

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA
E INFORMÁTICA INDUSTRIAL

LUCIANO URGAL PANDO

**AVALIAÇÃO DE TÉCNICAS DE ESTIMAÇÃO DA MATRIZ
ORIGEM-DESTINO DO TRÁFEGO DE VEÍCULOS EM
CIDADES**

DISSERTAÇÃO

CURITIBA

2018

LUCIANO URGAL PANDO

**AVALIAÇÃO DE TÉCNICAS DE ESTIMAÇÃO DA MATRIZ
ORIGEM-DESTINO DO TRÁFEGO DE VEÍCULOS EM
CIDADES**

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do grau de “Mestre em Ciências” – Área de Concentração: Engenharia de Computação.

Orientador: Ricardo Lüders

CURITIBA

2018

Dados Internacionais de Catalogação na Publicação

P189a Pando, Luciano Urgal
2018 Avaliação de técnicas de estimação da matriz origem-destino do tráfego de veículos em cidades / Luciano Urgal Pando.- 2018.
92 f.: il.; 30 cm.

Disponível também via World Wide Web.
Texto em português com resumo em inglês.
Dissertação (Mestrado) - Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. Área de Concentração: Engenharia de Computação, Curitiba, 2018.
Bibliografia: f. 86-90.

1. Pesquisas de tráfego de origem e destino. 2. Programação heurística. 3. Mínimos quadrados. 4. Programação linear. 5. Programação inteira. 6. Algoritmos genéticos. 7. Partículas (Física nuclear). 8. Inteligência artificial. 9. Trânsito - Fluxo. 10. Transporte urbano - Porto (Portugal). 11. Transporte urbano - Curitiba (PR). 12. Sistemas inteligentes de veículos rodoviários. 13. Planejamento urbano - Inovações tecnológicas. 14. Engenharia elétrica - Dissertações. I. Lüders, Ricardo, orient. II. Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. III. Título.

CDD: Ed. 22 -- 621.3

TERMO DE APROVAÇÃO DE DISSERTAÇÃO Nº 798

A Dissertação de Mestrado intitulada “**Avaliação de Técnicas de Estimação da Matriz Origem-Destino do Tráfego de Veículos em Cidades**” defendida em sessão pública pelo(a) candidato(a) **Luciano Urgal Pando**, no dia 14 de junho de 2018, foi julgada para a obtenção do título de Mestre em Ciências, área de concentração Engenharia de Computação, e aprovada em sua forma final, pelo Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial.

BANCA EXAMINADORA:

Prof(a). Dr(a). Ricardo Lüders - Presidente (UTFPR)

Prof(a). Dr(a). Aurora Trinidad Ramirez Pozo - (UFPR)

Prof(a). Dr(a). Marcelo de Oliveira Rosa - (UTFPR)

A via original deste documento encontra-se arquivada na Secretaria do Programa, contendo a assinatura da Coordenação após a entrega da versão corrigida do trabalho.

Curitiba, 14 de junho de 2018.

AGRADECIMENTOS

Ao Instituto Federal do Paraná, em especial ao Campus Campo Largo, por possibilitar o afastamento integral para realização do mestrado stricto sensu na UTFPR, através da Portaria n° 839 de 13/07/2016.

Ao Engenheiro Francisco Caron Malucelli, por fornecer dados do modelo de simulação de uma região da cidade de Curitiba, que modelou a serviço do IPPUC. Tais dados foram utilizados como base para experimentos em um dos estudos de caso.

Ao meu orientador, Professor Ricardo Lüders, que além de me orientar durante meu mestrado, foi também meu orientador para as disciplinas de metodologia científica e de trabalho de conclusão de curso em meu bacharelado em sistemas de informação na UTFPR.

RESUMO

URGAL PANDO, Luciano. AVALIAÇÃO DE TÉCNICAS DE ESTIMAÇÃO DA MATRIZ ORIGEM-DESTINO DO TRÁFEGO DE VEÍCULOS EM CIDADES. 93 f. Dissertação – Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial, Universidade Tecnológica Federal do Paraná. Curitiba, 2018.

O conhecimento do padrão de mobilidade urbana é importante tanto para oferta de serviços públicos de qualidade quanto para planejamento das cidades. Isso pode ser feito através de onerosas pesquisas de campo ou utilizando a enorme quantidade de dados provenientes da monitoração de serviços e do ambiente urbano nas cidades inteligentes. A estimação da matriz origem-destino tem por objetivo obter uma estimação de tráfego de veículos entre determinadas origens e destinos da cidade, a partir do tráfego observado nas vias públicas através de sensores ou veículos sonda. Este trabalho avalia e compara quatro técnicas de estimação da matriz origem-destino: mínimos quadrados, programação linear inteira mista (MILP), algoritmo genético e enxame de partículas (PSO). Duas cidades são consideradas como estudos de caso: Porto e Curitiba. A cidade do Porto em Portugal dispõe de dados de viagens de táxi utilizados como veículos sonda. A cidade de Curitiba dispõe de sensores de tráfego de veículos nas vias públicas. Com o uso de dados georreferenciados de mobilidade, são considerados também algoritmos de agrupamento espacial para estabelecer áreas de origem e destino e de "map matching" para caracterizar rotas associadas aos deslocamentos. Os resultados obtidos mostram ligeira superioridade para a estimação usando MILP e PSO, mas que dependem fortemente da quantidade e posição dos sensores nas vias.

Palavras-chave: Estimação de matriz origem-destino, Tráfego urbano, Metaheurística, Transporte, Cidade inteligente

ABSTRACT

URGAL PANDO, Luciano. EVALUATING ORIGIN-DESTINATION MATRIX ESTIMATION APPROACHES FOR URBAN TRAFFIC OF VEHICLES. 93 f. Dissertação – Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial, Universidade Tecnológica Federal do Paraná. Curitiba, 2018.

The knowledge of urban mobility patterns is important to maintain good public services as well as for city planning. These mobility patterns can be characterized by using expensive fieldwork or through the huge amount of data available from services and environmental monitoring in smart cities. The origin-destination matrix estimation aims to estimate the traffic of vehicles between two particular origin and destination areas in the city from traffic observed by sensors installed on roads or from probe vehicles. This work evaluates and compares four origin-destination matrix estimation techniques: least squares, mixed-integer linear programming (MILP), genetic algorithm and particle swarm optimization (PSO). Two cities are considered as case studies: OPorto and Curitiba. The city of OPorto in Portugal has data from taxi trips used as probe vehicles. Curitiba in Brazil has road traffic sensors. In addition, due to georeferenced spatial data, algorithms for clustering and map matching are considered to characterize areas of origin-destination and routes, respectively. The case studies show better results for MILP and PSO estimates. However, they strongly depend on the amount and position of sensors.

Keywords: Origin Destination matrix estimation, Urban traffic, Metaheuristic, Transportation, Smart city.

LISTA DE FIGURAS

FIGURA 11 – Grafo direcionado correspondente à matriz origem-destino X da Tabela 11.	16
FIGURA 12 – Grafo direcionado correspondente às rotas mais utilizadas da Matriz OD x, da Tabela 11.	16
FIGURA 13 – Diagrama da distribuição de tráfego através de pontos OD, rotas, fluxos e suas medições por sensores em arcos.	17
FIGURA 14 – <i>Map matching</i> como um mapeamento entre dados georreferenciados (Marcações em X) e a rede viária (linhas cinzas e amarelas. Pontos verdes).	23
FIGURA 17 – Esquema do inter-relacionamento entre as atividades desenvolvidas na dissertação.	36
FIGURA 18 – <i>Map matching</i> de um dado georreferenciado (x na cor preta) para um arco BC.	50
FIGURA 20 – Agrupamento espacial por densidade (13 <i>clusters</i> ou áreas).	56
FIGURA 21 – <i>Map matching</i> (à direita) a partir de dados georreferenciados (à esquerda) de uma corrida de táxi na cidade do Porto.	58
FIGURA 22 – MAE para diferentes técnicas e número de sensores.	69
FIGURA 23 – RMSE para diferentes técnicas e número de sensores.	70
FIGURA 24 – GEH para diferentes técnicas e número de sensores.	71
FIGURA 25 – R^2 ODM para diferentes técnicas e número de sensores.	72
FIGURA 26 – R^2 Arcos para diferentes técnicas e número de sensores.	72
FIGURA 27 – TDD para diferentes técnicas e número de sensores.	73
FIGURA 28 – Desvio padrão normalizado obtido com 30 rodadas do GA para diferentes métricas e número de sensores.	74
FIGURA 29 – Desvio padrão normalizado obtido com 30 rodadas do PSO para diferentes métricas e número de sensores.	74
FIGURA 30 – Mapa da região do estudo de caso de Curitiba.	76
FIGURA 31 – Gráficos de dispersão para o fluxo de veículos estimados e observados nos arcos da rede viária de Curitiba, usando GA, VISUM, MILP, PSO e LS.	81
FIGURA 32 – Representação no VISUM da matriz OD encontrada pelo PSO para Curitiba.	82

LISTA DE TABELAS

TABELA 11 –	Informações de fluxos OD entre nós A, B e C em veículos/min da Matriz origem-destino X	15
TABELA 13 –	Artigos com aplicações de diferentes técnicas para a estimação da matriz OD	27
TABELA 14 –	Bases de dados de táxis	33
TABELA 16 –	Métricas utilizadas para avaliar a qualidade da estimação da matriz OD	37
TABELA 17 –	Variáveis do modelo de programação linear inteira mista	45
TABELA 18 –	Restrições do modelo de programação linear inteira mista	46
TABELA 19 –	Características das técnicas de estimação da matriz OD aplicadas	48
TABELA 21 –	Detalhes dos estudos de casos da cidade do Porto e de Curitiba ..	53
TABELA 22 –	Atributos dos dados de corridas de táxi da cidade do Porto	54
TABELA 23 –	Informações sobre a clusterização espacial na cidade do Porto	57
TABELA 24 –	Resultados do <i>map matching</i> para o estudo de caso da cidade do Porto	59
TABELA 25 –	Seleção de parâmetros do GA, para a cidade do Porto	61
TABELA 26 –	Seleção de parâmetros do PSO, para a cidade do Porto	61
TABELA 27 –	Resultados do GA para diferentes funções <i>fitness</i> , para a cidade do Porto	63
TABELA 28 –	Resultados do PSO para diferentes funções <i>fitness</i> , para a cidade do Porto	64
TABELA 29 –	Resultados do GA para critérios de escolha de sensores, para a cidade do Porto	65
TABELA 30 –	Resultados do PSO para critérios de escolha de sensores, para a cidade do Porto	66
TABELA 31 –	Resultados do MILP para critérios de escolha de sensores, para a cidade do Porto	67
TABELA 32 –	Resultados do LS para critérios de escolha de sensores, para a cidade do Porto	67
TABELA 33 –	Comparação das técnicas de estimação da matriz OD para a cidade do Porto	68
TABELA 34 –	Seleção de parâmetros do GA, para Curitiba	77
TABELA 35 –	Seleção de parâmetros do PSO, para Curitiba	77
TABELA 36 –	Resultados do GA para diferentes funções <i>fitness</i> , para Curitiba ..	78
TABELA 37 –	Resultados do PSO para diferentes funções <i>fitness</i> , para Curitiba .	79
TABELA 38 –	Comparação das técnicas de estimação da matriz OD para Curitiba	80
TABELA 42 –	p-values para o teste de normalidade de Shapiro-Wilk no estudo de caso da cidade do Porto	92
TABELA 43 –	p-values para o teste de normalidade de Shapiro-Wilk no estudo de caso de Curitiba	92
TABELA 44 –	p-values para os testes t de Student	93

LISTA DE SIGLAS

OD	Origem-destino
ODM	<i>Origin Destination Matrix</i> (Matriz origem-destino)
ODME	<i>Origin Destination Matrix Estimation</i> (Estimativa de Matriz origem-destino)
STA	Static Traffic Assignment
TDS	<i>Total demand scale</i> (Escala total de demanda)
MPRE	<i>Maximum possible relative error</i> (Máximo erro relativo possível)
UE	<i>User Equilibrium</i> (Equilíbrio de usuário)
KF	<i>Kalman Filter</i> (Filtro de Kalman)
BI	<i>Bayesian inference</i> (Inferência bayesiana)
ML	<i>Maximum Likelihood</i> (Máxima verosimilhança)
GM	<i>Gravity model</i> (Modelo de gravidade)
PSO	<i>Particle swarm optimization</i> (Otimização por nuvem de partículas)
GLS	<i>General Least Squares</i> (Método dos mínimos quadrados)
MILP	<i>Mixed Integer Linear Programming</i> (Programação Linear Inteira Mista)
GA	<i>Genetic algorithm</i> (Algoritmo Genético)
NSLP	<i>Network sensor location problem</i> (Problema da alocação de sensores em uma rede)
MAE	<i>Mean Absolute Error</i> (Erro Absoluto médio)
RMSE	<i>Root Mean Square Error</i> (Erro quadrado médio)
TDD	<i>Total Demand Deviation</i> (Desvio total de demanda)
GEH	Métrica Geoffrey E. Havers
pGEH5	Porcentagem de arcos com GEH abaixo de 5
R^2	Coefficiente de determinação
R^2 Arcos	R^2 para arcos
R^2 ODM	R^2 para valores da matriz OD
iRace	Pacote de seleção de parâmetros <i>iterated Race</i>
OSM	Base geográfica <i>Open Street Map</i>
ANOVA	<i>Analysis of variance</i> (Análise de variância)
IPPUC	Instituto de Pesquisa e Planejamento Urbano de Curitiba

LISTA DE SÍMBOLOS

G	Grafo da rede viária
N	Conjunto de nós
E	Conjunto de arcos
$R \subseteq N$	Conjunto de nós origem e destino
$ R = o$	Cardinalidade do conjunto R
W	Conjunto de pares OD
$ W = n_p$	Cardinalidade do conjunto W
$w \in W$	Par origem-destino w pertencente a W
q_w	Fluxo conhecido de veículos para par OD w
\bar{q}_w	Fluxo estimado de veículos para par OD w
$Z \subseteq E$	Conjunto de sensores em arcos
$ Z = n_s$	Cardinalidade do conjunto Z
$z \in Z$	Arco z pertencente ao conjunto Z
v_z	Fluxo medido de veículos no arco z
\bar{v}_z	Fluxo estimado de veículos no arco z
pop	Parâmetro de tamanho da população do GA
t_{ga}	Parâmetro de número de gerações do GA
P_c	Parâmetro de probabilidade de crossover do GA
P_m	Parâmetro de probabilidade de mutação do GA
s	Parâmetro de dimensão do enxame do PSO
t_{pso}	Parâmetro de número de iterações do PSO
c1	Parâmetro de memória cognitiva do PSO
c2	Parâmetro de memória social do PSO
w_i	Parâmetro de inércia inicial do PSO
w_f	Parâmetro de inércia final do PSO

SUMÁRIO

1 INTRODUÇÃO	12
1.1 MOTIVAÇÃO	12
1.2 OBJETIVOS	13
1.2.1 Objetivo geral	13
1.2.2 Objetivos específicos	13
1.3 ORGANIZAÇÃO DA DISSERTAÇÃO	14
2 DESCRIÇÃO DO PROBLEMA	15
2.1 ESTIMAÇÃO DA MATRIZ OD	15
2.2 PROBLEMAS RELACIONADOS	20
2.2.1 Alocação de sensores na rede viária	20
2.2.2 <i>Map matching</i>	22
2.2.3 Equilíbrio de usuário	23
2.3 CONSIDERAÇÕES DO CAPÍTULO	24
3 REVISÃO BIBLIOGRÁFICA	25
3.1 MÉTODOS DE ESTIMAÇÃO DA MATRIZ OD	25
3.2 FONTES E DESCRIÇÃO DOS DADOS PARA ESTIMAÇÃO DA MATRIZ OD	30
3.2.1 Bases de dados de táxi	32
3.3 CONSIDERAÇÕES DO CAPÍTULO	35
4 METODOLOGIA	36
4.1 MEDIDAS DE DESEMPENHO	37
4.1.1 <i>Mean Absolute Error</i>	37
4.1.2 <i>Root Mean Square Error</i>	37
4.1.3 <i>Total Demand Deviation</i>	38
4.1.4 GEH Statistic	38
4.1.5 Coeficiente de determinação R^2	39
4.2 ALGORITMO DE ENXAME DE PARTÍCULAS	39
4.2.1 Seleção de parâmetros	41
4.2.2 Funções <i>fitness</i>	42
4.3 ALGORITMO GENÉTICO	43
4.4 MÉTODO DOS MÍNIMOS QUADRADOS	44
4.5 PROGRAMAÇÃO LINEAR INTEIRA MISTA	45
4.6 <i>T-FLOW FUZZY</i>	47
4.7 RESUMO DAS ABORDAGENS DE ESTIMAÇÃO DA MATRIZ OD UTILIZA-	
DAS	48
4.8 MAP MATCHING	49
4.9 CONSIDERAÇÕES DO CAPÍTULO	51
5 RESULTADOS	52
5.1 ESTUDO DE CASO DA CIDADE DO PORTO	53
5.1.1 Agrupamento espacial	55
5.1.2 <i>Map matching</i>	56
5.1.3 Configuração dos algoritmos GA e PSO	60

5.1.4	Influência da função <i>fitness</i> na estimação da matriz OD	61
5.1.5	Influência da localização de sensores na estimação da matriz OD	64
5.1.6	Comparação dos resultados de estimação da matriz OD	67
5.1.7	Influência do número de sensores na estimação da matriz OD	69
5.2	ESTUDO DE CASO DE CURITIBA	75
5.2.1	Configuração dos algoritmos GA e PSO	75
5.2.2	Influência da função <i>fitness</i> na estimação da matriz OD	78
5.2.3	Comparação dos resultados de estimação da matriz OD	79
5.3	CONSIDERAÇÕES DO CAPÍTULO	82
6	CONCLUSÃO	85
	REFERÊNCIAS	87
	Apêndice A – DETALHES DOS TESTES DE SHAPIRO-WILK E T DE	
	STUDENT	92

1 INTRODUÇÃO

Compreender o padrão de movimento humano é fundamental para o planejamento urbano, que visa gerenciar e planejar a infraestrutura de forma a fornecer melhores serviços públicos. Parte dessa informação é relativa ao fluxo de veículos ou pessoas entre diferentes pontos de origem e destino, sendo que esta informação pode ser representada pela chamada matriz origem-destino. Tradicionalmente, tal matriz é obtida através de pesquisas de campo, que são demoradas e têm alto custo, resultando em matrizes origem-destino que se tornam desatualizadas rapidamente.

O assunto já vem sendo estudado há algumas décadas, e existem diversas técnicas de estimação da matriz origem-destino. Um *survey* sobre a estimação de matrizes origem-destino pode ser encontrado em Bera e Rao (2011), no qual são citadas diversas técnicas de solução do problema, bem como métricas de desempenho dessa estimação.

Esta dissertação de mestrado compara quatro técnicas de estimação de matrizes OD: Mínimos quadrados, algoritmo genético, otimização por enxame de partículas e programação linear inteira mista. São estudados casos da cidade do Porto em Portugal e de Curitiba no Brasil. A partir deste estudo de casos é possível avaliar e caracterizar as diferentes técnicas em termos da qualidade da estimação obtida, e a influência dos diversos parâmetros que afetam essa estimação.

Parte dos resultados desta dissertação geraram uma publicação no ENIAC 2017, realizado em Uberlândia. (PANDO; LUDERS, 2017)

1.1 MOTIVAÇÃO

Atualmente, as cidades possuem sistemas de informação capazes de produzir uma enorme quantidade de dados a partir da monitoração de diversos serviços e condições do ambiente urbano. Este é o caso da monitoração do volume de tráfego nas vias públicas, captado por câmeras e dispositivos de medição de tráfego. Assim, é possível monitorar o deslocamento de veículos ao longo do dia e caracterizar o padrão de mobilidade no

ambiente urbano, identificando suas origens e destinos. Esse é um problema relevante, pois pode ser utilizado não apenas para a melhoria da circulação na cidade, mas também para planejamento futuro da urbanização.

Buscar alternativas de soluções para o problema é um tema de interesse, pois tais informações de fluxos entre origens e destinos são geralmente obtidas a partir de pesquisas de campo, que possuem custo elevado (CHEN et al., 2005). Por exemplo, uma pesquisa de campo para origem-destino foi iniciada em Curitiba em 2016, abrangendo 17 municípios da região metropolitana além da capital, ao custo aproximado de 6 milhões de reais. (CURITIBA, 2016)

Uma das formas de caracterizar esta mobilidade é através do mapeamento das rotas comumente utilizadas para o deslocamento de veículos na cidade e o respectivo volume de veículos em cada rota. Ao longo de décadas, métodos para obtenção deste conjunto de rotas foram desenvolvidos a partir dos dados então disponíveis, mas a situação deve ser reavaliada considerando novas fontes de dados para auxiliar o processo de estimação de matriz origem-destino.

Na estimação da matriz origem-destino, o conjunto de dados de entrada necessários para solução do problema se altera de acordo com a técnica utilizada, e mesmo a partir de um mesmo conjunto de dados de entrada, existem diversas técnicas disponíveis. É necessário investigar diferentes conjuntos de dados que podem ser utilizados, bem como comparar diferentes técnicas, uma vez que existem poucos trabalhos dedicados à comparação de técnicas de estimação da matriz origem-destino, e os resultados desta dissertação podem auxiliar no direcionamento de soluções para problemas futuros.

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

Avaliar diferentes técnicas de estimação da matriz origem-destino que fazem uso de dados de sensores de fluxo de veículos, a partir de métricas de desempenho existentes na literatura.

1.2.2 OBJETIVOS ESPECÍFICOS

- Comparar o algoritmo de enxame de partículas, algoritmo genético, mínimos quadrado e programação linear inteira mista, para solução do problema de estimação

da matriz origem-destino.

- Utilizar métricas da literatura para a avaliação das soluções obtidas;
- Realizar a configuração dos algoritmos de inteligência computacional utilizados (enxame de partículas e algoritmo genético);
- Avaliar a influência da qualidade e quantidade de informação proveniente dos sensores (número e posicionamento dos sensores);
- Identificar origens e destinos, assim como a distribuição do fluxo de veículos na rede viária, quando se utilizam dados de viagens de táxi para a estimação da matriz origem-destino;
- Considerar as cidades do Porto e Curitiba como estudos de caso, pois utilizam diferentes fontes de dados de fluxo de veículos para a estimação da matriz origem-destino.

1.3 ORGANIZAÇÃO DA DISSERTAÇÃO

O Capítulo 2 formaliza o problema da estimação da matriz-origem-destino a partir de dados de sensores de tráfego. Contém ainda a descrição de problemas relacionados: caracterização das probabilidades de escolha de rotas, localização dos sensores na rede viária e reconstrução de rotas a partir de dados georreferenciados de movimentação de veículos. O Capítulo 3 apresenta uma revisão bibliográfica sobre técnicas de estimação da matriz origem-destino encontradas na literatura. O Capítulo 4 define as métricas de desempenho utilizadas para avaliar os resultados, define as técnicas consideradas nesta dissertação para estimação da matriz origem-destino e o algoritmo de *map matching* usado para caracterizar a distribuição de fluxo de veículos na rede viária. No Capítulo 5 são apresentados os resultados de comparação das técnicas consideradas para dois estudos de caso: utilizando dados da cidade do Porto e de Curitiba. A conclusão da dissertação é apresentada no Capítulo 6.

2 DESCRIÇÃO DO PROBLEMA

Neste capítulo, o problema de estimação da matriz origem-destino (OD) é formalizado. Na Seção 2.1, a matriz OD é definida, bem como sua relação com a topologia da rede viária, e a distribuição do trânsito através desta rede. Na Seção 2.2, são abordados outros problemas relevantes ao problema de estimação da matriz OD. É o caso do problema de alocação de sensores considerado na Seção 2.2.1, do *map matching* da Seção 2.2.2, e do equilíbrio de usuário tratado na Seção 2.2.3.

