cada vez mais complexos procuram simular de acordo com a realidade esses padrões de viagens a fim de realizar estimativas com maior grau de exatidão.

Esses modelos baseados em novas tecnologias, como técnicas de mineração de dados ou de aprendizado de máquinas, apresentam uma série de vantagens frente aos modelos de estimativa tradicionais. A principal delas é a identificação de comportamentos que não possuem natureza linear (AGUIAR JÚNIOR, 2004; GONÇALVES, *et al.*, 2015).

Para Roma *et al.* (2017), as técnicas de aprendizado de máquinas têm a facilidade de identificar padrões de maneira ágil. Além disso, possuem maior flexibilidade na hora de usar qualquer tipo de variável de entrada devido ao fato de que essas técnicas não partem de suposições matemáticas rígidas

Mesmo que atualmente existam várias publicações acadêmicas que tentem vislumbrar o processo de tomada de decisão das pessoas no momento de realizar a viagem e como este é influenciado pelo entorno e condições próprias de cada pessoa, ainda existe muito a ser estudado.

Assim, pode-se observar que os modelos tanto desagregados como agregados possuem vantagens que possibilitam a estimativa de viagens em maior ou menor grau. Com a utilização de informações desses modelos, procura-se entender o comportamento nos deslocamentos de Medellín.

1.4. Delimitação da pesquisa

Esta pesquisa será orientada ao estudo da distribuição das viagens realizadas por transporte público coletivo dentro da zona urbana da cidade de Medellín – Colômbia. Para tal fim, serão utilizados os dados coletados na pesquisa origem-destino realizada no Vale de Aburrá em 2012.

Na realização da presente pesquisa, serão consideradas apenas as viagens realizadas por motivo de trabalho. Essa medida foi adotada porque se pretende utilizar a informação relativa ao número de empregos como oportunidade interveniente.

Perante a ausência de informação referente ao número de empregos fora da zona urbana, serão descartadas as viagens realizadas com origem e/ou destino nas zonas rurais. No entanto, os dados obtidos terão que ser submetidos a um processo de desagregação, procedimento que será desenvolvido no Capítulo 4.1.

Da mesma forma, no desenvolvimento da pesquisa só serão analisadas as viagens realizadas por meio do sistema de transporte público coletivo. Nesse sentido, serão considerados os deslocamentos feitos por ônibus, micro-ônibus, BRT e metrô como modo principal.

1.5. Estrutura da pesquisa

Esta dissertação será desenvolvida em seis capítulos, tal como se mostra a seguir:

INTRODUÇÃO: Este capítulo apresenta o contexto do tema de estudo, problema da pesquisa, os objetivos gerais e específicos, justificativa, os resultados esperados e a estrutura da pesquisa.

MODELOS DE DISTRIBUIÇÃO DE VIAGENS: Em seguida, é estabelecido o estado da arte dos modelos no marco da segunda etapa do modelo sequencial.

PROCEDIMENTO METODOLOGICO: Aborda o procedimento metodológico desenvolvido, sendo subdividido em cinco etapas, como a determinação da área de estudo, o levantamento dos dados, o tratamento dos dados, a construção dos modelos e finalmente a constituição da arquitetura da rede neural.

APLICAÇÃO DO PROCEDIMENTO METODOLOGICO: Serão seguidas as pautas estabelecidas no Capítulo 3. Além de realizar um estudo de caso na cidade de Medellín – Colômbia, serão apresentados os resultados obtidos a partir de cada modelo.

ANÁLISE DE RESULTADOS: Após a execução prática dos modelos, procede-se a compará-los tanto de maneira individual como coletiva, visando assim a validação dos objetivos apresentados anteriormente.

CONCLUSÕES E RECOMENDAÇÕES: Este capítulo apresenta as conclusões da pesquisa e as recomendações para trabalhos futuros.

2. MODELOS DE DISTRIBUIÇÃO DE VIAGENS

2.1. Modelos de distribuição de viagens

A seguir são descritos os principais modelos de distribuição de viagens. Embora existam diversas maneiras de agrupá-los, optou-se pela forma apresentada no trabalho de Mendonça (2008).

2.1.1. Agregados

Nos modelos agregados, as variáveis ou atributos inseridos no modelo representam um grupo de usuários, por exemplo, os tempos médios ou custos de todas as viagens entre duas zonas de tráfego, ou o número médio de carros pertencentes a famílias de uma determinada classe (CASCETTA, 2009).

Esse tipo de modelo, contudo, apresenta uma limitação de agregação espacial, no qual usualmente são ignoradas as viagens realizadas dentro da mesma zona de tráfego, denominadas intrazonais, podendo gerar problemas em etapas posteriores, como na alocação do tráfego na rede (MANOUT & BONNEL, 2018).

O Modelo Sequencial de Quatro Etapas (MSQE) possui uma grande relevância no planejamento de transportes, sendo um modelo muito utilizado (TEODOROVIC & JANIC, 2017). É composto por quatro fases (geração e distribuição de viagens, divisão modal e finalmente alocação de viagens).

Segundo McNally (2016), esse modelo provê um mecanismo para determinar o equilíbrio dos fluxos em um sistema tal como se apresenta na Figura 2.1. A seguir, cada etapa do modelo será aprofundada.

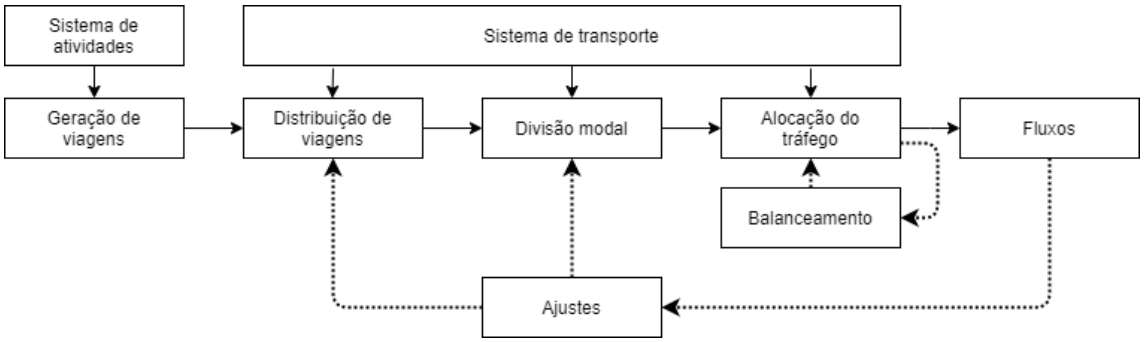

Figura 2.1. Estrutura MSQE.
Adaptado de McNally (2016)

2.1.1.1. Geração de viagens

A etapa de geração de viagens visa a estimar o número de viagens atraídas e produzidas em cada zona de tráfego de uma determinada região. Segundo Bruton (1979), a geração de viagens é influenciada por três fatores:

- Padrão do uso do solo e do desenvolvimento da área de estudo;
- Características socioeconômicas da população que realiza viagens dentro da área de estudo; e
- Natureza, tamanho e capacidade do sistema de transportes da área em estudo.

Usualmente no processo de geração de viagens são usados modelos baseados em regressão linear múltipla, classificação cruzada e fator de crescimento.

2.1.1.1.1. Regressão linear múltipla

Por meio deste modelo prevê-se encontrar uma função linear, ou não linear, que explique a geração futuras das viagens, tanto produzidas quanto atraídas, por meio da associação de fatores que influenciam a geração. A Equação 1 mostra de maneira geral a estrutura do modelo.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \quad (1)$$

Em que:

y é o número de viagens geradas (variável dependente);

x são os fatores que influenciam a geração de viagens podendo ser socioeconômicos e de uso do solo (variáveis independentes); e

a são a constante e coeficientes estimados.

Para a utilização desse tipo de modelo, deve-se considerar a hipótese da existência de multicolinearidade entre as variáveis independentes.

2.1.1.1.2. Classificação cruzada

O objetivo principal deste método é a análise de viagens produzidas por unidade familiar em função ao propósito da viagem, sendo primordial conhecer características como composição e condições econômicas do núcleo familiar (ORTUZAR & WILLUMSEN, 2011). Na Equação 2 se apresenta de maneira geral a estrutura do modelo.

$$t_{ij} = R_{ij} * H_{ij} \quad (2)$$

Em que:

t_{ij} é o total de viagens geradas por categoria *i, j*;

R_{ij} é a taxa de geração de viagens por categoria *i, j*; e

H_{ij} é o número de residências por categoria *i, j*.

2.1.1.1.3. Fator de crescimento

Este tipo de modelo parte do pressuposto que a estimativa de viagens em um horizonte de tempo dado pode ser expressa mediante o produto de um fator de crescimento com os dados referentes ao ano zero do projeto, desconhecendo as variações exógenas que possam se apresentar. A Equação 3 mostra a fórmula geral do modelo.

$$T_f = T_a * F_c \quad (3)$$

Em que:

T_f é o número de viagens futuras;

T_a é o número de viagens que ocorrem atualmente; e

F_c é o fator de crescimento calculado.

Por outro lado, o fator de crescimento pode ser calculado pela divisão direta do valor futuro sobre o atual da variável explicativa escolhida (ver Equação 4).

$$F_c = \frac{P^f}{P^a} \; ou \; \frac{E^f}{E^a} \quad (4)$$

Em que:

F_c é o fator de crescimento calculado;

P é a população da zona de estudo;

E é o número de empregos existentes na zona de estudo; e

a,f indicam valores atuais e futuros respectivamente.

Com base no conceito que dá origem a esse modelo, foram criados quatro submodelos: fator uniforme, fator médio, Fratar e Detroit.

2.1.1.2. Distribuição de viagens

A segunda etapa, denominada distribuição de viagens, conecta ou organiza as origens e destinos obtidos na etapa anterior em formato de tabela ou matriz (VUCHIC, 2005). Nessa fase, é possível prever um padrão de distribuição de viagens com base em três fatores: a geração de tráfego futura, a atração futura e o padrão de distribuição atual (FAGHRI *et al.*, 1997).

O principal exponente nesta etapa é o modelo gravitacional. No entanto, existem modelos complementares, como o modelo de oportunidades ou modelos de menor complexidade baseados em fatores de crescimento.

2.1.1.2.1. Modelo gravitacional

O modelo gravitacional e seus derivados é considerado modelo sintético. Esses modelos estimam viagens para cada célula da matriz sem usar diretamente o padrão de viagem observado. Provavelmente o primeiro uso rigoroso de um modelo baseado na lei gravitacional de Isaac Newton foi realizado por Casey (1955), que sugeriu tal abordagem para sintetizar viagens de compras e áreas de captação entre cidades de uma região (ORTUZAR & WILLUMSEN, 2011).

$$T_{ij} = \frac{\alpha P_i P_j}{d_{ij}^2} \tag{5}$$

Em que *P* é a população das cidades de origem e destino respectivamente, *d* é a distância que separa as cidades e α é um fator de proporcionalidade (Equação 5).

Wilson (1967) reformulou o modelo gravitacional incorporando fatores de balanceamento, os quais devem ser calculados por meio de iterações. Denominou-se modelo gravitacional duplamente restrito.

$$T_{ij} = A_i B_j O_i D_j \, f(d_{ij}) \tag{6}$$

Em que T_{ij} são os números de viagens produzidas na zona i e atraídas pela zona j; O_i são os números de viagens originados na zona i; e D_j são os números de viagens que convergem à zona j (Equação 6).

$$A_i = \left[\sum_j B_j D_j \, f(d_{ij})\right]^{-1} \tag{7}$$

$$B_j = \left[\sum_i A_i O_i \, f(d_{ij})\right]^{-1} \tag{8}$$

Desenhados como fatores de balanceamento, A_i e B_j garantem que as restrições foram satisfeitas. Além disso, foi incorporado o conceito de impedância *dij*, a qual pode ser medida como distância real, tempo de viagem ou custo generalizado ao transporte *dij* (Equações 7 e 8).

2.1.1.2.2. Modelo de oportunidades intervenientes

O conceito de oportunidades intervenientes foi introduzido por Stouffer em 1940. O qual estabelece que número de pessoas que se desloca a uma certa distância é diretamente proporcional ao número de oportunidades ofertadas naquele destino e inversamente proporcional ao número de oportunidades intervenientes (STOUFFER, 1940). No entanto, foi Schneider (1959) quem, baseado na teoria de Stouffer (1940), aperfeiçoou o modelo mediante a formulação da teoria de distribuição de viagens.

Schneider (1959) mostrou que a probabilidade de uma viagem ser finalizada em qualquer ponto da região é proporcional ao número de oportunidades existentes no ponto final da viagem. Em outras palavras, a viagem, ao ser vinculada intrinsicamente a um motivo, terá seu deslocamento realizado de maneira satisfatória, levando em consideração a acessibilidade relativa existente na zona, tal como se apresenta na Equação 9.

$$P\left[{}^{i}/_{j}\right] = k_i e^{-\lambda W_{ij}}\left(1 - e^{-\lambda S_j}\right) \quad (9)$$

Em que:

$P\left[{}^{i}/_{j}\right]$ é a probabilidade de uma viagem que se origina em i, terminar na zona j;

k_i é o fator de balanceamento;

λ é a constante de probabilidade de uma oportunidade ser aceita em um destino;

W_{ij} é o número de oportunidades que se interpõem entre as zonas i e j; e

S_j é o número de oportunidades na zona j.

Assim, dependendo da localização da oportunidade, esta pode ser escolhida ou rejeitada. Esse modelo parte do suposto que o indivíduo sempre vai realizar a viagem que permita atingir satisfatoriamente a oportunidade mais próxima da sua origem, caso contrário será considerada a oportunidade subsequente em função à proximidade da origem (OKABE, 1977).

De acordo com Cascetta *et al.* (2007), esse tipo de modelo introduz a importância de considerar e reproduzir o processo de ponderação de alternativas por parte do tomador de decisão ao realizar a viagem. No entanto, apresenta como deficiência a ausência explícita nos cálculos de fatores de impedância, como o custo generalizado vinculado ao transporte.

2.1.1.2.3. Modelo gravitacional de oportunidades

O modelo gravitacional de oportunidades surgiu com a visão de juntar as teorias gravitacional com o modelo desenvolvido. Assim, esse novo modelo consideraria a distância como impedância e como um fator de ponderação da ordem na qual as zonas de destino são consideradas pelos indivíduos (WILLS, 1986).

Segundo Lemos (2020), enquanto o modelo gravitacional de Wilson (1967) apresenta deficiências em efeitos intervenientes, o modelo proposto por Wills (1986) possui um problema ao omitir as impedâncias relacionadas com o transporte.

Gonçalves (1992) conseguiu uma versão do modelo gravitacional de oportunidades com menor complexidade na hora da calibração. Ainda segundo a autora, o modelo contempla também os aspectos comportamentais que influenciam na formação dos padrões espaciais de fluxos (ver Equação 10).

$$T_{ij} = A_i B_j O_i D_j e^{-\beta c_{ij} - \lambda w_{ij}} \quad (10)$$

Em que T_{ij} são os números de viagens produzidas na zona i e atraídas pela zona j; O_i são os números de viagens originados na zona i; e D_j são os números de viagens que convergem à zona j.

$$A_i = \left[\sum_j B_j D_j \, e^{-\beta c_{ij} - \lambda w_{ij}}\right]^{-1} \quad (11)$$

$$B_j = \left[\sum_i A_i O_i \, e^{-\beta c_{ij} - \lambda w_{ij}}\right]^{-1} \quad (12)$$

Por outro lado, A_i e B_j são fatores de balanceamento que operam em função do número de viagem atraída e produzida pela zona *i* e *j* respectivamente, sendo *c* o fator de impedância existente nelas e *w* as oportunidades intervenientes entre a zona de origem e destino (Equações 11 e 12).

Já Lemos (2020) incorporou um novo coeficiente α e Θ nas variáveis de produção e atração, avaliando a elipse como nova forma de zoneamento, assim como o aumento na distância por meio da alteração das dimensões do zoneamento δ (ver Equação 13).

$$T_{ij} = A_i B_j O_i^{\propto} D_j^{\Theta} e^{-\beta c_{ij} - \lambda w_{ij}} \quad (13)$$

Esses novos parâmetros são calculados mediante o uso do método de mínimos quadrados (Equações 14 e 15).

$$A_i = \left[\sum_j B_j D_j^{\Theta} \, e^{-\beta c_{ij} - \lambda w_{ij}}\right]^{-1} \quad (14)$$

$$B_j = \left[\sum_i A_i O_i^{\propto} \, e^{-\beta c_{ij} - \lambda w_{ij}}\right]^{-1} \quad (15)$$

2.1.1.3. Divisão modal

Segundo Chatterjee & Venigalla (2004), na etapa de divisão modal, é estabelecido o número de viagens obtido na etapa anterior por cada modo de transporte. Para tal, são estudados os fatores endógenos e exógenos que influenciam o usuário na escolha modal. São eles: características do usuário, da viagem e do sistema de transporte.

Chatterjee & Venigalla (2004) identificam dois tipos de modelos que auxiliam no cálculo da divisão modal. O primeiro deles, denominado modelo agregado, examina a escolha do modo dos viajantes e suas viagens em grupos, com base em características socioeconômicas e/ou de viagem semelhantes, usualmente representadas como curvas de desvios modal. O segundo grupo, denominado modelos comportamentais desagregados logísticos, contém modelos de tipo logístico. Esses modelos partem de duas premissas: a primeira é que a escolha do modo de viajante é tomada de maneira individual e está baseada no princípio denominado maximização da utilidade. A segunda é que a utilidade de se usar um meio de transporte para uma viagem pode ser estimada

O modelo *logit mulinomial* calcula a probabilidade de escolha de uma alternativa (variável dependente) mediante a análise das variáveis independentes sendo expressa pela Equação 16.

$$p(k) = \frac{e^{U_k}}{\sum_x e^{U_k}} \tag{16}$$

Em que:

$p(k)$ é a probabilidade de escolha do modo k;

k é o modo de transporte;

x são todos os modos concorrentes x = 1 ... k; e

U_k é a utilidade do modo k.

Quando as alternativas de transporte se reduzem a um sistema binário, o modelo *logit binomial* pode realizar as estimativas pertinentes, sendo apenas necessário o cálculo da probabilidade do uso de um modo. O resíduo ou complemento fornece a informação do segundo modo, tal como se observa na Equação 17.

$$p(A) = \frac{1}{1 + e^{U_B - U_A}}; p(B) = 1 - P(A) \tag{17}$$

Em que:

$p(A)$ é a probabilidade de escolha do modo A;

$p(B)$ é a probabilidade de escolha do modo B;

U_A é a utilidade do modo A; e

U_B é a utilidade do modo B.

De acordo com Caldas *et al.* (2019) e Hess *et al*. (2010), os modelos *logit multinomial* apresentam restrições matemáticas, como suposições de distribuição populacional, problemas de multicolinearidade e o fato de considerar a independência das alternativas irrelevantes

2.1.1.4. Alocação do tráfego

O processo de alocação do tráfego é atribuir os carregamentos, ou volumes de usuário, em cada segmento de uma rede de transporte, bem como os movimentos de conversão nas interseções da rede, podendo ser o número de veículos, o número do total pessoas, o número de passageiros em trânsito ou quaisquer outras unidades de demanda de viagem, descritas por uma origem e um destino (CHATTERJEE & VENIGALLA, 2004).

2.1.2. Desagregados.

2.1.2.1. Comportamentais

Os modelos comportamentais foram concebidos a partir da teoria do consumidor e buscam a relação entre as necessidade individuais dos usuários com seus deslocamentos dentro do sistema de transporte (MENDONÇA, 2008). Esses modelos partem da premissa que toda decisão tomada pelo individuo envolve fatores racionais e subjetivos. Entretanto, essas características perduram ao longo do tempo, podendo ser consideradas um padrão comportamental.

Visando a cumprir tal objetivo, as preferências dos usuários devem ser mensuradas quantitativamente. E para tal fim é usada uma função de utilidade vinculada às variáveis de custo, tempo, segurança e conforto, as quais representam o nível de serviço. O modelo é representado na Equação 18.

$$P_k = \frac{e^{U_k}}{\sum_{i=1}^{n} e^{U_l}} \tag{18}$$

Em que:

P_k é a probabilidade de o usuário decidir viajar pelo modo *k*;

U_k é a utilidade da alternativa *k*;

U_l é a utilidade da alternativa *l*, para *l* = 1 .... *n*; e

n é o número de alternativas do modo de transporte;

2.1.2.2. Atitudinais

Os modelos atitudinais, também conhecidos como de preferência declarada, visam a identificar possíveis mudanças dos usuários diante de alterações no sistema de transporte existente. Assim como buscam estabelecer relações entre as variáveis comportamentais e atitudinais das pessoas por meio da percepção e da preferência dos usuários (NOVAES, 1986). Esse modelo é representado pela Equação 19.

$$F_k = f(\propto_1 - \propto_2) \tag{19}$$

Em que F_k o número de passageiros do modo k e $(\propto_1 - \propto_2)$ a diferença entre os modos 1 e 2 no atributo α. Tanto os modelos comportamentais quanto os atitudinais são utilizados na distribuição de viagens e na divisão modal.

2.1.3. Técnicas de computação.

Nos últimos anos, a quantidade de dados gerados pela população aumentou consideravelmente em comparação ao século anterior. Devido à evolução das tecnologias

de computação, como algoritmos de aprendizado de máquina, é possível identificar padrões cada vez mais complexos e ocultos nos dados (BHAVSAR *et al.*, 2017).

Assim, essas técnicas estão sendo utilizadas no auxílio de processamento de dados para lidar com um grande número de variáveis, buscando uma melhor compreensão e representação do objeto modelado (PAIVA, 2011)[1].

Além da classificação dos modelos com base no nível de detalhe das informações anteriormente descrito, a seguir são apresentadas algumas técnicas de computação ou de aprendizado de máquinas que atualmente são utilizadas na área de transportes.

2.1.3.1. Árvore de decisão

A árvore de decisão permite a classificação de dados a partir da geração sucessiva de subconjuntos cada vez com maior grau de homogeneização. Essa técnica está composta pelo nó raiz, o qual contém o banco de dados, por nós filhos, que são o produto da classificação da subdivisão do nó raiz, e pelo nó folha ou terminal, o qual é gerado uma vez que a subdivisão cessa.

De acordo com Souza *et al.* (2017), o algoritmo se detém uma vez que algumas das regras a seguir são atendidas.

- Se o tamanho de um nó for menor que o valor mínimo especificado pelo usuário;
- Se a profundidade atual da árvore atingir o valor limite máximo especificado em profundidade;
- Se todos os casos em um nó tiverem valores idênticos da variável dependente;
- Se a divisão de um nó resultar em um nó filho cujo tamanho do nó for menor que o valor mínimo do tamanho do nó filho especificado pelo usuário;
- Se todos os casos em um nó tiverem valores idênticos para cada covariável.

[1] Este documento não foi publicado em revista científica, podendo ser consultado no seguinte link: https://www.sinaldetransito.com.br/artigos/novos_modelos_3.pdf

Existem vários tipos de algoritmos. Entre eles se encontra a árvore de classificação e regressão (CART por suas siglas em inglês) e o ID3. No caso do CART, a divisão e subdivisão é realizada por meio do índice Gini que mede a frequência com que um elemento escolhido arbitrariamente é classificado de maneira errada.

$$Gini(S) = 1 - \sum_{i=1}^{k} - (p_i)^2 \tag{20}$$

Onde P(t) é a probabilidade de S pertencer à classe i (Equação 20).

Na árvore ID3, a medida da quantidade de incerteza presente no banco de dados S é calculada por meio da entropia (ver Equação 21). Da mesma forma, a redução da incerteza que ocorre cada vez que o conjunto é subdividido no atributo A se denomina ganho de informação (ver Equação 22).

$$E(S) = - \sum_{c \in C} P(c)\ log_2 P(c) \tag{21}$$

Em que S o conjunto de dados para o qual está sendo calculada a entropia, C é o conjunto de classes e P(c) é a proporção do número de elementos em c com o número de elementos no conjunto S.

$$IG\ (A\ .S) = E(S) - \sum_{t \in T} P(t)\ E(t)\ ; S = \cup_{t \in T}\ t \tag{22}$$

Em que E(S) é a entropia do conjunto, S, T são os subconjuntos criados a partir do conjunto de divisão S pelo atributo A. Sendo P(t) a proporção do número de elementos em t com o número de elementos no conjunto S e E(t) é a entropia do subconjunto t.

Uma das grandes dificuldades desse tipo de técnica é a instabilidade, tendo em vista que variações mínimas no banco de dados podem produzir uma árvore totalmente diferente à original. Mesmo assim, a árvore de decisão apresenta uma desvantagem em relação aos modelos tradicionais, como o modelo gravitacional. Por ser de natureza não

paramétrica, não é possível estimar a significância dos coeficientes das variáveis explicativas (PITOMBO & GUIMARÃES, 2016).

2.1.3.2. Classificação Bayesiana

O Teorema de Bayes possui um viés diferenciado em comparação com as estatísticas baseadas em frequências. Essa última parte do pressuposto que os dados observados surgem a partir de processos aleatórios. O contrário ocorre com Bayes, pois o valor do parâmetro θ depende dos valores observados y, os quais são fixos (DOBSON & BARNETT, 2013).

A classificação Bayesiana se respalda em três parâmetros (ver Equação 23). Onde $P(\theta|y)$ é a probabilidade de observar θ com base nos dados observados que é y, sendo este termo denominado probabilidade *a posteriori*. A probabilidade condicional é representada pelo fator $P(\theta)$ da equação, que evidencia a probabilidade de ocorrência de θ ante a ausência de qualquer dado observado que é y. E finalmente a probabilidade condicional $P(y|\theta)$ representa a distribuição de probabilidade da variável y quando θ é a classe observada. O símbolo $\propto$ representa "proporcional à".

$$P(\theta|y) \propto P(y|\theta)P(\theta) \tag{23}$$

No planejamento de transportes, vários pesquisadores avaliaram o desempenho do teorema de Bayes mediante a incorporação de modelos complementares. Em 2012, Perrakis *et al.* (2012) incluíram os modelos Poisson-gama e binomial negativo, modelos regularmente usados na presença de dados altamente dispersos. Esse modelo demostrou resultados muito aproximados aos observados, sendo uma alternativa viável às primeiras fases do modelo sequencial de quatro etapas para os casos em que existem dados históricos de origem-destino.

2.1.3.3. Lógica Fuzzy

A Lógica Fuzzy simula a maneira de pensar dos seres humanos. Projetada para se comportar conforme o raciocínio dedutivo, no qual as conclusões são inferidas a partir de

conhecimentos ou experiências prévias, denominado método heurístico (GODOY & SHAW, 2007).

Diferentemente das técnicas de computação tratadas nesta dissertação, a lógica Fuzzy não possui um viés totalmente matemático, sendo possível a utilização de expressões linguísticas. Baseia-se em regras heurísticas de ação-reação: SE – ENTÃO (*IF – THEN* em inglês), tal como se mostra na Equação 24.

$$SE \ < condição > ENTÃO \ < consequência > \tag{24}$$

A Lógica Fuzzy é composta por quatro etapas. A primeira é denominada "fuzzificação", onde os dados de entrada são transformados em suas respectivas variáveis linguísticas. O segundo módulo, denominado base de conhecimento, contém todas as regras SE – ENTÃO do modelo. Na terceira, ocorre a inferência no qual o modelo simula o processo de raciocínio humano, apoiado nas regras condicionais anteriormente estipuladas. Por fim, na quarta etapa, os dados são "desfuzzificados", convertendo o valor da variável linguística para valores numéricos.

Esse tipo de técnica permite desenvolver modelos comportamentais reais, pois, ao se basear em métodos heurísticos, simula a tomada de decisão dos usuários do transporte. Da mesma forma, o modelo de lógica difusa contrasta dos modelos convencionais porque estes consideram a relação não linear existente entre as variáveis de entrada e saída (SALINI *et al.*, 2017).

Complementando a afirmação anterior, Sarkar (2012) destaca que a Lógica Fuzzy permite a modelagem de processos complexos de tráfego e transporte com características de subjetividade, ambiguidade, incerteza e imprecisão.

2.1.3.4. Redes Neurais Artificiais

As redes neurais são modelos matemáticos que foram concebidos a fim de simular o comportamento do cérebro humano visando a resolver problemas de maior

complexidade. Seu funcionamento consta da identificação de características e padrões a partir da repetição contínua de um grupo de dados.

Os modelos baseados em redes neurais usualmente estão compostos por quatro elementos que trabalham como um todo: neurônios, camadas, algoritmo de aprendizagem e função de ativação. Uma rede neural consiste na interligação de processadores chamados neurônios, os quais são ativados de maneira sequencial a partir da recepção de estímulos e cumprimento de determinadas condições (SCHMIDHUBER, 2015).

De maneira geral, uma rede neural está composta por três camadas: entrada, oculta ou intermediária e saída. Por sua vez estão conformadas por uma série de neurônios interligados entre si, funcionando como sinapses, tal como é mostrado na Figura 2.2.

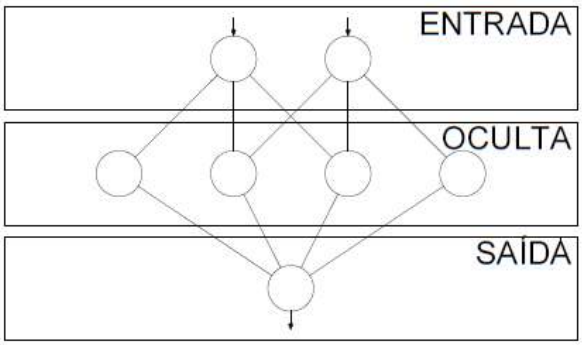

Figura 2.2. Estrutura da rede neural.

Por meio de arcos, os neurônios transmitem dados (x) em função do valor de peso (w) destinado a cada um deles (DOUGHERTY, 1995) (ver Figura 2.3). A Equação 25 apresenta o modelo linear proposto por McCulloch & Pitts (1943), que expressa matematicamente o trabalho efetuado pelo neurônio.

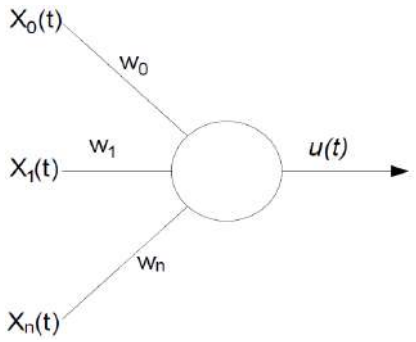

Figura 2.3. Diagrama do neurônio.

Por outro lado, a informação é transmitida à seguinte camada pelo valor de ativação (u(t)) incorporada ao mesmo. Cabe ressaltar que os dados, produtos do resultado do processamento do neurônio, convertem-se em dados de entrada para os neurônios pertencentes à camada subsequente.

$$u(t) = h(z(t)) = h\left(w_0 + \sum_{i=0}^{i=n} W_{ij}\, X_i\right) \tag{25}$$

Neste caso, h representa a função de ativação, sendo z(t) é o potencial de ativação. Quanto a w_0, se denomina bias, tendo um valor unitário fixo.

Ao construir o modelo, determinar o número de neurônios na camada oculta não é tarefa fácil. Segundo a pesquisa realizada por Gonçalves (2015), Lippmann (1987) sugeriu que o número de neurônios da camada intermediária pode ser calculado mediante o acréscimo de uma unidade na quantidade de neurônios na camada de entrada e multiplicado pelo número de neurônios da camada de saída.

Por sua vez, Hercht-Nielsen (1989) indicou que o resultado ao multiplicar duas vezes o número de neurônios da camada de entrada e a sua vez acrescenta-o em um, pode solucionar essa dicotomia.

Por último, Eberhart & Dobbins (1990) propuseram um método matemático que estabelece o número de neurônios da camada intermediária por meio do cálculo resultante da raiz quadrada e do somatório do número total de neurônios existente na camada de

entrada e do número de neurônios da camada de saída. Apesar do anterior, a maioria dos autores optam por determinar esse parâmetro por meio da criação e execução de cenários.

Deve-se considerar que um número insuficiente de neurônios pode sobrecarregar o modelo porque é forçado a realizar a tarefa de encontrar representações ou respostas ótimas levando muito tempo de processamento, tornando-o ineficiente.

Em contrapartida, a utilização excessiva de neurônios produz o fenômeno chamado *overfitting*, que consiste na memorização dos dados por parte do modelo em vez de extrair as características gerais de eles (PINTO FERREIRA *et al.*, 2016). Em consequência, a taxa de acertos diminui para entradas diferentes daquelas utilizadas para a aprendizagem (AGUIAR JÚNIOR, 2004).

O reconhecimento de padrões em modelos baseados em redes neurais deve-se à repetição contínua de um subconjunto de dados (geralmente concebido de maneira aleatória a partir do conjunto maior), gerando um processo de treinamento do sistema.

O processo de treinamento consiste, portanto, na aplicação de passos ordenados que sejam necessários para sintonização dos pesos sinápticos, tendo-se como objetivos final a generalização de soluções a serem produzidas pelas saídas. Esse conjunto de passos ordenados denomina-se algoritmo de aprendizagem (SILVA *et al.*, 2016).

Existem cinco tipos de treinamento que podem ser usados nas redes neurais. O primeiro se denomina treinamento supervisado, onde a partir de ações comparativa, objetiva-se chegar a um conjunto de resultados desejáveis previamente estabelecido.

O treinamento não-supervisionado se caracteriza pela ausência do conjunto de saída de resultados desejáveis. Nesse sentido, o modelo por meio da identificação de similaridades procede à criação de *clusters*, realizando uma tarefa de auto-organização.

Existe o treinamento com reforço, no qual o modelo opera sob a lógica de tentativa-erro. Esse tipo de treinamento baseia-se em métodos estocásticos, no qual suas ações de ajuste são selecionadas probabilisticamente.

A aprendizagem *offline* consiste no treinamento com base no uso de lotes de amostras (padrões), os quais ajustam os pesos sinápticos e limiares do modelo. Tudo a partir da apresentação de um conjunto de dados de treinamento que deve estar presente durante a execução do processo. Caso contrário ocorre com a aprendizagem *online*. Na qual os conjuntos de dados podem ser descartados quando usados.

Baseada no funcionamento do neurônio biológico, a função de ativação regula a passagem de informação de uma camada para outra a partir do nível de estimulação ao qual é sometido. Conforme o tipo de função, usualmente trabalha com valores flutuantes entre intervalos [0 a 1] ou [-1 a 1].

As funções de ativação se encontram classificadas em dois grupos de acordo com o domínio da função. O primeiro são funções parcialmente diferenciáveis encontrando-se degrau, de grau bipolar ou função sinal e rampa simétrica, aliás, o grupo de funções totalmente diferenciáveis está composto por sigmoidal (logística), gaussiana, linear e tangente hiperbólica. esta última tornou-se a ser a mais utilizada devido à natureza continua da função e de sua derivada (SILVA *et al.*, 2016).

Karlaftis & Vlahigianni (2011) compararam os métodos estatísticos com as redes neurais em pesquisas de transportes. Descobriram que, apesar de possuir várias características em comum, existem variações com relação ao enfoque, já que as redes neurais visam a fornecer uma representação eficiente dos dados.

Hirun (2016) percebe que as redes neurais se comportam como uma “caixa preta”, tendo em vista que não é possível determinar a relação existente entre variáveis dependentes e independentes, tornando impossível a formulação de equações.

2.2. Agregação espacial e sua interferência

Diversas pesquisas foram realizadas visando a deslumbrar os efeitos que causa a agregação espacial nos modelos de previsão. A seguir apresentam-se as conclusões obtidas por alguns autores consultados.

Para Cabrera Delgado & Bonnel (2016) deve existir um equilíbrio entre os dados disponíveis e o tamanho do zoneamento da região em estudo, já que a presença de uma grande quantidade de células com valor zero na matriz origem-destino pode afetar negativamente a qualidade da estimativa e a transferência temporária do modelo.

Outro fator a ser considerado, é a influência que exerce o nível de agregação da informação na escolha do tipo de modelo mais apropriado, sendo uma relação inversamente proporcional entre esses dois fatores. Segundo De Grange *et al.* (2010), modelos com maior nível de complexidade fornecem melhores resultados mediante o uso de dados desagregados. Em contrapartida, modelos simples produzem bons resultados quando a informação disponível é agregada.

Além disso, De Grange *et al.* (2010) determinaram que a agregação espacial alta pode produzir alterações significativas nos valores dos parâmetros, no caso de modelos, como o gravitacional, podendo alterar o seu sinal, razão a considerar quando for dada uma interpretação econômica. Por outro lado, para dados com nível de agregação baixo, os parâmetros flutuam dentro de faixas de valores estáveis.

Outro fator a considerar, é a existência e omissão das viagens realizadas dentro da mesma zona ou unidade espacial, sendo estreitamente ligado ao nível de agregação espacial dos dados. Para Manout & Bonnel (2019), além das consequências geradas no momento da modelagem do tráfego em uma rede, à medida que os dados possuem menor grau de agregação, a confiabilidade dos resultados dos modelos aumenta em virtude da mitigação do erro produzido pelo desconhecimento de viagens produzidas de maneira intrazonal.

Em contrapartida, Perrakis *et al.* (2012) afirmam que, ao ser realizada uma estimativa com informação com alta agregação espacial, o resultado é vantajoso porque essas estimativas são transferidas de maneira direta aos níveis inferiores (maior desagregação), situação que não acontece no sentido contrário.

Abdel-Aty *et al.* (2013) encontrou que o número de variáveis significativas varia em função das variáveis de resposta, assim como de acordo com o nível de agregação espacial ou unidades geográficas.

2.3. Análise e síntese

Com base na revisão bibliográfica realizada, pode-se observar as vantagens e desvantagens que cada modelo de distribuição de viagens possui. Os modelos agregados fornecem estimativas mediante o uso de informação vinculada a zonas de tráfego em detrimento do desconhecimento ou omissão dos indivíduos mediante a homogeneização das características em cada divisão espacial. Assim como, este tipo de modelo apresenta limitações quando os dados utilizados não possuem comportamentos lineares. Por outro lado, a necessidade de dados com menor nível de detalhe pode gerar economia no momento de obter a informação.

Os modelos desagregados demandam informação com maior grau de detalhamento. Assim como, o uso desses modelos torna-se mais complexo em relação aos modelos agregados, apesar de que tanto os modelos comportamentais e atitudinais fornecem informação importante, permitindo conhecer mais os padrões de tomada de decisão e ponderação de opções dos usuários.

No caso das técnicas de computação, estas fornecem melhores resultados, apresentando em alguns casos alta precisão, podendo ser um substituto potencial dos modelos estatísticos (GONÇALVES, *et al.*, 2015). No entanto, depois de treinadas é difícil interpretar seu funcionamento, sendo comparadas com uma caixa preta (AKAMINE, 2005).

Por serem técnicas não paramétricas, é impossível estabelecer as relações existentes entre a variável dependente e as independentes. Entretanto, estabelecem um nível de importância das covariáveis para previsão da variável dependente (ROCHA *et al.*, 2015), além de reconhecerem a não linearidade dos dados (RAIA JR, 2000).

Além das publicações consultadas na realização do presente capítulo, foram vistos arquivos complementares visando a identificar as variáveis usadas em cada um deles. Essa informação auxiliará na definição das variáveis utilizadas nos modelos a serem testados. Na Tabela é apresentada a relação de documentos consultados por variável utilizada.

Tabela 2.1. Relação de variáveis por documento pesquisado.

Autor	Modelo utilizado	Variáveis desagregadas														Variáveis agregadas														
		Sexo	Idade	Grau de instrução	Renda	Ocupação	Residência	Carros no domicílio	Frequência da viagem	Motivo da viagem	Composição familiar	Custo da viagem	Forma de pagamento	Distância	Modo de transporte	Taxa de emprego	Renda média *per capita*	Frequência escolar	Propriedade de carro	PIB *per capita*	PIB	População	Densidade populacional	Atração	Produção	Acessibilidade	Número de viagens	Tempo da viagem	Oportunidades intervenientes	Vagas de estacionamento	Uso do solo
Pitombo & Guimarães (2016)	DT	*	*	*	*	*	*	*	*	*		*	*	*			*			*								*			
Caldas *et al.* (2019)	ANN-MNL				*			*		*								*				*									
Salini *et al.*, (2017)	FL											*		*								*		*	*	*		*			
Arentze &Molin (2013)	DC	*	*	*		*					*																				
Khaki *et al.* (2009)	BI							*			*																				
Walker *et al.* (2010)	HCM-MNL	*		*	*			*				*			*													*			
Hess *et al.* (2010)	NL-MNL-COVNL											*			*																
Zhao *et al.* (2018)	DC-MNL	*			*							*			*													*			
Yang *et al.*, (2015)	GM									*								*						*	*						
Souza *et al.* (2017)	DT-MG	*	*	*	*	*		*	*	*		*			*						*	*						*			
Perrakis *et al.* (2012)	NBM-PM							*						*		*							*								
Aguiar Junior (2004)	GM-ANN													*								*									
Kompil Celik (2013)	ANN-FL									*																	*				
Lenormand *et al.* (2016)	GM-IO									*																	*	*	*		
Tadeu (2000)	GM-IO									*																	*	*	*		
Celik (2010)	GM									*																	*	*			
Tapkin & Akyilmaz (2004)	GM-ANN				*															*			*						*	*	*
Mozolim *et al.* (2000)	MLE-ANN																						*			*		*			*

Sendo: *DT: Decision tree; ANN: Artificial neural network; MNL: Multinomial logit; FL: Fuzzy logic; DC: Discrete choice; BI: Bayesian Interference; HCM: Hybrid choice model; NL: Nested logit; COVNL: Discrete mixture covariance; GM: Gravity model; NBM: Negative binomial model; PM: Poisson model; MLE: Maximum Likelihood Estimation.*

3. PROCEDIMENTO METODOLÓGICO

Este capítulo apresenta a sequência de passos que serão executados visando a atender os objetivos previamente estabelecidos. De maneira geral, este procedimento consta de seis etapas. São elas: (1) Área de estudo, (2) Levantamento dos dados, (3) Tratamento dos dados, (4) Construção dos modelos e (5) Arquitetura da rede neural. Na Figura 3.1., mostra-se em formato de diagrama de fluxo as atividades que compõem cada fase do estudo.

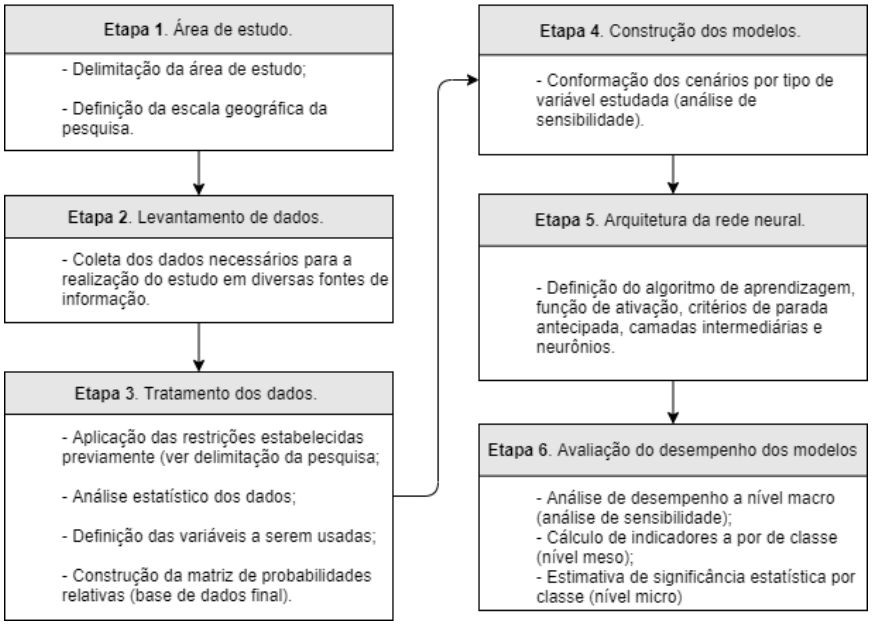

Figura 3.1. Procedimento proposto.
Elaboração própria.

A sexta etapa está centrada na análise de resultados da topologia que apresentou melhor desempenho. Nessa etapa, os modelos serão submetidos a uma série de testes que permitirão observar com maior detalhe seu comportamento perante os diferentes grupos de variáveis estudadas.

3.1. Área de estudo

A área de estudo será definida em função da disponibilidade de dados.

3.2. Levantamento de dados

Para desenvolvimento da pesquisa, são necessários dados relacionados com as características socioeconômicas do indivíduo, quantidade de empregos, tempo e modo de viagem para cada par origem-destino. Para isso, serão consultados estudos prévios na zona de estudo determinada e bases oficiais.

Cabe destacar que se propõe no presente trabalho um modelo híbrido que considere variáveis tanto desagregadas quanto agregadas, pois essas últimas representam atributos alusivos ao destino da viagem.

3.3. Tratamento dos dados

A definição das variáveis a serem utilizadas será realizada a partir de um ensaio estatístico que envolve a identificação de valores extremos *(outliers)*, uma análise de significância das variáveis.

Uma vez definidas, as informações que conformarão o banco de dados serão transformadas em uma matriz de probabilidades relativas, sendo necessária a criação de classes que agrupem as variáveis de tipo numérico. Essa nova base de dados auxiliará na definição da arquitetura da rede neural e análises posteriores.