2.1 ESTIMAÇÃO DA MATRIZ OD

A matriz OD (do inglês ODM, *Origin Destination Matrix*) é uma informação fundamental para planejar ou gerenciar sistemas de transporte. Matrizes OD geralmente fornecem informações de fluxo de tráfego de nós origem para nós destino, frequentemente localizados nas fronteiras de uma área de interesse (TEKNOMO; FERNANDEZ, 2014). A primeira definição de uma matriz OD data de um artigo publicado em 1955, sendo este artigo republicado em Voorhees (2013), pela grande relevância para a área de transportes, de acordo com o ITE (*Institute of Transportation Engineers*).

Na época de sua formulação, os métodos de obtenção da matriz OD eram através de pesquisas de campo ou por métodos de estimação da matriz OD que não envolviam sensores de tráfego, de modo que a matriz OD não apresenta informações sobre a topologia da rede viária, contendo apenas informação do fluxo de veículos entre determinadas origens e destinos, sem levar em conta as rotas utilizadas para estes fluxos.

Tabela 11: Informações de fluxos OD entre nós A, B e C em veículos/min da Matriz origem-destino X

	A	B	C
A		2	1
B	10		1
C	5	2	

A Tabela 11 contém informações de uma matriz origem-destino X , com 3 nós origem (A, B e C) que são também nós destino. Considerando as linhas como pontos de origem, e as colunas como pontos de destino, o nó A é origem de 3 veículos por minuto, sendo 2 para B, e 1 para C. Além disso, o nó A é destino de 15 veículos por minuto, sendo 10 originários de B, e 5 originários de C.

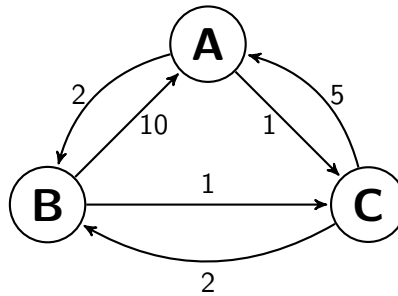


Figura 11: Grafo direcionado correspondente à matriz origem-destino X da Tabela 11.

A Figura 11 mostra o grafo direcionado correspondente à matriz origem-destino X apresentada na Tabela 11. Somente as informações contidas na matriz OD são apresentadas, sem que a topologia da rede viária seja considerada.

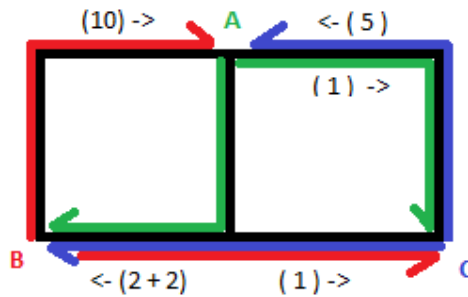


Figura 12: Grafo direcionado correspondente às rotas mais utilizadas da Matriz OD x , da Tabela 11.

A Figura 12 mostra uma possível distribuição de fluxo de veículos da mesma matriz origem-destino X da Tabela 11, quando existem diferentes rotas na rede viária para um mesmo par OD. É comum que existam mais de uma rota possível para um mesmo par OD, e também que arestas (conhecidas nesta área de pesquisa por arcos) contenham fluxos de vários pares OD. Por exemplo, no arco localizado a direita do nó B, ocorrem fluxos dos pares AB e CB, somando 4 veículos por unidade de tempo.

Para a maioria das técnicas utilizadas para estimativa de matriz origem-destino, dados de sensores de tráfego são utilizados, e desta forma, a topologia da rede viária passa

a fazer parte do problema, já que é preciso relacionar essa informação com a localização dos sensores, de forma que os valores estimados de tráfego nos arcos seja o mais próximo possível dos valores medidos pelos sensores. Segundo Hu e Liou (2014), a estimação de matrizes OD com base em dados de sensores de tráfego tem sua origem na década de 1980 pelo uso de métodos estatísticos e de otimização matemática (descritos na Seção 3.1).

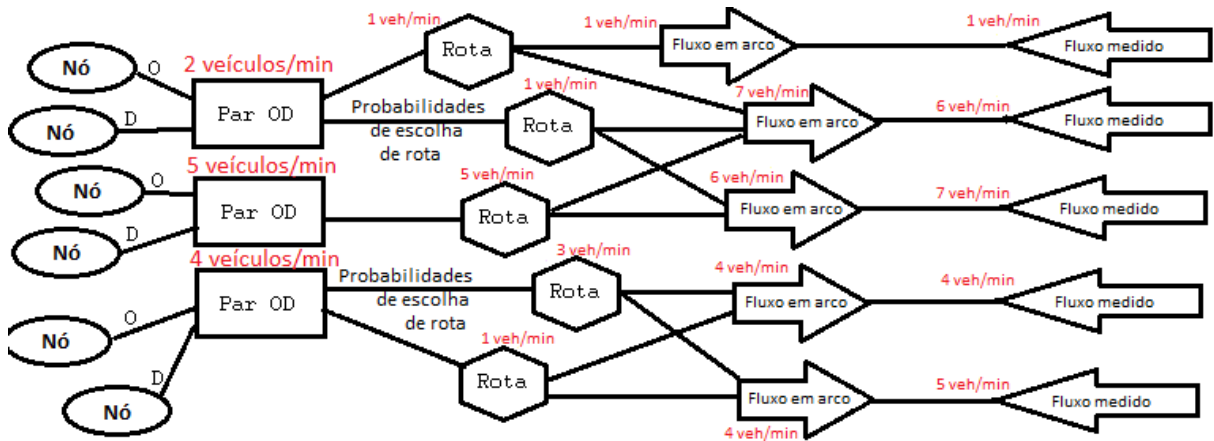


Figura 13: Diagrama da distribuição de tráfego através de pontos OD, rotas, fluxos e suas medições por sensores em arcos.

Um diagrama geral da estrutura do problema de estimação da matriz OD pode ser visto na Fig. 13, em que cada par OD é formado por um nó de origem e um de destino, e cada rota é uma sequência de arcos que liga um par OD, sendo escolhida de acordo com a probabilidade de escolha de rota. Por consequência, os arcos tem determinados fluxos de veículos, que podem ser medidos através de sensores. No entanto, somente com os sensores de tráfego nos arcos, as informações de fluxo por par OD, rotas por par OD e probabilidade de escolha de rota por par OD não podem ser observadas diretamente.

Além disso, conforme pode ser visto no exemplo da Fig. 13 na segunda, terceira e quinta linha de setas indicando sensores, podem existir divergências entre as medições dos sensores e o fluxo nos arcos, seja por falha dos sensores, ou por diferenças temporais. Quanto às diferenças temporais, uma vez que ocorra um tempo superior a zero para que um veículo saia do ponto de origem até algum sensor, e ocorram também variações nos fluxos entre pares origem-destino ao longo do tempo, o tráfego medido nos sensores não corresponde somente a veículos que tenham saído dentro de um determinado intervalo de tempo, mas também a alguns veículos que são de um intervalo de tempo anterior, em que o fluxo pode ser diferente. Da mesma forma, alguns veículos do intervalo considerado não chegam a tempo nos sensores para serem contabilizados neste mesmo intervalo. Com o objetivo de simplificar a modelagem do problema, pode-se considerar que o fluxo está

em regime, tratando o problema como estático, e então, desconsiderando a existência de erros e variações temporais.

Com relação à topologia e à localização de sensores na rede viária, pode-se considerar o mapa de uma cidade ou região como um grafo conexo e direcionado $G = (N, E, g)$, sendo N o conjunto de n nós que representam junções entre trechos de vias, como cruzamentos ou curvas. Os trechos de vias são representadas pelo conjunto E de arcos do grafo G , sendo $a \in E$ um arco, sendo g uma função que associa os arcos aos nós que correspondem às extremidades desses arcos. Além disso, uma rota é um caminho direcionado no grafo.

As seguintes definições são necessárias para formalizar o problema de estimação da matriz OD, ou ODME (do inglês, *Origin Destination Matrix Estimation*):

- $R \subseteq N$ é o conjunto de nós origem e destino; $|R| = o$ (o grafo tem o nós origem ou destino); $r \in R$ é um nó origem ou destino;
- $W = \{(i, j), i, j \in R, i \neq j\}$ é o conjunto de pares OD; $|W| = n_p$ (o grafo tem $n_p = o \cdot (o - 1)$ pares OD); w é um par $w \in W$;
- Fluxo real q_w para $w \in W$;
- Fluxo estimado \bar{q}_w para $w \in W$;
- p_{wa} é a probabilidade de uso do arco $a \in E$ pelo par OD w ;
- $Z \subseteq E$ é o conjunto de arcos com sensores; $|Z| = n_s$ (o grafo tem n_s sensores); $z \in Z$ é um arco com sensor;
- Fluxo medido v_z para $z \in Z$;
- Fluxo estimado \bar{v}_z para $z \in Z$;

O artigo Bierlaire (2002) considera que a estimação da matriz OD é o inverso do problema STA (*Static Traffic Assignment*). No STA, o fluxo de veículos nos pares OD são conhecidos, e deseja-se obter a distribuição deste fluxo pelas arcos da rede viária, conhecendo-se a probabilidade de uso das diversas rotas. Desta forma, na estimação da matriz OD, têm-se os valores de fluxos em determinados arcos (medidos por sensores), e deseja-se obter os fluxos por pares OD.

Podemos considerar que existe um número limitado de nós no mapa que podem ser tanto de origem como de destino, e a partir da combinação destes, formar um conjunto

W de $n_p = o(o - 1)$ pares OD. Para cada um destes n_p pares OD, há uma demanda de fluxo q_w e um conjunto de rotas (sendo que uma rota é um conjunto de arcos pelas quais se transita para chegar do nó de origem para o de destino) de tamanho maior ou igual a um. Existem também probabilidades de escolha entre essas rotas.

De acordo com a distribuição do tráfego pelos n_p pares OD e posteriormente pelas rotas de acordo com suas probabilidades de escolha de rota, há quantidades determinadas de veículos passando em cada um dos arcos do grafo. A fim de obter a distribuição do tráfego nos arcos a partir de uma matriz OD, as informações acerca de rotas e escolha destas rotas podem ser representadas por uma matriz com n_p colunas (pares OD), e n_s linhas (arcos), com valores reais variando entre 0 e 1, que é a probabilidade p_{wa} de uso do arco a para cada fluxo q_w de origem-destino. Essa matriz é denominada por (*traffic assignment matrix*).

Basicamente, a distribuição do tráfego na rede viária pode ser resumida através de (11), em que indica-se que o fluxo medido de um arco v_a é igual ao somatório dos n_p fluxos q_w , sendo cada fluxo multiplicado pela probabilidade de uso do arco p_{wa} por cada um dos pares OD. (12) é uma restrição de não negatividade de fluxos q_w , uma vez que é impossível existir um fluxo negativo de veículos.

$$v_a = \sum_{w \in W} q_w \cdot p_{wa} \quad (11)$$

$$q_w \geq 0 \quad (12)$$

O problema de estimar os fluxos q_w pelos n_p pares OD usando dados de sensores passa então pela necessidade de identificar a *assignment matrix*. É possível observar somente as diferenças entre valores estimados e medidos para fluxo em arcos, especificamente arcos com sensores (do conjunto Z), de forma que se busca uma situação próxima de (13), em que os fluxos estimados dos arcos sejam iguais aos fluxos medidos pelos sensores, com a finalidade de encontrar uma situação próxima de (14), em que os fluxos estimados para os pares OD seja próximo do real. Desta forma, o problema de estimação da matriz OD pode ser abordado como um problema de otimização buscando (13).

$$\forall z (v_z \approx \bar{v}_z) \quad (13)$$

$$\forall w(q_w \approx \bar{q}_w) \quad (14)$$

Para Bera e Rao (2011), é surpreendente que existam poucas abordagens realistas na literatura para aplicação em redes grandes, e as técnicas desenvolvidas precisam ser analisadas quanto à aplicabilidade em redes grandes e reais. Esta dissertação se utiliza de dois estudos de caso de redes viárias reais, sendo uma delas da segunda maior cidade de Portugal.

2.2 PROBLEMAS RELACIONADOS

Durante o desenvolvimento desta dissertação, foi necessário considerar problemas relacionados à estimação da matriz OD. A definição da quantidade e localização dos sensores na rede viária é discutida na Seção 2.2.1. Quando se usam dados georreferenciados de localização dos veículos (GPS) para reconstruir a rota utilizada na rede viária, é necessário utilizar uma técnica de *map matching*. Isso é tratado na Seção 2.2.2. A caracterização das probabilidades de uso das rotas para cada par OD, que é uma informação fundamental para estimação da matriz OD, é tratada na Seção 2.2.3.

2.2.1 ALOCAÇÃO DE SENSORES NA REDE VIÁRIA

Conhecido por *Network Sensor Location Problem* ou *NSLP*, o problema consiste em encontrar a quantidade e distribuição dos sensores na rede viária de forma a obter, por exemplo, uma melhor estimação da matriz OD. (YANG, 1998)(CHEN et al., 2005)

Segundo Hu e Liou (2014), o *NSLP* pode ser resolvido através de otimização matemática, abordagem algébrica ou teoria dos grafos (ao procurar por arcos críticos), e sensores ativos (como de identificação de placas veiculares) podem ser incorporados para melhorar a qualidade da solução encontrada para o *NSLP*. A codificação de uma solução do *NSLP* é simples, podendo ser utilizado um vetor de variáveis binárias, com uma posição para cada uma das arcos, indicando a presença ou ausência de um sensor no arco correspondente. Em Yang (1998), um problema de *NSLP* é resolvido para posterior estimação da matriz OD. Neste mesmo artigo, 4 regras para definição da localização dos sensores são definidas.

Ye e Wen (2017) apresenta uma boa revisão dos principais trabalhos que abordaram o problema do *NSLP* relacionado ao problema de estimação da matriz OD, a fim

de propor um método de resolução do NSLP que evita formar equações redundantes. Segundo Ye e Wen (2017), o problema de alocação de sensores é NP difícil e é tradicionalmente visto como um subproblema da estimação da matriz OD, existindo diversas abordagens para resolver o problema. Existem as medidas de TDS e MPRE para direcionar a solução, mas também, um conjunto de nove regras demonstradas no artigo que podem ser aplicadas a fim de definir o conjunto de sensores, conforme a lista abaixo:

- R1, cobertura de pares OD: Todos pares OD devem ser cobertos por sensores;
- R2, fração de fluxo OD em relação ao arco: Para cada par OD coberto, a fração dos fluxos deste par OD no arco, em relação ao fluxo total dos arco, deve ser maximizada;
- R3, cobertura de rotas: Todas as rotas devem ser cobertas por sensores;
- R4, maximização de fluxo: Deve-se escolher arcos com maior fluxo;
- R5, fração de fluxo OD em relação a rotas: Similar á regra R2, porém busca encontrar arcos nos quais o fluxo de um par OD seja maximizado em relação à demanda total deste mesmo par OD, que é distribuída em diferentes rotas;
- R6, maximização de rotas: Escolher arcos que apresentam a passagem do maior número de rotas;
- R7, maximização de pares OD: Escolher arcos que apresentam a passagem do maior número de pares OD;
- R8, independência dos arcos: Escolher um conjunto de arcos que evite informações redundantes;
- R9, minimização do número de sensores: Buscar o menor número de sensores, ao passo em que seja possível uma boa estimação da matriz OD.

O artigo Ye e Wen (2017) Realiza experimentos de estimação da matriz OD em uma rede no centro de Guangzhou na China, contendo 93 arcos direcionados, 38 nós e 210 pares OD. Utiliza do *equilíbrio de usuário* (Ver Seção 2.2.3) para definir as escolhas de rotas. Considera as regras R1, R8 e R9 para todos os casos, realiza comparações do método proposto, em relação a utilizar a regra R2, e ao utilizar as regras R4 e R7 em conjunto. É possível observar nos resultados que utilizar das regras R4 e R7 levou a menores erros em relação à utilização da regra R2. O método proposto supera o uso das

regras em casos específicos. Ainda em Ye e Wen (2017), são especificados as regras foram utilizadas para alguns artigos de estimação da matriz OD.

Nesta dissertação, o estudo de caso de Curitiba possui sensores com localização já definida pelo motivo de serem sensores físicos, enquanto que para o estudo de caso da cidade do Porto, devido à utilização de uma base de dados de corridas de táxi para definição de um problema de estimação da matriz OD, apresenta a necessidade de definir a localização de contadores de tráfego. O NSLP é utilizado também com o objetivo de avaliar como as diferentes técnicas de estimação da matriz OD são influenciadas pela quantidade de informação disponível.

A aplicação do NSLP neste trabalho é direcionada para as regras R1 (regra da cobertura), R4 (regra de máximo fluxo), R7 (regra de maior número de pares OD) e R9 (regra da minimização do número de sensores). Observa-se que as nove regras não devem ser obedecidas simultaneamente em uma mesma rede, podendo ser contraditórias entre si.

2.2.2 MAP MATCHING

Com os avanços da computação móvel e de localização, há uma quantidade crescente de dados georreferenciados. Estes dados, que podem ser processados usando diversas técnicas para obtenção de trajetórias, são genericamente denominados *spatial trajectory data*. Para esta obtenção de trajetórias, o *map matching*, também conhecido por *geo-matching*, é uma parte da chamada *trajectory data mining*, que, especificamente, busca detectar os segmentos de vias em que os sinais de GPS foram originados, criando uma correspondência entre o dado georreferenciado e as vias em um mapa.

Na Fig. 18 existem diversos segmentos de vias, representados por arcos (linhas cinza ou amarelas), e cruzamentos, representados por nós (junções das arcos). Através de dados georreferenciados (marcações em X), são realizados *matchings* entre tais pontos e a topologia da rede, a fim de se obter uma sequência de arcos ou nós no mapa que corresponda ao trajeto realizado (linhas amarelas e vermelhas para arcos identificadas, e pontos verdes para nós identificados).

Tal assunto é abordado devido à utilização de uma base de dados de corridas de táxi para definição de um problema de estimação da matriz OD na cidade do Porto, ao mesmo tempo em que é um assunto relacionado a dados georreferenciados de veículos sonda, que podem ser utilizados no auxílio da estimação da matriz OD. Veículos sonda

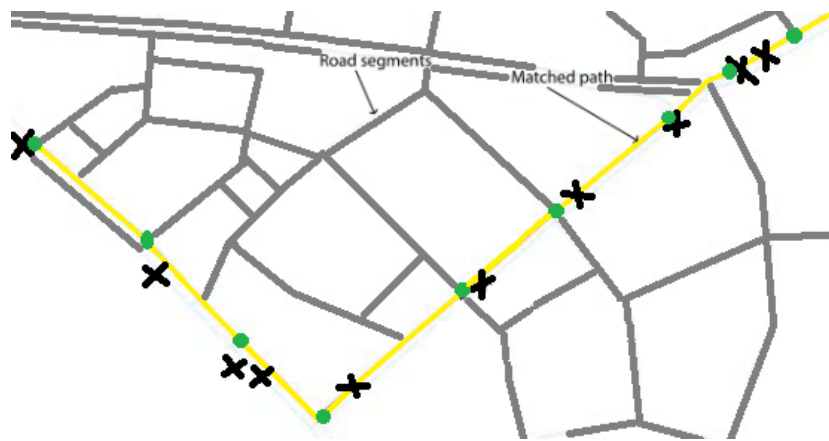


Figura 14: *Map matching* como um mapeamento entre dados georreferenciados (Marcações em X) e a rede viária (linhas cinzas e amarelas. Pontos verdes).

Fonte: (XIE et al., 2017)

correspondem a uma frota de veículos cujos dados de deslocamento são disponíveis, e então, permitem micro representações do tráfego em geral. Pela não disponibilidade de dados de veículos sonda para Curitiba, o problema não é abordado para o estudo de caso de Curitiba.

2.2.3 EQUILÍBRIO DE USUÁRIO

O conjunto e as probabilidades de escolha de rota por par OD dificilmente estão disponíveis, de forma que para redes viárias mais complexas, em que existe mais de uma possibilidade de rota utilizada por pares OD, um processo de estimativa do conjunto de rotas e suas probabilidades de escolha se faz necessário. Considerando a topologia da rede viária, velocidade máxima, capacidade, quantidade de veículos e degradação de desempenho de arcos, e possível utilizar técnicas de *equilíbrio de usuário*, ou UE (do inglês, *User Equilibrium*), para obter estes dados.

O UE pode ser calculado com auxílio de funções logísticas, seguindo o primeiro princípio de WARDROP e WHITEHEAD (1952), em que todas as rotas com fluxos positivos de um par OD devem ter custos iguais, e o custo não deve exceder o de nenhuma rota não utilizada para esse par OD. Basicamente, isso significa que os motoristas, individualmente, buscam utilizar rotas que minimizem seus respectivos tempos de viagem, e a partir disso, as rotas mais rápidas são degradadas por um maior fluxo, até que o custo dessas rotas se iguale ao custo das demais rotas utilizadas, originalmente mais lentas.

No entanto, o UE considera que os motoristas conhecem os tempos de viagem

em todos os arcos, o que faz com que o processo possa ser considerado enviesado, pois tal informação ideal pode não ocorrer no ambiente real. Dada esta característica, buscar alternativas ao UE é um tema de interesse. Conforme Teknomo e Fernandez (2014), a utilização de outros tipos de dados neste contexto não é recorrente, pois tecnologias de *tracking* são recentes, e ainda não há uma padronização acerca do compartilhamento destes dados.

O Equilíbrio de Usuário está relacionado com este trabalho, pois, para a rede viária de Curitiba, o conjunto e a distribuição de escolhas de rotas não são conhecidos. Para a rede viária da cidade do Porto, o problema de UE não é abordado, uma vez que são obtidas as informações de escolha de rota através do *map matching*.

2.3 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo foi apresentado o contexto geral da estimação da matriz OD, a partir do grafo da rede viária, fluxos por par OD, fluxos em arcos, e probabilidades de escolha de arco por par OD. O problema de estimação da matriz OD foi formalizado como um problema de otimização cujas técnicas de solução serão apresentadas no Capítulo 3.

Também foram descritos subproblemas fundamentais para a estimação da matriz OD. O primeiro sobre a definição da localização dos sensores de tráfego na Seção 2.2.1, que influencia os dados de entrada. Na Seção 2.2.2, é definido o problema de *map matching*, que é abordado neste trabalho para correspondência de dados georreferenciados e a topologia de uma rede viária, contribuindo para a definição do estudo de caso da cidade do Porto. Por fim, é discutida a distribuição de demandas de pares OD através de uma rede na Seção 2.2.3, seguindo princípios de equilíbrio de usuário, que são dependentes da topologia da rede e demandas de pares OD.

No próximo capítulo, será feita uma revisão bibliográfica das técnicas consideradas nesta dissertação para a estimação da matriz OD e uma discussão sobre os problemas da Seção 2.2 com trabalhos da literatura.

3 REVISÃO BIBLIOGRÁFICA

Na Seção 3.1 são citados diversas abordagens de resolução da estimação da matriz OD encontradas na literatura, incluindo mas não se limitando às aplicadas nos experimentos do Capítulo 5.

Na Seção 3.2 é analisada a questão inerente ao uso de dados para estimação da matriz OD, considerando uma possível nova tendência de resolução do problema, devido às novas possibilidades de obtenção de dados de mobilidade, como o uso dados adicionais provenientes de veículos sonda. Nesta mesma seção, alguns trabalhos interessantes que se utilizam de dados de mobilidade são apresentados.