3.4. Construção dos modelos

Os modelos a serem testados serão construídos a partir da divisão do banco de dados estabelecido na etapa anterior em função ao tipo de variável (socioeconômica, demográfica, impedância, uso do solo, entre outras), tendo como objetivo determinar a influência que exercem na escolha do destino da viagem.

3.5. Arquitetura da rede neural

Nesta parte, será definida a estrutura com a qual vai ser elaborada a rede neural, sendo indispensável delinear o algoritmo de aprendizado, a função de ativação, a quantidade de camadas ocultas e o número de neurônios que comporão essas camadas.

A definição do algoritmo de aprendizado e a função de ativação se basearão na revisão bibliográfica realizada do capítulo anterior. Por outro lado, o número de camadas intermediárias e neurônios será definido por meio de iteração, avaliação e ajuste dos modelos, sendo testadas várias configurações, iniciando com modelos básicos e aumentado sua complexidade até obter medidas de desempenho aceitáveis. No entanto, como ponto de referência, serão utilizadas algumas das heurísticas encontradas na literatura.

Na Figura 3.2., é apresentado o processo de estruturação dos modelos.

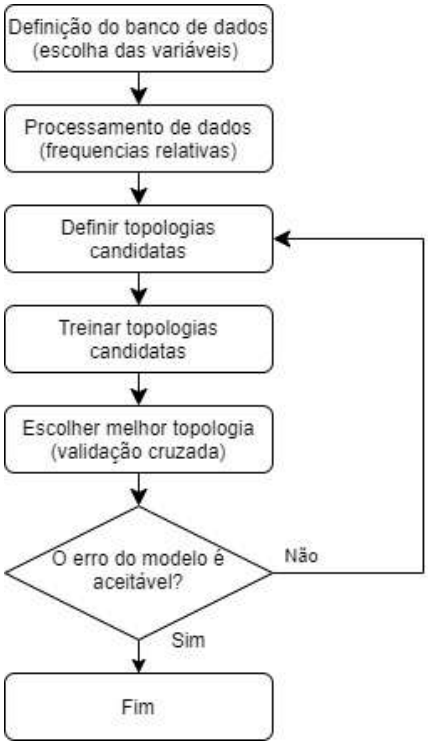

Figura 3.2. Processo de construção dos modelos.
Adaptação de Silva *et al.* (2016)

O desempenho das redes neurais será avaliado mediante o uso da validação cruzada, especificamente por meio da técnica *K-folds*, na qual o banco de dados é dividido em k subgrupos onde um conjunto de dados terá a função de grupo de teste e o restante (k-1) será usado para o treinamento do modelo. A particularidade dessa técnica é que o grupo de teste alterna de acordo com o ciclo, onde, ao final, todas as partições serão avaliadas, tal como se mostra na Figura 3.3.

Figura 3.3. Validação cruzada – *K-folds*

Como resultado da validação cruzada, são obtidas métricas de rendimento do modelo, o que auxilia na determinação do desempenho dos modelos. Inicialmente os indicadores são calculados por meio da geração da matriz de confusão que compara as classes observadas *versus* as classes preditas, sendo arbitrariamente denominadas positivas e negativas (Tabela 3.1).

Tabela 3.1. Matriz de confusão.

	C_1 – Predita	C_2 - Predita
C_1 – Observada	Verdadeiro Positivo	Falso Negativo
C_2 – Observada	Falso Positivo	Verdadeiro Negativo

Elaboração própria.

Quando o número de registros for adequadamente classificado pelo modelo na classe correspondente ($C_1 = C_1$ ou $C_2 = C_2$), os valores são considerados acertos do modelo sendo etiquetados como verdadeiros, sejam positivos ou negativos. Em contrapartida, quando o modelo classifica de maneira errônea os registros na classe incorreta ($C_1 \neq C_1$ ou $C_2 \neq C_2$), os desacertos do modelo são denominados falsos.

A primeira avaliação dos modelos, macro análise, será realizada por meio de indicadores de desempenho globais como a acurácia, a taxa de erro e o coeficiente Kappa. Por um lado, a acurácia pondera o número de elementos corretamente classificados no modelo (SKANSI, 2018). Ver Equação 26.

$$ACC = \frac{VP+VN}{VP+FP+VN+FN} ; 0 \leq \text{ACC} \leq 1 \quad (26)$$

Por outro lado, a taxa de erro indica a proporção de elementos classificados de maneira incorreta (BRAMER, 2016), tal como se mostra na Equação 27.

$$ERR = \frac{FP + FN}{VP + FP + VN + FN} ; \ 0 \ \leq \ ERR \ \leq \ 1 \quad (27)$$

Em que:

ACC é a acurácia do modelo;

VP são os verdadeiros positivos;

FP são os falsos positivos;

VN são os verdadeiros negativos;

FN são os falsos negativos; e

ERR é o erro global do modelo.

O coeficiente de concordância de Kappa mede a concordância entre as classes previstas e observadas de um conjunto de dados (WITTEN & FRANK, 2005). Em 1977, Landis & Koch (1977) elaboraram uma tabela de classificação que permite interpretar de maneira padronizada o resultado dado por esse indicador. Embora a divisão foi realizada de maneira arbitrária, a classificação é amplamente usada (ver Tabela 3.2).

Tabela 3.2. Interpretação do coeficiente *Kappa*.

Estatística *Kappa*	Grau de concordância
< 0,00	Pobre
0,00 - 0,20	Ligeira
0,21 - 0,40	Justa
0,41 - 0,60	Moderada
0,61 - 0,80	Substancial
0,81 - 1,00	Quase perfeita

Adaptado de Landis & Koch (1997).

Em matrizes desbalanceadas, alguns elementos possuem um maior peso em comparação às outras. É necessário realizar uma análise no nível de classe (meso análise),

sendo recomendado o cálculo do F-Score, que é uma medida de desempenho para classificadores (ver Equação 28).

$$F(C_n) = \frac{2 * PRE_{Cn} * REC_{Cn}}{PRE_{Cn} + REC_{Cn}} \quad (28)$$

Em que:

$F(C_n)$ é a medida F da classe *n;*

$PRE(C_n)$ é a precisão da classe *n;* e

$REC(C_n)$ é a sensibilidade da classe *n*.

O F-Score é a média harmônica dos indicadores precisão e *recall* do modelo. A precisão mensura a capacidade de predição do modelo (ver Equação (29) Já a recuperação, ver Equação 30, mede a capacidade do modelo em reconhecer os registros da classe correspondente (MANNING *et al.*, 2008).

$$PRE(C_n) = \frac{VP_{Cn}}{VP_{Cn} + VN_{Cn}} \quad (29)$$

Em que:

$PRE(C_n)$ é a precisão da classe *n*;

VP_{Cn} são os verdadeiros positivos da classe *n*; e

VN_{Cn} são os verdadeiros negativos da classe *n*.

$$REC(C_n) = \frac{VP_{Cn}}{VP_{Cn} + FN_{Cn}} \quad (30)$$

Em que:

$REC(C_n)$ é a sensibilidade da classe *n*;

VP_{Cn} são os verdadeiros positivos da classe *n*; e

FN_{Cn} são os falsos negativos da classe *n*.

No nível micro da análise, pretende-se usar o método desenvolvido por Gedeon (1997), que visa a determinar o significado comportamental de neurônios ocultos. Em outras palavras, pondera as variáveis envolvidas no modelo por meio do cálculo da significância estatística.

Finalmente, optou-se pela utilização do método de parada antecipada (*early stopping*) a fim de evitar o sobre-ajuste dos modelos (*overfitting*). Essa técnica detém o treinamento do modelo, uma vez que certos critérios previamente estabelecidos são atendidos.

Para efeitos deste trabalho, determinaram-se como critérios de parada antecipada os parâmetros usados como padrão pelo algoritmo do pacote h2o do *software* R, sendo eles:

- Métrica de parada: Perda logarítmica, entropia cruzada ou *logloss*, o qual avalia o quão próximos os valores estimados de um modelo estão do valor alvo observado. Este tipo de métrica é usada em classificadores, tal como é o caso do projeto;
- Tolerância de parada: o qual é o espectro ou limite permitido que o erro pode aumentar sem deter o processo. Nesse caso, tem um valor de 0,001;
- Rodadas de detenção: Quando o erro de validação supera o valor estabelecido como tolerância por três rodadas seguidas, irá parar evitando o sobre-ajuste.

4. APLICAÇÃO DO PROCEDIMENTO METODOLÓGICO

4.1. Área de estudo

O estudo de caso será realizado com dados da cidade de Medellín na Colômbia, pois na última década tem desenvolvido duas pesquisas de origem-destino, as quais contém informações detalhadas sobre o entrevistado, assim como sobre seu padrão de deslocamentos.

Medellín é a segunda cidade colombiana mais importante, depois de Bogotá, caracterizando-se por sua diversidade modal quando se refere a transporte público coletivo, além de ser referente no nível local com relação a transporte sustentável. Sendo a capital do departamento de Antioquia, Medellín possui uma extensão de 376,4 km^2, sendo que 111,61 km^2 se denominam solo urbano, 263,04 km^2 solo rural e 1,75 km^2 como solo de expansão (ALCALDÍA DE MEDELLÍN, 2020).

O município está composto por 21 subdistritos, dos quais 16 são denominados distritos (sub-regiões de índole urbano) e 5 corregimentos. Esses últimos formam a área rural do município, tal como se mostra na Figura 4.1.

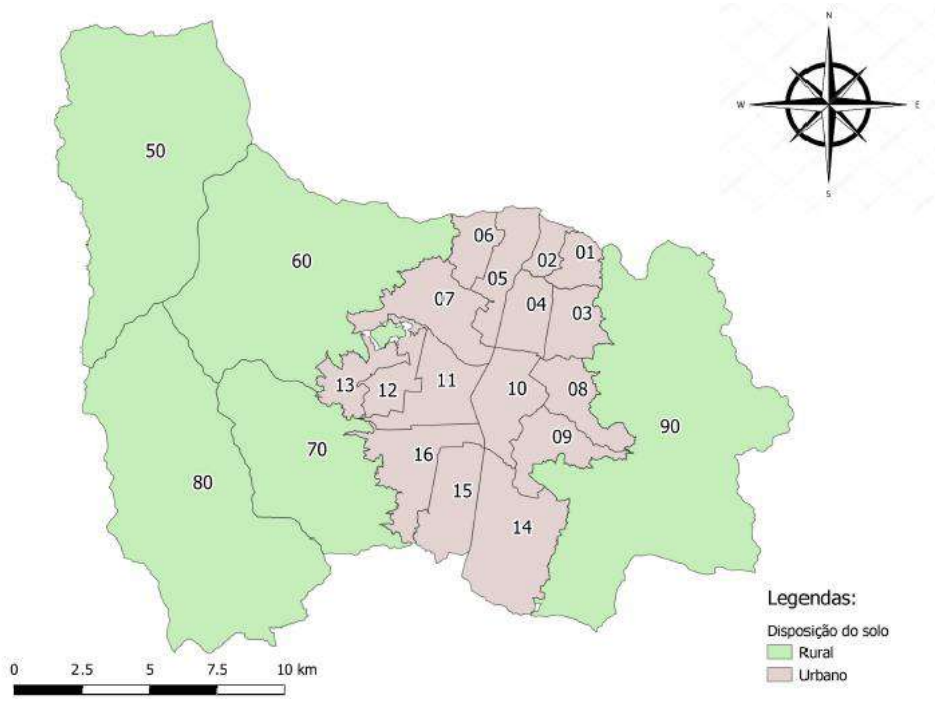

Figura 4.1. Divisão política do município de Medellín.
Elaboração própria com base nos dados da Prefeitura de Medellín.

O planejamento do uso do solo está regulamentado com o Acordo 48 de 2014 emitido pela Prefeitura Municipal e faz referência ao Plano de Ordenamento Territorial – POT, pautando o planejamento e a gestão do uso do solo da cidade de Medellín.

O Artigo 242 do acordo supracitado faz referência à categorização do uso do solo, classificando-o principalmente em seis grupos, sendo residencial, comercial, serviços, indústria, "dotacional", espaços públicos existentes e espaços públicos projetados.

Por outro lado, os artigos 243, 244 e 245 definem as áreas de corredores de baixa, meia e alta mistura (ver Figura 4.2a e Figura 4.2b), onde o uso do solo é definido em função das atividades econômicas e de prestação de serviços públicos (ALCALDIA DE MEDELLÍN, 2014). Entre elas encontram-se:

- **Área de atividade econômica em transformação:** São áreas com predominância de atividades produtivas, industriais e terciárias que resultam de forma relevante na geração de empregos, tendo como objetivo promover o desenvolvimento de atividades produtivas de alto valor agregado e conhecimento que dinamize a economia da cidade;
- **Área predominantemente residencial:** Espaço no qual a existência de moradias ocupa a maior parte do terreno, no entanto, há o uso misto das atividades econômicas de uso cotidiano (em menor proporção);
- **Centralidades com predominância econômica:** Espaços definidos pelo intercâmbio de bens, serviços e centro de empregos. Caracteriza-se pela proximidade com outras áreas e pela maior acessibilidade;
- **Centralidades e corredores de alta mistura:** Áreas que tendem a incentivar o uso misto do solo devido à adjacência com vias arteriais;
- **Área de uso misto:** Caracterizam-se por ser espaços que reúnem atividades econômicas de meia intensidade em um bairro. Zonas de transição que compreendem espaços entre as áreas com alta intensidade de uso misto e as zonas predominantemente residenciais. As centralidades dotacionais concentram equipamentos básicos sociais e comunitários, como centros educativos, unidades de pronto atendimento, espaços recreativos, entre outros. Os agrupamentos

comerciais e de serviço do bairro são locais que fornecem serviços em áreas de abrangência limitadas aos bairros.

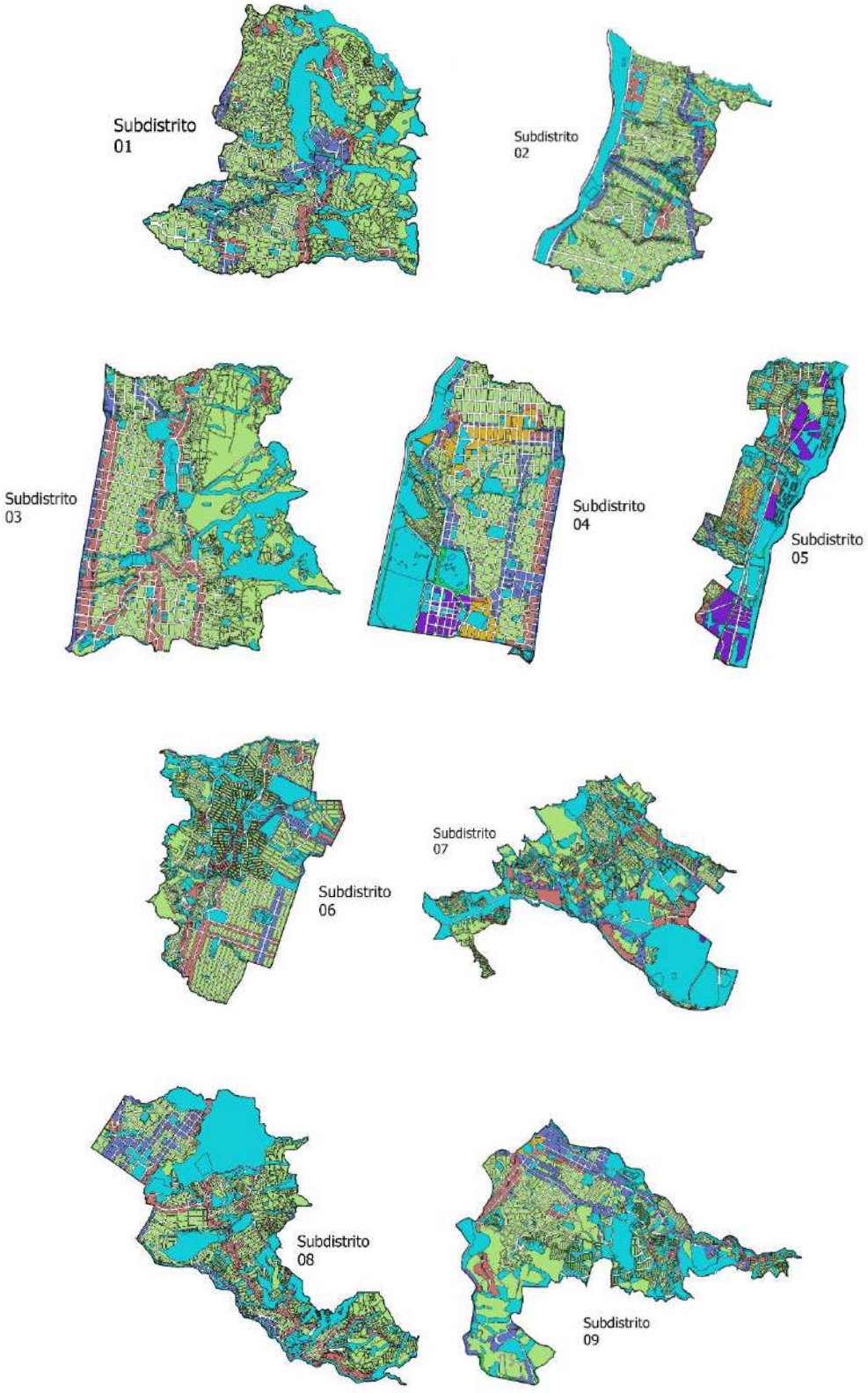

Figura 4.2a. Uso do solo do município por subdistrito.
Adaptado da Prefeitura de Medellín (Alcaldía de Medellín, 2014)

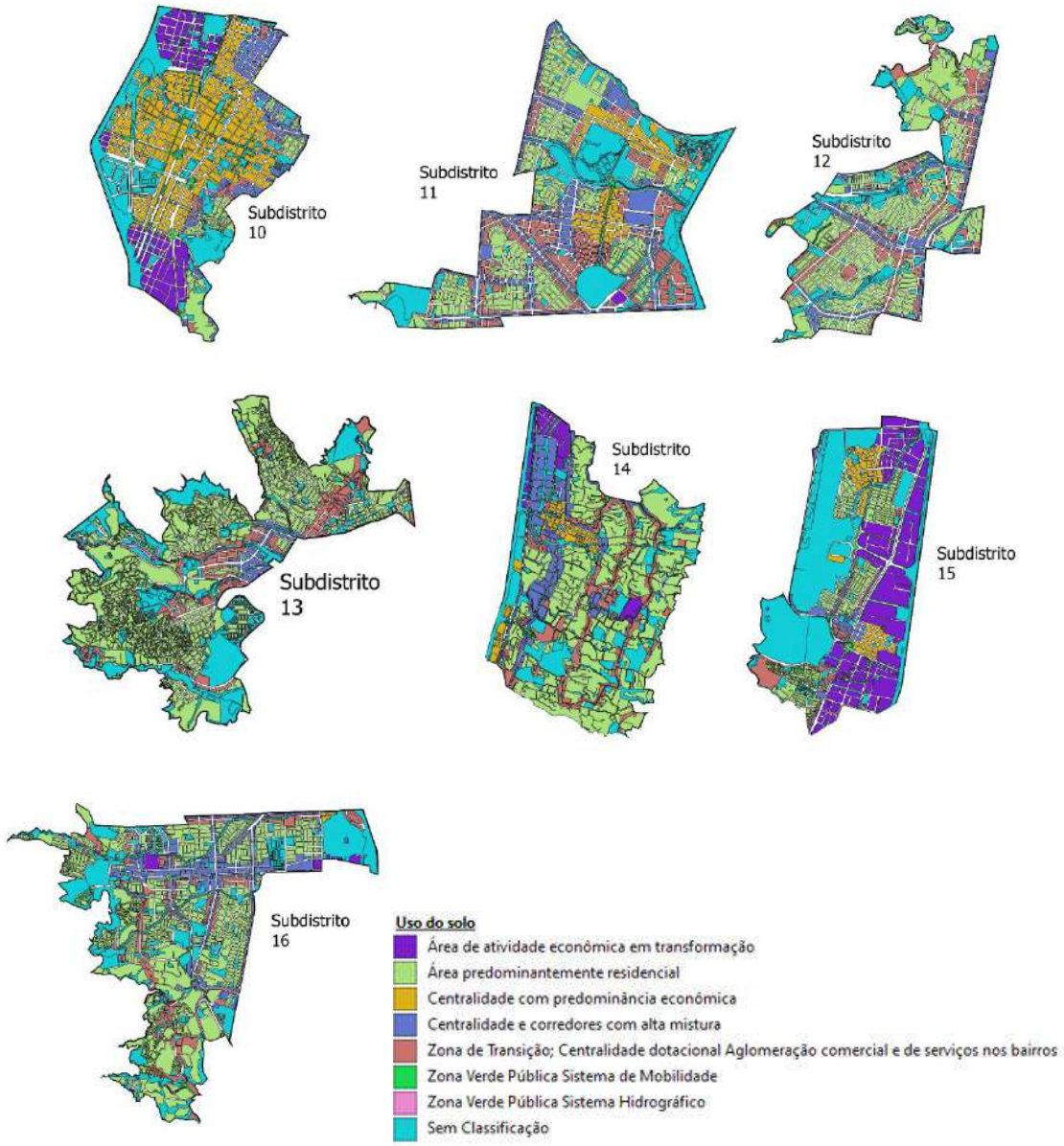

Figura 4.2b. Uso do solo do município por subdistrito.
Adaptado da Prefeitura de Medellín (Alcaldía de Medellín, 2014)

A Figura 4.2a e Figura 4.2b mostram a distribuição do uso do solo por subdistrito, permitindo perceber a localização das principais zonas atratoras da cidade, sendo destacadas pelas cores laranja e roxo.

Ao confrontar informações relacionadas ao uso do solo com o número de empregos, pode-se contatar que os subdistritos 10, 11, 14 e 15 atraem maior número de pessoas devido à presença destas nos centros de atividade econômicas, como comercio, indústrias e escritórios (ver Figura 4.3).

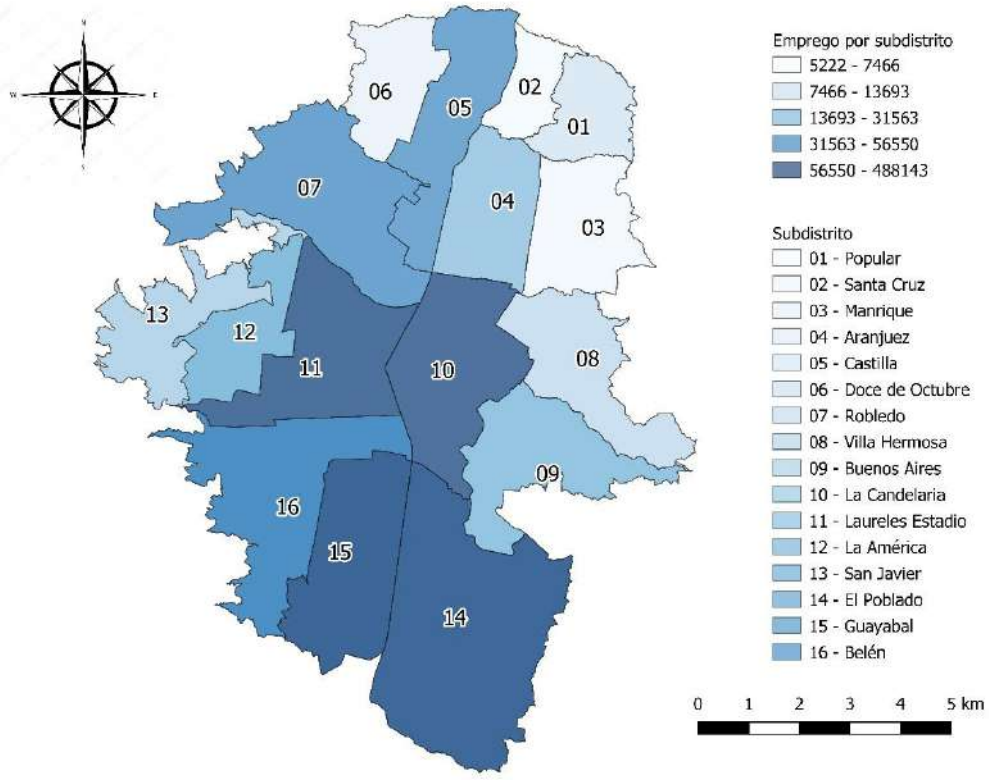

Figura 4.3. Número de empregos por subdistrito.
Elaboração própria com base nos dados da Prefeitura de Medellín.

4.2. Levantamento dos dados

A informação coletada para este estudo provém de duas fontes. A primeira é uma pesquisa origem-destino realizada no ano 2012 pela Prefeitura Municipal em parceria com a Universidade Nacional da Colômbia – Sede Medellín e a Área Metropolitana do Vale de Aburrá - AMVA. Para isso, foram realizadas entrevistas tanto domésticas quanto relacionadas ao transporte de carga.

Embora tenha sido realizada uma versão dessa pesquisa em 2017, optou-se pelo uso dos dados de 2012, pois foi possível obter a planilha de cálculo em formato *MS-Excel* com toda a informação pertinente aos hábitos relacionados aos deslocamentos da população entrevistada, assim como suas características socioeconômicas.

O total da base de dados da pesquisa evidenciava que foram realizadas 22.693 entrevistas domiciliares na região metropolitana do Vale de Aburrá, formada pelos municípios de Medellín, Barbosa, Girardota, Copacabana, Bello, Envigado, Itagüi,

Sabaneta, La Estrella e Caldas. Destas, 11.456 correspondem a viagens realizadas dentro da cidade de Medellín.

Esse banco de dados foi consolidado por meio da realização de pesquisas residenciais, permitindo a obtenção de respostas individuais. Dessa forma, cada linha que compõe a base de dados representa uma pessoa entrevistada.

Cabe ressaltar que o lar ou a moradia podem ser compostos por uma ou várias pessoas que moram juntas, mas que ao se locomoverem realizam deslocamentos diferentes. Sendo assim, o número de identificação da pesquisa pode conter um ou mais registros dependendo da estrutura familiar.

Na Tabela 4.1, é apresentada parte da base de dados concernentes às informações socioeconômicas dos entrevistados. Pode-se observar que as respostas, com exceção da idade, são categóricas, permitindo sua padronização.

Tabela 4.1. Estrutura da base de dados – Informações socioeconômicas.

# Pesq	Estrato	Nome Morador	Id Parentesco	Idade	Sexo	Id Estudo	Id Ocupação	Id Deficiência
24737	2	ESTEFANY	3	15	2	3	3	0
24737	2	WILIAN	1	43	1	2	4	0
24789	2	ROBERTINA	1	68	2	1	1	0
24765	1	LUZ MARINA	2	60	2	1	1	0
24765	1	JUAN BAUTISTA	1	62	1	1	4	0
13601	1	BEATRIZ	2	34	2	3	1	0
13601	1	CESAR	1	53	1	1	4	0
24816	2	JHON	1	50	1	3	5	1
24816	2	AURA	2	47	2	3	1	0
95091	1	ESTEFANIA	4	20	2	3	4	0
95091	1	YURI	1	23	2	3	4	2
95091	1	LUCAS	3	6	1	1	3	0
95084	1	CAMILO	3	22	1	2	5	0
95084	1	CARLOS	1	50	1	2	4	0
95084	1	MARTHA	2	43	2	3	4	0
95084	1	ANA	3	18	2	2	3	1
95084	1	MAYERLI	3	12	1	2	3	0
95096	2	KARLA	3	4	2	1	9	0
95096	2	CARLOS	1	44	2	2	1	0

Adaptado da pesquisa AMVA (2012).

Informações relacionadas à origem e destino da viagem foram referenciadas com as quatro escalas geográficas-administrativas, fornecendo um grau de agregação espacial variável. Observa-se, na Tabela 4.2., os códigos de identificação dos municípios, subdistritos, bairros e zonas de tráfego nos quais os entrevistados responderam ter-se deslocado no dia anterior.

Tabela 4.2. Estrutura da base de dados – Origem e destino das viagens.

Id Muni o	Id Subdis o	Id Bairro o	Id Sit o	Id Muni d	Id Subdis d	Id Bairro d	Id Sit d	Id Motivo Viagem
10	2	7	1980	10	1	1	3470	4
10	1	1	3470	10	2	7	1980	3
10	14	22	691	10	1	1	3470	4
10	1	1	3470	10	14	22	691	1
10	1	12	3550	10	1	1	3490	4
10	1	1	3490	10	1	12	3550	8
10	4	15	2212	10	10	6	30	11
10	10	6	30	10	4	15	2212	4
10	10	6	30	10	4	15	2212	4
10	4	15	2212	10	10	6	30	11
10	1	1	3510	35	8	0	5505	10
10	1	1	3510	35	8	0	5505	12
10	1	1	3460	10	10	13	60	1
10	1	4	3380	10	1	1	3460	4
10	1	4	3380	10	1	1	3460	4
10	1	1	3460	10	1	4	3380	12
10	1	1	3460	10	1	4	3380	1
10	10	13	60	10	1	1	3460	4
10	1	4	3380	10	1	1	3460	4

Adaptado da pesquisa AMVA (2012).

Já informações relativas à viagem *per se* no banco de dados permite conhecer a hora de início e fim da viagem e consequentemente o tempo de viagem demonstrado em minutos, além do modo de transporte principal usado ao se locomover (ver Tabela 4.3).

Tabela 4.3. Estrutura da base de dados – Informações da viagem.

Hora início	Hora fim	Tempo de viagem (minutos)	Modo Principal	FEV
18:00	19:30	90	10	60
11:40	12:45	65	10	60
17:30	19:30	120	07	61
05:00	06:10	70	07	60
12:00	12:30	30	10	69
10:00	10:30	30	10	69
11:30	12:00	30	01	121
12:05	13:00	55	01	190
12:05	13:00	55	01	190
11:30	12:00	30	01	121
14:00	15:00	60	01	87
14:00	15:00	60	01	87
14:00	15:00	60	04	37
23:00	23:30	30	04	37
08:40	09:20	40	04	37
07:05	08:30	85	04	37
14:00	15:00	60	04	37
23:00	23:30	30	04	37
13:30	14:30	60	04	37

Adaptado da pesquisa AMVA (2012).

Na seção 4.3 desta dissertação, será detalhado o processo de agregação realizado no tempo de viagem declarado pelo entrevistado. Por fim, a última coluna da Tabela 4.3. indica o fator calculado pelos realizadores da pesquisa origem-destino com o propósito de expandir os resultados da amostra à população

Com relação às oportunidades intervenientes utilizadas nos modelos, essas informações foram obtidas diretamente da Prefeitura de Medellín. Como se observa na Tabela 4.4., foram recebidas as informações relacionadas ao número de emprego para o ano 2011 por subdistrito, material com um alto grau de agregação.

Tabela 4.4 Número de empregos por subdistrito.

Subdistrito / Comuna	Total	Empregos Formais	Empregos Informais	Empregos Formais %	Empregos Informais %
Total	1.024.055	545.207	478.848	53,24	46,76
1. Popular	7.736	3.494	4.242	45,17	54,83
2. Santa Cruz	6.189	2.814	3.375	45,47	54,53
3. Manrique	7.156	3.395	3.760	47,45	52,55
4. Aranjuez	23.401	11.345	12.056	48,48	51,52
5. Castilla	32.491	16.603	15.888	51,1	48,9
6. Doce de Octubre	5.222	2.694	2.528	51,59	48,41
7. Robledo	43.902	24.049	19.852	54,78	45,22
8. Villa Hermosa	7.543	3.467	4.075	45,97	54,03
9. Buenos Aires	15.472	8.319	7.153	53,77	46,23
10. La Candelaria	488.143	239.190	248.953	49,00	51,00
11. Laureles Estadio	94.766	62.167	32.600	65,6	34,4
12. La América	30.944	19.386	11.558	62,65	37,35
13. San Javier	11.024	5.670	5.354	51,43	48,57
14. El Poblado	145.824	90.513	55.311	62,07	37,93
15. Guayabal	57.247	33.060	24.187	57,75	42,25
16. Belén	46.996	27.028	19.969	57,51	42,49

Adaptado da Prefeitura de Medellín (2012).

Os dados relativos a emprego foram submetidos a um processo de desagregação da informação até alcançar a escala de zona de tráfego, tema que será aprofundado na seção 4.3 desta dissertação.

4.3. Tratamento dos dados

Considerando as pautas estabelecidas na seção 1.4, a base de dados foi reduzida inicialmente para 10.960 entrevistas com origem e destino a algum dos 16 subdistritos de Medellín, das quais 7.067 tinham como motivo da viagem o trabalho. Sabe-se que 4.051 foram realizadas por meio do sistema de transporte público coletivo da cidade (modo principal da viagem).

Com base nos tempos de viagens declarados pelos entrevistados, foi determinada a amplitude para cada modal, sendo que para o modo ônibus foram considerados o menor e o maior tempo de viagem, 5 e 630 minutos respectivamente. No caso do micro-ônibus, 10 e 120 minutos; BRT, 5 e 90 minutos e metrô, 2 a 240 minutos. Por meio do método baseado na amplitude interquartil (IQR), procedeu-se à identificação dos *outliers* para

cada modalidade, sendo necessário o cálculo de indicadores tal como se mostra Tabela 4.5.

Tabela 4.5. Parâmetros de identificação de *outliers* por modo.

	Ônibus	Micro-ônibus	BRT	Metrô
Média	38,53	45,56	41,60	49,08
Quartil 1	30,00	30,00	30,00	30,00
Quartil 3	50,00	60,00	50,00	60,00
IQR	20,00	30,00	20,00	30,00
Limite superior	68,53	90,56	71,60	94,08
Limite inferior	8,53	0,56	11,60	4,08

Elaboração própria.

No caso das viagens realizadas por ônibus, foram identificadas duas viagens que excedem as 10 horas de tempo de viagem quando comparadas com viagens similares. Quanto a lugar de origem e destino e hora do início da viagem, se determinou o descarte dos dados. Essa análise foi realizada para os quatros modos de transporte. Na Figura 4.4. pode ser observar graficamente os *outliers* identificados para cada modo de transporte.

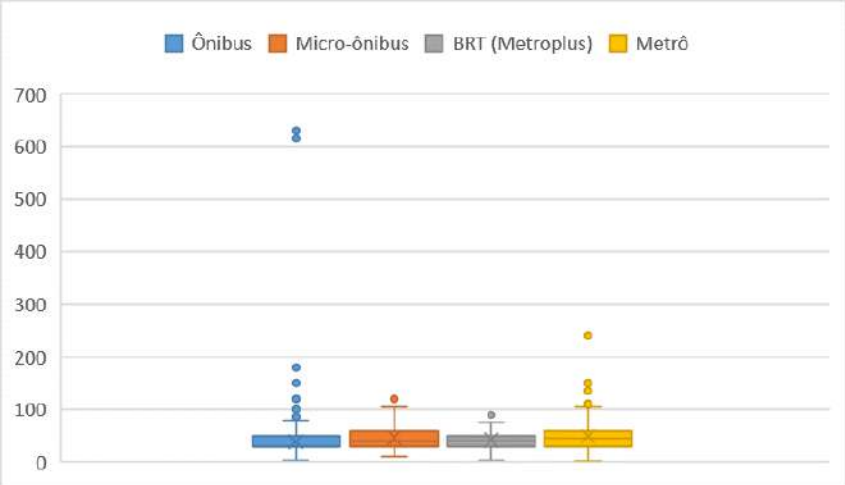

Figura 4.4. Outliers por modo de transporte – *Boxplot.*
Elaboração própria.

Das 1.023 entrevistas realizadas por motivo de trabalho (quantidade de viagens antes da aplicação do fator de expansão), foram descartadas 13, sendo 10 viagens por ônibus, 1 por micro-ônibus e 2 por metrô. Vale a pena destacar que o número de entrevistados por pesquisa realizada oscila entre 1 e 21 pessoas, dependendo do número de lares e moradores.

A partir da depuração do banco de dados, determinou-se como linha base do estudo o total de 1.010 pesquisas, que equivalem, uma vez usado o fator de expansão, a 288.392 viagens. Cabe ressaltar que o fator de expansão foi calculado AMVA (2012) da pesquisa citada anteriormente, baseados na relação da estratificação socioeconômica e divisão zonal.

Obteve-se uma base de dados equivalente a 288.392 viagens, empregando 15 variáveis das quais 12 eram categóricas. Visando a estimar a correlação existente entre as variáveis, optou-se pela transformação das variáveis numéricas para categóricas aplicando o método de Sturges, que permite o agrupamento de valores por meio da criação de intervalos de igual tamanho (ver Equação 31). Consequentemente o banco de dados foi avaliado com a criação de tabelas de contingência ou tabelas cruzadas.

$$k = 1 + 3{,}322(log_{10} n)\ ; w = \frac{R}{k} \tag{31}$$

Em que:

k é o número ótimo de classes ou intervalos;

n é o total de observações da amostra;

w o tamanho da classe; e

R é a diferencia entre o limite superior e inferior dos dados observados.

Com base no método de Sturges, foram categorizadas as variáveis, originalmente escalares e numéricas: idade, tempo de viagem e o número de empregos no destino da viagem. Tal como se mostra na Tabela 4.6.

Tabela 4.6. Variáveis categorizadas.

No. De identificação da classe	Intervalos		
	Idade	T. de Viagem	OI – Empregos
1	4-10	5-11	Até 32509
2	11-17	12-18	32510 – 65019
3	18-24	19-25	65020 – 97529
4	25-31	26-32	97530 – 130039
5	32-38	33-39	130040 – 162549
6	39-45	40-46	162550 – 195059
7	46-52	47-53	195060 – 227569
8	53-59	54-60	227570 – 260079
9	60-66	61-67	260080 – 292589
10	67-73	68-74	292590 – 325099
11	74-80	75-81	325100 – 357609
12	81-87	82-88	357610 – 390119
13	88-94	89-95	390120 – 422629
14	-	-	422630 – 455139
15	-	-	455140 – 487649
16	-	-	487650 – 520159
17	-	-	520160 – 552669

Elaboração própria.

A correlação entre variáveis categóricas é realizada a partir da elaboração de tabelas de contingências que cruzam a informação de duas variáveis categóricas por vez, relacionando as frequências relativas existentes entre as duas (ver Figura 4.5.). Posteriormente é gerada a matriz de probabilidade conjunta para cada célula a partir da divisão de seu conteúdo sobre a soma total (ver Equação 32).

		Variável 2						∑ linhas
Variável 1	Classes	C_1	C_2	C_3	C_4	C_5	C_6	
	C_1	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	TL_1
	C_2	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	f_{26}	TL_2
	C_3	f_{31}	f_{32}	f_{33}	f_{34}	f_{35}	f_{36}	TL_3
	C_4	f_{41}	f_{42}	f_{43}	f_{44}	f_{45}	f_{46}	TL_4
∑ colunas		TC_1	TC_2	TC_3	TC_4	TC_5	TC_6	∑ Matriz

Figura 4.5. Conceito da tabela de contingência.

Sendo

C_n: a classe n da variável 1 ou 2;

f_{ij}: a frequência de cada par de classes;

TC_n: a soma da coluna n;

TL_n: a soma da linha n;

p_{ij}: a probabilidade conjunta de cada par de classes; e

$\sum$ Matriz: a soma total da matriz.

$$p_{ij} = \frac{f_{ij}}{\sum matriz} \tag{32}$$

Uma vez calculada a probabilidade conjunta de cada célula, calculam-se os graus de liberdade do modelo e finalmente a correlação entre as variáveis. Para determinar se as variáveis estão relacionadas entre si, é pertinente saber o número de ocorrências cruzada estimadas e o tamanho da ocorrência mínima na matriz.

O teste Chi quadrado de Pearson (ver Equação 34), permite estimar se duas variáveis estão relacionadas, verificando a independência entre o par de variáveis sempre que essas variáveis sejam de caráter numérico. Esse teste mede a discrepância entre os dados observados e estimados.

$$X^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right] \tag{33}$$

Em que:

X^2 é o coeficiente Chi quadrado de Pearson;

O_i é o valor observado de i; e

E_i é o valor estimado de i.

Quando a frequência esperada em alguma das células que compõem a matriz é menor que cinco, o teste exato de Fisher é mais apropriado do que o teste Chi quadrado (LEON, 1998).

O teste exato de Fisher permite determinar a existência de associação significativa entre variáveis categóricas. Embora esse modelo tenha sido concebido para matrizes de 2

x 2, recursos computacionais permitem seu uso em matrizes de maior tamanho. Na Equação 34 se apresenta a formulação do modelo em sua concepção original.

$$P_{\alpha} = \frac{(a+b)! * (c+d)! * (a+c)! * (b+d)!}{a!\, b!\, c!\, d!\, N!} \tag{34}$$

Em que:

a, b, c, d são as frequências individuais da tabela de contingência de 2 x 2; e

N é a frequência total da matriz.

Na construção das tabelas de contingência, foram cruzadas as informações de cada uma das variáveis independentes (nas filas) com as zonas de tráfego de destino da viagem (nas colunas). Essa última foi catalogada como variável dependente. Por meio do pacote *"stats"* do *software* R, foi calculado o nível de significância para cada variável por meio do teste exato de Fisher. Os resultados são apresentados na Tabela 4.7.

Tabela 4.7. Significância das variáveis.

Nome	Significância
Sito	-
Sexo	0,0004
Idade	0,0184
Estrato	0,0004
Tipo moradia	0,1059
Modo Transp próprio	0,0549
IdEstudo	0,0004
IdOcupação	0.0079
IdTemprego	0,0059
IdLugartrab	0,0004
IdDeficiencia	0,8716
Ndeficiencia	0,8561
Modo Principal	0,0004
TotalViajeMinutos	0,0004
Emprego	-

Elaboração própria.

Com base na Tabela 4.7, fica evidente que as variáveis tipo de moradia, modo de transporte próprio, deficiência e natureza da deficiência possuem um grau de significância

maior a 0,05. Essa forma, estas não serão consideradas na construção dos modelos. Essa análise foi feita apenas com os dados extraídos da pesquisa origem-destino. Por isso, a variável emprego não foi incluída.

Cabe ressaltar que parte deste estudo visa a determinar a influência da impedância na realização da viagem. O tempo de viagem é uma variável desagregada que indica o tempo declarado de deslocamento medido desde a saída da residência até a chegada ao lugar de trabalho.

Com relação às oportunidades intervenientes e diante da inexistência de registros oficiais referentes ao número de empregos na escala de zonas de tráfego, esse valor foi calculado a partir da ponderação do número de empregos por subdistrito e da atração que cada um deles possui.

Para isso, foi necessária a criação de um novo código que permitisse a identificação única das zonas de tráfego. Observou-se que alguns códigos eram usados de maneira duplicada para identificar bairros de diferente subdistrito ou, em alguns casos, utilizava-se o mesmo código da zona de tráfego para células de diferente bairro (ver Tabela 4.8).

.

Tabela 4.8. Códigos de identificação por escala espacial.

Id Município	Id Subdistrito	Id Barrio	Id Sit
10	8	5	270
10	8	1	270
10	8	8	280
10	8	4	280
10	8	11	300
10	8	7	300
10	8	14	390
10	8	13	390
10	8	12	390
10	8	19	400
10	8	15	400

Elaboração própria.

Esse novo código foi criado a partir da combinação dos códigos de subdistrito, bairro e SIT, tal como se mostra na Figura 4.4. Posteriormente esse código foi numerado

de maneira sequencial, garantindo uma identificação única a cada zona de tráfego. No ANEXO 2 são relacionados os códigos de – para por ZT.

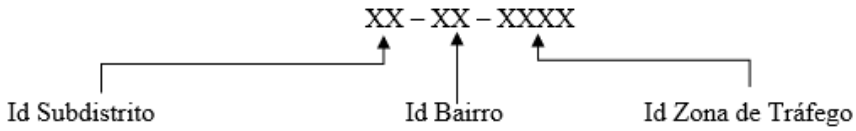

Figura 4.4. Estrutura da codificação das origens e destinos.
Elaboração própria.

Uma vez recodificadas as zonas de tráfego, determinou-se o número de viagens atraídas por cada uma delas. Posteriormente foram calculados o peso de cada zona dentro do subdistrito, permitindo a ponderação do número de empregos de uma escala macro (subdistritos) para uma escala micro (ZT) mediante interpolação.

Na Tabela 4.9. são caracterizadas as variáveis que conformam o estudo. Cabe destacar que foram calculadas as frequências relativas para cada uma delas, com exceção da zona de origem e destino da viagem.

Tabela 4.9. Caracterização das variáveis.

Nome	Natureza	Tipo (antes da conversão)
SITO	Uso do solo	Categórica
Sexo	Demográfica	Categórica
Idade	Demográfica	Numérica
Estrato	Socioeconômica	Categórica
IdEstudo	Demográfica	Categórica
IdOcupação	Socioeconômica	Categórica
IdTemprego	Socioeconômica	Categórica
IdLugartrab	Socioeconômica	Categórica
Modo Principal	Infraestrutura de transporte	Categórica
TotalViajeMinutos	Impedância	Numérica
SITD	Uso do solo	Categórica
Emprego	Uso do solo	Numérica

Elaboração própria.

Finalmente a quantidade de empregos por zona de tráfego foi transformada em oportunidade interveniente seguindo a metodologia desenvolvida por Stouffer (1940) e aperfeiçoada por Schneider (1959).