3.1 MÉTODOS DE ESTIMAÇÃO DA MATRIZ OD

De acordo com Chen et al. (2005), as matrizes OD geralmente são obtidas a partir de pesquisas de campo de custo elevado, realizadas em residências ou no trânsito. Porém, as informações obtidas nessas pesquisas tornam-se obsoletas rapidamente. Portanto, pesquisas vêm sendo desenvolvidas nas últimas décadas com o objetivo de obter matrizes OD de forma rápida e barata. Basicamente, as maiores dificuldades da estimação da matriz OD são:

- Trata-se de um problema subespecificado, com grande quantidade de soluções válidas e mínimos locais devido ao elevado número de variáveis e reduzido número de equações independentes entre si;
- A localização dos sensores influencia bastante o resultado, conforme visto na Seção 2.2.1. Além disso, existe um grau de erro associado aos dados dos sensores;
- Algumas técnicas podem utilizar uma matriz OD semente ou priori advinda de pesquisa de campo ou um processo de estimação da matriz OD anterior, para direcionar a solução. Ainda assim, existe incerteza nesta matriz, e muitas vezes tal matriz é indisponível;

- A escolha das rotas envolve um grau de incerteza. Eventualmente, a fim de simplificar o problema, considera-se que cada par OD possui somente uma rota possível (geralmente a rota de menor custo). Porém, aplicações de técnicas de equilíbrio de usuário (Seção 2.2.3) para a definição das probabilidades de escolha das rotas são comuns na literatura.

Segundo Bierlaire (2002), o problema de estimação da matriz OD é complexo pelo fato de que um efeito possui múltiplas causas. Neste problema, geralmente existe um número de incógnitas (fluxo de veículos por par OD) muito superior ao número de equações disponíveis (fluxo de veículos medido em cada arco) para a resolução, o que faz com que a estimação da matriz OD possua inúmeras soluções válidas. Em alguns métodos de estimação da matriz OD, é utilizada uma matriz de viagens OD semente (*A Priori OD Matrix*), para que, dentre as soluções possíveis, busque-se uma que seja mais próxima desta matriz de viagens OD de referência. Esta matriz de referência é obtida através de uma pesquisa anterior, ou definida arbitrariamente com valores iniciais zero ou um.

A estimação da matriz OD é uma área de pesquisa que já possui algumas décadas. Inicialmente, foram desenvolvidas soluções considerando métodos baseados em modelos de demanda de viagens, como modelos de gravidade (GM). Posteriormente, foram utilizados métodos fazendo uso de dados de sensores de tráfego. Métodos estatísticos (GLS, KF, BI e ML) de otimização (MILP, GA, PSO). Mais recentemente, algumas abordagens são direcionadas à ciência dos dados, considerando o crescente volume de dados gerados por sistemas de informação oriundos da monitoração de serviços e condições ambientais das cidades. As siglas GM, GLS, KF, BI, ML, MILP, GA e PSO são explicadas posteriormente nesta mesma seção.

Na Tabela 13 é apresentada uma lista não exaustiva de aplicações de técnicas de estimação da matriz OD, sendo a(s) técnica(s) utilizada(s) explicitadas na segunda coluna. A terceira coluna indica se o artigo utiliza cálculos de *equilíbrio de usuário* (Seção 2.2.3) para definição da *assignment matrix*. A quarta coluna indica se o problema envolve uma matriz OD estática ou dinâmica, sendo que uma matriz OD dinâmica apresenta fluxos variantes nos pares OD ao longo do tempo, sendo incluída então uma dimensão temporal ao problema. A quinta coluna mostra a dimensão da rede viária utilizada nos artigos. Na última coluna, indica-se que métricas de desempenho os artigos utilizaram para avaliar os resultados. Maiores detalhes sobre as métricas de desempenho podem ser obtidos na Seção 4.1.

Em especial, os artigos Maher (1983), Cascetta (1984), Spiess (1987), Bell (1991)

Tabela 13: Artigos com aplicações de diferentes técnicas para a estimação da matriz OD

Artigo	Técnica(s) utilizada(s)	Aborda o equilíbrio de usuário	Matriz OD dinâmica	Dimensão do problema	Métrica(s) de desempenho para ODME
Voorhees (2013)	GM	Não	Não	Rede viária não considerada	-
Maher (1983)	BI	Não	Não	6 pares OD, 4 arcos	-
Cascetta (1984)	GLS	Não	Não	Matriz OD 5x5. 14 arcos	MSE
Nihan (1987)	KF	Não	Não	Matriz OD 3x3	RMSE
Spiess (1987)	ML	Não	Não	Matriz OD 4x4	RMSE
Cascetta e Nguyen (1988)	GLS, ML e BI	Não	Não	14 arcos, 6 nós	MSE
Bell (1991)	GLS	Não	Não	Matriz OD 2x2. 5 arcos	-
Yang et al. (1992)	GLS	Sim	Não	24 arcos, 9 nós	MAE e RMSE
Sherali e Hobeika (1994)	MILP, ML	Sim	Não	Matriz OD 3x6. 18 arcos	RMSE
Sherali e Park (2001)	GLS	Sim	Sim	15 arcos, 10 nós	MAE
Lo e Chan (2003)	ML	Sim	Não	Matriz OD 4x4. 14 arcos	RMSE
Baek et al. (2004)	GA	Sim	Não	16 arcos, 9 nós	MAE
Yun et al. (2005)	GA	Sim	Sim	3364 arcos, 1465 nó. 57 regiões de OD	SSE
Hazelton (2008)	BI	Sim	Sim	50 arcos, 21 nós. 85 pares OD	-
Du e Wei (2009)	PSO	Não	Sim	Matriz OD 4x4	MSE
Parry e Hazelton (2012)	ML	Sim	Sim	50 arcos, 21 nós. 85 pares OD	RMSE

e Sherali e Hobeika (1994) são referências recorrentes nos artigos da área. É possível notar na Tabela 13 que a maioria dos artigos realiza experimentos em redes viárias de pequeno porte, e as métricas RMSE e MSE são as mais utilizadas (veja Seção 4.1). Assim como o uso de cálculos de *equilíbrio de usuário* para definição da *assignment matrix*, matrizes OD dinâmicas são tratadas somente em artigos mais recentes. A maioria dos artigos aplica somente uma técnica de estimação da matriz OD, e apenas uma métrica de desempenho.

O artigo Voorhees (2013), inicialmente publicado em 1955, apresenta a primeira aplicação de um modelo de gravidade (sigla GM. Outros artigos utilizando da técnica podem encontradas no *survey* de Bera e Rao (2011)). Basicamente, a estimação da matriz OD a partir de modelos de gravidade não considera a topologia da rede ou dados de sensores de tráfego, mas sim dados territoriais, dados demográficos, fórmulas próprias e parâmetros para calcular as demandas OD. Alguns exemplos de dados utilizados são a quantidade de famílias residentes em uma região, a área de lojas em metros quadrados, ou a distância de uma zona residencial em relação a uma área comercial.

Na maximização de verossimilhança (ML), o modelo tem como objetivo estimar os parâmetros de processo Poisson que ocorre na formação da matriz de viagens OD, considerando que cada par OD é gerado através de um processo de Poisson independente. Aplicar a técnica para esse problema significa encontrar a matriz OD que a possui maior probabilidade de ser observada, de acordo com as distribuições de Poisson. (SPIESS, 1987)

No artigo Nihan (1987) sobre filtro de Kalman (KF), é descrita uma abordagem utilizando um método de mínimos quadrados recursivo. Pelo motivo de repetidas inversões

de matriz ser um trabalho computacional ineficiente, os autores passam a utilizar, com resultados similares, da técnica do filtro de kalman. É uma técnica útil para se atualizar matrizes OD dinâmicas, pela possibilidade de colocar pesos maiores para dados mais recentes.

Segundo Maher (1983), o método proposto de Inferência Bayesiana (BI) possui uma flexibilidade na confiança entre os valores encontrados na matriz OD semente (ou *a priori*) e os valores medidos pelos sensores nos arcos. Em casos de pouca confiança na matriz OD semente, o resultado é próximo dos métodos de maximização de entropia.

Descrito em Cascetta (1984) e Bell (1991), o método de mínimos quadrados, ou GLS, considera explicitamente a presença de erros de medição e a variabilidade no ambiente ao longo do tempo, neste caso através de matrizes de covariância de fluxos. O objetivo do *GLS* é encontrar um conjunto de valores para as variáveis tal que o somatório do quadrado das diferenças entre os valores estimados e observados seja minimizado. Além disso, o método de GLS pode também considerar diferenças em relação á matriz OD semente, a fim de que a matriz OD resposta não seja tão diferente da matriz OD semente. Tal característica é importante para casos em que há um bom grau de confiança na matriz OD semente ou *priori*.

A Programação linear inteira mista (*Mixed integer linear programming* ou MILP) é uma variação da *integer linear programming*, sendo a diferença o fato de que algumas variáveis assumem valores não inteiros. É uma técnica de otimização matemática, em que a função *fitness* e restrições são lineares, sendo o algoritmo *simplex* uma solução popular. (CORMEN et al., 2001)

O problema de estimação da matriz OD pode ser formulado como um problema de otimização, conforme Bera e Rao (2011) e Sherali e Hobeika (1994). Erros nos sensores que levam a inconsistências são tratados no modelo MILP através de variáveis de folga, para que haja soluções válidas para o problema, mesmo com erros nos dados. Algumas restrições precisam ser incorporadas para a concepção do modelo como, por exemplo, a regra de que todo par OD precisa estar coberto por sensores, ou caso contrário, a variável de quantidade de veículos do par OD em questão tenderia ao infinito. Partindo da mesma formulação do problema de estimação da matriz OD com um problema de otimização, outras técnicas de otimização podem ser aplicadas.

Um exemplo de aplicação do GA para solução do problema de estimação da matriz OD é em Baek et al. (2004), em que os autores usam do GA para um problema considerando contadores de tráfego capazes de identificar diferentes tipos de veículos. A

partir dos experimentos, conclui que a técnica produz melhores resultados em relação a algumas técnicas de estimação de matriz OD de dois níveis que realizam iterações entre estimação da matriz OD e cálculos de *equilíbrio de usuário*.

Para problemas de estimação da matriz OD em que os efeitos do congestionamento são levados em conta, existem aplicações de otimização de dois níveis, em que o problema de *equilíbrio de usuário* é tratado a um nível inferior em relação a estimação da matriz OD. Por exemplo, em Yang et al. (1992) é demonstrado que pode ocorrer uma combinação de métodos para resolver o problema, ao utilizar a técnica de mínimos quadrados em conjunto com o cálculo de equilíbrio de usuário. Neste caso, a otimização de dois níveis visa à interdependência encontrada entre os problema de estimação da matriz OD e o problema de *equilíbrio de usuário*, a fim de se chegar a uma convergência. Em Sherali e Hobeika (1994) ocorre uma otimização de dois níveis de MILP para estimação da matriz OD, em conjunto com o procedimento de *equilíbrio de usuário* a um nível inferior.

Outro exemplo de otimização de dois níveis ocorre em Hu e Liou (2014), mas envolvendo o NSLP. Foi utilizado o *solver* comercial LINGO 11 para resolver o problema da estimação da matriz OD utilizando programação não linear, enquanto a um nível superior, se utilizou programação linear inteira mista para resolver o problema do NSLP (descrito na seção 2.2.1), a fim de obter melhores resultados na estimação da matriz OD.

O artigo Cascetta e Nguyen (1988) compara resultados de três diferentes métodos estatísticos de estimação da matriz OD. Os métodos utilizados são mínimos quadrados, inferência bayesiana e máxima verossimilhança em diferentes cenários, considerando erros nos dados dos sensores. A técnica de máxima verossimilhança se saiu levemente superior à técnica de mínimos quadrados, mas considerando que o esquema de amostragem favorecia a técnica de máxima verossimilhança, isso indica robustez por parte do estimador de mínimos quadrados. As técnicas de inferência bayesiana e máxima verossimilhança se comportaram de forma similar.

Em Alexander et al. (2015), os autores utilizam dados de telefones celulares, técnicas de pré-processamento, agrupamento espacial e inferência da atividade realizada para estimação da matriz OD na cidade de Boston, em contraposição às técnicas tradicionais de estimação da matriz OD, sendo o foco nos motivos que levam aos deslocamentos, sem incluir sensores em arcos ou a rede viária. Dados de localização dessa natureza são identificados através de triangulação sem grande precisão. Assim, as posições identificados possuem uma margem de erro de 200 a 300 metros, sendo a topologia da rede viária e rotas não consideradas, tampouco dados de sensores de tráfego. Os resultados são va-

lizados através de uma comparação com censos locais. O artigo conclui que tais dados podem ser complementares para realização de censos de mobilidade.

Em Yang et al. (2017), os autores utilizam dados de veículos sonda de três formas diferentes para estimar a matriz OD. A primeira é baseada nas proporções de veículos sonda por arco para determinar a matriz diretamente, a segunda ao utilizar os veículos sonda como matriz *a priori*, e a terceira ao utilizar os dados para definir a proporção de uso de arcos por par origem-destino. A conclusão é que os dois últimos métodos tem bom desempenho em um primeiro cenário de homogeneidade de quantidade de veículos sonda em relação aos veículos comuns, mas que para um cenário de não homogeneidade, a terceira técnica produz melhores resultados, desde que haja uma quantidade suficiente de veículos sonda em relação ao trânsito geral.

Nesta dissertação, as técnicas definidas para serem comparadas são de mínimos quadrados, programação linear inteira mista, algoritmo genético e otimização por nuvem de partículas. A escolha conjunta do PSO, GA e MILP é vantajosa pelo motivo de que todas estas abordagens solucionam a estimação da matriz OD como um problema de otimização utilizando funções *fitness*. O uso dos mínimos quadrados se faz especialmente interessante por ele ser uma das técnicas comparadas em Cascetta e Nguyen (1988), além de ser uma abordagem utilizada em vários artigos que solucionam a estimação da matriz OD, conforme visto na Tabela 13.

3.2 FONTES E DESCRIÇÃO DOS DADOS PARA ESTIMAÇÃO DA MATRIZ OD

Embora existam problemas clássicos na área de transporte extensivamente estudados, tais como roteamento e localização de facilidades, atualmente é necessário unir conhecimentos de TI, transporte e infraestrutura para obter melhores soluções utilizando dados espaço-temporais. De acordo com Zwitter (2014), os governos mudarão progressivamente a perspectiva de soluções fornecidas por especialistas regionais para uma perspectiva de estratégias baseadas em algoritmos e dados globais. As cidades inteligentes (*smart cities*) deverão se tornar sistemas de sistemas, de forma que exista uma integração de dados de várias fontes independentes, criando assim uma visão holística da cidade (CAVALCANTE et al., 2016). Este novo paradigma pode ser benéfico para o gerenciamento e planejamento de transportes, entre outros objetivos. Porém, existem alguns desafios como o fato de que as possíveis fontes de dados possuem formatos próprios, e estão sob controle de diferentes agentes. Em Elwood (2010), é considerada uma mudança de paradigma de dados espaciais produzidos pelo governo para dados espaciais produzidos

por múltiplos agentes, o que requer uma mudança na forma de se trabalhar, considerando os problemas políticos inerentes ao compartilhamento de dados. Além disso, as áreas de transporte e infraestrutura podem ser assuntos distantes para um profissional de TI, da mesma forma que sistemas de sistemas e ética do uso de *big data* podem ser distantes para um profissional dos transportes, exigindo colaboração multidisciplinar.

Planejadores urbanos estão embutidos no conceito de utilizadores de *big data* de Zwitter (2014), e eventualmente, podem também ser coletores de *big data*, para o caso de sensores de tráfego, por exemplo. Entretanto, a maior parte dos dados de trânsito coletados permanece sob controle de empresas privadas que não necessariamente está disposta a fornecer estes dados. Por outro lado, a comunidade *open-source* é uma força contrária às aplicações proprietárias. (ELWOOD, 2010).

Ainda não existe uma conclusão sobre quais dados devam ser compartilhados com quem. A ética de *big data* é apenas uma parte do problema, já que informações sigilosas como endereços residenciais devem ser omitidas, e até certo ponto, informações acerca de grupos (CAVALCANTE et al., 2016). Em Xiong et al. (2006), os autores citam um caso em que os endereços residenciais de voluntários de um projeto podiam ser identificados através do uso de informação pública. Outro problema do compartilhamento de dados é que conflitos e relações políticas são inerentes a isso, criando uma situação complexa. (CAVALCANTE et al., 2016)(ELWOOD, 2010)

Segundo Hu e Liou (2014), outras informações começaram a ser incluídas, principalmente a partir da década de 2000, ao processo de estimação da matriz OD com fim de aumentar a exatidão. Foram acrescentadas informações acerca de:

- Sensores infravermelhos;
- Identificação de veículos;
- Dados de veículos sonda;
- Dados de celulares;
- Reconhecimento de placas veiculares;
- Probabilidades de conversão em cruzamentos.

Alguns exemplos de veículos sonda seriam ônibus, veículos de entrega e táxis. Considerando as características das viagens, os táxis seriam veículos sonda mais adequados como dados complementares para estimação da matriz OD, pois não possuem rotas

definidas previamente, possuem bom grau de penetração em relação aos veículos comuns, e existe maior homogeneidade de distribuição de viagens por par OD em relação a veículos de entrega.

Tradicionalmente, os dados utilizados para resolução do problema de estimação da matriz OD são:

- Topologia da rede, grafo G
- Fluxo de veículos em determinados pontos da rede
- Probabilidades de uso de rota ou arco por par OD
- Matriz OD *a priori* ou semente
- Modelos de comportamento dos motoristas

No entanto, com o número crescente de possibilidades de obtenção de dados através de veículos sonda, surge a oportunidade de uma nova abordagem para o problema, reduzindo a incerteza em relação à definição das rotas, ao fluxo de veículos em determinados pontos da rede, à matriz OD semente e em relação ao comportamento dos motoristas. O custo de manter e instalar um grande número de sensores físicos é proibitivo, enquanto a obtenção de dados através de veículos sonda apresenta um bom custo benefício. Porém, o acesso a estes dados pelos planejadores urbanos pode envolver um problema político com os coletores destes dados.

Dentre os diversos tipos de veículos que circulam no ambiente urbano, a movimentação dos táxis parece fornecer dados mais relevantes a estimação da matriz OD, uma vez que apresenta um bom número e diversidade de rotas para cada par OD. A seção a seguir discute alguns aspectos do uso de dados de táxis na avaliação da mobilidade urbana.

3.2.1 BASES DE DADOS DE TÁXI

Taxistas são especialistas em transitar pela cidade, podendo indicar boas escolhas de rotas para diferentes origens e destinos em diferentes horários, com alguma semelhança com o trânsito em geral. (YANG et al., 2017)(XIE et al., 2017)

Na Tabela 14, têm-se exemplos de bases de dados de mobilidade de táxis. Porém, muitas possuem dados que não incluem dados de trajetória dos táxis, que é uma in-

formação relevante para a estimação da matriz OD. Em Yang et al. (2017), são apresentados resultados de estimação da matriz OD com e sem informações de trajetória. Os resultados com informações de trajetória são melhores.

Tabela 14: Bases de dados de táxis

Cidade	Ano	Acesso	Dados de trajetória	Dimensão aproximada
Porto	2014	Livre	Sim	1.710.671 corridas
Chigaco	2013 - Atual	Livre	Não	27.000.000 corridas
São Francisco	2008	Restrito	Sim	500 táxis, 30 dias
Nova York	2009 - Atual	Livre	Não	600.000.000 corridas
Roma	2014	Restrito	Sim	320 táxis, 30 dias
Shenzhen	2008	Livre	Sim	10.357 táxis, 7 dias
Singapura	Tempo real	Restrito	Não	5.000 táxis
Hangzhou	Tempo real	Restrito	Sim	11.000 táxis

Apesar de existirem alguns *datasets* disponíveis, a quantidade de fontes de dados ainda deixa a desejar, principalmente quando são necessários dados de trajetória. Trabalhos recentes como Yang et al. (2017) utilizam dados sintéticos, uma vez que dados reais de mobilidade dificilmente estão disponíveis. A maior disponibilidade deste tipo de dados permitiria um maior número de aplicações, incluindo, mas não limitando-se à estimação da matriz OD. Outro ponto a se considerar é que para que os dados de veículos sonda sejam representativos para determinadas aplicações, é necessário que haja uma quantidade suficiente de veículos sonda em relação ao trânsito em geral para garantir a representatividade. Esta característica é denominada grau de penetração do veículo sonda.

Dados de mobilidade, especificamente de táxis, possibilitam uma grande gama de aplicações, com alguns exemplos a seguir. Em Matias et al. (2016), dados abertos de corridas de táxi são utilizados para inicialmente dividir a cidade em regiões, através de um agrupamento espacial em densidade utilizando *Half Space Trees*. Esse agrupamento espacial é realizado com o objetivo de criar uma estimativa do tempo de viagem entre regiões de interesse, sendo que a ideia básica é considerar os *clusters* como áreas densas.

Em Jiang et al. (2015), os autores utilizaram uma base de dados não relacional em *Hadoop* com dados de trajetórias de táxis. A cada dia, mais de 400.000 novas trajetórias de táxis são adicionadas. O objetivo é possibilitar consultas e viabilizar uma ferramenta visual que permita compreender os padrões de movimento humano entre diferentes regiões da cidade. A estrutura de dados utilizada é apropriada para utilizar algoritmos de mineração de dados, e implementação semelhante poderia ser utilizada para outras aplicações. Essa aplicação demonstra como os dados dos veículos de sonda

são ricos, possibilitando diferentes aplicações.

No artigo Geroliminis e Daganzo (2008), os autores mostram que existe uma correspondência entre o número de táxis que entram e saem de uma região e a quantidade total de veículos que entram e saem dessa mesma região. Isso é um indicativo de que o comportamento dos táxis é semelhante ao do trânsito em geral. Geroliminis e Daganzo (2008) também observa que sensores físicos não são abundantes nas vias, enquanto localizadores GPS são cada vez mais presentes.

Em Barann et al. (2017), um *dataset* de 5 milhões de corridas de táxis da cidade de Nova York foi utilizado para desenvolver uma solução de compartilhamento de corrida de táxi ponto a ponto, que agrupa passageiros com origens e destinos similares. Através desta aplicação, o compartilhamento de corrida de táxi ponto a ponto reduz o custo para os usuários, e permite que as empresas de táxis façam melhor uso de seus recursos, ao mesmo tempo em que são reduzidas as emissões de gases estufa. O resultado da abordagem proposta agrupou até 48,34% de todas as corridas de táxi, economizando 2.892.036 Km de distância viajada, 231.362,89 litros de combustível, e 532.134,64 Kg de emissões de gás carbônico por semana.

Em Liu et al. (2012), a mobilidade urbana é investigada por meio de dados da trajetória de táxis coletados de Xangai, China. O conjunto de dados registra mais de 6.600 veículos de uma companhia de táxi anônima e abrange sete dias consecutivos de junho de 2009. Para cada táxi, informações sobre posição, velocidade e se clientes estão sendo transportados são coletados automaticamente aproximadamente a cada 10 segundos aproximadamente.

No artigo de Xie et al. (2017), partindo do pressuposto de que as corridas de táxi podem ser consideradas como uma microrrepresentação do trânsito em geral, os autores procuram validar a condição de proporcionalidade de escolha de rotas usando mais de 27 milhões de observações de escolha de rota ao minerar a base de dados de corridas de táxi da cidade de Shenzhen, China. A condição da proporcionalidade tem sido amplamente utilizada para produzir um conjunto de rotas e escolhas de rotas única para o cálculo do *equilíbrio de usuário*, ligado ao problema de assinalação de tráfego (STA), partindo do princípio de que os motoristas buscam minimizar seus tempos de viagem. A condição de proporcionalidade foi testada através de regressão linear e testes chi-quadrado em mais de 300 segmentos alternativos pareados (chamados no artigo de PAS, representando rotas alternativas para um mesmo par OD). Os resultados demonstram que a maioria dos PAS testados (até 85%) satisfaz a condição de proporcionalidade com um nível relevante de

significância estatística.

O artigo Biem et al. (2010) mostra o uso do software comercial IBM *infoSphere Streams* para *map matching* em tempo real, identificando o tempo médio de viagem em cada arco de uma rota. São utilizados dados de táxis e caminhões, com mais de 1000 leituras de GPS por segundo, e um mapa com mais de 600000 arcos. O resultado é uma ferramenta que permite consultas sobre velocidade média em cada arco de uma rota, ou tempo estimado de viagem entre dois pontos da cidade.

Desde 2014, a prefeitura de Curitiba disponibiliza, através de um portal aberto de dados (PMC, 2018), diversos dados públicos, incluindo dados relativos ao transporte público, de forma que eles possam ser reutilizados para produzir novas informações e aplicações digitais para a sociedade. Porém, dados de táxis como veículos sonda não estão inclusos nesta base de dados. O principal problema relativo ao compartilhamento de dados de mobilidade de táxis na cidade de Curitiba é que houve uma terceirização do gerenciamento do sistema de localização de táxis em 6 das 7 empresas, e a empresa responsável pelo sistema fornece somente acessos pontuais aos dados através de uma interface *web*, sem viabilidade prática para acessar um grande volume de dados. Existe também um receio por parte dos profissionais do ramo envolvendo privacidade dos dados. Porém, este problema pode ser resolvido com anonimização temporal como em UCI (2015) ou espacial, como é feito em Biem et al. (2010).

3.3 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo foram mencionados várias técnicas utilizadas ao longo das últimas décadas para a estimação da matriz OD. Além disso, foram discutidos aspectos relacionados à fonte e descrição dos dados para estimação da matriz OD. Particularmente, dados de viagens de táxis que serão utilizados nesta dissertação.

No próximo capítulo, será apresentada a metodologia utilizada nesta dissertação, detalhando as técnicas utilizadas para a estimação da matriz OD, bem como as métricas de desempenho para avaliação destas técnicas.

4 METODOLOGIA

Este capítulo aborda a metodologia adotada nesta dissertação. Embora o objetivo principal deste trabalho seja a comparação de técnicas para a estimação da matriz OD, outras atividades foram necessárias para atingir os objetivos específicos. Um esquema do inter-relacionamento entre as atividades desenvolvidas é mostrado na Fig. 17. Porém, nem todas as atividades estão presentes em ambos os estudos de caso considerados (Porto e Curitiba - descritos no Capítulo 5. Os retângulos em cinza são restritos ao estudo de caso da cidade do Porto, enquanto que o retângulo tracejado é restrito ao estudo de caso de Curitiba). Conforme a Fig. 17, a definição dos dados de entrada foi realizada através de *map matching* (Seção 4.8) e agrupamento espacial quando dados de GPS estavam disponíveis, para o estudo de caso da cidade do Porto, e foi realizada importação do software VISUM para o estudo de caso de Curitiba.

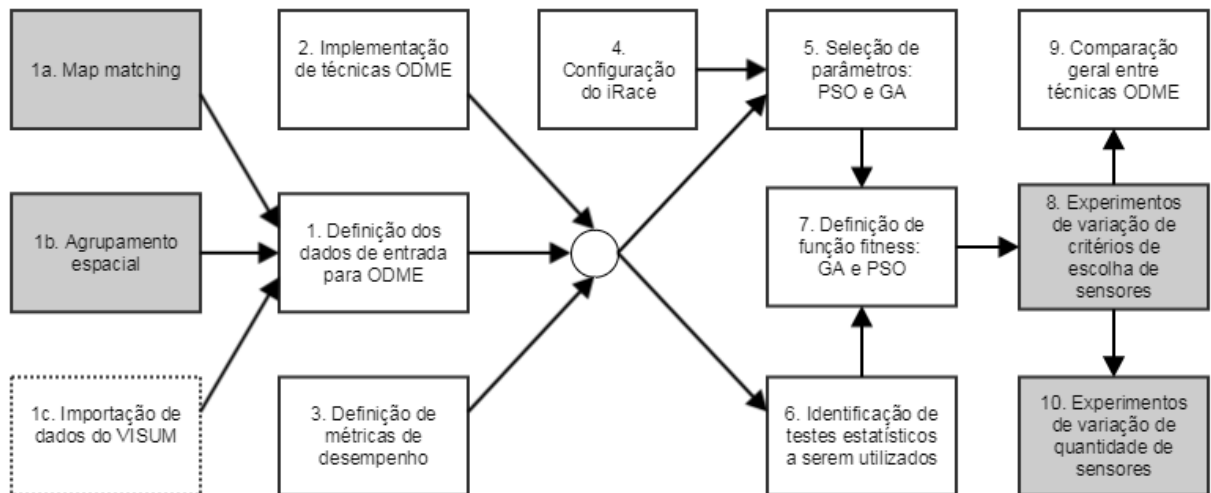


Figura 17: Esquema do inter-relacionamento entre as atividades desenvolvidas na dissertação.

As medidas de desempenho utilizadas são descritas na Seção 4.1, para que seja possível realizar comparações entre diferentes técnicas, e também para diferentes configurações dentro de uma mesma técnica. Quatro técnicas de estimação da matriz OD (PSO, GA, LS e MILP) foram avaliadas e descritas entre as Seções 4.7 e 4.5. São defi-

nidos detalhes das implementações das técnicas utilizadas no trabalho, como seleção de parâmetros, definição de funções *fitness*, restrições e variáveis utilizadas.

4.1 MEDIDAS DE DESEMPENHO

A Tabela 16 apresenta um resumo das métricas utilizadas neste trabalho, cujas descrições mais detalhadas podem ser encontradas nas Seções 4.1.1 para MAE, 4.1.2 para RMSE, 4.1.3 para TDD, 4.1.4 para GEH e pGEH5, e 4.1.5 para R^2 ODM e R^2 Arcos.

Tabela 16: Métricas utilizadas para avaliar a qualidade da estimação da matriz OD

	MAE	RMSE	R^2 ODM	TDD	GEH	pGEH5	R^2 Arcos
Objetivo	Minimizar	Minimizar	Maximizar	Minimizar	Minimizar	Maximizar	Maximizar
Variável analisada	Matriz OD	Matriz OD	Matriz OD	Soma da Matriz OD	Arcos	Arcos	Arcos
Valores	$[0, +\infty)$	$[0, +\infty)$	$[0, 1]$	$[0, +\infty)$	$[0, +\infty)$	$[0, 1]$	$[0, 1]$

Basicamente, são utilizadas quatro métricas (MAE, RMSE, R^2 ODM e TDD) para medir o desempenho da estimação da matriz OD em relação ao valor conhecido da matriz OD, e outras três métricas para avaliar a adequação dos fluxos de veículos nos arcos. Além disso, quatro métricas (GEH, MAE, RMSE e TDD) são medidas de erro, e portanto, devem ser minimizadas. Três métricas (pGEH5, R^2 Arcos e R^2 ODM) derivam de correlações entre os valores e estimados e observados, e portanto, devem ser maximizadas.

4.1.1 MEAN ABSOLUTE ERROR

A métrica MAE só pode ser aplicada caso a matriz OD seja conhecida *a priori*, e mede o quanto a solução estimada está próxima da real em termos de fluxo de veículos por par OD (q_w), conforme (17). Um valor MAE menor indica menores desvios entre o valor real e o estimado para o fluxo de veículos em cada par OD.

$$MAE = \frac{\sum_{w \in W} |\bar{q}_w - q_w|}{n_p} \quad (17)$$

4.1.2 ROOT MEAN SQUARE ERROR

De forma semelhante a MAE, a métrica de RMSE só pode ser aplicada caso a matriz OD seja conhecida *a priori*, e mede o quanto a solução estimada está próxima

da real em termos de fluxo de veículos por par OD (q_w). Dados w pares OD, a métrica RMSE é calculada segundo 18.

$$RMSE = \frac{\sqrt{\frac{1}{n_p} \sum_{w \in W} (\bar{q}_w - q_w)^2}}{\sum_{w \in W} q_w} \quad (18)$$

A métrica RMSE é muito utilizada nos artigos de estimação da matriz OD, conforme a Tabela 13. Uma métrica similar também muito utilizada nos artigos de estimação da matriz OD (mas não aplicada neste trabalho) é a MSE, que é basicamente o dividendo da métrica RMSE, sem o uso da raiz quadrada.

4.1.3 TOTAL DEMAND DEVIATION

Assim como MAE e RMSE, a métrica TDD só pode ser obtida caso a matriz OD resposta seja conhecida *a priori*, e também mede o quanto a solução estimada está próxima da real em termos de fluxo de veículos nos pares OD (q_w), segundo (19).

$$TDD = \frac{|\sum_{w \in W} \bar{q}_w - \sum_{w \in W} q_w|}{\sum_{w \in W} q_w} \quad (19)$$

No entanto, a métrica não analisa desvios individuais de fluxo de cada par OD, mas sim o somatório de fluxos em todos os pares OD. O valor TDD pode ser zero, ao mesmo tempo em que RMSE e MAE não sejam, pois erros de estimação nos fluxos individuais podem compensar uns aos outros, resultando em somatório de fluxo em todos os pares OD muito próxima, ou até igual a real.

4.1.4 GEH STATISTIC

A medida GEH (Geoffrey E. Havers) é utilizada para mensurar a adequação da estimativa de fluxos dos arcos, avaliando o quanto o fluxo estimado de veículos está próximo do observado pelos sensores (MA; ABDULHAI, 2002). Isso é especialmente útil em casos nos quais não se conhece a matriz OD.

A métrica $GEH(z)$ para um arco z pode ser descrita por (20).

$$GEH(z) = \frac{\sqrt{2(\bar{v}_z - v_z)^2}}{\bar{v}_z + v_z} \quad (20)$$

Neste trabalho, com a finalidade de considerar um valor médio do GEH nas redes

para comparar diferentes técnicas, é utilizada (21).

$$GEH = \frac{\sum_{z \in Z} GEH(z)}{n_s} \quad (21)$$

Segundo Ciuffo et al. (2008) e Hourdakis e Michalopoulos (2003), $GEH(z) < 5$ em 85% dos arcos monitorados por sensores, é considerado adequado. Por esse motivo, outro ponto a ser analisado é a porcentagem de arcos com GEH abaixo de 5. Pela necessidade de se analisar este ponto, neste trabalho foi definida uma métrica adicional, batizada por pGEH5.

4.1.5 COEFICIENTE DE DETERMINAÇÃO R^2

Um modelo de regressão linear pode ser obtido entre as variáveis estimadas e observadas, sendo o coeficiente de determinação R^2 é um valor no intervalo $[0, 1]$, indica quanto da variação na variável dependente pode ser explicada de acordo com a variação na variável independente, podendo ser compreendido como uma medida de ajuste de um modelo estimado em relação ao modelo observado.

Neste trabalho, o R^2 é utilizado para avaliar a relação entre os valores estimados e observados, tanto para os fluxos v_z nos arcos, quanto para os valores q_w representando as demandas da matriz OD.

Para cálculo do coeficiente R^2 , é utilizada a (22), em que x e y correspondem aos n estimados e observados (no caso de sensores) ou estimados e reais (no caso da matriz OD conhecida).

$$R^2 = \left(\frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)}} \right)^2 \quad (22)$$

Nesta dissertação, para o R^2 relacionado aos valores estimados e observados nos arcos, é utilizada a sigla R^2 Arcos, enquanto que para os valores estimados e reais da matriz OD, é utilizada a sigla R^2 ODM.

4.2 ALGORITMO DE ENXAME DE PARTÍCULAS

O Algoritmo de enxame de partículas, ou PSO (do inglês, *Particle Swarm Optimization*) é um algoritmo não determinístico bio-inspirado, proposto em Kennedy e Eberhart (1995). Um *survey* contendo diversas aplicações bem-sucedidas do PSO pode

ser encontrado em Poli et al. (2007).

O funcionamento básico do PSO foi inspirado no comportamento de aves em bando. Uma quantidade s de soluções (denominadas partículas) se movimentam em um espaço de busca multidimensional, sendo essa dimensão equivalente ao número de variáveis livres a serem otimizadas para o problema. Considerando um problema com n_p variáveis livres, cada partícula do PSO é representada por:

- Um vetor de n_p elementos, correspondente à posição atual de uma partícula;
- Um vetor de n_p elementos ($best_p$), correspondente à memória da melhor posição já encontrada de uma partícula;
- Um vetor de n_p elementos, correspondente à velocidade atual de uma partícula.

O pseudocódigo do PSO utilizado nesta dissertação é dado pelo Algoritmo 1. Resumidamente, o algoritmo PSO inicia com um enxame aleatório, com posições e velocidades aleatórias. Dentro do limite de iterações, as velocidades e posições são sucessivamente atualizadas a cada iteração, e então é analisado se cada partícula superou sua melhor posição anterior, bem como se surgiu uma nova solução que é mais adequada que a melhor solução anterior ($best_{global}$) do enxame. Quando um máximo de t_{ps0} iterações é alcançado, $best_{global}$ representa a melhor solução encontrada pelo enxame.

Algorithm 1 Pseudo-código do PSO

```

1: Inicializa o enxame com  $s$  partículas aleatórias
2: Avalia a adequação de cada partícula de acordo com a função fitness
3:  $best^{global}$  é a partícula de melhor adequação
4:  $t=1$ 
5: while  $t \leq t_{ps0}$  do
6:   for Cada partícula  $\mathbf{p}^t$  do
7:     Calcula velocidade da partícula  $\mathbf{p}^t$  (Equação 23)
8:     Desloca partícula  $\mathbf{p}^t$  (Equação 24)
9:     Reavalia a adequação da partícula  $\mathbf{p}^t$ 
10:    Atualiza a melhor posição da partícula (Melhor entre  $best_p$  e  $\mathbf{p}^t$ )
11:    Atualiza a melhor posição global (Melhor entre  $best_{global}$  e  $best^t$ )
12:  end for
13:   $t = t + 1$ 
14: end while

```

O cálculo da velocidade de uma partícula j é dado por (23), a fim de definir a trajetória no tempo discreto t . Para tal, são consideradas informações sobre o melhor desempenho individual, e o melhor desempenho de todo o enxame. A cada iteração, cada

uma das partículas tende a se aproximar destes pontos, a partir de parâmetros de memória cognitiva, direcionada para a melhor posição já alcançada pela própria partícula (c_1), memória social, direcionada para a melhor posição já encontrada pelo enxame (c_2), inércia de movimento (w), considerando velocidades na iteração anterior (v_j^t) e multiplicadores aleatórios (r_1 e r_2).

$$v_j^{t+1} = w \cdot v_j^t + (c_1 r_1) \cdot (best_j - p_j^t) + (c_2 r_2) \cdot (best_{global} - p_j^t) \quad (23)$$

Uma vez que a velocidade de uma partícula j foi definida por (23), a posição da partícula é atualizada por (24). A adequação da partícula é então recalculada por uma função *fitness*. Por fim, é verificado se esta nova posição é a melhor posição já alcançada por essa partícula, e se é a melhor posição já encontrada em todo o enxame.

$$\mathbf{p}^{t+1} = \mathbf{p}^t + \mathbf{v}^{t+1}. \quad (24)$$

O processo se repete até que o número máximo de iterações seja atingido. Para evitar que o PSO fique preso em mínimos locais e aumentar a diversidade do enxame, foi utilizado um procedimento de reinício das posições das partículas após determinado número de iterações, mantendo a memória das partículas. Além disso, foi utilizado uma variação linear no coeficiente inercial w . O artigo Zheng et al. (2003) apresenta aumento do peso inercial ao longo da execução, enquanto que Shi e Eberhart (1998) apresenta redução do peso. Neste trabalho, a decisão entre a redução ou aumento do valor inercial ao longo das iterações ocorre na seleção de parâmetros, que é descrita na Seção 4.2.1.

Nesta aplicação do PSO, cada partícula codifica uma estimativa da matriz OD candidata a solução, contendo n_p valores reais.

4.2.1 SELEÇÃO DE PARÂMETROS

Tanto GA quanto PSO, apesar de funcionamentos diferentes, têm a característica de possuírem conjuntos parâmetros que influenciam no desempenho e tempo de processamento. O método de seleção de parâmetros foi o iRace (*Iterated Race*, disponível como um pacote da linguagem de programação R).

Com o iRace, é possível escolher um conjunto de parâmetros a serem ajustados, podendo ser valores ordinais ou contínuos. O iRace executa uma quantidade configurável de execuções do algoritmo cujos parâmetros devem ser ajustados, selecionando ao final o

conjunto de parâmetros melhor ajustados, ou seja, que obtém o melhor valor de adequação. (LOPEZ-IBANEZ et al., 2016)

Os seguintes parâmetros do PSO foram selecionados através do iRace:

- Tamanho do enxame, s
- Número de iterações, t_{pso}
- Coeficiente cognitivo, $c1$
- Coeficiente social, $c2$
- Valor inicial do coeficiente de inércia, w_i
- Valor final do coeficiente de inércia, w_f
- Intervalo de iterações para que ocorra uma reinicialização aleatória da posição das partículas, $reset_t$

4.2.2 FUNÇÕES *FITNESS*

Este trabalho utiliza três diferentes funções *fitness* para PSO e GA, dadas por (25), (26) e (27). No Capítulo 5 são detalhados os resultados obtidos para a estimação da matriz OD com estas funções *fitness*.

$$\min_{q_w} \sum_{z \in Z} |v_z - \bar{v}_z| \quad (25)$$

$$\min_{q_w} \sum_{z \in Z} (v_z - \bar{v}_z)^2 \quad (26)$$

$$\min_{q_w} \sum_{z \in Z} GEH(z) \quad (27)$$

O termo $(v_i - \bar{v}_i)$ de (25) e (26) é a diferença entre o tráfego estimado e medido em cada arco do conjunto Z (que possui sensores), enquanto que em (27), o objetivo é minimizar o somatório de GEH de todos os arcos.

A partir da restrição de (12), uma solução é considerada inválida se algum dos fluxos q_w é menor que zero. Portanto, foi implementado ainda um procedimento de reparação da solução para valores negativos, que são alterados para zero na codificação

da solução, antes mesmo do procedimento de cálculo de adequação da solução através da função *fitness*.

4.3 ALGORITMO GENÉTICO

O algoritmo genético (GA) é um algoritmo não determinístico e bio-inspirado. No GA, existe um conjunto de soluções candidatas denominado população de tamanho *pop*, em que cada variável a ser otimizada age como um cromossomo. O algoritmo tem características inspiradas na seleção natural, hereditariedade, mutação e recombinação, tendo o comportamento influenciado por uma série de parâmetros de execução. O objetivo é encontrar melhores soluções ao longo das iterações, denominadas gerações.

O pseudocódigo do GA utilizado nesta dissertação é dado pelo Algoritmo 2. As soluções iniciais são geradas aleatoriamente, e em seguida avaliadas. O processo iterativo começa em seguida, no qual existe um processo de seleção de pais para geração de novas soluções (filhos), através do método da roleta.

Uma vez que $(pop/2)$ pares de pais são escolhidos, de acordo com o parâmetro de probabilidade de *crossover*, cada par de pais é re combinado em duas novas soluções a partir de um *crossover* de dois pontos, ou duplicados sem recombinação, fazendo com que a população passe de *pop* para $(2 \cdot pop)$. Os novos indivíduos passam então por um processo de mutação, de acordo com o parâmetro de probabilidade de mutação, em que alguns cromossomos sofrem modificações.

Após todo este processo, as novas soluções são avaliadas, e a seleção natural faz com que a população deixe de ser $(2 \cdot pop)$ e volte para (pop) . Findo este processo, ocorre uma nova geração começando novamente pela seleção de pais, até que o limite de gerações seja atingido.

Algorithm 2 Pseudo-código do GA

- 1: Inicializa a população com *pop* indivíduos aleatórios
 - 2: Avalia a adequação de cada indivíduo de acordo com a função *fitness*
 - 3: $t=1$
 - 4: **while** $t \leq t_{ga}$ **do**
 - 5: Seleção de pais
 - 6: Realiza *crossover*. População passa a ser $(2 \cdot pop)$
 - 7: Realiza mutação dos novos indivíduos
 - 8: Avalia a adequação dos novos indivíduos
 - 9: Somente os (pop) indivíduos de melhor adequação sobrevivem
 - 10: $t = t + 1$
 - 11: **end while**
-

O método de seleção de pais para *crossover* utilizado é o da roleta, com objetivo de maximização do inverso do erro, uma vez que o método da roleta é de maximização. Após o *crossover*, que é realizado utilizando dois pontos como pivôs, existe uma probabilidade de mutação P_m para cada alelo, e também um parâmetro proposto $varMut$ que define um percentual máximo de mutação nos alelos. Os indivíduos que continuam na próxima geração são escolhidos pela adequação, independente da iteração em que foram gerados, sendo então uma seleção de sobreviventes baseada em adequação.

Nesta aplicação do GA, cada indivíduo corresponde a uma matriz OD candidata a solução, sendo um vetor de n_p valores reais, de maneira similar ao PSO. Por ocasião do *crossover*, temporariamente existem ($2 \cdot pop$) indivíduos possuindo n_p valores reais cada um.

O método de seleção de parâmetros para o GA utiliza também o pacote iRace. Os seguintes parâmetros do GA são selecionados:

- população, pop
- número de gerações, t_{ga}
- probabilidade de *crossover*, P_c
- probabilidade de mutação de cromossomo, P_m
- valor percentual máximo de variação em um cromossomo por ocorrência de mutação, $vMut$

As funções *fitness* avaliadas pelo GA são as mesmas do PSO, dadas por (25), (26) e (27). Ocorre somente uma diferença em questão de implementação: Pelo método de seleção de pais ser o método de roleta, e tal método visar a maximização, o valor inverso de (25), (26) e (27) é considerado como valor de *fitness* das soluções.

4.4 MÉTODO DOS MÍNIMOS QUADRADOS

A estimação da matriz OD pode ser representada por um sistema de equações lineares (28), cuja solução de mínimos quadrados (sigla LS no Capítulo 5) é dada por (29) usando a pseudo-inversa $(A^T A)^{-1}$.

Para tanto, é necessário definir as Equações (28) e (29):

- Um grafo da rede viária com n_s sensores (arcos do conjunto Z);
- Uma matriz OD com n_p elementos - exceto a diagonal (pares OD do conjunto W);
- Uma matriz A ($n_s \times n_p$) representando a *assignment matrix*, contendo probabilidades de uso de cada arco do conjunto Z para cada par OD do conjunto W ;
- Um vetor coluna x de dimensão n_p contendo a contagem de veículos de cada um dos n_p pares OD (incógnitas do problema);
- Um vetor coluna b de dimensão n_s com as contagens de veículos de cada um dos n_s sensores.

$$Ax = b \quad (28)$$

$$x = (A^T A)^{-1} A^T b \quad (29)$$

Com a aplicação do mínimos quadrados, alguns fluxos q_w (fluxos por pares OD) podem apresentar valores negativos, o que, em via de regra, faz com que soluções através desta técnica sejam soluções invalidas. Devido a esse fato, para calcular corretamente os valores das métricas de desempenho, neste trabalho, os valores negativos são transformados em zero antes da avaliação de acordo com as métricas, de acordo com (12).

4.5 PROGRAMAÇÃO LINEAR INTEIRA MISTA

Para resolução do modelo MILP, foi utilizado o software GUSEK 0.2.24. Abaixo são descritas as variáveis, a função *fitness* e as restrições do modelo.

A Tabela 17 mostra os conjuntos de variáveis utilizados no modelo. A solução é dada pelas n_p variáveis de decisão correspondentes ao fluxo de veículos por par OD.