A seguir são descritas as variáveis desagregadas utilizadas no trabalho, pertencentes às bases de dados pesquisadas:

- *SITO:* Código de identificação da origem da viagem na zona do Sistema Integrado do Transporte;
- *Sexo:* Sexo do entrevistado, sendo um (1) para masculino e dois (2) para feminino;
- *Idade:* Idade do entrevistado;
- *Estrato (Classe social):* Classificação dos imóveis de acordo com o acesso que este tem aos serviços ofertados pela cidade, variando de um (1) a seis (6). Segundo a Lei 689 de 2001, os imóveis residenciais são classificados como 1, baixo-baixo; 2, baixo; 3, meio-baixo; 4, meio; 5, meio-alto; e 6, alto. (CONGRESO DE LA REPÚBLICA DE COLOMBIA, 2001);
- *IdEstudo:* Último nível de instrução que obteve o entrevistado, sendo um (1), nenhum; dois (2), ensino fundamental; três (3), ensino médio; quatro (4), educação não formal; cinco (5), técnico; seis (6), tecnológico; sete (7), bacharel; e oito (8), pós-graduação;
- *IdOcupação:* Código da atividade à qual o entrevistado se dedica atualmente, sendo um (1), dona de casa; dois (2), aposentado; três (3), estudante; quatro (4), trabalhador; cinco (5), sem emprego; seis (6), aposentado e estudante; sete (7), dona de casa e estudante; oito (8), trabalhador e estudante; e nove (9), nenhum ou economicamente inativo;
- *IdTemprego:* Tempo dedicado ao trabalho realizado pelo entrevistado, sendo um (1), tempo integral; dois (2), tempo parcial; e três (3), voluntariado não remunerado;
- *IdLugartrab:* Lugar de trabalho do pesquisado, sendo um (1), casa; dois (2), escritório ou estabelecimento; três (3), casa e escritório; quatro (4), na rua; e cinco (5), nenhum;
- *Modo Principal:* Modo de transporte escolhido para cada viagem de acordo com a resposta dada pelo entrevistado e na tabela de ponderação dos modos;
- *TotalViajeMinutos:* Tempo total da viagem do pesquisado incluindo todas as suas etapas;
- *SITD:* Código de identificação do destino da viagem na zona do Sistema Integrado do Transporte;

- *Emprego:* Número de empregos fornecidos da zona do Sistema Integrado do Transporte de destino da viagem.

Adicionalmente considerou-se pertinente a inserção de informações agregadas com as oportunidades intervenientes (impedância), assim como a atração e a produção de cada lugar de origem e destino da cidade. Essas últimas permanecem na base de dados como variáveis numéricas.

Uma característica desta pesquisa é a maneira em que foi estruturada a base de dados que contou com duas grandes alterações. A primeira delas foi a categorização das variáveis numéricas, como foi citado anteriormente. A segunda foi a construção da matriz de frequência relativa.

Exemplificando, na Tabela 4.10. mostram-se 4 das 12 variáveis escolhidas. Como se pode observar, a variável Sexo foi dividida em duas colunas por não possuir o mesmo número de classes. Sendo que S1 refere-se a masculino e S2, feminino. No caso das viagens originadas na zona de tráfego 3 com destino à zona de tráfego 192, a probabilidade relativa que esta seja realizada por um homem é de 67% e 36% por mulher.

Tabela 4.10. Estrutura do banco de dados – Exemplo.

SO	S1	S2	ES1	ES2	ES3	ES4	ES5	ES6	M3		SD_C
1	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	...	15
1	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	...	49
1	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	...	61
1	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	...	184
1	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	182
1	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	297
1	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	180
1	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	320
2	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	176
2	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	207
2	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	219
2	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	308
2	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	...	320
2	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	192
2	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	241
2	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	284
2	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	...	305
2	0.50	0.50	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	193
3	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	...	64
3	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	172
3	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	...	174
3	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	...	175
3	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	...	177
3	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	...	196
3	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	...	202
3	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	...	193
3	0.67	0.33	0.00	0.67	0.33	0.00	0.00	0.00	0.00	...	192

Elaboração própria.

Tal como se observou na tabela anterior para cada classe que compõe as variáveis, foi calculada a probabilidade de ocorrência. Por este motivo, o banco de dados, que antes apresentava 14 colunas (uma por variável), passou a apresentar 83 colunas por 288.382 linhas.

4.4. Construção dos modelos

Foram criados seis modelos baseados em redes neurais artificiais que auxiliarão na estimativa da influência das variáveis na escolha do destino (ver Tabela 4.11.). O modelo 0 ou padrão é composto pelas variáveis usadas em modelos tradicionais, como o modelo gravitacional. Por sua vez, no M1 é substituída a impedância no tempo de viagem pelo número de emprego em formato de oportunidades intervenientes.

Tabela 4.11. Composição dos cenários ANN

Variável	Tipo	M0	M1	M2	M3	M4	M5
Atração	Numérica - Ordinal	*	*	*	*	*	*
Emprego	Numérica - Decimal		*			*	*
Estrato	Numérica - Decimal			*	*	*	*
Idade	Numérica - Decimal			*	*	*	*
IdEstudo	Numérica - Decimal			*	*	*	*
IdLugartrab	Numérica - Decimal			*	*	*	*
IdOcupação	Numérica - Decimal			*	*	*	*
IdTemprego	Numérica - Decimal			*	*	*	*
Modo Principal	Numérica - Decimal			*	*	*	*
Produção	Numérica - Ordinal	*	*	*	*	*	*
Sexo	Numérica - Decimal			*	*	*	*
SITD	Categórica	*	*	*	*	*	*
SITO	Categórica	*	*	*	*	*	*
TotalViajeMinutos	Numérica - Decimal	*		*			*

Elaboração própria

Em síntese, na Tabela 4.12. são relacionadas as diferenças existentes entre os modelos. Pode-se observar que a partir do segundo modelo – M2 são incorporadas as variáveis socioeconômicas visando a testar sua influência. É importante ressaltar que no modelo M0 embora esteja relacionado ao modelo gravitacional não foi usada a metodologia deste se não as variáveis associadas a ele.

Tabela 4.12. Diferenças entre modelos baseados em ANN

Modelo	Resumo
C0	padrão - modelo gravitacional
C1	Padrão sem tempo com oportunidades intervenientes
C2	socioeconômicas com tempo sem oportunidades intervenientes
C3	socioeconômicas sem tempo sem oportunidades intervenientes
C4	socioeconômicas sem tempo com oportunidades intervenientes
C5	socioeconômicas com tempo com oportunidades intervenientes

Elaboração própria.

Além de determinar a influência das variáveis individuais na escolha do destino, busca-se observar a interferência da impedância nos modelos. Assim, do modelo 2 ao 5 é alternada a presença desses fatores de dissuasão.

4.5. Arquitetura da rede neural

Para a construção e estruturação dos modelos de rede neural, foi utilizada a linguagem de programação R no *software RStudio* devido a sua versatilidade e sua natureza de *software* livre, permitindo assim o uso da ferramenta sem nenhum tipo de restrição. Na Tabela 4.13a e Tabela 4.13b são relacionados os pacotes que auxiliaram na elaboração do algoritmo.

Tabela 4.13a. Relação dos pacotes usados.

Nome	Versão	Descrição	Autor(es)
amelia	1.7.6	Um programa para dados ausentes	Honaker, King e Blackwell (2011)
caret	6.0-86	Classificação e treinamento de regressão.	Kuhn (2020)
caretensemble	2.0.1	Conjuntos de modelos de Caret	Zachary A., Deane-Mayer e Knowles (2019)
data.table	1.13.0	Extensão do 'data.frame'	Dowle e Srinivasan (2020)
e1071	1.7-3	Funções diversas do Departamento de Estatística	Meyer, Dimitriadou, Hornik, Weingessel e Leisch. (2019)
h2o	3.30.0.1	Plataforma escalonável de aprendizado de máquina	LeDell, et al. (2020)
mice	3.11.0	Imputação multivariada por equações encadeadas em R	Van Buuren e Groothuis-Oudshoorn (2011)

Elaboração própria

Tabela 4.13b. Relação dos pacotes usados.

Nome	Versão	Descrição	Autor(es)
readxl	1.3.1	Leitura de arquivos Excel	Wickham e Bryan (2019)
rpart	4.1-15	Particionamento recursivo e árvores de regressão	Therneau e Atkinson (2019)
stats	4.0.2	Pacote estatístico	R Core Team and contributors worldwide

Elaboração própria

Principalmente foi usado o pacote h2o, que está baseado na arquitetura *feedfoward* de camadas múltiplas especificamente do tipo *Perceptron* de múltiplas camadas. Entre as características dessa classe de arquitetura está o uso do algoritmo de aprendizado denominado retro-propagação do erro ou *backpropagation* no treinamento da rede, sendo o treinamento realizado de maneira supervisionada.

O h2o possui diversas funções de ativação, destacando-se Tangente hiperbólica, função retificadora e *Maxout*. Baseado na revisão bibliográfica, observou-se que a maioria de documentos pesquisados utilizaram a função Tanh. No entanto, de maneira experimental e a fim de fornecer um melhor desempenho do modelo, foram comparadas as funções Tangente hiperbólica e a retificadora *(Rectifier)*.

A prova foi realizada por meio da base de dados do projeto e testada mantendo a quantidade de camadas e neurônios constante, de forma igual aos critérios de parada antecipada. Cabe destacar que, para esse teste, os neurônios e as camadas foram estabelecidos de maneira aleatória, tendo duas camadas intermediarias. A primeira delas com 40 neurônios e a segunda com 100.

Como resultado, obteve-se uma acurácia maior para cada dobra e, consequentemente, um menor erro com a função retificadora, tal como se mostra na Figura 4.7. Portanto, optou-se por seu uso em detrimento da função Tanh.

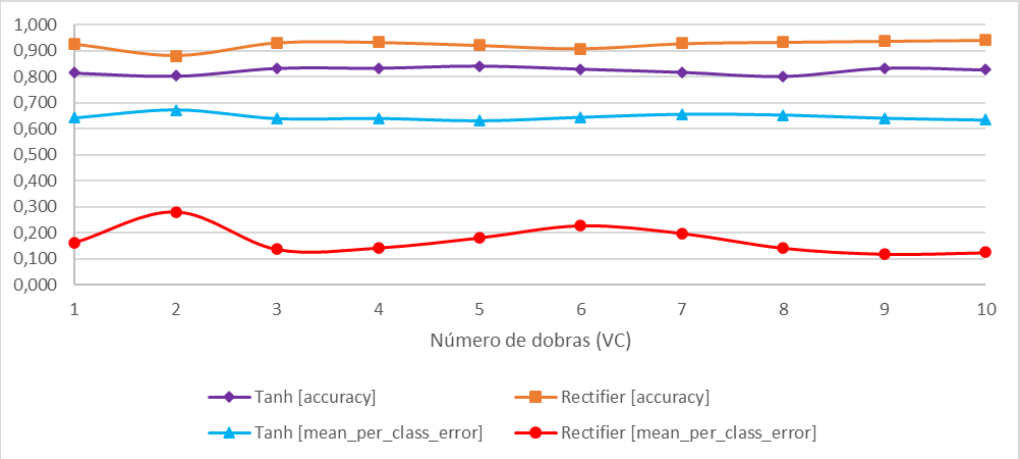

Figura 4.7. Avaliação da função de ativação.
Elaboração própria.

Os indicadores de desempenho mostrados na Figura 4.7 foram calculados pelo método de validação cruzada – VC, que divide o banco de dados em n número de dobras, destinando uma delas para o teste do modelo e k-1 para o treinamento. No caso do algoritmo do pacote h2o, a validação cruzada gera um modelo adicional.

De acordo com LeDell *et al.* (2020), o banco de dados é dividido da seguinte forma: 80% para treinamento e 20% para teste. Para a base de treinamento foram criadas dobras que, para efeitos do presente trabalho, são 10. A partir dessa subdivisão, se inicia a alternância de dobras entre teste e treinamento do modelo onde, ao final de cada cenário, são calculadas as respectivas métricas de desempenho (ver Figura 4.8.).

MP	Treinamento [80%]									Teste [20%]
Ciclos n = 10	TE	TR	TR	TR	TR	TR	TR	TR	TR	TR
	TR	TE	TR	TR	TR	TR	TR	TR	TR	TR
	TR	TR	TE	TR	TR	TR	TR	TR	TR	TR
	TR	TR	TR	TE	TR	TR	TR	TR	TR	TR
	TR	TR	TR	TR	TE	TR	TR	TR	TR	TR
	TR	TR	TR	TR	TR	TE	TR	TR	TR	TR
	TR	TR	TR	TR	TR	TR	TE	TR	TR	TR
	TR	TR	TR	TR	TR	TR	TR	TE	TR	TR
	TR	TR	TR	TR	TR	TR	TR	TR	TE	TR
	TR	TR	TR	TR	TR	TR	TR	TR	TR	TE
	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10

Figura 4.8. Validação cruzada algoritmo h2o.
Elaboração própria.

Adicionalmente foi criado o modelo principal, que possui a divisão original do *dataset*, destinando 80% dos dados originais. Já o desempenho deste é calculado levando em consideração a união das previsões dos submodelos (partições da validação cruzada) e contrastando-os com 20% do total global da testagem.

A comparação das métricas geradas de cada partição permite verificar a estabilidade na previsão do modelo principal (LeDell *et al.* 2020).

Considerou-se dividir o banco de dados em 10 dobras porque os testes realizados em vários conjuntos de dados, utilizando diferentes técnicas de aprendizagem, mostraram que esse método fornece melhores estimativas de erro (WITTEN & FRANK, 2005). Além de representar a prática mais usual em modelos *Perceptron* multicamada (ANDERSEN &MARTINEZ, 1999).

Por ser um banco de dados de grandes dimensões (288.382 linhas x 83 colunas), optou-se por manter o número de épocas padrão (*epoch* = 10), já que seu incremento demandaria mais recursos computacionais, tornando inviável a execução de todos os cenários e testes.

Em relação à topologia da rede existem diversas heurísticas na literatura que auxiliam na determinação do número de neurônios por camada intermediária. Sheela & Deepa (2013) avaliaram oito desses modelos. Além disso, propuseram um modelo (ver Equação (*35*), sendo proposto nesse trabalho e o que apresentou menor Erro médio quadrático (EMQ). No entanto, ao ser testado com a base de dados desta dissertação, essa heurística apresentou acurácias muito menores a 0,45 e EMQ próximo de 0,60, como se aprecia na Tabela 4.14.

Tabela 4.14. Número de neurônios na camada oculta por método heurístico.

Modelo	n	N_h	Acurácia	EMQ
M0	17	4	0,4127	0,5945
M1	21	4	0,3944	0,6022
M2	66	4	0,3481	0,6438
M3	53	4	0,3681	0,6213
M4	70	4	0,3508	0,6335
M5	83	4	0,3990	0,6775

Elaboração própria

Sendo assim, decidiu-se estimar o número de neurônios de maneira iterativa por meio de prova–erro. Para isso, foram testadas 13 topologias para duas camadas intermediárias.

$$N_h = \frac{4n^2 + 3}{n^2 - 8} \tag{35}$$

Em que:

N_h é o número de neurônios da camada intermediária; e

n é o número de neurônios da camada de entrada.

A testagem iniciou-se com a implantação de duas camadas ocultas. A primeira delas com apenas um neurônio e a segunda com dez neurônios (1,10). Contudo, observaram-se acurácias inferiores a 0,35. Razão pela qual o número de neurônios da primeira camada foi aumentado em um em cada rodada até alcançar a configuração (9,10) (ver Tabela 4.15).

Tabela 4.15. Acurácia e Erro global.

Config.	M0		M1		M2		M3		M4		M5	
	Acc	Err	Acc	Err	Acc	Err	Acc	Err	Acc	Err	Acc	Err
1-10	0,3283	0,6717	0,3259	0,6741	0,2691	0,7309	0,2524	0,7476	0,2184	0,7816	0,2202	0,7798
2-10	0,4570	0,5430	0,4437	0,5563	0,4541	0,5459	0,4467	0,5533	0,3964	0,6036	0,3825	0,6175
3-10	0,4877	0,5123	0,4416	0,5584	0,4633	0,5367	0,4344	0,5656	0,4642	0,5358	0,4341	0,5659
4-10	0,5216	0,4784	0,4983	0,5017	0,4575	0,5425	0,4831	0,5169	0,4947	0,5053	0,4338	0,5662
5-10	0,5270	0,4730	0,5015	0,4985	0,4656	0,5344	0,4977	0,5023	0,4577	0,5423	0,4377	0,5623
6-10	0,5306	0,4694	0,5159	0,4841	0,4798	0,5202	0,4961	0,5039	0,4621	0,5379	0,4548	0,5452
7-10	0,5306	0,4694	0,5235	0,4765	0,4738	0,5262	0,5058	0,4942	0,4541	0,5459	0,4691	0,5309
8-10	0,5520	0,4480	0,5649	0,4351	0,4597	0,5403	0,5131	0,4869	0,4805	0,5195	0,4763	0,5237
9-10	0,5144	0,4856	0,5443	0,4557	0,5036	0,4964	0,5916	0,4084	0,6005	0,3995	0,4994	0,5006
10-100	0,6310	0,3690	0,6519	0,3481	0,5738	0,4262	0,5916	0,4084	0,6005	0,3995	0,5951	0,4049
20-100	0,7238	0,2762	0,7087	0,2913	0,7275	0,2725	0,7332	0,2668	0,7144	0,2856	0,8089	0,1911
30-100	0,8085	0,1915	0,7665	0,2335	0,8918	0,1082	0,8312	0,1688	0,8926	0,1074	0,9093	0,0907
40-100	0,7954	0,2046	0,8359	0,1641	0,9436	0,0564	0,9100	0,0900	0,9534	0,0466	0,9657	0,0343

Elaboração própria.

Nota-se que sob a configuração (9,10), o erro global diminuiu nos modelos, oscilando entre 0,3995 e 0,5006, o que é considerado um desempenho muito baixo, sendo

necessário modificar novamente a topologia da rede neural. Dessa vez, aumentou-se em dez vezes a quantidade de neurônios da segunda camada oculta, permitindo a continuação do processo de aprendizado dos modelos.

Consequentemente geraram-se três topologias adicionais, tendo 10, 20 e 30 neurônios na primeira camada oculta e mantendo constante 100 neurônios na segunda camada intermediária. O experimento foi interrompido quando quatro dos seis modelos alcançaram acurácias próximas a 0,90.

5. ANÁLISE DE RESULTADOS

À medida que são incorporados novos neurônios nas camadas intermediárias, o desempenho dos modelos melhora. No entanto, como se observa na Figura 5.1., inicialmente os modelos com menor número de varáveis (M0 e M1) apresentaram menor erro global, situação que foi revertida a partir da configuração 20-100.

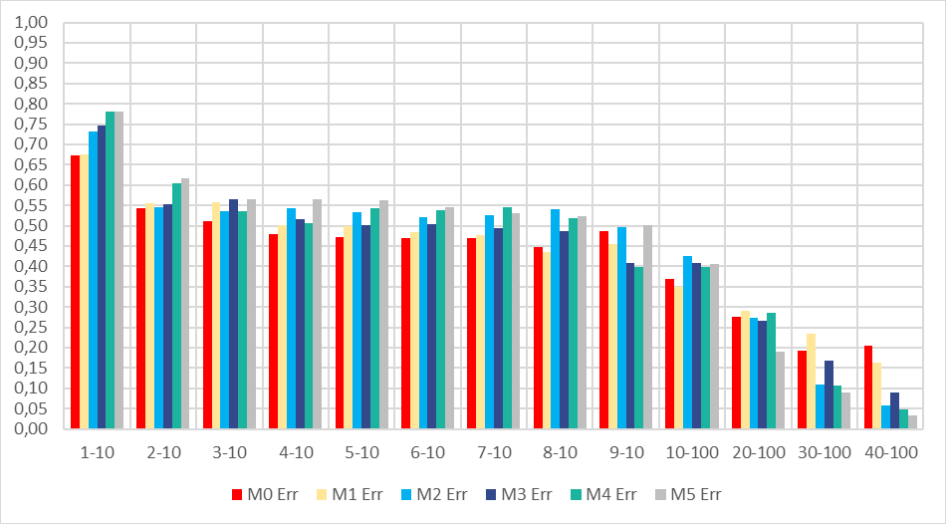

Figura 5.1. Evolução do erro global nos modelos.
Elaboração própria.

A análise baseada na acurácia e no erro global tornou-se insuficiente nas configurações, pois, à medida que um maior número de neurônios era utilizado, essas métricas diminuíam, porém, os resultados eram muito próximos um do outro, impossibilitando a comparação direta.

Decidiu-se realizar uma análise mais profunda por meio dos resultados obtidos e do cálculo de indicadores como Precisão, *Recall* e F1 (nível meso). Em primeiro lugar, observou-se que cada modelo era capaz de reconhecer um número de destinos. Em outras palavras, a matriz original é composta por 320 origens e 271 destinos. No entanto, nos cenários com baixo número de neurônios, o total de viagens era distribuído no máximo em 99 destinos estimados, ou seja, em 36,53% dos destinos reais.

Tabela 5.1. Destinos identificados por topologia.

Config.	M0		M1		M2		M3		M4		M5	
	D. I.	%.	D. I.	%.	D. I.	%.	D. I.	%	D. I.	%.	D. I.	%
1-10	19	7,01%	17	6,27%	17	6,27%	18	6,64%	19	7,01%	18	6,64%
2-10	31	11,44%	27	9,96%	0	0,00%	37	13,65%	35	12,92%	31	11,44%
3-10	31	11,44%	37	13,65%	30	11,07%	29	10,70%	34	12,55%	34	12,55%
4-10	41	15,13%	43	15,87%	31	11,44%	28	10,33%	34	12,55%	35	12,92%
5-10	34	12,55%	34	12,55%	31	11,44%	43	15,87%	38	14,02%	29	10,70%
6-10	37	13,65%	38	14,02%	31	11,44%	38	14,02%	31	11,44%	32	11,81%
7-10	39	14,39%	35	12,92%	33	12,18%	37	13,65%	32	11,81%	36	13,28%
8-10	46	16,97%	46	16,97%	40	14,76%	52	19,19%	37	13,65%	41	15,13%
9-10	41	15,13%	39	14,39%	34	12,55%	99	36,53%	90	33,21%	43	15,87%
10-100	84	31,00%	93	34,32%	78	28,78%	99	36,53%	90	33,21%	91	33,58%
20-100	124	45,76%	105	38,75%	162	59,78%	154	56,83%	159	58,67%	224	82,66%
30-100	165	60,89%	145	53,51%	249	91,88%	213	78,60%	251	92,62%	263	97,05%
40-100	140	51,66%	170	62,73%	261	96,31%	250	92,25%	265	97,79%	268	98,89%

Elaboração própria.