Tabela 17: Variáveis do modelo de programação linear inteira mista

Variável	Tipo	Símbolo	Quantidade
Fluxo estimado por par OD w	$\mathbb{R} \geq 0$	q_w	n_p pares OD
Fluxo estimado no arco z	$\mathbb{R} \geq 0$	\bar{v}_z	n_s sensores
Folga superior no arco z	$\mathbb{R} \geq 0$	F_{z+}	n_s sensores
Folga inferior no arco z	$\mathbb{R} \geq 0$	F_{z-}	n_s sensores
Total			$(n_p + 3 \cdot n_s)$

A variável \bar{v}_z representa o conjunto de valores estimados de fluxo de veículos nos arcos, que devem ser comparados com os valores medidos v_z , dos sensores.

As variáveis de folga F_{z+} e F_{z-} são necessárias para obter soluções factíveis, uma vez que o ajuste do fluxo de veículos em um arco pode desajustar o fluxo em outra.

A função *fitness* é minimizar o somatório das folgas nos sensores, conforme (30).

$$\min_{q_w} \sum_{z \in Z} (F_{z+} + F_{z-}) \quad (30)$$

As possibilidades de formulação de função *fitness* do modelo MILP são mais restritas, uma vez que determinadas operações matemáticas não podem ser realizadas porque causariam não linearidades, o que impede a resolução do problema através desta técnica. Portanto, não existem tantas opções de função *fitness* para o modelo MILP quando comparado a soluções com os algoritmos PSO e GA, e somente (30) foi utilizada, sendo esta similar a (25) testada no PSO e GA.

Na Tabela 18 é apresentado um resumo sobre as restrições utilizadas no modelo. Basicamente, existe a necessidade de igualar o valor estimado de fluxos de veículos ao valor observado, e então, as variáveis de fluxo por par OD são multiplicadas pelos valores P_{wz} de probabilidade de uso de arco por par OD para se chegar aos valores \bar{v}_z , de fluxo estimado em cada arco.

Tabela 18: Restrições do modelo de programação linear inteira mista

Restrição	Motivação	Equação	Quantidade
Estimação de fluxo em arcos	(11)	(31)	n_s sensores
Adequação do tráfego em arcos	(30)	(32)	n_s sensores
Não negatividade de fluxos	(12)	(33)	n_p pares OD
Não negatividade de folga inferior	(30)	(34)	n_s sensores
Não negatividade de folga superior	(30)	(35)	n_s sensores
Total			$(n_p + 4 \cdot n_s)$

A restrição (31) determina o fluxo estimado nos arcos que possuem sensores, a partir dos valores de probabilidade de uso de arco por par OD, e as variáveis livres q_w , que estão sendo estimadas. É necessário uma restrição deste tipo para cada sensor considerado.

$$\bar{v}_z = \sum_{w \in W} q_w \cdot p_{wz} \quad (31)$$

A restrição (32), por sua vez, cria a obrigação de igualar os fluxos estimados e medidos nos sensores, porém utilizando das variáveis de folga F_{z+} para quando o valor estimado é inferior ao medido, e F_{z-} para quando o valor estimado é superior ao valor medido. Isso, em conjunto com (31), e a função *fitness* (30), faz com que, de acordo com as probabilidades de uso de arco dadas, o modelo encontre os valores mais próximos quanto possível entre \bar{v}_z e v_z , ao ajustar os valores de q_w .

$$v_z = (\bar{v}_z + F_{z+} - F_{z-}) \quad (32)$$

A restrição (33) é baseada no princípio de não negatividade de fluxo dos pares OD, anteriormente especificada em (12), sendo fundamental para a validade da solução encontrada. É necessária uma restrição deste tipo para cada par OD.

$$q_w \geq 0 \quad (33)$$

Em (34) e (35), é indicado que as variáveis de folga estão restritas a valores não negativos. A situação ideal para a estimativa de fluxo de veículos em um sensor é de que o estimado é igual o medido, e então, ambas as variáveis assumem valores zero. É necessária uma restrição (34) para cada um dos sensores, e o mesmo vale para (35).

$$F_{z+} \geq 0 \quad (34)$$

$$F_{z-} \geq 0 \quad (35)$$

4.6 T-FLOW FUZZY

O T-Flow Fuzzy foi inicialmente proposto na dissertação de Rosinowski (1994), e se utiliza da teoria de conjuntos *fuzzy* para considerar incertezas nas medições de sensores. Está incorporado no software de simulação de tráfego PTV VISUM (AG, 2018) como uma ferramenta em que o uso de uma matriz OD semente é obrigatório, sendo o foco principalmente na atualização de uma matriz OD.

De maneira similar à otimização de dois níveis, o *T-Flow Fuzzy* foi utilizado neste

trabalho em um procedimento iterativo entre estimação da matriz OD e de *equilíbrio de usuário*, em uma abordagem similar a encontrada em Yousefikia et al. (2013). A quantidade de iterações t é configurável, bem como as folgas para cada um dos arcos. O pseudocódigo pode ser observado no Algoritmo 3.

Algorithm 3 Pseudo-código da aplicação do *T-Flow Fuzzy*

- 1: Matriz OD resposta é inicialmente a matriz OD semente
 - 2: Procedimento de *equilíbrio de usuário* para definir a *assignment matrix*
 - 3: $t=1$
 - 4: **while** $t \leq 5$ **do**
 - 5: Atualização da matriz OD resposta através do *T-Flow Fuzzy*
 - 6: Procedimento de *equilíbrio de usuário* para redefinir a *assignment matrix*
 - 7: Reduz os desvios máximos aceitos nos sensores
 - 8: $t = t + 1$
 - 9: **end while**
-

Para a utilização do cálculo de *equilíbrio do usuário*, foram utilizados dados das vias, tais como capacidade e velocidade máxima. Existem uma série de parâmetros a serem ajustados para que o procedimento de *equilíbrio de usuário* embutido ao software PTV VISUM (AG, 2018) obtenha resultados mais precisos. O número máximo de iterações foi fixado em 5 pois valores superiores a este ocasionavam falhas de execução.

4.7 RESUMO DAS ABORDAGENS DE ESTIMAÇÃO DA MATRIZ OD UTILIZADAS

Na Tabela 19, são exemplificadas as características básicas das técnicas testadas neste trabalho. Com exceção do *T-Flow Fuzzy*, que é utilizado somente para a rede de Curitiba, as demais técnicas são utilizadas em ambas as redes.

Tabela 19: Características das técnicas de estimação da matriz OD aplicadas

	PSO	GA	LS	MILP	T-Flow Fuzzy
Plataforma	Java	Java	Java	GUSEK/GLPK	PTV Visum
Determinismo	Não	Não	Sim	Sim	Sim
Parâmetros de execução	Sim	Sim	Não	Não	Sim
Tempo de execução	Configurável	Configurável	Fixo	Fixo	Configurável
Liberdade de definição de objetivos	Alta	Alta	Baixa	Média	Baixa

Conforme Tabela 19, o T-Flow Fuzzy é embutido no simulador PTV VISUM, o modelo MILP necessita do kit de programação linear GLPK para execução, e as demais

técnicas podem ser codificadas em qualquer linguagem de programação, tendo sido codificadas em Java 8 neste trabalho. Com exceção do PSO e do GA, todas as técnicas utilizadas apresentam resultados determinísticos, ou seja: Para um mesmo conjunto de dados de entrada, o resultado será sempre o mesmo, então, uma execução destas técnicas determinísticas é suficiente para cada cenário.

As técnicas MILP e LS não necessitam de parâmetros de execução, tendo tempo de execução fixo. Para GA e PSO, existem os parâmetros clássicos que podem ser selecionados, impactando no desempenho e no tempo de execução, necessitando então de um processo de seleção de parâmetros para definir essas configurações. A estimação da matriz OD pelo VISUM possui também algumas configurações que podem ser realizadas, como número de iterações e desvio máximo nos arcos, porém por ser integrado ao VISUM, sua operabilidade não se mostrou muito prática, já que dependendo da configuração de parâmetros, o VISUM gera uma mensagem de erro e não resolve o problema.

Quanto à liberdade de definição de objetivos das técnicas, o LS e o VISUM são os mais restritos. LS por ser uma operação puramente matemática, e o VISUM por ser uma plataforma fechada. Para o MILP, é possível mudar a função *fitness*, caso o modelo seja reconstruído com este objetivo, mas é preciso atenção para não criar situações de não linearidade. Para o GA e PSO, a formulação da função *fitness* é bastante livre, embora possam ocorrer penalidades de desempenho a depender da função *fitness* utilizada. Dadas estas características, é possível uma maior exploração na questão de escolha de função *fitness* somente para GA e PSO.

4.8 MAP MATCHING

O artigo Zheng (2015) apresenta um panorama geral acerca da área de pesquisa de *trajectory data mining*, e cita que existem algoritmos geométricos, topológicos, probabilísticos e de técnicas avançadas para realizar a tarefa de *Map matching*. Ainda segundo Greenfeld (2002), as abordagens mais intuitivas para o *map matching* são relacionar os pontos de GPS aos nós ou arcos mais próximos, mas tais abordagens podem ocasionar erros, devido a ruídos inerentes aos dados. Para tanto, existem métodos que consideram conectividade, adjacência ou proximidade da rede, chamados de métodos de *map matching* topológicos.

O *map matching* é utilizado neste trabalho para caracterizar, a partir de dados georreferenciados de viagens de táxis, a rota utilizada dentro da malha viária.

Foram utilizadas duas técnicas de *map matching*. Uma está baseada na identificação de nós mais próximos, considerando um raio limite rm . Essa técnica é a mais simples, escolhendo entre os nós existentes dentro do raio limite rm , qual o mais próximo, e que deve ser adicionado na sequência de nós da rota.

A segunda baseada na identificação de arcos mais próximas, com uma distância máxima do arco dm , mostrada na Fig. 18.

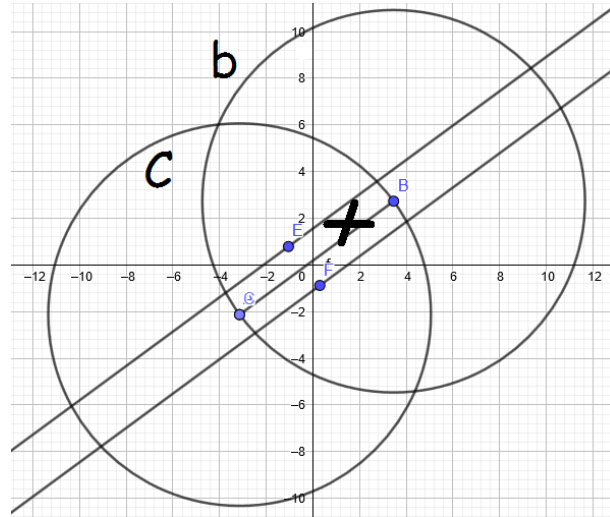


Figura 18: *Map matching* de um dado georreferenciado (x na cor preta) para um arco BC.

Conforme a Fig. 18, na segunda técnica de *map matching*, para que um ponto (x na cor preta) seja considerado como pertencente a um arco de extremidades B e C, é necessário que:

- Esteja no interior das duas circunferências b e c, que possuem centros nos pontos B e C, e raios de tamanho BC;
- Esteja entre as retas paralelas que cruzam os pontos E e F, sendo estas retas paralelas à reta BC, com uma distância de dm metros da reta BC.

Uma vez que um ponto é identificado como pertencente ao arco BC, é calculado se está mais próximo de B ou C. O nó mais próximo é então adicionado à sequência de nós da rota.

Duas métricas são utilizadas neste trabalho para a qualidade do *map matching*:

- Quantidade média de nós identificadas por trajetória;

- Grau de conectividade médio das trajetórias.

Desta forma, os objetivos são identificar uma quantidade grande de nós e arcos por trajetória (para que haja maiores contagens nos arcos), ao mesmo tempo em que as trajetórias identificadas formem um única componente conexa (sendo este um indicativo de boa identificação da rota na malha viária, sem trechos perdidos).

Por exemplo, para um grau de conectividade de 80%, considerando uma trajetória contendo 101 nós e então, um número máximo de 100 arcos, verifica-se que, ao comparar a sequência de nós identificados com a rede viária, 80 pares de nós (sendo um nó n e um nó $n + 1$, dentro dos 101 nós) da sequência possuem um arco correspondente na rede viária, de início em n , e término em $n + 1$. Nos 20 pares restantes, um nó é seguido por outro na sequência identificada, mas sem que exista um arco correspondente entre estes nós na rede viária, indicando um trecho perdido.

4.9 CONSIDERAÇÕES DO CAPÍTULO

Este Capítulo apresentou a metodologia e sequência de atividades do trabalho para obtenção dos resultados apresentados no Capítulo 5. Inicialmente foram definidas as medidas de desempenho para avaliar a qualidade da estimação da matriz OD na Seção 4.1, e assim, possibilitar a comparação entre diferentes técnicas de estimação da matriz OD, bem como analisar consequências de variações no problema ao utilizar uma mesma técnica. Foram definidas quatro métricas para matriz OD, e três métricas para arcos.

A partir da Seção 4.2 foram apresentadas as técnicas de estimação da matriz OD testadas neste trabalho, abrangendo tanto seu funcionamento geral, como detalhes de implementação. Por fim, foi detalhada a metodologia utilizada para *map matching* na Seção 4.8.

5 RESULTADOS

Este capítulo é dividido em duas seções. A Seção 5.1, apresenta os resultados dos experimentos realizados com a rede viária e dados de viagens de táxis da cidade do Porto. A Seção 5.2 mostra os resultados dos experimentos realizados com a rede viária da cidade de Curitiba.

Os estudos de caso foram desenvolvidos e analisados através de duas diferentes metodologias. Para o estudo de caso da cidade do Porto, existem resultados de *map matching*, agrupamento espacial e de definição de critérios de escolha dos arcos da rede viária a serem observados. A rede de Curitiba, por outro lado, foi importada de um modelo de simulação de uma região estudada pelo IPPUC.

Na Tabela 21, são mostradas as características das redes. De forma geral, são redes bastante distintas. A rede viária da cidade do Porto possui um grande número de nós e arcos (153840 e 221446 respectivamente) distribuídos por uma grande área, enquanto que a rede viária de Curitiba possui 191 nós e 594 arcos. As características do estudo de caso de Curitiba já estão pré-definidas de acordo com dados recebidos do IPPUC, enquanto que as do estudo de caso da cidade do Porto necessitam de resultados adicionais, utilizando-se do OSM e de uma base de dados de viagens de táxi.

Em relação aos dados sensores, são utilizados dados de sensores reais de Curitiba. Para a cidade do Porto, os sensores são virtuais, ou seja, identificam o fluxo de veículos no *dataset* UCI (2015) através de *map matching* (Seção 5.1.2), e utilizando diferentes critérios de escolha de sensores (Seção 5.1.5).

A rede de Curitiba tem 29 pontos de OD definidos nas fronteiras da rede, enquanto que a rede da cidade do Porto tem pontos de OD definidos como áreas no mapa, através de agrupamento espacial, na Seção 5.1.1.

De forma semelhante, para a *assignment matrix* utilizada nos experimentos com a rede de Curitiba, foi utilizada a informação utilizada pelo IPPUC, a fim de manter a mesma *assignment matrix* para todos os casos. Para a cidade do Porto, a *assignment*

Tabela 21: Detalhes dos estudos de casos da cidade do Porto e de Curitiba

	Curitiba (Brasil)	Porto (Portugal)
Fontes dos dados	IPPUC	OSM e dataset UCI (2015)
<i>Assignment matrix</i>	Equilíbrio de usuário (VISUM)	Agrupamento espacial (Seção 5.1.1) e <i>map matching</i> (Seção 5.1.2)
Nós	191	153840
Arcos	594	221446
Sensores	71	Agrupamento espacial (Seção 5.1.1), <i>map matching</i> (Seção 5.1.2) e NSLP (Seção 5.1.5)
Pontos OD	29, nas fronteiras da região	Agrupamento espacial (Seção 5.1.1)
Total de pares OD	841	Agrupamento espacial (Seção 5.1.1)
Pares OD com fluxo superior a zero	183	Agrupamento espacial (Seção 5.1.1)
Valores não zerados da <i>assignment matrix</i>	574	Agrupamento espacial (Seção 5.1.1)
Latitude	[-49,2570526; -49,2756589]	[41,3142094; 41,0374522]
Longitude	[-25,4358037; -25,4375282]	[-8,3749317; -8,7314937]

matrix é gerada a partir do cruzamento de informações de *clusters* com o *map matching*, gerando relações de probabilidade de uso de arcos para cada par OD.

5.1 ESTUDO DE CASO DA CIDADE DO PORTO

Para o estudo de caso da cidade do Porto, são utilizados dados de corridas de táxis, e dados topológicos da cidade. A topologia da rede é obtida da base geográfica aberta OSM (OpenStreetMap contributors, 2018), enquanto que o *dataset* aberto de corridas de táxi (UCI, 2015), contém informações de mobilidade da cidade do Porto em Portugal.

As principais características são:

- Viagens de 442 táxi coletadas entre 07/01/2013 e 06/30/2014;
- Posicionamento dos táxis em intervalos de 15 segundos;
- 1.710.671 registros de corridas;
- Dados em um arquivo no formato *csv* de 1,8 GB.

Os oito atributos do *dataset* são apresentados na Tabela 22. Os atributos de POLYLINE, TIMESTAMP e DATATYPE identificam as rotas no espaço-tempo e se ocorrem em dias de semana ou feriados. Os demais atributos (TRIP ID, CALL TYPE,

Tabela 22: Atributos dos dados de corridas de táxi da cidade do Porto

Atributo	Tipo	Descrição
TRIP ID	String	Identificador da corrida
CALL TYPE	Caractere	A (corrida iniciada através da central de táxis) B (corrida iniciada em ponto de táxi) C (Outros casos)
ORIGIN CALL	Inteiro	ID do ponto de táxis, se CALL TYPE = B
TAXI ID	Inteiro	ID do motorista
TIMESTAMP	Inteiro	Início da viagem (segundos).
DATATYPE	Caractere	B (Feriado ou outro dia especial) C (dia anterior a um dia tipo B) A (demais dias)
MISSING DATA	Booleano	Indicação de que a corrida apresenta todo a trajetória registrada
POLYLINE	String	Coordenadas GPS a cada 15 segundos para a corrida, no formato WGS84.

ORIGIN CALL e TAXI ID) são informações administrativas ou privadas não utilizadas neste trabalho.

A Seção 5.1.1 apresenta resultados do agrupamento espacial, utilizado para definir os pares OD, e a Seção 5.1.2 apresenta resultados de *map matching* a fim de definir a *assignment matrix* da rede. A Seção 5.1.3 mostra a seleção de parâmetros dos algoritmos GA e PSO. A Seção 5.1.4 apresenta experimentos com GA e PSO com diferentes funções *fitness*. A Seção 5.1.5 apresenta experimentos com diferentes critérios de escolha de localização sensores virtuais, enquanto que a Seção 5.1.6 apresenta os resultados de experimentos utilizando as diferentes técnicas de estimação da matriz OD. Por fim, a Seção 5.1.7 contém uma avaliação de como a quantidade de sensores afeta o desempenho destas técnicas.

Por se tratar de uma grande rede viária, há um interesse em analisar o tráfego OD entre as regiões da cidade, e não entre locais específicos como nós e arcos. Origens e destinos de chamadas de táxi gravados no conjunto de dados são agrupados por posição geográfica, reduzindo, consequentemente, o número de pares OD. Esse processo é realizado se utilizando de um algoritmo de agrupamento espacial (MATIAS et al., 2016).

Para realizar o agrupamento espacial, um novo conjunto de dados precisa ser formado a partir dos dados originais. Este novo conjunto de dados contém apenas coordenadas GPS (latitude e longitude). Para cada instância (corrida) do conjunto de dados original, uma instância (localização do início ou fim da corrida) é gerada para origem e outra para destino da corrida, sendo que este novo conjunto de dados possui somente dois atributos: latitude e longitude. Devido a limitações de memória na utilização do software

Weka, o agrupamento com todos os pontos no conjunto de dados não pôde ser executado, e então, o conjunto de dados foi reduzido para 10% do tamanho original. Outra adaptação foi feita como um pós-processamento, com a exclusão de pequenos *clusters* (quando menores que $0,25km^2$), a fim de *tornar a matriz OD menos esparsa* sem perda considerável de informações, unindo tais *clusters* pequenos ao *clusters* vizinho com densidade mais próxima possível.

Outra questão abordada é a quantidade ideal de *clusters*, que neste estudo de caso serão considerados como regiões de origem e destino. Para isso, o método iterativo *CascadeSimpleKMeans* do WEKA foi usado, que identifica um número ideal de *clusters*. O método realiza uma série de testes utilizando o algoritmo de agrupamento *k-means* de Lloyd (1982), e seleciona o melhor valor de *k*, de acordo com o critério definido por Calinski e Harabasz (1974), que basicamente se utiliza de um teste F a partir de uma formula que faz referência ao número de instâncias, número de *clusters*, e soma de erros quadrados intra e inter *clusters*. Os resultados de tal processo são mostrados na Seção 5.1.1, sendo lá definidos os pontos OD considerados.

5.1.1 AGRUPAMENTO ESPACIAL

Através do método *CascadeSimpleKMeans*, a quantidade ideal de *clusters* foi identificada como $k = 14$ *clusters*. Em seguida, o método de agrupamento baseado em densidade *MakeDensityBasedClusterer* do WEKA foi definido com $k = 17$ *clusters*, para que o número de *clusters* fosse então reduzido para $k = 13$, depois de descartar os três menores *clusters* em área, e também o maior *cluster* em área, mas com baixa densidade. Os centroides dos *clusters* numerados *c0* a *c12* podem ser vistos na Fig. 20.

A matriz OD resposta utilizada, bem como a *assignment matrix* correspondem a um período de 1h, para uma agregação de corridas de táxis de diferentes dias, entre 8 e 9h, para dias úteis. Tal horário foi definido por ser um horário de pico, ao mesmo tempo em que o período de 1h é usual para estudos envolvendo modelagem de tráfego.

A Tabela 23 resume algumas informações dos *clusters* obtidos, tais como a área total do *cluster*, e o número de corridas que se iniciaram e terminaram em cada *cluster*. É interessante notar que pequenas áreas podem originar ou serem destinos de muitas viagens, assim como grandes áreas podem concentrar poucas viagens.

Uma vez que os *clusters* estão definidos, as corridas de táxi podem ser agregadas de acordo com as localizações de origem e destino, dentre um total de 169 pares OD. Na

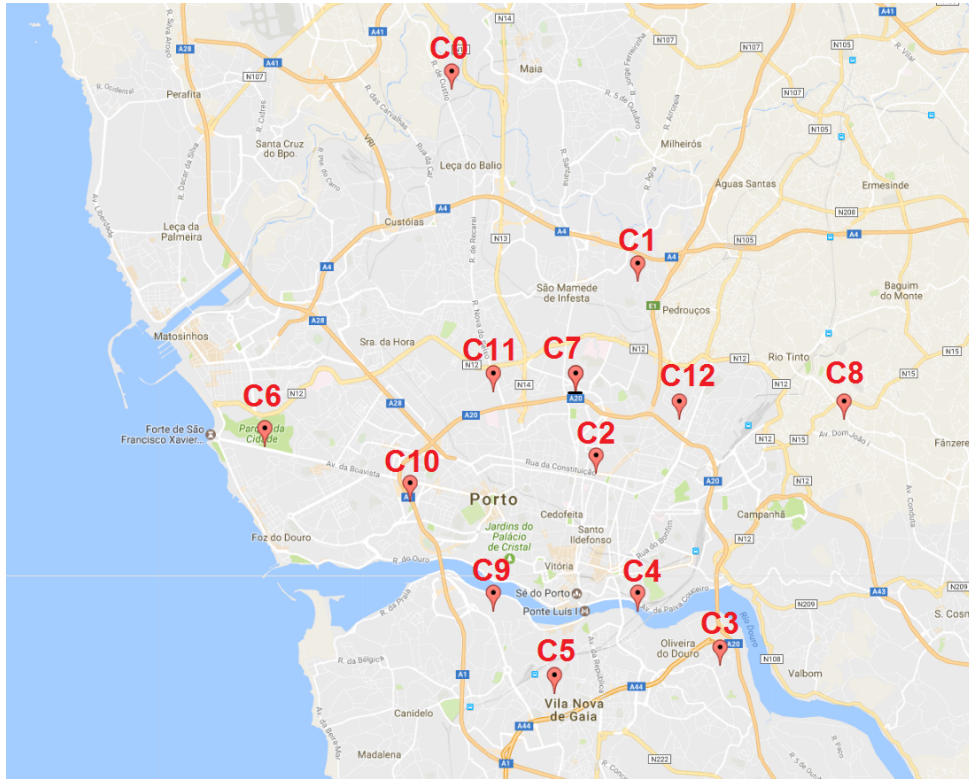


Figura 20: Agrupamento espacial por densidade (13 clusters ou áreas).

Seção 5.1.2, as rotas utilizadas são mapeadas, a fim de obter as rotas utilizadas para cada um dos pares OD.

5.1.2 MAP MATCHING

Neste trabalho, a necessidade de se utilizar o um algoritmo de *map matching* surgiu a partir do estudo da rede viária da cidade do Porto. Os dados disponibilizados em UCI (2015) têm a característica de *spatial trajectory data*, ou seja, são pontos geográficos sequenciais, com informação de horário em cada um deles. Para devida utilização dos dados contidos no *dataset*, é necessário realizar um *map matching*, para compreender por que vias os táxis efetivamente transitaram, e então computar fluxo e probabilidades de uso de segmentos de via.

O conjunto de dados contém o posicionamento dos táxi para cada viagem a cada 15 segundos contido no atributo POLYLINE. Com base nessa informação parcial, uma rota pode ser construída para cada corrida através de *map matching*. No entanto, um algoritmo de caminho mínimo (considerando distância em metros) foi usado para preencher lacunas, já que o intervalo de 15 segundos não é suficiente para a reconstrução completa dos trajetos, e alguns pontos possuem ruídos que impossibilitam a identificação de nós

Tabela 23: Informações sobre a clusterização espacial na cidade do Porto

Cluster	Área (km^2)	# de viagens iniciadas	# de viagens finalizadas
c0	47	2258	12146
c1	12	42494	25530
c2	3	60532	108054
c3	43	30738	41511
c4	6	115528	83467
c5	15	375361	245454
c6	36	101930	77865
c7	3	28781	42170
c8	120	4649	12987
c9	9	154907	217807
c10	17	154685	209749
c11	7	60114	69462
c12	7	87707	72484

ou arcos ao ponto. A rota no mapa da cidade é inferida através de um grafo obtido a partir de uma base aberta de dados cartográficos de OpenStreetMap contributors (2018). Ao final do processo, a informação da rota é composta por nós e arcos em um gráfico direcionado pelos quais os táxis passam em cada uma das corridas. Uma vez que as rotas foram identificadas, é possível associar contagens de tráfego a cada arco como sensores de tráfego virtual (HERRERA et al., 2010).

Uma vez que o agrupamento espacial (Seção 5.1.1) definiu os pontos OD, e o *map matching* (Seção 5.1.2) identificou por quais arcos cada corrida de táxi do *dataset* passou utilizando do conhecimento do intervalo de 15 segundos para saber em que momento o táxi passou por cada arco, é possível combinar as duas informações para obter a *assignment matrix* do problema, bem como a matriz OD de viagens de táxis, que é utilizada como matriz OD resposta, sendo estes dados fundamentais para realização dos demais experimentos para o estudo de caso da cidade do Porto.

A base de dados geográficos de OpenStreetMap contributors (2018) é aberta, enquanto os dados de tráfego podem ser obtidos por sensores, e de certa forma, a partir da combinação de dados da rede viária e de veículos sonda, como táxis, a partir do *Map Matching*. Para a rede da cidade do Porto, foi utilizada a combinação de dados de veículos sonda e geográficos para encontrar contagens de veículos em arcos, criando dados semelhantes aos que seriam obtidos por sensores, em menor escala.

A Fig. 21, ilustra o resultado do *map matching*, para uma das corridas de táxi do *dataset* da cidade do Porto. Na parte da esquerda, tem-se a sequência de dados espaciais (*polyline*) da corrida de táxi sobre o mapa da cidade. Na parte da direita, após o *map*

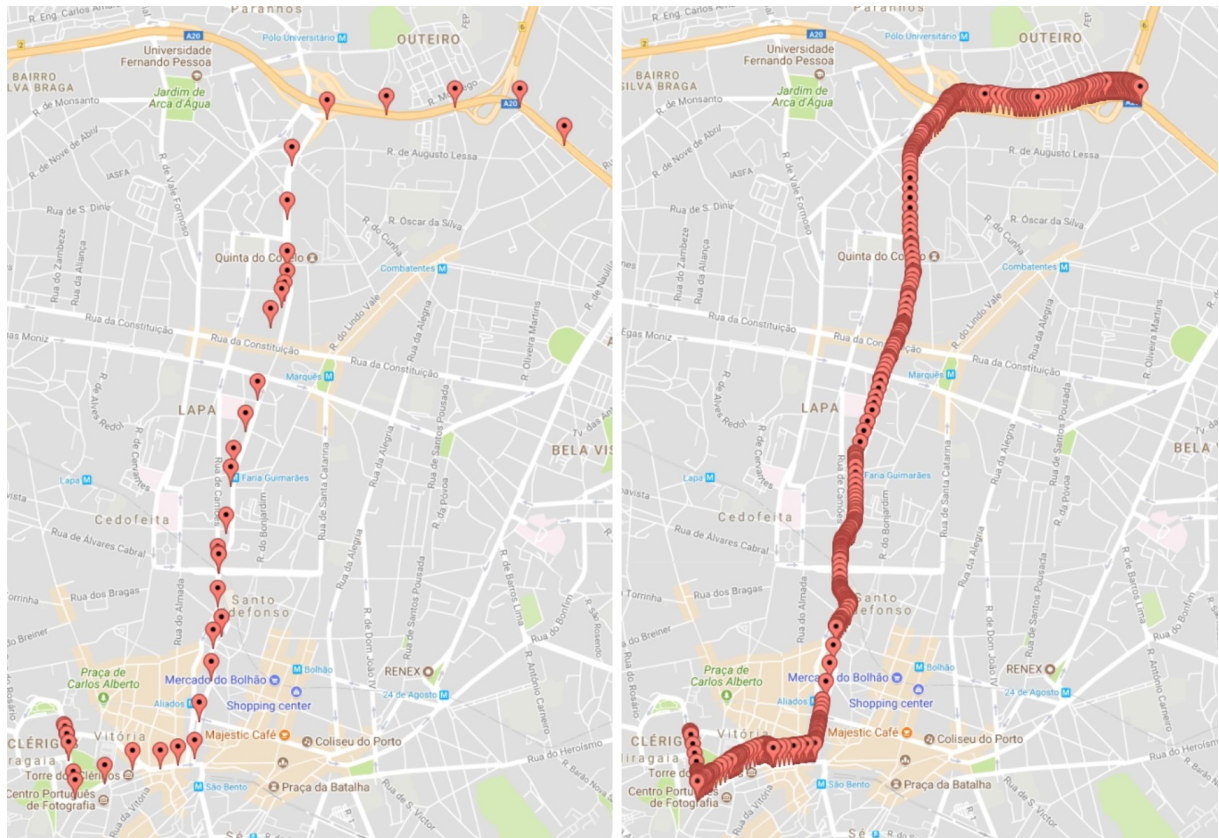


Figura 21: *Map matching* (à direita) a partir de dados georreferenciados (à esquerda) de uma corrida de táxi na cidade do Porto.

matching, uma sequência de nós e arcos da rede viária pelos quais o táxi passou foi identificada mostrada no mapa. No canto superior direito é possível observar o último ponto da *polyline* não identificado como nó ou arco, de tal forma que o trajeto identificado é um pouco mais curto que o original.

Existe certa incerteza na trajetória, como os intervalos de tempo entre localizações são de 15 segundos. Esse é um problema que é discutido em Zheng (2015) sendo que podem ser utilizados dados de outras trajetórias para completar uma trajetória com menos informações, por exemplo. Tal incerteza pode ser inserida nos dados para proteger a privacidade dos usuários, ou reduzir a quantidade de dados transmitidos.

Neste trabalho, para completar as falhas (falta de informação) encontradas nas trajetórias, foi utilizado um algoritmo de caminho mínimo, considerando distância em metros e cada corrida de táxi individualmente. Uma abordagem semelhante para completar caminhos é também utilizada por Biem et al. (2010), que utiliza dados de localização em intervalos de 60 segundos.

Com dados de localização obtidos a cada 15 segundos e considerando uma velo-

cidade máxima de 60 km/h, um veículo se desloca até 250 metros a cada novo ponto de localização. Considerando que o tamanho médio de um arco da rede viária da cidade do Porto é de 30, 60 metros, tem-se que um veículo pode percorrer até 8 arcos entre um ponto e outro na sequência de localizações GPS da trajetória. Para que seja possível completar as lacunas nas trajetórias, ao mesmo tempo em que se evite reconstruir trajetos que não ocorreram, Foi utilizado uma profundidade máxima de busca no grafo de 18 arcos, para busca por caminho mínimo em metros a fim de completar os trajetos.

Foram utilizados métodos de identificação por nós e por arcos, utilizando diferentes distâncias limites (veja Seção 4.8), a fim de encontrar um equilíbrio entre conectividade da sequência de nós e quantidade média de nós identificados por corrida de táxi.

Tabela 24: Resultados do *map matching* para o estudo de caso da cidade do Porto

Distância Limite (m)	Por nós		Por arcos	
	Conectividade	Média de nós identificados	Conectividade	Média de nós identificados
8	91,71%	48,80	82,67%	12,31
9	91,97%	51,80	82,48%	13,04
10	92,83%	54,49	82,12%	12,73
11	92,72%	56,53	82,13%	13,49
12	92,05%	58,42	82,30%	13,95
13	92,27%	61,28	83,73%	14,18
14	92,06%	62,18	83,64%	14,42
15	91,80%	63,53	84,01%	14,52
16	91,80%	64,64	83,97%	15,09
17	91,50%	65,17	84,65%	15,08
18	91,64%	66,30	84,35%	15,05
19	91,40%	67,42	84,46%	15,07
20	91,54%	69,43	84,53%	14,83

Na Tabela 24 são mostrados os resultados do *map matching*. Para cada método (nós ou arcos) tem-se a média de nós identificados e conectividade. Quanto maior a distância limite considerada, maior o tamanho da rota identificada, para ambas as técnicas. No entanto, o melhor valor de conectividade de rota foi 92,83% para nós e 84,53 para arcos, encontrado com 10 e 20 m de distância máxima, respectivamente

Buscando um equilíbrio entre os critérios escolhidos, a técnica de *map matching* definida foi através de nós, com uma distância máxima de 14 metros, com uma conectividade de 92,06%, e média de 62,18 nós por rota ou corrida de táxi.

Uma possível explicação para o método utilizando nós ter se saído superior ao método por arcos pode ser o fato de que aparelhos modernos de GPS possuem uma

precisão entre poucos metros e polegadas, o que diminui consideravelmente o ruído encontrado nos dados de trajetória. Outro ponto é que as distâncias médias e medianas entre nós do mapa importado do *OpenStreetMap* são relativamente curtas, o que facilita a identificação das rotas através dos nós.

Na Seção 5.1.1 foram definidos os pares OD para o estudo de caso da cidade do Porto, bem como foi calculada a matriz OD real. A informação das áreas de origem e destino foi então combinada com as rotas obtidas por *map matching*, obtendo as probabilidades de uso de arco por par OD (*assignment matrix*). Com isso, os dados necessários para realizar um procedimento de estimação da matriz OD estão disponíveis, permitindo os experimentos das seções seguintes do estudo de caso da cidade do Porto.

5.1.3 CONFIGURAÇÃO DOS ALGORITMOS GA E PSO

Nesta Seção são apresentados os resultados da seleção de parâmetros realizada pelo iRace. Foram utilizadas quantidades máximas de partículas (parâmetro s) do PSO ou indivíduos do GA de acordo com a limitação de 16Gb de memória RAM disponível, considerando execução em 8 instâncias paralelas. Para a cidade do Porto, o número de sensores utilizado foi de 1000, sendo este mesmo valor utilizado posteriormente nos experimentos das Seções 5.1.4, 5.1.5 e 5.1.6. A métrica de desempenho utilizada é (27), com critério de escolha de sensores de utilizar como sensores os arcos com mais rotas.

Na execução do iRace, foi utilizado nível de confiança de 95% para o F-Test, variações nos parâmetros até a segunda casa decimal e um limite de 1000 experimentos, sendo necessárias cerca de 12 horas de processamento para cada um dos algoritmos. A leitura e escrita de dados em disco demonstrou ser um fator limitante de desempenho, de forma que optou-se por emular disco em memória. As seguintes configurações de hardware e software foram utilizadas para ambos os algoritmos:

- Processador AMD Ryzen 7 1700 3 Ghz de 8 núcleos
- Memória 16Gb DDR4 2966Mhz
- Sistema operacional Windows 10 64 bits
- Execução através de *Bash on Ubuntu on Windows*
- Pacote R versão 3.2.3
- Pacote iRace versão 1.07.1202

- Compilação de código Java com JDK 1.8.0.131
- Uso do software MSI RAMDisk para acesso a arquivos do iRace diretamente na memória, utilizando 256 Mb

Tabela 25: Seleção de parâmetros do GA, para a cidade do Porto

Parâmetro	Intervalo avaliado	Valor selecionado
pop	[20;1500]	515
t_{ga}	$3000000/pop$	5825
P_c	[0,01;1,0]	0,28
P_m	[0,01;1,0]	0,98
vMut	[0,01;1,0]	0,01

Na Tabela 25, são apresentadas os intervalos de valores avaliados, bem como os valores selecionados pelo iRace. A fim de não melhorar o desempenho às custas de maior tempo de processamento, a limitação de execução é o número de avaliações fixado em 3000000, criando uma dependência entre o tamanho s do enxame e número de iterações t_{pso} .

Tabela 26: Seleção de parâmetros do PSO, para a cidade do Porto

Parâmetro	Intervalo avaliado	Valor Selecionado
s	[20;2500]	862
t_{pso}	$3000000/s$	3480
c1	[0,01;1,0]	0,13
c2	[0,01;1,0]	1,00
w_i	[0,01;1,0]	0,85
w_f	[0,01;1,0]	0,25
$reset_t$	[20,500]	85

Na Tabela 26, são apresentadas os intervalos avaliados, bem como os valores selecionados pelo iRace para os parâmetros de execução. Foi utilizado o mesmo limite de 3000000 avaliações.

5.1.4 INFLUÊNCIA DA FUNÇÃO *FITNESS* NA ESTIMAÇÃO DA MATRIZ OD

Para fins de padronização, todos os experimentos realizados com técnicas de estimação da matriz OD não determinísticas (GA e PSO) neste capítulo são de 30 rodadas, ou seja, 30 execuções independentes, sendo analisada a média destas rodadas. É inicialmente necessário descobrir se as distribuições numéricas dos resultados obtidos nas

diferentes métricas correspondem à distribuição normal, através do teste de normalidade de Shapiro-Wilk (CONOVER, 1971), para selecionar os testes estatísticos aplicáveis. Foi utilizado um nível de significância de 5% em todo o Capítulo 5. Valores *p-value* superiores a este não rejeitam a H_0 , e para tal caso, assume-se que o conjunto pertence a uma distribuição normal. Em todos os experimentos realizados nesta dissertação utilizando GA e PSO, os testes de Shapiro-Wilk retornaram valores superiores ao nível de significância para todas as métricas analisadas, de forma que pode-se assumir que os conjuntos pertencem a distribuições normais. Os valores p-value são exibidos na Tabela 43.

Para comparação de desempenho nas métricas, podem ser utilizados os testes estatísticos t de Student, ANOVA, Mann-Whitney e Kruskal-Wallis. Para comparação de somente 2 conjuntos de dados, as opções são t de Student e Mann-Whitney, e os demais para 3 ou mais. Deve-se utilizar t de Student ou ANOVA quando as condições para se utilizar um teste paramétrico (conjuntos numéricos pertencentes a distribuições normais, conforme o teste de Shapiro-Wilk) são atingidas, ficando os demais como alternativa para quando não for possível utilizar testes paramétricos. Todos os testes tem como saída um valor p-value no intervalo $]0, 1[$.

Neste Capítulo, é considerado um nível de significância α para tais testes estatísticos (sendo utilizados testes ANOVA e t de Student, devido aos resultados dos testes de Shapiro-Wilk), de forma que para valores de saída *p-value* $< \alpha$, rejeita-se a hipótese nula, e então, considera-se que as diferenças observadas são estatisticamente significativas. Os testes estatísticos são realizados com relação às médias encontradas, sendo que os resultados para testes t de Student estão concentrados na Tabela 44.

Uma vez que a função *fitness* escolhida direcionará a evolução das soluções nos algoritmos heurísticos, realizar experimentos com as Equações (25), (26) e (27) se faz necessário. O objetivo desta Seção não é comparar o GA e PSO, mas sim, avaliar a escolha da função *fitness* para cada algoritmo. Os valores em negrito representam os melhores valores obtidos para cada métrica. Para os conjuntos nos quais o teste ANOVA indicou de existirem grupos distintos, foi realizado teste pareado t de Student. O critério de escolha de localização de sensores foi o de escolher os arcos com passagem de maior número de pares OD, segundo regra R7 (Ver Seção 2.2.1).

Na última coluna das Tabelas 27 e 28 são mostrados os p-values do teste estatístico ANOVA para as diferentes métricas, a partir dos experimentos com diferentes funções *fitness*, considerando um nível de significância de 5%. Na Tabela 27, a H_0 é rejeitada para as métricas de MAE, GEH e R^2 Arcos. Nestes três casos, é possível concluir que

Tabela 27: Resultados do GA para diferentes funções *fitness*, para a cidade do Porto

Métrica	Função objetivo	Mínimo	Média	Mediana	Máximo	Desvio padrão	p-value ANOVA
MAE	Equação 25	10,86767	13,30929	12,76822	19,53625	1,945	0,017
	Equação 26	11,14313	14,31201	13,92296	19,17522	2,086	
	Equação 27	10,34039	12,8993	12,54277	16,7681	1,723	
RMSE	Equação 25	97,79227	145,27042	131,43525	295,12815	47,618	0,416
	Equação 26	99,9748	153,97209	146,645	260,3932	44,895	
	Equação 27	89,36184	138,50392	122,89202	248,24941	42,793	
GEH	Equação 25	0,30809	0,32847	0,32479	0,38935	0,018	0,000
	Equação 26	0,34937	0,3728	0,37158	0,40176	0,011	
	Equação 27	0,28768	0,30949	0,30504	0,36069	0,018	
R^2 Arcos	Equação 25	0,98996	0,99136	0,99146	0,99228	$5,8 \cdot 10^{-4}$	0,000
	Equação 26	0,99384	0,99415	0,99415	0,99441	$1,2 \cdot 10^{-4}$	
	Equação 27	0,98878	0,99068	0,9908	0,9915	$6,1 \cdot 10^{-4}$	
R^2 ODM	Equação 25	0,43062	0,77613	0,81555	0,88717	0,113	0,485
	Equação 26	0,51011	0,75684	0,77293	0,88481	0,105	
	Equação 27	0,53249	0,79023	0,82798	0,90281	0,102	
TDD	Equação 25	0,20769	8,6391	7,84774	31,85538	6,484	0,948
	Equação 26	0,81897	9,10822	7,17805	27,46652	6,886	
	Equação 27	0,12278	8,62535	6,82644	25,82952	6,158	

existem diferenças estatisticamente significativas na escolha da função *fitness*. As métricas nas quais a H_0 não foi rejeitada estão com fundo cinza, indicando que não podem ser comparadas por não existir diferença estatisticamente significativa nestes casos.

É possível observar que, para diferentes funções *fitness*, não há diferenças estatisticamente significativas para RMSE, R^2 ODM e TDD. As diferenças estatisticamente significativas ocorrem para GEH e R^2 Arcos. Melhores resultados da Equação 27 para a métrica GEH, e melhores resultados para a métrica R^2 Arcos com o uso da Equação 26. Não há diferença estatisticamente significativa entre a Equação 25 e Equação 27 para MAE.

Na Tabela 28, é possível observar que não existe diferença estatisticamente significativa na escolha de função *fitness* para a maioria das métricas, sendo a escolha determinada através do GEH e R^2 Arcos. O comportamento dos algoritmos GA e PSO foi semelhante para estas duas métricas: Melhores resultados da Equação 27 para a métrica GEH, e melhores resultados para a métrica R^2 Arcos com o uso da Equação 26 Nas métricas de GEH e R^2 Arcos.

A partir destes resultados, a função *fitness* a ser utilizada no estudo de caso da cidade do Porto é dada por (27) tanto para o GA quanto para o PSO, ficando (26) como uma segunda opção, pelo melhor desempenho para R^2 Arcos.

Tabela 28: Resultados do PSO para diferentes funções *fitness*, para a cidade do Porto

Métrica	Função objetivo	Mínimo	Média	Mediana	Máximo	Desvio padrão	p-value ANOVA
MAE	Equação 25	9,36434	11,63505	10,81801	21,98346	2,486	0,321
	Equação 26	10,73506	12,04275	11,52271	15,26974	1,320	
	Equação 27	8,75937	11,23562	10,19799	17,25993	2,195	
RMSE	Equação 25	84,07547	151,39068	123,79863	475,95861	78,009	0,647
	Equação 26	108,79722	137,97824	116,72612	267,79211	45,614	
	Equação 27	88,36997	152,44179	117,49723	339,31932	72,466	
GEH	Equação 25	0,27226	0,27808	0,27745	0,28872	0,004	0,000
	Equação 26	0,33203	0,33877	0,33847	0,34838	0,003	
	Equação 27	0,26371	0,26762	0,26743	0,27409	0,002	
R^2 Arcos	Equação 25	0,9925	0,99301	0,99297	0,99352	$2,8 \cdot 10^{-4}$	0,000
	Equação 26	0,99467	0,99474	0,99474	0,99478	$2,2 \cdot 10^{-4}$	
	Equação 27	0,99113	0,9916	0,99155	0,99222	$2,7 \cdot 10^{-4}$	
R^2 ODM	Equação 25	0,23103	0,78012	0,84005	0,91617	0,149	0,605
	Equação 26	0,50981	0,80535	0,85714	0,86741	0,102	
	Equação 27	0,37854	0,77007	0,85442	0,91178	0,161	
TDD	Equação 25	0,12483	6,69107	4,23159	41,0173	8,025	0,595
	Equação 26	0,89732	5,05938	4,95665	14,014	3,148	
	Equação 27	0,04891	5,89024	3,41138	25,68239	6,343	

É interessante notar que, em quase todas as métricas, os piores casos costumam ser bem mais distantes da mediana que os melhores casos em relação também à mediana, levando a mediana a ter melhores valores nas métricas em relação à média. Esse comportamento se repete nos demais resultados para o estudo de caso da cidade do Porto, nas seções seguintes.

5.1.5 INFLUÊNCIA DA LOCALIZAÇÃO DE SENSORES NA ESTIMAÇÃO DA MATRIZ OD

Nesta seção, são realizados experimentos para avaliar a localização dos sensores. A função *fitness* utilizada para o GA e para o PSO é (27), definida na seção anterior.