Na Tabela 5.1. é apresentada a relação de destinos identificados (D.I.) por topologia utilizada. Nota-se que no caso de M0 na configuração 30-100 distribuiu o número de viagens observadas em 60,89% dos destinos. Por outro lado, modelos com variáveis socioeconômicas encontraram seu máximo desempenho na configuração 40-100.

Mostra-se na Figura 5.2. o comportamento da medida F para cada modelo e topologia, evidenciando que, no nível de classe, os modelos M2 e M4, com a configuração 30-100, e M2, M3, M4 e M5, com 40-100 neurônios, na primeira e segunda camada intermediária, realizam ao menos 75% das predições corretamente.

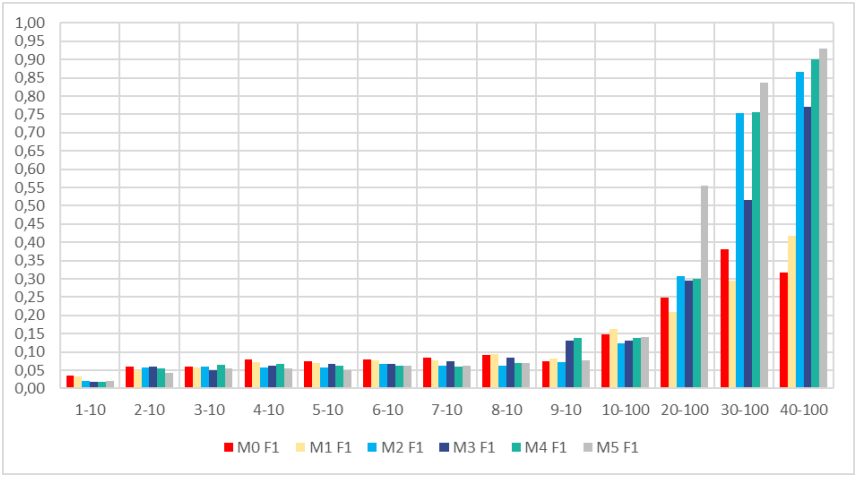

Figura 5.2. Evolução do F1 nos modelos.
Elaboração própria.

Nos modelos baseados no modelo gravitacional M0 e M1, observa-se que, ainda na configuração 40-100, cada classe em média tinha 31,69% e 45,80% de chances de acerto (ver Tabela 5.2.). Já nos modelos que incorporaram variáveis socioeconômicas, o modelo que apresentou menor taxa de acertos no nível de classe é o M3, modelo que se caracteriza por não considerar nenhum tipo de impedância (ver Tabela 5.3.).

Tabela 5.2. Métricas de classe de M0 a M2.

Config.	M0				M1				M2			
	Precisão	Recall	F1	Kappa	Precisão	Recall	F1	Kappa	Precisão	Recall	F1	Kappa
1-10	0,0304	0,0554	0,0348	0,3050	0,0270	0,0529	0,0318	0,2984	0,0189	0,0356	0,0211	0,2351
2-10	0,0549	0,0853	0,0592	0,4425	0,0420	0,0838	0,0512	0,4295	0,0569	0,0743	0,0567	0,4396
3-10	0,0599	0,0834	0,0599	0,4737	0,0572	0,0853	0,0572	0,4268	0,0605	0,0749	0,0601	0,4481
4-10	0,0846	0,1005	0,0803	0,5086	0,0793	0,0934	0,0709	0,4846	0,0596	0,0731	0,0569	0,4430
5-10	0,0730	0,0958	0,0741	0,5143	0,0680	0,0871	0,0684	0,4890	0,0578	0,0702	0,0578	0,4518
6-10	0,0803	0,0975	0,0785	0,5187	0,0830	0,0912	0,0767	0,5025	0,0693	0,0774	0,0674	0,4666
7-10	0,0855	0,0962	0,0835	0,5164	0,0820	0,0917	0,0770	0,5114	0,0669	0,0760	0,0619	0,4599
8-10	0,0945	0,1106	0,0919	0,5394	0,1048	0,1094	0,0951	0,5539	0,0664	0,0756	0,0624	0,4450
9-10	0,0821	0,0894	0,0756	0,5016	0,0852	0,1025	0,0827	0,5323	0,0790	0,0848	0,0723	0,4908
10-100	0,1596	0,1717	0,1478	0,6222	0,1849	0,1848	0,1632	0,6436	0,1512	0,1438	0,1235	0,5636
20-100	0,2746	0,2718	0,2477	0,7173	0,2374	0,2286	0,2104	0,7017	0,3412	0,3517	0,3083	0,7210
30-100	0,4255	0,4093	0,3797	0,8040	0,3347	0,3188	0,2958	0,7609	0,7907	0,7643	0,7534	0,8893
40-100	0,3646	0,3455	0,3169	0,7905	0,4489	0,4580	0,4173	0,8320	0,8893	0,8769	0,8673	0,9423

Elaboração própria.

De maneira geral, pode-se dizer que os modelos M0 e M1 apresentam os piores resultados tanto globalmente quanto na análise por classe. Com relação aos modelos socioeconômicos, apresentam um F-score entre 0,7705 e 0,9309, sendo M3 o modelo com desempenho mais baixo, com F = 0,5161.

Tabela 5.3. Métricas de classe de M3 a M5.

Config.	M3				M4				M5			
	Precisão	Recall	F1	Kappa	Precisão	Recall	F1	Kappa	Precisão	Recall	F1	Kappa
1-10	0,0160	0,0363	0,0193	0,2208	0,0125	0,0366	0,0174	0,1926	0,0198	0,0393	0,0197	0,1847
2-10	0,0598	0,0797	0,0599	0,4317	0,0588	0,0726	0,0539	0,3784	0,0495	0,0572	0,0429	0,3650
3-10	0,0470	0,0641	0,0501	0,4192	0,0659	0,0812	0,0648	0,4502	0,0594	0,0672	0,0551	0,4181
4-10	0,0625	0,0797	0,0631	0,4689	0,0693	0,0828	0,0661	0,4813	0,0631	0,0676	0,0558	0,4189
5-10	0,0717	0,0877	0,0678	0,4849	0,0653	0,0742	0,0623	0,4442	0,0534	0,0643	0,0508	0,4225
6-10	0,0801	0,0850	0,0678	0,4820	0,0630	0,0744	0,0628	0,4474	0,0663	0,0726	0,0624	0,4401
7-10	0,0757	0,0854	0,0747	0,4932	0,0665	0,0684	0,0591	0,4398	0,0682	0,0721	0,0630	0,4559
8-10	0,0927	0,0981	0,0836	0,5005	0,0699	0,0804	0,0696	0,4673	0,0769	0,0771	0,0692	0,4627
9-10	0,1550	0,1561	0,1319	0,5819	0,1642	0,1573	0,1380	0,5909	0,0858	0,0826	0,0758	0,4875
10-100	0,1550	0,1561	0,1319	0,5819	0,1642	0,1573	0,1380	0,5909	0,1621	0,1551	0,1407	0,5854
20-100	0,3261	0,3277	0,2945	0,7269	0,3351	0,3346	0,2993	0,7075	0,5941	0,5762	0,5555	0,8044
30-100	0,5545	0,5572	0,5161	0,8272	0,7834	0,7748	0,7551	0,8901	0,8673	0,8439	0,8368	0,9071
40-100	0,8053	0,7912	0,7705	0,9079	0,9136	0,9100	0,9000	0,9523	0,9363	0,9390	0,9309	0,9649

Elaboração própria.

Nos modelos 2 e 4, são testados os efeitos que podem gerar a presença de apenas uma impedância, sendo alternados o tempo de viagem e as oportunidades intervenientes. Com base nas Tabelas 5.2 e 5.3, percebe-se uma leve diferença de 0,0327 a favor do modelo 4, que considera as oportunidades intervenientes na topologia 40-100.

Esses mesmos modelos apresentam métricas F quase idênticas na configuração 30-100, sendo 0,7534 para M2 e 0,7551 para M4, e próximas nas topologias mais básicas. Portanto, não é relevante o tipo de impedância usada apenas para garantir a presença de alguma delas.

O modelo 5 contém os dois tipos de impedâncias avaliadas no presente trabalho, sendo o que apresenta os melhores resultados, com um F 0,9309 e erro global de 0,0343. No entanto, considerado com M4 (F 0,9000 e erro de 0,0466), não houve um ganho significativo no modelo.

É necessário apontar que quanto maior o número de variáveis existentes nos modelos, mais recursos computacionais são requeridos e nesse caso M4 torna-se mais viável.

Por fim, foi calculado o coeficiente de concordância Kappa (ver Tabela 5.2. e Tabela 5.3.), que mede a concordância associada entre dois grupos de dados. De acordo com Landis & Koch (1977), M0 e M1 apresentam uma concordância forte entre os dados observados e os previstos, a partir da tipologia 10-10. M5 teve uma concordância quase perfeita a partir da topologia 20-100 e os demais modelos a alcançaram em 30-100.

Na Figura 5.3. é apresentado o número de viagens realizadas por intervalo de tempo de viagem, nesta observa-se que o Modelo com melhor desempenho é M4 pois a curva desenhada por ele se assemelha aos dados observados.

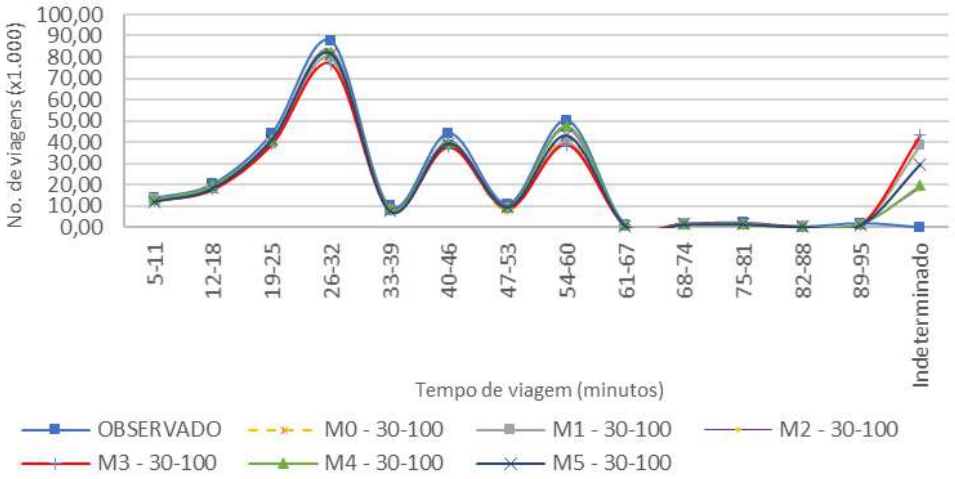

Figura 5.3. Relação nº de viagens x Tempo de viagem (30-100).
Elaboração própria.

Destaca-se na Figura 5.3. a classe "indeterminado", que faz referência quando o modelo previu um novo par origem–destino, o qual não existe no banco de dados original e, consequentemente, o tempo de viagem desse par é desconhecido.

Decidiu-se realizar uma análise no nível micro, mediante a execução de um teste de significância das variáveis de entrada envolvidas nos modelos, especificamente naqueles com topologia 30-100, pois demostraram obter bons resultados.

O método desenvolvido por Gedeon (1997), usado no pacote h2o, analisa magnitudes e medidas funcionais dos dados de entrada em modelos baseados em redes neurais. Assim, por meio dessas medidas, é determinada a influência que exerce as variáveis de entrada sobre os modelos

Devido à estrutura da base de dados usada no estudo de caso, o método Gedeon pontuará as 10 primeiras classes com base na sua significância estatística, em vez de ranquear em função das variáveis. Consequentemente será apresentado um "perfil" das pessoas com maior influência para cada modelo.

Nas Figuras 5.4 a 5.9, são apresentadas as principais classes que influenciam os modelos, tendo-se como denominador comum a atração da zona de tráfego de destino na primeira posição.

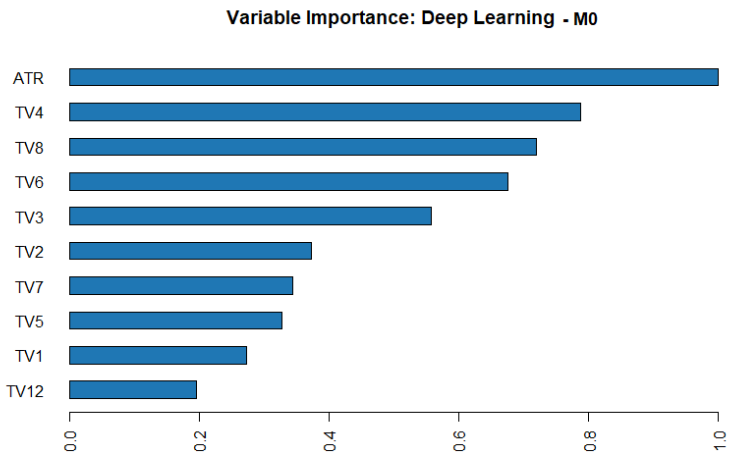

Figura 5.4. Significância das variáveis em M0 (30-100).
Elaboração própria.

No modelo 0 observa-se que 9 das 13 classes pertencentes à variável tempo de viagem aparecem no *ranking*. Portanto essa variável é primordial em M0 (ver Figura 5.4.). Sendo assim, nesse modelo são priorizadas as viagens com tempo de deslocamento menores que 60 minutos.

Perante a substituição do tipo de impedância, o modelo 1 possui um comportamento similar em comparação com M0, tendo grande importância as variáveis de dissuasão de viagens (ver Figura 5.5.).

Tanto M0 quanto M1 utilizam variáveis agregadas em sua estrutura, sendo evidente que para cada um desses modelos a atração de viagens da zona de destino e a impedância são relevantes na construção dos modelos. Por outro lado, nota-se que a produção de viagens tem um nível de significância menor que 20%, portanto, não aparece nas Figuras 5.4. e Figura 5.5.

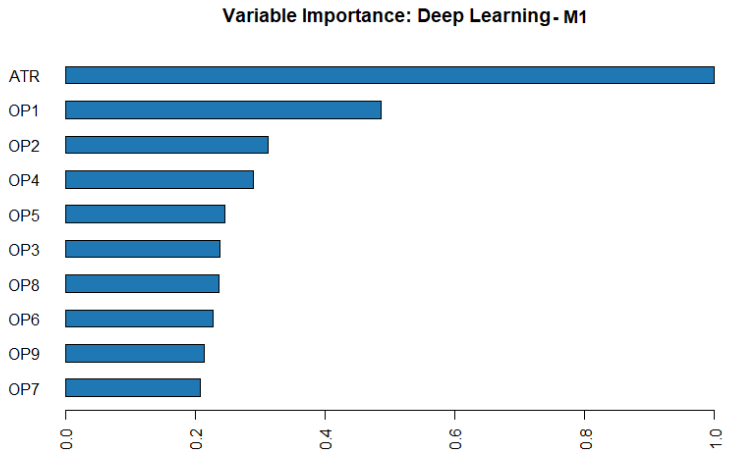

Figura 5.5. Significância das variáveis em M1 (30-100).
Elaboração própria.

O modelo 2 é o primeiro a incorporar variáveis agregadas e desagregadas na sua estrutura. O tempo de viagem (agregada) ocupa vários lugares na classificação de classes, seguida pelo nível de instrução, porém, com exceção da atração, a significância estatística dessas variáveis não ultrapassa o limite de 40%.

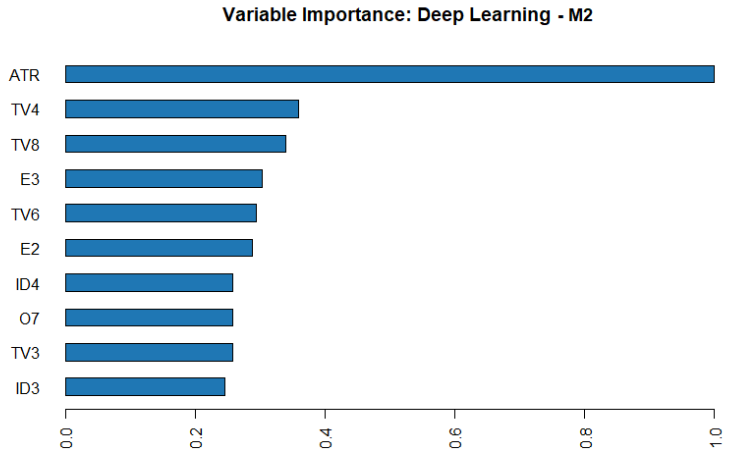

Figura 5.6. Significância das variáveis em M2 (30-100).
Elaboração própria.

Com base nas informações da Figura 5.6., pode-se inferir que as viagens realizadas por pessoas entre 19 e 32 anos, que concluíram principalmente o ensino fundamental e em alguns casos o ensino médio, e que ocupam a dupla função de dona de casa e estudantes são mais relevantes para o modelo. Essas viagens não ultrapassam uma hora de duração.

A ausência de impedâncias em M3 e, consequentemente, a utilização de variáveis próprias do indivíduo permitiram que a idade se destaque no modelo, mostrando 5 das 13 classes que o compõem (ver Figura 5.7.).

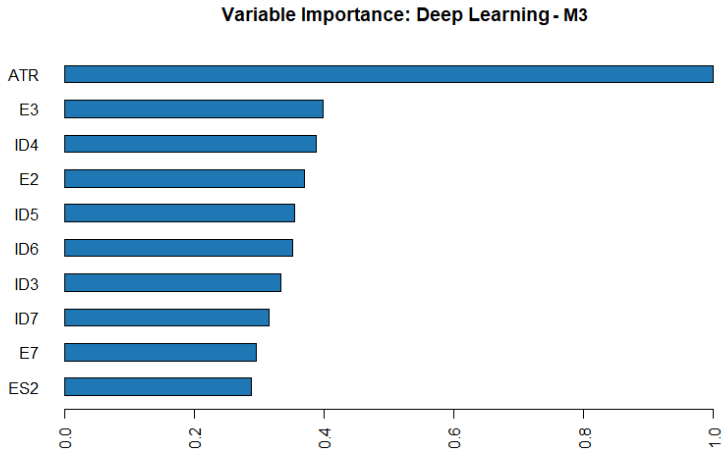

Figura 5.7. Significância das variáveis em M3 (30-100).
Elaboração própria.

Assim, a partir da Figura 5.7, pode-se determinar o perfil identificado pelo modelo, sendo importantes as viagens realizadas por pessoas dentro do intervalo de idade entre 18 e 52 anos, com um perfil acadêmico de ensino fundamental e médio e pertencentes à classe socioeconômica 2 (baixo nível socioeconômico).

Indiferentemente do tipo de impedância utilizada nos modelos, observa-se a importância que esse tipo de informação agregada possui. Como se mostra na Figura 5.8, as oportunidades intervenientes estão presentes no escalão das classes seguidas pelo nível de instrução e pelaa ocupação das pessoas.

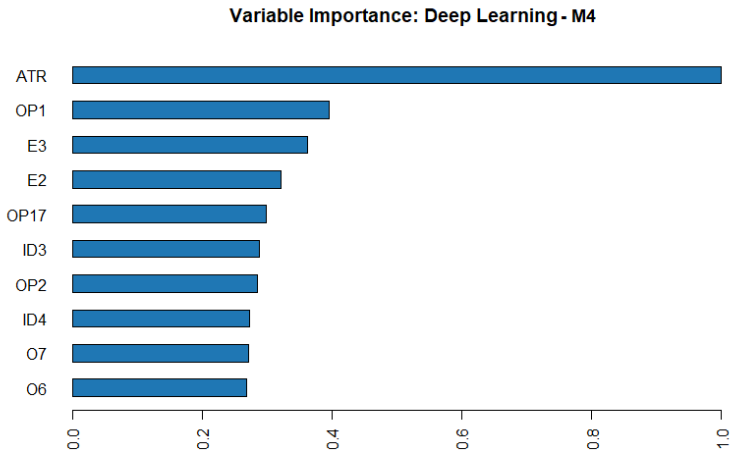

Figura 5.8. Significância das variáveis em M4 (30-100).
Elaboração própria.

Assim sendo, a perfilação mostra que para esse modelo as viagens realizadas por adultos jovens (entre 18 e 31 anos) que desempenham papeis simultâneos na área acadêmica e laboral são relevantes (ver Figura 5.8.).

Na Figura 5.9. nota-se em M5 que, embora a significância do tempo de viagem oscile entre 0,4 e 0,5, essa variável aporta maior número de classe ao modelo em comparação com as oportunidades intervenientes. Com relação a essa última, a única categoria presente no *ranking* obedece àquela que agrupa o menor número de oportunidades possíveis (até 32.509).

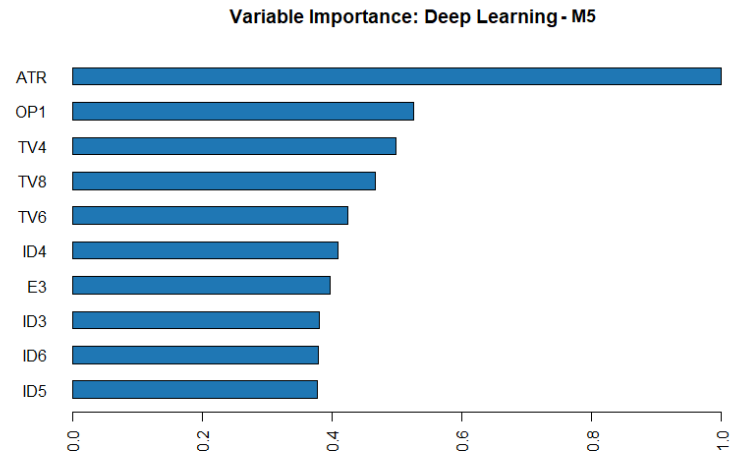

Figura 5.9. Significância das variáveis em M5 (30-100).
Elaboração própria.

A partir dos *rankings* classificatórios apresentados nas Figuras 5.4 a 5.9, pode-se concluir que:

- As variáveis de tipo agregado possuem maior relevância nos modelos, destacando-se a atração da zona e as impedâncias testadas;
- Com exceção de M5 e da atração do destino da viagem, não há outra variável com uma significância estatística maior que 40%;
- Variáveis desagregadas como sexo, modo de transporte, tempo de emprego, lugar de trabalho e produção de viagens na zona de origem possuem uma relevância menor que 0,2 em todos os casos. Razão pela qual não estão presentes nos gráficos figuras;
- No quinto modelo observa-se que só as variáveis de impedância alcançam uma significância estatística próxima a 50%;
- A fim de determinar a influência da impedância nos modelos, determinou-se que o tempo de viagem possui uma maior relevância do que as oportunidades intervenientes, sendo evidente nas Figuras 5.4., Figuras 5.5. e Figuras 5.9.;
- O método de Gedeon determinou que a atração das zonas de tráfego é a informação de entrada de maior relevância nos modelos.