O resultado é mostrado na Tabela 29 para o GA, Tabela 30 para o PSO, Tabela 31 para o MILP e Tabela 32 para o LS. Uma vez que as viagens dos táxis são conhecidas, e existem registros de passagem de veículos em mais de 10.500 arcos, é necessário definir não somente uma quantidade razoável de sensores virtuais, como também um critério para a escolha destes. Foram levantadas 4 possibilidades, com base nas regras citadas Seção 2.2.1:

- Escolher os arcos com maior fluxo de veículos (regra R4)
- Escolher os arcos com passagens de mais pares OD (regra R7)

- Escolher os arcos com passagens de menos pares OD (regra R2)
- Escolher os arcos com maior fluxo de veículos e maior número de passagens de pares OD no mesmo arco (regras R4 e R7 em conjunto, indicado por R4+R7)

Nesta seção são avaliadas somente as métricas relacionadas a erros na matriz OD de MAE, RMSE, R^2 ODM e TDD. O conjunto de arcos muda de acordo com o critério de escolha de sensores, e neste caso, não é possível realizar uma boa comparação utilizando as métricas relativas aos próprios arcos (GEH, pGEH5 e R^2 Arcos). A quantidade de sensores utilizados foi reduzida de 1000 para 500 para o GA e PSO, a fim de que as diferenças entre os critérios de escolha se tornem mais evidentes nestes algoritmos.

Na última coluna das Tabelas 29 e 30 são mostrados os p-values do teste estatístico ANOVA. O teste é feito sobre as médias com nível de significância 5%. Quando há diferença estatisticamente significativa entre as médias, são realizados testes pareados t de Student. Os melhores valores estão em negrito, e células em cinza indicam que não existe diferença estatisticamente significativa entre tais células.

Tabela 29: Resultados do GA para critérios de escolha de sensores, para a cidade do Porto

Métrica	Critério de escolha	Mínimo	Média	Mediana	Máximo	Desvio padrão	p-value ANOVA
MAE	R2	27,23145	36,39587	36,27701	46,98064	4,864	0,000
	R7	11,79929	15,77185	15,49369	19,22322	1,845	
	R4	16,30832	20,71626	20,78453	26,61558	2,634	
	R4+R7	15,50764	19,55686	19,72661	23,58588	2,094	
RMSE	R2	303,92081	480,2834	493,34652	576,73003	63,522	0,000
	R7	102,98694	177,31775	164,22885	289,20565	53,922	
	R4	124,39427	192,42068	181,78114	315,44415	57,291	
	R4+R7	124,55914	196,08512	188,12155	285,37277	46,352	
R^2 ODM	R2	0,05679	0,4518	0,466	0,60146	0,1171	0,000
	R7	0,48968	0,71405	0,75139	0,88191	0,1190	
	R4	0,39989	0,64524	0,65722	0,81269	0,1344	
	R4+R7	0,42259	0,63228	0,64519	0,81823	0,1140	
TDD	R2	32,37266	76,66466	76,3006	123,15196	21,750	0,000
	R7	4,31151	13,43672	11,1808	30,55638	7,533	
	R4	7,15596	20,46598	20,26572	46,73333	10,101	
	R4+R7	7,78381	19,14043	18,31176	33,36232	7,769	

Na Tabela 29 é possível observar que valores ANOVA encontrados determinam que existe diferença estatisticamente significativa nos resultados a partir da definição do critério de definição de sensores para o GA. Segundo testes t de Student, existe um empate estatístico na métrica RMSE entre os critérios R7, R4 e R4+R7. Na métrica TDD, há um empate estatístico entre os critérios R7 e mais R4+R7. Na métrica R^2 ODM, há um empate estatístico entre os critérios R4 e mais R4+R7.

Segundo a Tabela 29, o critério R7 apresenta melhores resultados para as métricas MAE e R^2 ODM, e nenhum critério tem melhor desempenho que esse nas métricas RMSE e TDD, sendo definido o critério R7 como o mais adequado para o GA.

Tabela 30: Resultados do PSO para critérios de escolha de sensores, para a cidade do Porto

Métrica	Critério de escolha	Mínimo	Média	Mediana	Máximo	Desvio padrão	p-value ANOVA
MAE	R2	31,24651	59,12506	58,23624	86,31717	13,718	0,000
	R7	12,12361	14,78203	14,25724	20,61597	2,311	
	R4	16,75332	20,35754	19,63718	28,49099	2,872	
	R4+R7	14,81377	17,63872	17,30716	26,88226	2,422	
RMSE	R2	379,58746	669,74103	668,25899	944,04799	130,592	0,000
	R7	112,84665	181,44643	170,48155	295,58913	56,788	
	R4	140,51206	213,81266	195,85559	411,41716	64,261	
	R4+R7	122648	174,45751	161,32096	356,88692	49,090	
R^2 ODM	R2	0,00545	0,13692	0,13621	0,31036	0,0924	0,000
	R7	0,47976	0,71559	0,73953	0,8644	0,1230	
	R4	0,28703	0,62146	0,66189	0,79637	0,1288	
	R4+R7	0,32859	0,70333	0,73241	0,83126	0,1109	
TDD	R2	20,16757	133,34241	129,7074	241,98817	54,618	0,000
	R7	0,07096	6,82785	5,72099	27,73376	6,050	
	R4	0,08912	7,95179	5,36515	38734	9,115	
	R4+R7	0,08509	4,93811	2,50303	27,2007	5,810	

Na Tabela 30, é possível observar que os valores ANOVA determinam que existe diferença estatisticamente significativa nos resultados a partir da definição do critério de definição de sensores para o PSO. Segundo testes t de Student, existe um empate estatístico nas métricas RMSE e R^2 ODM entre os critérios R7 e R4+R7. Existe também um empate estatístico entre R4, R7 e R4+R7 para a métrica TDD.

Nenhum critério é melhor que R7 nas métricas de RMSE, R^2 ODM e TDD, e o critério R7 obteve o melhor resultado para MAE. Segundo a Tabela 30, o critério R7 é o que apresenta os melhores resultados para o PSO, sendo então também definido como o mais adequado para o PSO.

Na Tabela 31, é possível observar que o critério R7 apresenta os melhores resultados na métricas de MAE e R^2 ODM, enquanto que o critério R2 tem o melhor valor TDD, e o critério R4 tem o melhor valor RMSE, embora próximo neste ponto do critério R7.

Pelo motivo de obter melhores resultados para duas das quatro métricas avaliadas, e estar próximo do melhor resultado encontrado de uma terceira métrica, considerou-se que o critério de escolha de sensores mais adequado para o MILP é também o R7.

Na Tabela 32, demonstra-se que o critério R7 apresentou os melhores resultados

Tabela 31: Resultados do MILP para critérios de escolha de sensores, para a cidade do Porto

	R2	R7	R4	R4+R7
MAE	16,378	9,16	10,86	10,773
RMSE	376,61	99,003	97,506	107,435
R^2 ODM	0,738	0,894	0,892	0,875
TDD	0,691	2,839	5,459	2,635

para a métrica RMSE e TDD, enquanto o critério R2 apresentou os melhores resultados para as métricas MAE e R^2 ODM.

Tabela 32: Resultados do LS para critérios de escolha de sensores, para a cidade do Porto

	R2	R7	R4	R4+R7
MAE	13,671	16,259	24,589	19,651
RMSE	238,184	164,606	327,058	191,925
R^2 ODM	0,809	0,753	0,437	0,694
TDD	8,731	2,096	4,899	2,616

Para LS, pelo motivo de que há um empate de dois para dois sobre qual o critério mais adequado, e então a escolha é menos óbvia neste caso. Porém, pode-se também considerar que o critério R7 é o segundo colocado nas métricas de MAE e R^2 ODM, enquanto o critério R2 possui o pior resultado em TDD, e o segundo pior em RMSE. Desta forma, considerou-se que o critério mais adequado para o LS é também o R7.

Assim, será considerado que os sensores serão posicionados nos arcos segundo a regra R7, que seleciona arcos que tem passagem de maior número de pares OD.

5.1.6 COMPARAÇÃO DOS RESULTADOS DE ESTIMAÇÃO DA MATRIZ OD

Foram executadas 30 rodadas do PSO e GA, e apenas uma do MILP e LS, uma vez que estes últimos são determinísticos. A função *fitness* utilizada pelo GA e PSO foi (27), e o critério de escolha de localização de sensores foi o de escolher os arcos com passagem de maior número de pares OD, segundo regra R7.

Na Tabela 33, é demonstrado o desempenho das diferentes técnicas de estimação da matriz OD, de acordo com as métricas previamente definidas na Seção 4.1. Na última coluna, são demonstrados os valores ANOVA para comparar PSO, GA, LS e MILP. Considerando um nível de significância de 5%, a H_0 é rejeitada para todas as métricas avaliadas, e portanto é possível concluir que existem diferenças estatisticamente significativas entre a estimação da matriz OD entre diferentes técnicas, levando a necessidade de se verificar

Tabela 33: Comparação das técnicas de estimação da matriz OD para a cidade do Porto

Métrica	Técnica	Mínimo	Média	Mediana	Máximo	Desvio padrão	p-value ANOVA
MAE	PSO	8,75937	11,23562	10,19799	17,25993	2,195	0,000
	GA	10,34039	12,8993	12,54277	16,7681	1,723	
	LS	16,25877	16,25877	16,25877	16,25877	0	
	MILP	9,15974	9,15974	9,15974	9,15974	0	
RMSE	PSO	88,36997	152,44179	117,49723	339,31932	72,466	$9,7 \cdot 10^{-8}$
	GA	89,36184	138,50392	122,89202	248,24941	42,793	
	LS	164,60557	164,60557	164,60557	164,60557	0	
	MILP	99,00264	99,00264	99,00264	99,00264	0	
GEH	PSO	0,26371	0,26762	0,26743	0,27409	0,002	0,000
	GA	0,28768	0,30949	0,30504	0,36069	0,018	
	LS	0,97954	0,97954	0,97954	0,97954	0	
	MILP	0,27001	0,27001	0,27001	0,27001	0	
R^2 Arcos	PSO	0,99113	0,9916	0,99155	0,99222	$2,7 \cdot 10^{-4}$	0,000
	GA	0,98878	0,99068	0,9908	0,9915	$6,1 \cdot 10^{-4}$	
	LS	0,95856	0,95856	0,95856	0,95856	0	
	MILP	0,99328	0,99328	0,99328	0,99328	0	
R^2 ODM	PSO	0,37854	0,77007	0,85442	0,91178	0,161	$2,0 \cdot 10^{-7}$
	GA	0,53249	0,79023	0,82798	0,90281	0,102	
	LS	0,75296	0,75296	0,75296	0,75296	0	
	MILP	0,89398	0,89398	0,89398	0,89398	0	
TDD	PSO	0,04891	5,89024	3,41138	25,68239	6,343	$9,3 \cdot 10^{-8}$
	GA	0,12278	8,62535	6,82644	25,82952	6,158	
	LS	2,09616	2,09616	2,09616	2,09616	0	
	MILP	2,83939	2,83939	2,83939	2,83939	0	

par a par através de testes t de Student. Para os resultados da Tabela 33, são realizados testes t de Student entre os casos considerados de melhor desempenho e os demais. Em caso de não haver diferença estatística significativa, as linhas dos conjuntos estaticamente semelhantes tem cor de fundo cinza.

Em relação aos resultados médios em cada métrica, o MILP obteve os melhores resultados valores nas métricas de MAE, RMSE, R^2 Arcos e R^2 ODM, enquanto que o PSO obteve um melhor GEH, e o LS, um melhor valor TDD. As métricas MAE, RMSE e R^2 ODM observam desvios individuais nos fluxos dos pares OD, e o MILP obteve melhores resultados nestas três métricas, e então, é um vencedor claro neste quesito.

Quanto ao melhor desempenho do PSO na métrica GEH, nota-se na Tabela 37 que o PSO só é capaz de superar o MILP neste quesito se a função *fitness* pelo PSO seja (27), que trata diretamente da métrica GEH. Porém, para métrica de R^2 Arcos, que também trata da adequação dos fluxos nos arcos, o MILP obteve os melhores resultados, levando a uma dúvida de qual técnica seria mais adequada na questão dos arcos. Tal

situação demonstra que o uso de uma quantidade insuficiente de métricas pode enviesar análises de resultados.

A partir dos resultados da Tabela 33, a técnica de estimação da matriz OD de melhor desempenho foi a MILP na maioria das métricas. O uso do GA não se revela interessante pelo motivo de que o PSO obtém melhor desempenho em relação ao GA, ao mesmo tempo em que possui a mesma liberdade de formulação de função *fitness*. O LS é o melhor para TDD, sendo pouco satisfatório nas demais métricas em relação às outras técnicas. A principal vantagem do LS fica pela facilidade de uso, que consiste em procedimentos descritos na Seção 4.4.

5.1.7 INFLUÊNCIA DO NÚMERO DE SENSORES NA ESTIMAÇÃO DA MATRIZ OD

Uma vez que os resultados variam de acordo com a quantidade de sensores utilizados, esta seção apresenta experimentos com diferentes quantidades de sensores. De acordo com os resultados obtidos nas Seções 5.1.4 e 5.1.5, foi definido que o critério de escolha de sensores como R7, enquanto que a função *fitness* utilizada para GA e PSO foi (27). Para GA e PSO, o valor exibido nas Figuras é o da média.

Foram realizados experimentos de 200 até 2000 sensores em incrementos em intervalos de 100 sensores entre 200 e 500 sensores, e incrementos de 250 sensores entre 500 e 2000 sensores. Para cada caso testado, foram realizadas 30 rodadas para os algoritmos não determinísticos.

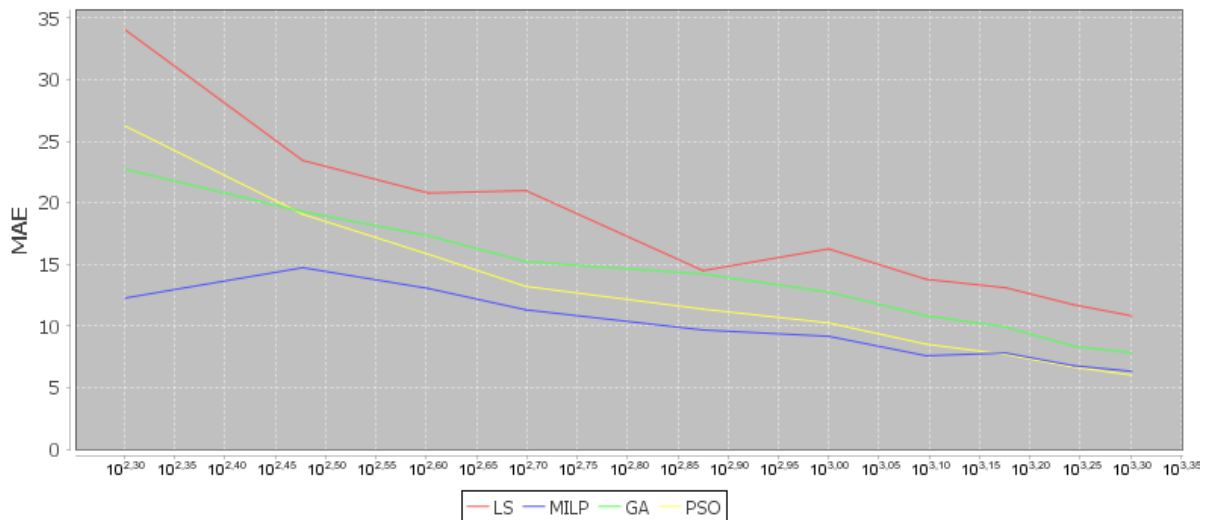


Figura 22: MAE para diferentes técnicas e número de sensores.

Na Fig. 22 da métrica MAE, o comportamento geral observado é que para mais

sensores, a estimação da matriz OD se aproxima mais do valor real, qualquer seja a técnica aplicada. Tal comportamento demonstra que conforme mais sensores são adicionados, efetivamente ocorre um acréscimo de informação para as técnicas, levando a melhores resultados nesta métrica.

A técnica LS apresenta o pior desempenho em todos os casos, com valor MAE de 34 para 200 sensores e próximo de 11 para 2000 sensores. A técnica de melhor desempenho é o MILP, para até 1500 sensores. A partir daí, o PSO tem desempenho muito próximo. A maior diferença entre MILP e PSO ocorre com 200 sensores, reduzindo conforme o número de sensores aumenta. O MILP é também a técnica com menor variação nos resultados em relação ao número de sensores. O GA supera o PSO com 200 sensores e é muito próximo com 300 sensores, mas tem desempenho entre LS e PSO a partir de 400 sensores.

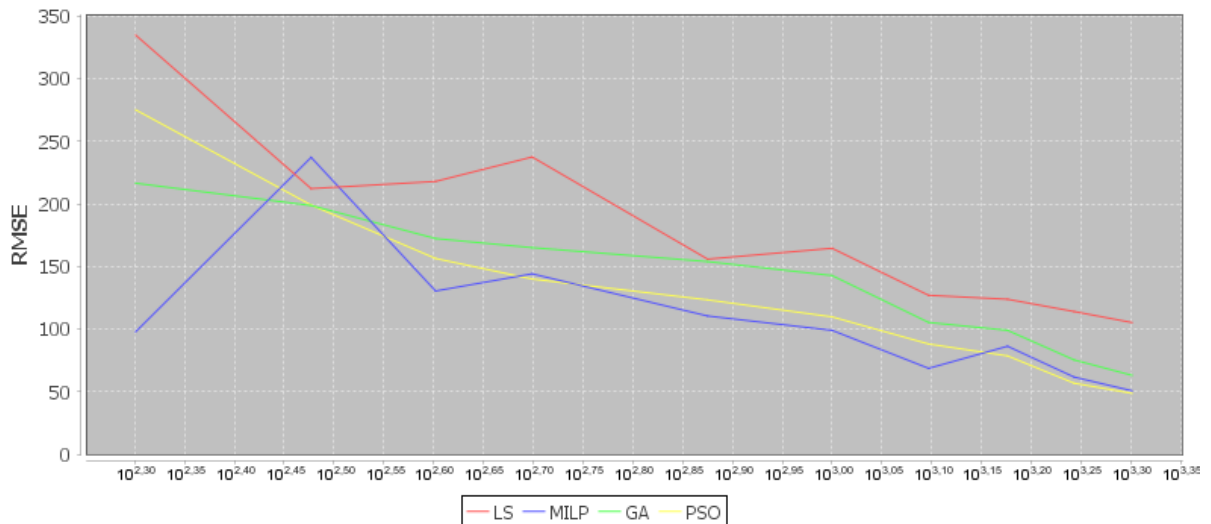


Figura 23: RMSE para diferentes técnicas e número de sensores.

Na Fig. 23 usando a métrica RMSE, o comportamento geral é similar ao da Fig. 22. O RMSE reduz conforme o número de sensores aumenta. O MILP tem desempenho ruim com 300 sensores, melhorando rapidamente com 400 sensores ou mais. Nesta métrica, o PSO se aproxima do MILP a partir de 1500 sensores. O GA só supera o PSO até 400 sensores, tendo desempenho entre o PSO e o LS nos demais casos.

Para o comportamento geral da Fig. 24, é possível perceber que no geral, para maiores quantidades de arcos considerados como sensores, o valor médio de GEH tende a subir. Tal comportamento é compreensível, já que na prática, embora a quantidade de variáveis livres e espaço de busca se mantenham, a quantidade de variáveis de observação aumenta, dificultando o processo de otimização.

Na Fig. 24, mostra-se que GA, PSO e MILP tem desempenhos muito próximos

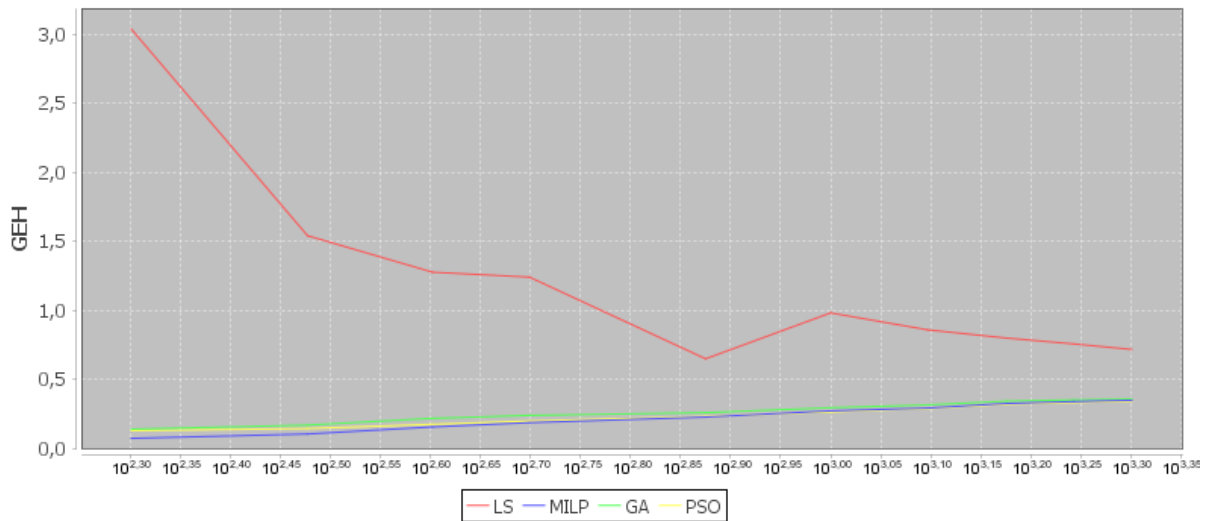


Figura 24: GEH para diferentes técnicas e número de sensores.

em todos os casos testados, havendo quase uma sobreposição das linhas no gráfico, e sendo afetados pela variação do número de sensores. Para estas três técnicas, o GEH é inferior a 0,2 até 300 sensores, aumentando gradualmente conforme mais sensores são adicionados, chegando próximo a 0,4 em 2000 sensores. O LS apresenta um comportamento praticamente oposto, com maiores valores de GEH para menores quantidades de sensores, chegando a 3,0 para 300 sensores, e reduzindo-se para abaixo de 0,8 para 2000 sensores.

Conforme a Seção 4.1.4, uma métrica também a ser considerada é pGEH5. A única técnica que apresentou dificuldades nesse critério foi a LS, especificamente para 200 sensores, em que o valor encontrado foi de somente 76%. A partir de 300 sensores, todas as técnicas obtiveram valores superiores ao mínimo necessário, com as técnicas de GA, PSO e MILP obtendo resultados iguais ou próximos de 100% para quaisquer número de sensores.

Na Fig. 25, em que é analisado o coeficiente R^2 entre as matrizes OD estimadas e reais, o comportamento geral também é similar as Figuras 22 e 23, porém considerando a maximização de R^2 ODM. O LS apresenta os piores resultados para praticamente qualquer quantidade de sensores, com PSO superando MILP somente para 300 sensores, com MILP tendo melhor desempenho em outros casos. Somente acima de 1500 sensores o PSO se aproxima mais do MILP. O GA tem um desempenho intermediário entre PSO e LS.