6. CONCLUSÕES E RECOMENDAÇÕES

Nesta dissertação, foi proposto um procedimento que permite a estruturação de modelos híbridos baseados em redes neurais artificiais na distribuição de viagens. A arquitetura da rede neural foi definida mediante o teste e cálculo de indicadores visando à obtenção de melhores resultados.

Os resultados obtidos nestes modelos foram calculados para viagens com motivo trabalho e que tem sido realizadas por médio do sistema de transporte público coletivo. Considera-se pertinente que em pesquisas futuras sejam considerados outros motivos de viagem a fim de avaliar o desempenho e a estabilidade das topologias apresentadas.

Nesse processo, encontrou-se uma dificuldade para determinar a topologia adequada para cada modelo, pois não existe um procedimento estabelecido para esse fim. Sendo necessária a utilização de iterações.

Na literatura consultada, existem diversos trabalhos acadêmicos que buscam, de maneira estruturada, dar solução ao problema da topologia anteriormente descrito. Foi encontrada uma quantidade inumerável de heurísticas, porém, observa-se como principal dificuldade que essas heurísticas possuem uma capacidade nula de generalização. Consequentemente, quando utilizadas em bancos de dados diferentes às quais foram desenhadas, apresentam baixos níveis de desempenho.

A partir da realização das análises nos níveis macro, meso e micro, além da execução do cálculo de métricas, pode-se se concluir que os modelos nos quais encontram-se ausentes as características socioeconômicas do indivíduo possuem baixos níveis de precisão na estimativa, requerendo maior número de neurônios em busca de uma convergência.

Assim, essas informações de caráter desagregado influenciam na realização da viagem e, consequentemente, não podem desconsiderados no processo de predição dos deslocamentos na zona de estudo.

Com base no método Gedeon, observou-se que as principais variáveis que influenciam na origem da viagem são a idade, a classe social (condição socioeconômica), o nível de instrução e a ocupação. Dessa forma, a origem da viagem não depende do modo de transporte ou do sexo da pessoa.

Comparando as variáveis desagregadas e agregadas, observa-se que estas últimas possuem uma maior significância estatística. Tendo em vista que, de acordo com o método Gedeon, esse tipo de variável ocupou os primeiros lugares na classificação das categorias.

Embora seja possível realizar modelos exclusivamente com dados agregados, determinou-se que a inclusão de informações referentes ao indivíduo complementa e melhora os resultados de desempenho globais, permitindo ter uma estimativa das viagens com menor índice de erro.

Da mesma forma, evidenciou-se que fatores de dissuasão, como tempo de viagem ou oportunidades intervenientes, possuem um papel importante na compreensão dos padrões de deslocamentos. Em M3 foi evidente que a falta desses elementos faz com que o modelo seja ineficiente, apresentando baixos níveis de acerto.

Na avaliação das impedâncias, observou-se que as oportunidades intervenientes têm uma menor significância estatística em comparação com o tempo de viagem. Esse resultado pode ter sido influenciado pela forma como foi tratada a informação, convertendo a quantidade de empregos do subdistrito para zona de tráfego. Recomenda-se avaliar outros métodos de desagregação de informação.

Com relação aos questionamentos apresentados na introdução, pode-se concluir que o uso de variáveis agregadas e desagregadas permitem a obtenção de informações relevantes sobre como os diversos fatores influenciam na realização da viagem.

Acerca do grau em que esse tipo de informações influência nos padrões de viagens, não foi possível encontrar um padrão nos resultados obtidos que permitisse determinar o peso que o grupo de variáveis agregadas e desagregadas possui nos modelos.

No ANEXO 3 é apresentado o algoritmo desenvolvido para esta dissertação, contendo comentários do que deve ser alterado dependendo do modelo.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDEL-ATY, M., LEE, J., SIDDIQUI, C. "Geographical unit based analysis in the context of transportation safety planning", **Transportation Research Part A: Policy and Practice**, v. 49, p. 62–75, 2013. DOI: 10.1016/j.tra.2013.01.030. Disponível em: http://dx.doi.org/10.1016/j.tra.2013.01.030.

AGUIAR JÚNIOR, S. R. de. "Modelo RAPIDE : uma aplicação de mineração de dados e redes neurais artificiais para a estimativa da demanda por transporte rodoviário interestadual de passageiros no Brasil", **Universidade Católica de Brasília**, p. 140, 2004.

AKAMINE, A. "Explorando alternativas para construção de modelos neurais de interação espacial", **Universidade de São Paulo**, 2005.

ALCALDIA DE MEDELLÍN. "Acuerdo 48 de 2014", **Gazeta Oficial**, v. 4267, p. 1–877, 2014.

ALCALDÍA DE MEDELLÍN. **Datos generales de Municipio**. 2020. Disponível em: https://www.medellin.gov.co/irj/portal/medellin?NavigationTarget=navurl://6488ef50a6787e1fdbc4e42e62a46a67. Acesso em: 7 abr. 2020.

ALCALDÍA DE MEDELLÍN. **Usos generales del Suelo Urbano POT 2014 - 2017**. 2014. Datos abiertos. Disponível em: https://www.datos.gov.co/Ordenamiento-Territorial/Usos-Generales-del-Suelo-Urbano-POT-2014-2027-Muni/fzyj-2xtv. Acesso em: 9 abr. 2020.

ÁREA METROPOLITANA DO VALE DE ABURRÁ - AMVA. ENCUESTA ORIGEN DESTINO DE HOGARES Y DE CARGA PARA EL VALLE DE ABURRÁ - INFORME FINAL. 2012. 347 f. Área Metropolitana do Vale de Aburrá - AMVA, Medellín, 2012.

ANDERSEN, T., MARTINEZ, T. "Cross Validation and MLP Architecture Selection", **Proceedings of the IEEE International Joint Conference on Neural Networks IJCNN'99**, 1999.

ARENTZE, T. A., MOLIN, E. J. E. "Travelers' preferences in multimodal networks: Design and results of a comprehensive series of choice experiments", **Transportation Research Part A: Policy and Practice**, v. 58, p. 15–28, 2013. DOI: 10.1016/j.tra.2013.10.005.

BAZZAN, A. L. C., KLÜGL, F. "Sistemas Inteligentes de Transporte e Tráfego : uma Abordagem de Tecnologia da Informação", p. 2296–2337, 2005.

BHAVSAR, P., SAFRO, I., BOUAYNAYA, N. "Machine Learning in Transportation Data Analytics". **Data Analytics for Intelligent Transportation Systems**, [S.l.], Elsevier Inc., 2017. p. 283–308. DOI: 10.1016/B978-0-12-809715-1.00012-2. Disponível em: http://dx.doi.org/10.1016/B978-0-12-809715-1.00012-2.

BRAMER, M. **Introduction to Data Mining**. Third edit ed. [S.l.], Springer, 2016.

BRUTON, M. J. **Introdução ao planejamento dos transportes**. [S.l.], Editora Interciência, 1979.

CABRERA DELGADO, J., BONNEL, P. "Level of aggregation of zoning and temporal transferability of the gravity distribution model: The case of Lyon", **Journal of**

Transport Geography, v. 51, p. 17–26, 2016. DOI: 10.1016/j.jtrangeo.2015.10.016. Disponível em: http://dx.doi.org/10.1016/j.jtrangeo.2015.10.016.

CALDAS, M. U. de C., FAVERO, R., PITOMBO, C. S. "Uso de redes neurais artificiais para abordagem desagregada de distribução de viagens urbanas com dados de preferência revelada", **33 Congresso de Pesquisa e Ensino em Transporte da ANPET**, 2019. DOI: 10.1017/CBO9781107415324.004.

CASCETTA, E. **Transportation Systems Analysis: Models and Applications**. Second ed. [S.l.], Springer, 2009.

CASCETTA, E., PAGLIARA, F., PAPOLA, A. "Alternative approaches to trip distribution modelling: A retrospective review and suggestions for combining different approaches", **Papers in Regional Science**, v. 86, n. 4, p. 597–620, 2007. DOI: 10.1111/j.1435-5957.2007.00135.x.

CELIK, H. M. "Sample size needed for calibrating trip distribution and behavior of the gravity model", **Journal of Transport Geography**, v. 18, n. 1, p. 183–190, 2010. DOI: 10.1016/j.jtrangeo.2009.05.013. Disponível em: http://dx.doi.org/10.1016/j.jtrangeo.2009.05.013.

CHATTERJEE, A., VENIGALLA, M. M., "TRAVEL DEMAND FORECASTING FOR URBAN TRANSPORTATION PLANNING". **HANDBOOK OF TRANSPORTATION ENGINEERING**, [S.l.], McGraw-Hill, 2004. .

CONGRESO DE LA REPÚBLICA DE COLOMBIA. "Ley 689 de 2001", **Diario Oficial**, n. 44, 2001.

DE GRANGE, L., FERNÁNDEZ, E., DE CEA, J. "A consolidated model of trip distribution", **Transportation Research Part E: Logistics and Transportation Review**, v. 46, n. 1, p. 61–75, 2010. DOI: 10.1016/j.tre.2009.06.001. Disponível em: http://dx.doi.org/10.1016/j.tre.2009.06.001.

DOBSON, A. J., BARNETT, A. G. **An Introduction to Generalized Linear Models**. Third Edit ed. [S.l.], CRC Press Taylor& Francis Group, 2013. v. 53.

DOUGHERTY, M. "A review of neural networks applied to transport", **Transportation Research Part C**, v. 3, n. 4, p. 247–260, 1995. DOI: 10.1016/0968-090X(95)00009-8.

DOWLE, MATT; SRINIVASAN, A. **data.table: Extension of "data.frame".** . [S.l: s.n.]. Disponível em: https://cran.r-project.org/package=data.table%0A. , 2020.

EBERHART, R. C., DOBBINS, R. W. **Neural Network PC Tools - A Practical Guide**. Maryland, Academic Press, 1990.

FAGHRI, A., MARTINELLI, D., DEMETSKY, M. J. "Neural network applications in transportation engineering", **Artificial Neural Networks for Civil Engineers: Fundamentals and Applications**, v. 1, p. 137–159, 1997.

FERREIRA, R. P., MARTINIANO, A., FERREIRA, A. "Study on Daily Demand Forecasting Orders using Artificial Neural Network", **IEEE Latin America Transactions**, v. 14, n. 3, p. 1519–1525, 2016. DOI: 10.1109/TLA.2016.7459644.

GEDEON, T. D. "Data mining of inputs: analysing magnitude and functional measures.", **International journal of neural systems**, v. 8, n. 2, p. 209–218, 1997. DOI: 10.1142/S0129065797000227.

GHASRI, M., HOSSEIN RASHIDI, T., WALLER, S. T. "Developing a disaggregate travel demand system of models using data mining techniques", **Transportation Research Part A: Policy and Practice**, v. 105, n. June 2016, p. 138–153, 2017. DOI: 10.1016/j.tra.2017.08.020. Disponível em: http://dx.doi.org/10.1016/j.tra.2017.08.020.

GODOY, S., SHAW, I. S. **Controle e modeloagem fuzzy**. São Paulo, Blucher: FAPESP, 2007.

GONÇALVES; D. N. S. **Procedimento para aplicação de Redes Neurais Artificiais na estimativa de matrizes Origem - Destino de carga**. 2015. Instituto Militar de Engenharia, 2015.

GONÇALVES, D. N. S., SILVA, M. A. da, D'AGOSTO, M. de A. "Procedimento para uso de Redes Neurais Artificiais no planejamento estratégico de fluxo de carga no Brasil", **Journal of Transport Literature**, v. 9, n. 1, p. 45–49, 2015. DOI: 10.1590/2238-1031.jtl.v9n1a9.

GONÇALVES, M. B. "Desenvolvimento e teste de um novo modelo gravitacional - de oportunidades para distribuição de viagens", **Universidade Federal de Santa Catarina**, 1992.

HESS, S., BOLDUC, D., POLAK, J. W. "Random covariance heterogeneity in discrete choice models", **Transportation**, v. 37, n. 3, p. 391–411, 2010. DOI: 10.1007/s11116-009-9255-3.

HIRUN, W. "Evaluation of interregional freight generation modelling methods by using nationwide commodity flow survey data", **American Journal of Engineering and Applied Sciences**, v. 9, n. 3, p. 625–634, 2016. DOI: 10.3844/ajeassp.2016.625.634.

HONAKER, JAMES; KING, GARY; BLACKWELL, M. **Amelia II: A Program for Missing Data.** . [S.l.], Journal of Statistical Software, 45(7), 1-47. Disponível em: http://www.jstatsoft.org/v45/i07/. , 2011.

INSTITUTE OF TRANSPORTATION ENGINEERS. **Transportation Planning Handbook**. Fourth ed. [S.l.], WILEY, 2016.

KARLAFTIS, M. G., VLAHOGIANNI, E. I. "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights", **Transportation Research Part C: Emerging Technologies**, v. 19, n. 3, p. 387–399, 2011. DOI: 10.1016/j.trc.2010.10.004. Disponível em: http://dx.doi.org/10.1016/j.trc.2010.10.004.

KHAKI, A. M., AFANDIZADEH, S., MOAYEDFAR, R. "Developing the composed probability model to predict household trip production (a case 4142study of isfahan city)", **Transport**, v. 24, n. 1, p. 30–36, 2009. DOI: 10.3846/1648-.2009.24.30-36.

KOMPIL, M., CELIK, H. M. "Modelling trip distribution with fuzzy and genetic fuzzy systems", **Transportation Planning and Technology**, v. 36, n. 2, p. 170–200, 2013. DOI: 10.1080/03081060.2013.770946.

KUHN, M. **caret: Classification and Regression Training.** . [S.l: s.n.]. Disponível em: https://cran.r-project.org/package=caret. , 2020.

LANDIS, J. R., KOCH, G. G. "The Measurement of Observer Agreement for Categorical Data", **Biometrics**, v. 33, n. 1, p. 159–174, 1977.

LEDELL, ERIN;GILL, NAVDEEP;AIELLO, SPENCER; FU, ANQI; CANDEL,

ARNO;CLICK, CLIFF; KRALJEVIC, TOM;NYKODYM, TOMAS; ABOYOUN, PATRICK; KURKA, M. M. M. **h2o: R Interface for the "H2O" Scalable Machine Learning Platform.** . [S.l: s.n.]. Disponível em: https://cran.r-project.org/package=h2o. , 2020.

LEMOS, B. M. **EXPLORAÇÃO DE MODELOS AGREGADOS PARA DISTRIBUIÇÃO DE VIAGENS URBANAS: UMA ABORDAGEM BASEADA NAS OPORTUNIDADES INTERVENIENTES**. 2020. 230 f. Universidade Federal do Rio de Janeiro, 2020.

LENORMAND, M., BASSOLAS, A., RAMASCO, J. J. "Systematic comparison of trip distribution laws and models", **Journal of Transport Geography**, v. 51, p. 158–169, 2016. DOI: 10.1016/j.jtrangeo.2015.12.008. Disponível em: http://dx.doi.org/10.1016/j.jtrangeo.2015.12.008.

LEON, A. C., "3.12 Descriptive and Inferential Statistics". **Neuroscience and Biobehavioral Psychology**, [S.l.], Elsevier Inc., 1998. p. 243–285. DOI: 10.1016/B0-12-369398-5/00145-6.

MANNING, C. D., RAGHAVAN, P., SCHUTZE, H. **Introduction to Information Retrieval**. [S.l.], Cambridge University Press, 2008.

MANOUT, O., BONNEL, P. "The impact of ignoring intrazonal trips in assignment models: a stochastic approach", **Transportation**, v. 46, n. 6, p. 2397–2417, 2018. DOI: 10.1007/s11116-018-9951-y. Disponível em: https://doi.org/10.1007/s11116-018-9951-y.

MCCULLOCH, W. S., PITTS, W. "A logical calculus of the ideas immanent in nervous activity", **The Bulletin of Mathematical Biophysics**, v. 5, n. 4, p. 115–133, 1943. DOI: 10.1007/BF02478259.

MCNALLY, M. G., "Chapter 3: The Four-Step Model". **Handbook of Transport Modelling**, [S.l: s.n.], 2016.

MENDONÇA, A. C. de; **Desenvolvimento de um Modelo de Previsão da Demanda de Passageiros do Transporte Rodoviário Interestadual Utilizando Regressão com Efeitos Espaciais Locais**. **Universidade de Brasília**. [S.l: s.n.]. , 2008.

MEYER, DAVID; DIMITRIADOU, EVGENIA; HORNIK, KURT; WEINGESSEL, ANDREAS;LEISCH, F. **e1071: Misc Functions of the Department of Statistics**. . [S.l.], , Probability Theory Group (Formerly: E1071), TU Wien. Disponível em: https://cran.r-project.org/package=e1071. , 2019.

MOZOLIN, M., THILL, J. C., LYNN USERY, E. "Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation", **Transportation Research Part B: Methodological**, v. 34, n. 1, p. 53–73, 2000. DOI: 10.1016/S0191-2615(99)00014-4.

NAZEM, M. **Contributions à la modélisation des déplacements en transport collectif**. 2014. 188 f. 2014.

NOVAES, A. G. **Sistemas de Transporte Volume 1: Análise da demanda**. Edgar Blüc ed. São Paulo, [s.n.], 1986.

OKABE, A. "Formulation of the Intervening Opportunities Model for Housing Location Choice Behavior", **Journal of Regional Science**, v. 17, n. 1, p. 31–40, 1977. DOI:

10.1111/j.1467-9787.1977.tb00470.x.

ORTUZAR, J. de D., WILLUMSEN, L. G. **Modelling Transport**. Fourth edi ed. [S.l: s.n.], 2011.

PAIVA, C. "MODELOS DE TRANSPORTE E TRÁFEGO NOVAS TECNOLOGIAS", **Universidade Pontifícia Católica de São Paulo**, 2011.

PERRAKIS, K., KARLIS, D., COOLS, M.."A Bayesian approach for modeling origin-destination matrices", **Transportation Research Part A: Policy and Practice**, v. 46, n. 1, p. 200–212, 2012. DOI: 10.1016/j.tra.2011.06.005. Disponível em: http://dx.doi.org/10.1016/j.tra.2011.06.005.

PINTO FERREIRA, R., MARTINIANO, A., FERREIRA, A. "Study on Daily Demand Forecasting Orders using Artificial Neural Network", **IEEE Latin America Transactions**, v. 14, n. 3, p. 1519–1525, 2016. DOI: 10.1109/TLA.2016.7459644.

PITOMBO, C. S., DE SOUZA, A. D., LINDNER, A. "Comparing decision tree algorithms to estimate intercity trip distribution", **Transportation Research Part C: Emerging Technologies**, v. 77, p. 16–32, 2017. DOI: 10.1016/j.trc.2017.01.009. Disponível em: http://dx.doi.org/10.1016/j.trc.2017.01.009.

PITOMBO, C. S., GUIMARÃES, H. S. "Uso de técnica de mineração de dados no auxílio à modelagem de distribuição de viagens intermunicipais Cira Souza Pitombo , Henrique Stramandinoli Guimarães", **Universidade de São Paulo**, v. 52, n. Md, p. 45–56, 2016.

RAIA JR, A. A. "Acessibilidade e Mobilidade na Estimativa de um Índice de Potencialde Viagens Utilizando Redes Neurais Artificiais e Sistema de Informações Geográfica", **Doutorado**, p. 212, 2000.

RICHARDS, M. G. "Disaggregate simultaneous urban travel demand models: A brief introduction", **Transportation**, v. 3, n. 4, p. 335–342, 1974. DOI: 10.1007/BF00167964.

ROCHA, S. S., PIANUCCI, M. N., PITOMBO, C. S. "Uso de Redes Neurais para previsão de produção de viagens : uma análise agregada", **ResearchGate**, n. November, p. 13, 2015.

SALINI, P. S., KEDIA, A., DHULIPALA, S. "Spatial distribution of urban trips in recently expanded Surat city through Fuzzy Logic with various clustering Techniques: A case study of typical metropolitan city in India", **Transportation Research Procedia**, v. 25, p. 2400–2411, 2017. DOI: 10.1016/j.trpro.2017.05.245. Disponível em: http://dx.doi.org/10.1016/j.trpro.2017.05.245.

SARKAR, A. "Application of Fuzzy Logic in Transport Planning", **International Journal on Soft Computing**, v. 3, n. 2, p. 1–21, 2012. DOI: 10.5121/ijsc.2012.3201. Disponível em: http://www.airccse.org/journal/ijsc/papers/3211ijsc01.pdf.

SCHMIDHUBER, J. "Deep Learning in neural networks: An overview", **Neural Networks**, v. 61, p. 85–117, 2015. DOI: 10.1016/j.neunet.2014.09.003. Disponível em: http://dx.doi.org/10.1016/j.neunet.2014.09.003.

SCHNEIDER, M. **Gravity Models and Trip Distribution Theory**. **Papers in Regional Science**. [S.l: s.n.]. , 1959.

SHEELA, K. G., DEEPA, S. N. "Review on Methods to Fix Number of Hidden Neurons in Neural Networks", **Hindawi**, v. 2013, 2013.

SILVA, I. N. da, SPATTI, D. H., FLAUZINO, R. A. **Redes Neurais Artificiais para engenharia e ciências aplicadas**. 2. ed. São Paulo, Artliber Editora Ltda., 2016.

SKANSI, S. **Introduction to deep learning: From Logical Calculus to Artificial Intelligence**. [S.l.], Springer US, 2018.

SOUZA ROMA, A. D. de;, GUIMARÃES, H. S., PITOMBO, C. S. "ANÁLISE DE DESEMPENHO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINAS PARA ANÁLISE DESAGREGADA DE VIAGENS INTERMUNICIPAIS", **Journal of Chemical Information and Modeling**, 2017. DOI: 10.1017/CBO9781107415324.004.

STOUFFER, S. A. "Intervening Oportunities: A Theory Relating Mobility And Distance", **American Sociological Review**, v. 5, n. 2, p. 845–867, 1940.

STOUFFER, Samuel A. "Intervening Opportunities: A Theory Relating Mobility and Distance", **American Sociological Review**, v. 5, n. 6, p. 845–867, 1940.

TADEU, E. "Um Estudo Sobre Os Procedimentos de Calibração De Alguns Modelos de Distribução De Viagens", **Universidade Federal de Santa Catarina**, p. 152, 2000.

TAPKIN, S., AKYILMAZ, O. "A recommended neural trip distribution model", **MIDDLE EAST TECHNICAL UNIVERSITY**, 2004.

TEODOROVIC, D., JANIC, M. **Transportation Engineering: Theory, Practice, and Modeling**. [S.l.], Elsevier B.V., 2017.

THERNEAU, TERRY; ATKINSON, B. **rpart: Recursive Partitioning and Regression Trees. R package**. . [S.l: s.n.]. Disponível em: https://cran.r-project.org/package=rpart. , 2019.

THOMPSON, B. K. **TRANSPORTATION ENGINEERING**. First ed. [S.l.], McGraw-Hill, 2019.

VAN BUUREN, STEF ; GROOTHUIS-OUDSHOORN, K. **mice: Multivariate Imputation by Chained Equations in R.** . [S.l.], Journal of Statistical Software. Disponível em: https://www.jstatsoft.org/v45/i03/. , 2011.

VEENSTRA, S. A., THOMAS, T., TUTERT, S. I. A. "Trip distribution for limited destinations: A case study for grocery shopping trips in the Netherlands", **Transportation**, v. 37, n. 4, p. 663–676, 2010. DOI: 10.1007/s11116-010-9274-0.

VUCHIC, V. R. **Urban Transit: Operations, Planning and Economics**. [S.l.], JOHN WILEY & SONS, INC., 2005. v. 1.

WALKER, J., LI, J., SRINIVASAN, S. "Travel demand models in the developing world: correcting for measurement errors", **Transportation Letters**, v. 2, n. 4, p. 231–243, 2010. DOI: 10.3328/tl.2010.02.04.231-243.

WICKHAM, H., BRYAN, J. **readxl: Read Excel Files.** . [S.l: s.n.]. Disponível em: https://cran.r-project.org/package=readxl. , 2019.

WILLS, M. J. "A flexible gravity-opportunities model for trip distribution", **Transportation Research Part B**, v. 20, n. 2, p. 89–111, 1986. DOI: 10.1016/0191-2615(86)90001-9.

WILSON, A. G. "A statistical theory of spatial distribution models", **Transportation Research**, v. 1, n. 3, p. 253–269, 1967. DOI: 10.1016/0041-1647(67)90035-4.

WITTEN, I. H., FRANK, E. **Dara Mining. Practical Machine Learning Tools and Techniques**. Second Edi ed. [S.l.], ELSEVIER, 2005.

YANG, F., JIN, P. J., CHENG, Y. "Origin-Destination Estimation for Non-Commuting Trips Using Location-Based Social Networking Data", **International Journal of Sustainable Transportation**, v. 9, n. 8, p. 551–564, 2015. DOI: 10.1080/15568318.2013.826312.

ZACHARY, A.; DEANE-MAYER; KNOWLES, J. E. **caretEnsemble: Ensembles of Caret Models.** [S.l: s.n.]. Disponível em: https://cran.r-project.org/package=caretEnsemble. , 2019.

ZHAO, Y., PAWLAK, J., POLAK, J. W. "Inverse discrete choice modelling: theoretical and practical considerations for imputing respondent attributes from the patterns of observed choices", **Transportation Planning and Technology**, v. 41, n. 1, p. 58–79, 2018. DOI: 10.1080/03081060.2018.1402745.

ANEXOS

ANEXO 1: RELAÇÃO SUBDISTRITO – BAIRRO DE MEDELLÍN.

Subdistrito	Bairro
1 – Popular	Santo Domingo Savio Nº 1
	Santo Domingo Sabio Nº 2
	Popular
	Granizal
	Moscú Nº 2
	Villa Guadalupe
	San Pablo
	Aldea Pablo VI
	La Esperanza Nº 2
	El Compromiso
	La Avanzada
	Carpinelo
2 - Santa cruz	La Isla
	El Playón de Los Comuneros
	Pablo VI
	La Frontera
	La Francia
	Andalucía
	Villa del Socorro
	Villa Niza
	Moscú Nº 1
	Santa Cruz
	La Rosa
3 – Manrique	La Salle
	Las Granjas
	Campo Valdes Nº 2
	Santa Inés
	El Raizal
	El Pomar
	Manrique Central No. 2
	Manrique Oriental
	Versalles Nº 1
	Versalles Nº 2
	La Cruz
	Oriente
	Maria Cano – Carambolas
	San José La Cima Nº 1
	San José La Cima Nº 2
4 – Aranjuez	Berlín
	San Isidro
	Palermo
	Bermejal - Los Álamos
	Moravia
	Sevilla

Subdistrito	Bairro
	San Pedro
	Manrique Central Nº 1
	Campo Valdes Nº 1
	Las Esmeraldas
	La Piñuela
	Aranjuez
	Brasilia
	Miranda
5 – Castilla	Toscana
	Las Brisas
	Florencia
	Tejelo
	Boyacá
	Héctor Abad Gómez
	Belalcazar
	Girardot
	Tricentenario
	Castilla
	Francisco Antonio Zea
	Alfonso López
	Caribe
	El Progreso
6 - Doce de Octubre	Santander
	Doce de Octubre Nº 1
	Doce de Octubre Nº 2
	Pedregal
	La Esperanza
	San Martín de Porres
	Kennedy
	Picacho
	Picachito
	Mirador del Doce
	Progreso Nº 2
	El Triunfo
7 – Robledo	Cerro El Volador
	San Germán
	Barrio Facultad de Minas Universidad Nacional
	La Pilarica
	Bosques de San Pablo
	Altamira
	Córdoba
	López de Mesa
	El Diamante
	Aures Nº 1
	Aures Nº 2
	Bello Horizonte
	Villa Flora
	Palenque
	Robledo

Subdistrito	Bairro
	Cucaracho
	Fuente Clara
	Santa Margarita
	Olaya Herrera
	Pajarito
	Monteclaro
	Nueva Villa de La Iguaná.
8 - Villa Hermosa	Villa Hermosa
	La Mansión
	San Miguel
	La Ladera
	Batallón Girardot
	Llanaditas
	Los Mangos
	Enciso
	Sucre
	El Pinal
	Trece de Noviembre
	La Libertad
	Villatina
	San Antonio
	Las Estancias
	Villa Turbay
	La Sierra (Santa Lucía - Las Estancias)
	Villa Lilliam.
9 - Buenos Aires	Juan Pablo II
	Barrios de Jesús
	Bombona Nº 2
	Los Cerros El Vergel
	Alejandro Echavarría
	Barrio Caicedo
	Buenos Aires
	Miraflores
	Cataluña
	La Milagrosa
	Gerona
	El Salvador
	Loreto
	Asomadera Nº 1
	Asomadera Nº 2
	Asomadera Nº 3
	Ocho de Marzo
10 - La Candelaria	Prado
	Jesús Nazareno
	El Chagualo
	Estación Villa
	San Benito
	Guayaquil
	Corazón de Jesús
	Calle Nueva
	Perpetuo Socorro

Subdistrito	Bairro
	Barrio Colón
	Las Palmas
	Bomboná Nº 1
	Boston
	Los Ángeles
	Villa Nueva
	La Candelaria
	San Diego
11 - Laureles – Estadio	Carlos E. Restrepo
	Suramericana
	Naranjal
	San Joaquín
	Los Conquistadores
	Bolivariana
	Laureles
	Las Acacias
	La Castellana
	Lorena
	El Velódromo
	Estadio
	Los Colores
	Cuarta Brigada
	Florida Nueva
12 - La América	Ferrini
	Calasanz
	Los Pinos
	La América
	La Floresta
	Santa Lucia
	El Danubio
	Campo Alegre
	Santa Mónica
	Barrio Cristóbal
	Simón Bolívar
	Santa Teresita
	Calasanz Parte Alta
13 - San Javier	El Pesebre
	Blanquizal
	Santa Rosa de Lima
	Los Alcázares
	Metropolitano
	La Pradera
	Juan XIII - La Quiebra
	San Javier Nº 2
	San Javier Nº 1
	Veinte de Julio
	Belencito
	Betania
	El Corazón
	Las Independencias
	Nuevos Conquistadores

Subdistrito	Bairro
	El Salado
	Eduardo Santos
	Antonio Nariño
	El Socorro
14 - El Poblado	Barrio Colombia
	Simesa
	Villa Carlota
	Castropol
	Lalinde
	Las Lomas Nº 1
	Las Lomas Nº 2
	Altos del Poblado
	El Tesoro
	Los Naranjos
	Los Balsos Nº 1
	San Lucas
	El Diamante Nº 2
	El Castillo
	Los Balsos Nº 2
	Alejandría
	La Florida
	El Poblado
	Manila
	Astorga
	Patio Bonito
	La Aguacatala
	Santa María de Los Ángeles
15 – Guayabal	Tenche
	Trinidad
	Santa Fé
	Parque Juan Pablo II
	Campo Amor
	Noel
	Cristo Rey
	Guayabal
	La Colina
	El Rodeo
16 – Belén	Fátima
	Rosales
	Belén
	Granada
	San Bernardo
	Las Playas
	Diego Echevarria
	La Mota
	La Hondonada
	El Rincón
	La Loma de Los Bernal
	La Gloria
	Altavista
	La Palma

Subdistrito	Bairro
	Los Alpes
	Las Violetas
	Las Mercedes
	Nueva Villa de Aburrá
	Miravalle
	El Nogal - Los Almendros
	Cerro Nutibara
50 - San Sebastián de Palmitas	
60 - San Cristóbal	
70 – Altavista	52 "veredas"
80 - San Antonio de Prado	
90 - Santa Elena	

Fonte: Elaboração própria com base em informação da Prefeitura de Medellín.

ANEXO 2: ATRAÇÃO E PRODUÇÃO POR ZONA DE TRÁFEGO.

SIT	COD SIT	ATRAÇÃO	EMPREGOS	SIT	COD SIT	ATRAÇÃO	EMPREGOS
1-1-3460	1	0	0	10-5-22	171	18	3553
1-1-3470	2	1	198	10-6-141	172	119	23489
1-1-3480	3	1	198	10-6-142	173	40	7896
1-1-3490	4	0	0	10-6-30	174	53	10462
1-1-3500	5	1	198	10-7-40	175	52	10264
1-1-3510	6	1	198	10-7-50	176	43	8488
1-2-3520	7	0	0	10-8-151	177	15	2961
1-3-3400	8	2	397	10-8-152	178	16	3158
1-3-3410	9	0	0	10-11-181	179	9	1777
1-3-3420	10	1	198	10-11-182	180	24	4737
1-3-3430	11	0	0	10-12-190	181	7	1382
1-3-3440	12	0	0	10-12-520	182	165	32569
1-3-3450	13	0	0	10-13-200	183	23	4540
1-4-2300	14	4	793	10-13-60	184	115	22700
1-4-3380	15	4	793	10-14-210	185	4	790
1-4-3390	16	0	0	10-15-220	186	3	592
1-5-2280	17	4	793	10-16-231	187	43	8488
1-6-2270	18	0	0	10-16-232	188	6	1184
1-7-2290	19	6	1190	10-17-240	189	2	395
1-9-3560	20	1	198	10-18-81	190	59	11646
1-10-3530	21	11	2182	10-18-82	191	89	17568
1-11-3540	22	2	397	10-19-11	192	223	44018
1-12-3550	23	0	0	10-19-12	193	524	103432
2-1-1940	24	1	200	10-19-71	194	295	58230
2-2-1930	25	1	200	10-19-72	195	65	12830
2-3-1930	26	0	0	10-20-510	196	138	27240
2-4-1930	27	1	200	10-50-170	197	157	30990
2-5-3320	28	0	0	10-51-161	198	32	6316
2-5-3330	29	0	0	10-51-162	199	4	790
2-5-3350	30	1	200	10-52-110	200	24	4737
2-6-3300	31	0	0	11-1-1400	201	17	3435
2-6-3310	32	1	200	11-2-1380	202	59	11922
2-6-3340	33	11	2196	11-2-1390	203	6	1212
2-7-1980	34	8	1597	11-3-1370	204	31	6264
2-7-3360	35	2	399	11-4-1110	205	23	4647
2-7-3370	36	1	200	11-5-1101	206	21	4243
2-8-1970	37	0	0	11-5-1102	207	31	6264
2-9-2000	38	0	0	11-7-1120	208	23	4647
2-10-2012	39	2	399	11-8-1081	209	29	5860
2-11-2011	40	2	399	11-8-1130	210	9	1819
3-1-2250	41	0	0	11-9-1061	211	51	10305
3-1-2260	42	2	409	11-10-1040	212	3	606
3-2-2230	43	3	613	11-10-1062	213	5	1010
3-2-2240	44	0	0	11-11-1140	214	9	1819
3-3-2190	45	3	613	11-12-1350	215	4	808

SIT	COD SIT	ATRAÇÃO	EMPREGOS	SIT	COD SIT	ATRAÇÃO	EMPREGOS
3-4-2180	46	2	409	11-13-1430	216	27	5456
3-5-2160	47	1	204	11-14-1440	217	15	3031
3-6-2170	48	3	613	11-15-1410	218	3	606
3-7-2130	49	3	613	11-17-1360	219	84	16973
3-8-2140	50	16	3271	11-50-1090	220	13	2627
3-9-2150	51	2	409	11-51-1420	221	6	1212
3-10-2150	52	0	0	11-51-1430	222	0	0
3-13-3572	53	0	0	12-1-1450	223	1	201
3-14-3571	54	0	0	12-2-1460	224	21	4220
3-15-3571	55	0	0	12-3-1340	225	49	9846
4-1-2030	56	6	1265	12-3-1450	226	1	201
4-2-2021	57	2	422	12-4-1330	227	20	4019
4-3-2022	58	1	211	12-5-1320	228	17	3416
4-4-2060	59	0	0	12-6-1312	229	6	1206
4-5-2070	60	15	3162	12-7-1170	230	1	201
4-7-2100	61	17	3584	12-7-1300	231	0	0
4-8-2110	62	26	5481	12-8-1180	232	4	804
4-9-2120	63	15	3162	12-9-1170	233	6	1206
4-10-2200	64	11	2319	12-10-1160	234	4	804
4-11-2220	65	0	0	12-11-1150	235	23	4622
4-12-2040	66	2	422	12-12-1050	236	1	201
4-13-2050	67	5	1054	13-1-1481	237	0	0
4-14-2211	68	1	211	13-2-1481	238	1	212
4-15-2212	69	1	211	13-3-1470	239	0	0
4-50-2081	70	1	211	13-4-1311	240	5	1060
4-51-2082	71	3	632	13-6-1280	241	9	1908
4-52-2090	72	5	1054	13-7-1270	242	2	424
5-1-1920	73	1	203	13-8-1290	243	2	424
5-2-1910	74	1	203	13-9-1250	244	7	1484
5-3-1882	75	5	1015	13-10-1240	245	2	424
5-4-1881	76	1	203	13-11-1190	246	8	1696
5-5-1890	77	16	3249	13-12-1200	247	6	1272
5-7-1900	78	1	203	13-13-1200	248	0	0
5-8-1860	79	1	203	13-13-1220	249	0	0
5-9-1870	80	12	2437	13-14-1230	250	0	0
5-10-1860	81	17	3452	13-15-1210	251	6	1272
5-11-1850	82	2	406	13-16-1220	252	1	212
5-13-1840	83	5	1015	13-17-1220	253	0	0
5-14-1830	84	4	812	13-18-1260	254	3	636
5-17-1801	85	35	7107	13-19-1260	255	0	0
5-17-1802	86	7	1421	14-1-530	256	29	6192
5-19-1600	87	5	1015	14-3-540	257	6	1281
5-19-1610	88	3	609	14-3-550	258	3	641
5-50-1810	89	31	6295	14-4-560	259	7	1495
5-51-1790	90	2	406	14-5-570	260	0	0
5-52-1820	91	11	2234	14-6-580	261	30	6405
5-53-1900	92	0	0	14-7-590	262	8	1708
6-1-1720	93	1	209	14-8-610	263	3	641
6-2-1700	94	1	209	14-9-621	264	11	2349
6-2-1710	95	0	0	14-9-622	265	2	427

SIT	COD SIT	ATRAÇÃO	EMPREGOS	SIT	COD SIT	ATRAÇÃO	EMPREGOS
6-3-1740	96	7	1462	14-10-721	266	11	2349
6-4-1730	97	5	1044	14-10-722	267	6	1281
6-5-1760	98	0	0	14-11-730	268	2	427
6-6-1750	99	2	418	14-11-740	269	5	1068
6-7-1770	100	1	209	14-12-751	270	14	2989
6-8-1680	101	7	1462	14-12-752	271	0	0
6-9-1680	102	1	209	14-13-760	272	1	214
6-10-1690	103	0	0	14-14-704	273	8	1708
6-11-1690	104	0	0	14-15-701	274	8	1708
6-12-1690	105	0	0	14-16-630	275	12	2562
7-3-1570	106	33	6866	14-16-710	276	0	0
7-6-1620	107	8	1665	14-17-641	277	18	3843
7-7-1630	108	21	4369	14-17-642	278	1	214
7-8-1630	109	17	3537	14-18-650	279	179	38217
7-9-1780	110	10	2081	14-19-660	280	57	12170
7-11-1640	111	1	208	14-20-670	281	64	13664
7-12-1670	112	3	624	14-21-681	282	38	8113
7-13-1660	113	2	416	14-21-682	283	16	3416
7-14-1650	114	5	1040	14-22-691	284	37	7900
7-15-1650	115	4	832	14-22-692	285	87	18575
7-16-1531	116	4	832	14-23-770	286	20	4270
7-16-1540	117	6	1248	15-2-3680	287	22	4613
7-17-1511	118	13	2705	15-3-3600	288	0	0
7-17-3150	119	8	1665	15-3-3610	289	3	629
7-18-1520	120	0	0	15-3-3620	290	2	419
7-19-1512	121	0	0	15-3-3640	291	0	0
7-20-1500	122	2	416	15-3-3660	292	3	629
7-20-1501	123	1	208	15-4-3670	293	6	1258
7-23-1520	124	2	416	15-4-830	294	1	210
7-24-1532	125	0	0	15-4-840	295	25	5242
7-50-1550	126	10	2081	15-7-800	296	13	2726
7-51-1560	127	0	0	15-7-810	297	9	1887
7-52-1580	128	5	1040	15-7-830	298	24	5033
7-53-1590	129	56	11652	15-9-791	299	27	5662
8-1-270	130	12	2662	15-9-792	300	37	7759
8-2-260	131	7	1553	15-10-780	301	47	9856
8-3-250	132	4	887	15-11-951	302	13	2726
8-4-280	133	1	222	15-50-821	303	40	8388
8-5-270	134	0	0	15-51-940	304	1	210
8-6-310	135	0	0	16-1-3710	305	27	5821
8-7-300	136	0	0	16-1-3740	306	0	0
8-8-280	137	0	0	16-2-3750	307	8	1725
8-8-290	138	2	444	16-2-3770	308	11	2371
8-9-330	139	2	444	16-3-3780	309	0	0
8-10-320	140	2	444	16-3-3820	310	1	216
8-11-300	141	1	222	16-3-3830	311	5	1078
8-12-390	142	0	0	16-3-3840	312	18	3880
8-13-390	143	1	222	16-4-3850	313	0	0
8-14-390	144	0	0	16-4-3860	314	3	647
8-15-400	145	0	0	16-5-3870	315	3	647

SIT	COD SIT	ATRAÇÃO	EMPREGOS	SIT	COD SIT	ATRAÇÃO	EMPREGOS
8-16-410	146	0	0	16-5-3880	316	2	431
8-17-410	147	0	0	16-5-3890	317	5	1078
8-19-400	148	2	444	16-6-920	318	11	2371
9-1-420	149	0	0	16-7-931	319	12	2587
9-2-420	150	0	0	16-8-932	320	26	5605
9-3-430	151	3	673	16-9-952	321	0	0
9-4-370	152	1	224	16-10-960	322	2	431
9-5-380	153	0	0	16-11-970	323	5	1078
9-6-340	154	13	2915	16-12-980	324	17	3665
9-7-350	155	4	897	16-12-990	325	1	216
9-8-360	156	3	673	16-13-990	326	1	216
9-9-440	157	0	0	16-13-1000	327	5	1078
9-10-440	158	2	448	16-14-3790	328	10	2156
9-11-450	159	4	897	16-14-3800	329	3	647
9-12-460	160	7	1570	16-14-3810	330	16	3449
9-13-470	161	15	3363	16-15-1010	331	8	1725
9-14-500	162	1	224	16-16-3900	332	0	0
9-15-490	163	15	3363	16-16-3920	333	0	0
9-16-480	164	1	224	16-16-3930	334	1	216
9-17-420	165	0	0	16-17-1030	335	1	216
10-1-90	166	24	4737	16-18-1070	336	3	647
10-1-100	167	9	1777	16-19-1070	337	4	862
10-3-120	168	13	2566	16-20-1082	338	4	862
10-3-21	169	20	3948	16-50-3690	339	5	1078
10-4-130	170	40	7896	-	-	-	-

ANEXO 3: ALGORITMO DOS MODELOS – *SOFTWARE R.*

```
#Loading required packages

library(Amelia)
library(caret)
library(caretEnsemble)
library(data.table)
library(e1071)
library(h2o)
library(mice)
library(psych)
library(readxl)
library(rpart)
library(stringr)

#Reading data into R

setwd("D:/Dissertacao/Modelos/Versao 19-03-2020")
data <- read_excel("Planilha_R_ANN.xlsx")

# How many replicates you want of each row
duptimes <- round(data$`FEV-R`)

# Create an index of the rows you want with duplication
idx <- rep(1:nrow(data), duptimes)

# Use that index to generate your new data frame
data <- data[idx,]

data = data[,-84]
rm(duptimes,idx)
data$SO = factor(data$SO,
                 levels = as.character(unique(data$SO)),
                 labels = as.numeric(unique(data$SO)))
data$SD_C = factor(data$SD_C,
                   levels = as.character(unique(data$SD_C)),
                   labels = as.numeric(unique(data$SD_C)))

## Sensitivity analysis

##data = data[,-c(2:50)] ## M0
##data = data[,-c(15:31)] ## M0
##data = data[,-c(2:63)] ## M1
##data = data[,-c(64:80)] ## M2
##data = data[,-c(51:80)] ## M3
##data = data[,-c(51:63)] ## M4
## M5 uses the full database

#Studying the structure of the data
```

```
str(data)
head(data)
describe(data)

## Model building

topology = c(30,100)

h2o.init(nthreads = -1)
neural.network = h2o.deeplearning(y = 'SD_C', training_frame = as.h2o(data), nfolds = 10, hidden = topology, activation = "RECTIFIER",
                                  stopping_metric = "logloss",stopping_rounds = 3, stopping_tolerance = 1e-3, score_each_iteration = TRUE )

neural.network
neural.network.predictions = predict(neural.network, newdata = as.h2o(data[,-83]),type = "raw") ##modify the column number

## Variable importance

h2o.varimp_plot(neural.network)

## Construction of the confusion matrix
x = as.data.frame(neural.network.predictions)
u = union(x$predict, data$SD_C)
h = table(factor(x$predict, u), factor(data$SD_C, u))
j = confusionMatrix(h)

j[["overall"]][["Accuracy"]]
j[["overall"]][["Kappa"]]

write.csv2(j[["byClass"]], str_c(as.character(topology[1]),"-", as.character(topology[2]), ".csv"))
h2o.exportFile(neural.network.predictions, path = "D:/Dissertacao/Modelos/Versao 19-03-2020/predict.csv", sep = ";")
h2o.shutdown ()
```