Na Fig. 26, é demonstrado o valor de R^2 Arcos de acordo com número de sensores. É possível observar que PSO, GA e MILP apresentam valores elevados de R^2 Arcos independente do número de sensores. LS apresenta os menores valores, com um R^2 Arcos

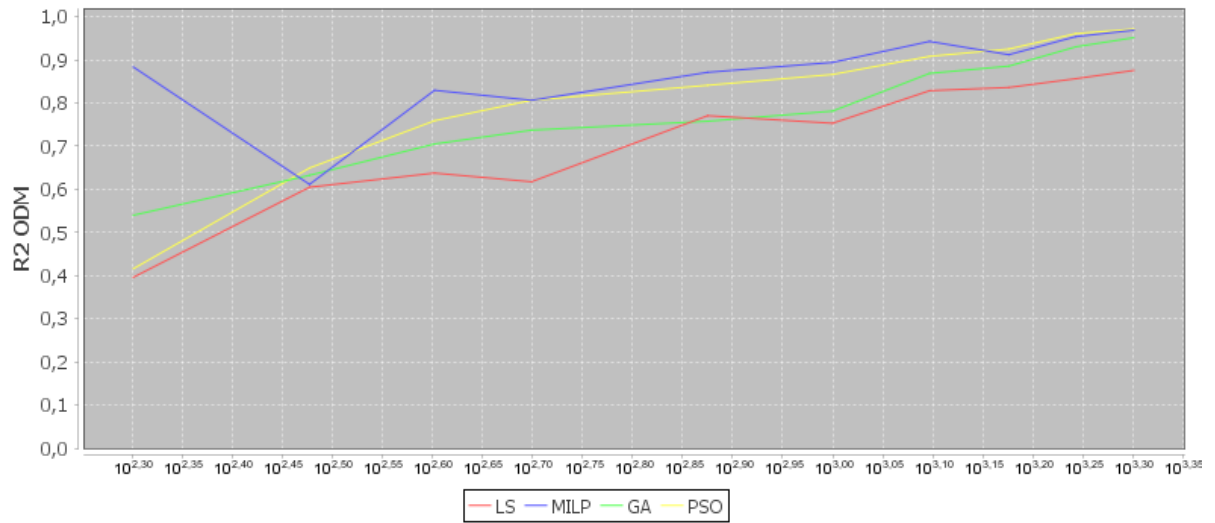


Figura 25: R^2 ODM para diferentes técnicas e número de sensores.

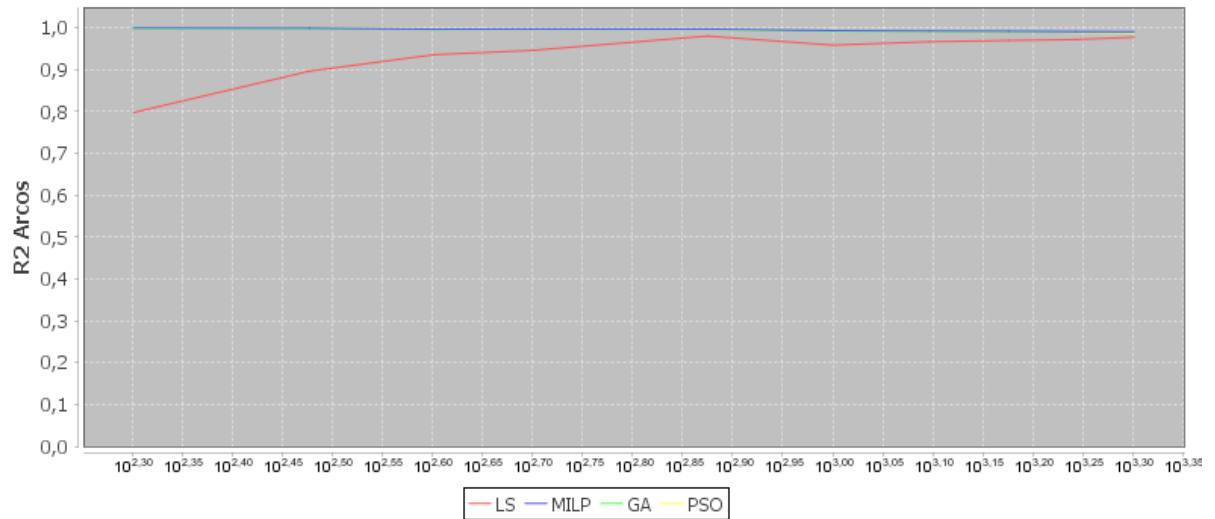


Figura 26: R^2 Arcos para diferentes técnicas e número de sensores.

de 0,8 com 200 sensores, com R^2 Arcos crescente conforme o número de sensores aumenta.

Para a métrica TDD, o comportamento geral observado é de que LS e MILP tem um comportamento no qual o desempenho é ruim para pequena quantidade de sensores, melhorando até um ponto mínimo em 500 sensores. O TDD volta a subir conforme o número de sensores é aumentado a partir deste ponto. O LS tem comportamento semelhante ao MILP, com ponto mínimo em 750 sensores, chegando a obter os melhores resultados entre 750 e 1250 sensores. O PSO tem resultado mais regular, com ponto mínimo também em 750 sensores. O GA tem comportamento mais diferenciado, diminuindo o TDD desde 200 sensores até um mínimo em 1750 sensores. A partir de 1500 sensores, é a técnica de melhor desempenho avaliado pelo TDD, apesar de ser a pior entre

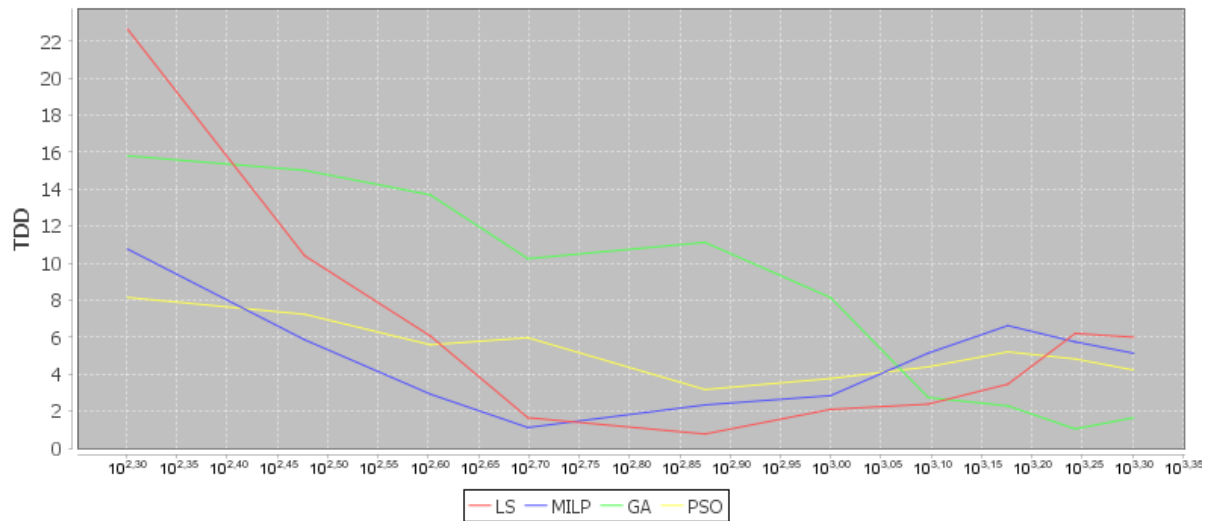


Figura 27: TDD para diferentes técnicas e número de sensores.

300 e 1000 sensores.

Considerando o contexto geral de todas as métricas avaliadas, parece existir uma vantagem para MILP com um menor número de sensores, que se reduz quando o número de sensores é aumentado. Em resumo, MILP aparenta ser mais adequado para menores quantidades de sensores, enquanto PSO tem desempenho muito próximo ou até superior para quantidades superiores de sensores, em que mais informação está envolvida. As técnicas são afetadas de formas diferentes conforme o número de sensores varia.

Nas Figuras 28 e 29, é mostrada a variação do desvio padrão das métricas MAE, RMSE, GEH, R^2 ODM, R^2 Arcos e TDD em função do número de sensores para GA e PSO, usando 30 rodadas para cada algoritmo. O desvio padrão é normalizado em relação aos maiores valores obtidos em cada uma das métricas entre os experimentos com diferentes números de sensores, a fim de que as técnicas possam ser comparadas.

Para ambas as figuras, a métrica MAE é representada pela linha vermelha, a RMSE pela linha azul escuro, GEH pela linha verde, R^2 ODM pela linha amarela, R^2 Arcos pela linha rosa, e TDD pela linha azul claro.

Na Fig. 28, o desvio padrão da métrica GEH diminui entre 200 e 500 sensores, existindo um pico em 750 sensores, com leve aumento do desvio padrão para GEH entre 1000 e 1500 sensores. Após esse ponto, o desvio padrão para a métrica aumenta novamente. O desvio padrão para as demais métricas (MAE, RMSE, R^2 ODM, R^2 Arcos e TDD) apresenta um comportamento semelhante: Há menos desvio conforme aumenta o número de sensores, entre 200 e 1500 sensores, com um pico para 1750, seguido de redução

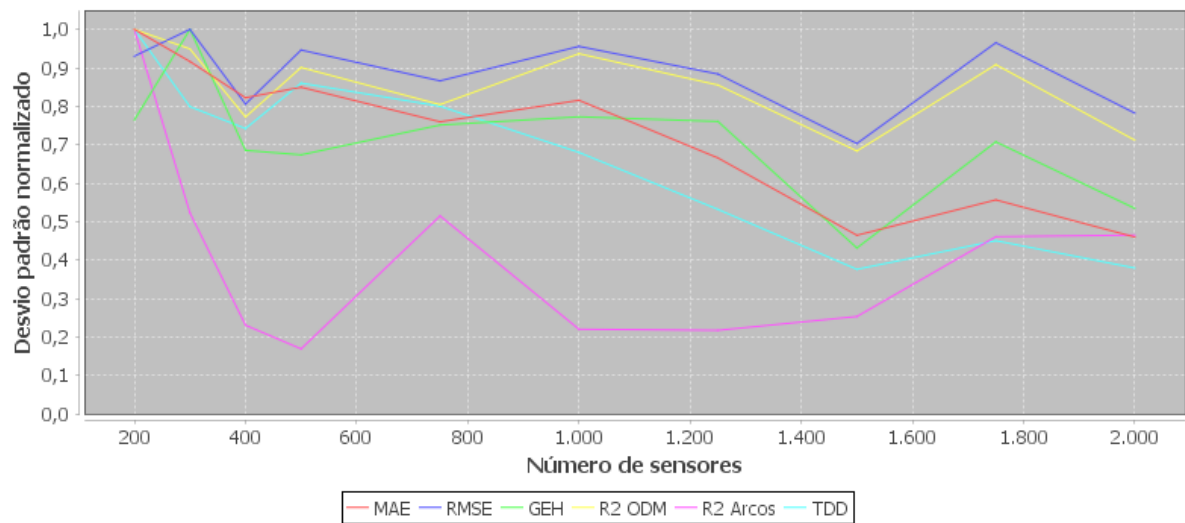


Figura 28: Desvio padrão normalizado obtido com 30 rodadas do GA para diferentes métricas e número de sensores.

em 2000 sensores, mas ainda com desvios maiores em 2000 do que em 1500.

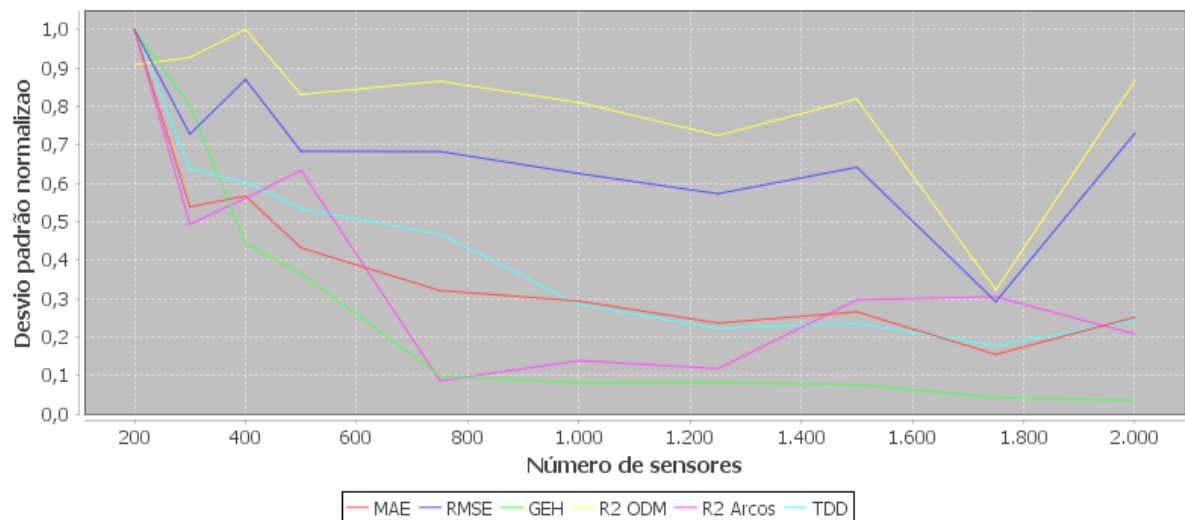


Figura 29: Desvio padrão normalizado obtido com 30 rodadas do PSO para diferentes métricas e número de sensores.

Na Fig. 29, o desvio padrão das métricas de GEH, MAE e TDD apresentam um tendência de redução conforme o número de sensores aumenta, principalmente entre 200 e 750 sensores. O desvio padrão da métrica R^2 arcos tem comportamento semelhante ao desvio padrão de GEH, MAE e TDD até 1250 sensores, quando então o desvio padrão passa a aumentar. O desvio padrão das métricas de RMSE e R^2 ODM apresentam uma redução mais leve conforme o número de sensores aumenta, com menores valores em 1750, seguido de um aumento no desvio padrão para 2000 sensores.

5.2 ESTUDO DE CASO DE CURITIBA

O estudo de caso de Curitiba deriva de um estudo já em curso no IPPUC, utilizando um modelo simulado de tráfego desenvolvido no software comercial PTV VISUM. Este software permite a exportação dos dados simulados sobre nós, arcos, contagens de tráfego, desempenho GEH obtido após estimação da matriz OD e *assignment matrix*, constituindo a base de dados utilizada neste estudo de caso.

Os sensores da rede têm sua localização definida de acordo com a localização dos sensores físicos da rede viária de Curitiba. Outra questão é que a matriz OD real não é conhecida, e portanto, são utilizadas somente métricas que avaliam arcos. As métricas MAE, RMSE, R^2 ODM e TDD não foram utilizadas. Assim, as métricas consideradas no caso de Curitiba são R^2 Arcos, GEH e pGEH5, que não necessitam do conhecimento da matriz OD real.

Uma vez que não existem informações acerca das probabilidades de uso de arco por par OD para esta região, é necessário caracterizar a *assignment matrix*. Optou-se por utilizar a mesma formulação utilizada no VISUM, ou seja, um método de equilíbrio de usuário que se utiliza de uma função logística para gerar a *assignment matrix*, utilizando-se as mesmas capacidades, fluxos registrados e tempos de viagem para cada arco em situações ideais que as utilizadas no estudo da IPPUC.

A rede da Fig. 30 é a representação no VISUM da rede viária considerada no estudo de caso, entre os bairros Centro e Rebouças. Contém 29 nós de origem e destino localizados na fronteira da região. Os dados de sensores importados do VISUM para este trabalho são medições reais de 71 sensores de tráfego, sendo 29 deles localizados nas extremidades da região.

A Seção 5.2 é dividida em três seções: Inicialmente, há a seleção de parâmetros do GA e PSO na Seção 5.2.1. Na Seção 5.2.2, são realizados experimentos com o GA e PSO, a fim de definir a função *fitness* de melhor desempenho de acordo com as métricas de estimação da matriz OD. Na Seção 5.2.3 é realizado um comparativo entre as técnicas de GA, PSO, MILP, LS e VISUM para estimação da matriz OD na rede de Curitiba.

5.2.1 CONFIGURAÇÃO DOS ALGORITMOS GA E PSO

Nesta Seção são apresentados os resultados da seleção de parâmetros realizada pelo iRace. Foram utilizadas quantidades máximas de partículas do PSO ou indivíduos

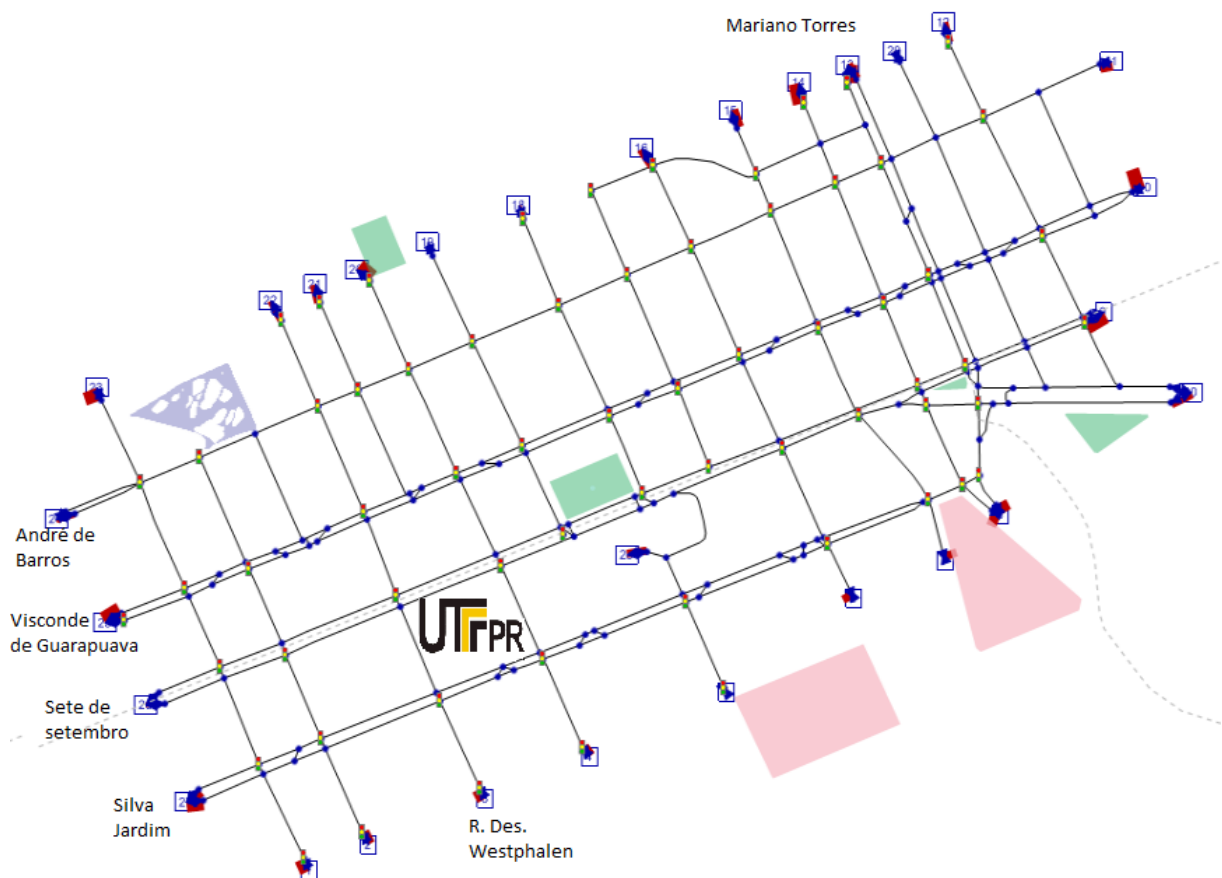


Figura 30: Mapa da região do estudo de caso de Curitiba.

do GA de acordo com a limitação de 16Gb de memória RAM disponível, considerando a execução em 8 instâncias paralelas. A função *fitness* utilizada foi (26).

Na execução do iRace, foi utilizado nível de confiança de 95% para o F-Test, variação nos parâmetros até a segunda casa decimal, e um limite de 1000 experimentos, o que levou a cerca de 12 horas de processamento para cada um dos algoritmos. A leitura e escrita de dados em disco demonstrou ser um fator limitante de desempenho, de forma que optou-se por emular disco em memória. As seguintes configurações de hardware e software foram utilizadas para ambos os algoritmos:

- Processador AMD Ryzen 7 1700 3Ghz de 8 núcleos
- Memória 16Gb DDR4 2966Mhz
- Sistema operacional Windows 10 64 bits
- Execução através de *Bash on Ubuntu on Windows*

- Pacote R versão 3.2.3
- Pacote iRace versão 1.07.1202
- Compilação de código Java com JDK 1.8.0.131
- Uso do software MSI RAMDisk para leitura e escrita de arquivos do iRace diretamente na memória, utilizando 256 Mb.

Tabela 34: Seleção de parâmetros do GA, para Curitiba

Parâmetro	Intervalo avaliado	Valor selecionado
pop	[20;1500]	50
t_{ga}	$3000000/pop$	40000
P_c	[0,01;1,0]	0,85
P_m	[0,01;1,0]	0,34
vMut	[0,01;1,0]	0,10

Na Tabela 34, são apresentadas os intervalos de valores testadas, bem como os valores selecionados pelo iRace. A limitação de execução foi o número de avaliações a serem realizadas fixado em 3000000, criando uma dependência entre o parâmetro de população pop e de gerações t_{ga} .

Em relação a Tabela 25, P_c apresenta um valor quase 3 vezes superior, enquanto P_m apresenta um valor quase 3 vezes inferior, sendo P_m compensado por $vMut$, que é 10 vezes superior para Curitiba em relação a cidade do Porto. A população escolhida foi inferior em relação ao estudo de caso da cidade do Porto, aumentando o número de iterações.

Tabela 35: Seleção de parâmetros do PSO, para Curitiba

Parâmetro	Intervalo avaliado	Valor selecionado
s	[20;2500]	1404
t_{pso}	$3000000/s$	2136
c1	[0,01;1,0]	0,44
c2	[0,01;1,0]	0,89
w_i	[0,01;1,0]	0,90
w_f	[0,01;1,0]	0,07
$reset_t$	[20;500]	71

Na Tabela 35, são apresentadas os intervalos de valores testadas, a limitação de execução foi o número de avaliações a serem realizadas fixado em 3000000. Em relação à

Tabela 26, os valores de $reset_t$, w_i e w_f se mantiveram relativamente próximos, com um enxame 63% maior. $c2$ foi reduzido em 11%, ao mesmo tempo em que $c1$ aumento de 0,13 para 0,44.

5.2.2 INFLUÊNCIA DA FUNÇÃO *FITNESS* NA ESTIMAÇÃO DA MATRIZ OD

Da mesma forma como foi analisada a escolha de função *fitness* na Seção 5.1.4, foram realizados experimentos com as diferentes Equações 25, 26 e 27 na rede Curitiba. Para os experimentos a seguir, foram realizadas 30 rodadas do PSO ou GA para cada função *fitness*, com o objetivo de avaliar a escolha da função *fitness* para cada algoritmo.

Tabela 36: Resultados do GA para diferentes funções *fitness*, para Curitiba

Métrica	Função objetivo	Mínimo	Média	Mediana	Máximo	Desvio padrão	p-value ANOVA
GEH	Equação 25	3,28577	3,90071	3,84033	5,29251	0,377	0,000
	Equação 26	2,31036	2,40437	2,39622	2,4810	0,051	
	Equação 27	3,03151	3,61156	3,52315	4,89015	0,417	
pGEH5	Equação 25	0,71831	0,77418	0,77465	0,83099	0,025	0,000
	Equação 26	0,85915	0,88357	0,88732	0,90141	0,011	
	Equação 27	0,70423	0,76714	0,77465	0,8169	0,033	
R^2 Arcos	Equação 25	0,82842	0,88432	0,88854	0,91597	0,021	0,000
	Equação 26	0,9572	0,95914	0,95947	0,96017	$8,0 \cdot 10^{-4}$	
	Equação 27	0,82246	0,88262	0,88887	0,9090	0,21	

Na última coluna das Tabelas 36 e 37, são demonstrados os valores ANOVA para comparar PSO e GA sob as 3 diferentes funções *fitness* propostas. Considerando um nível de significância de 5%, a H_0 é rejeitada para todas as métricas avaliadas, e portanto é possível concluir que existem diferenças estatisticamente significativas. Para os resultados das Tabelas 36 e 37, são realizados testes t de Student pareados. Em caso de não haver diferença estatística significativa, as linhas dos conjuntos estaticamente semelhantes tem cor de fundo cinza.

Na Tabela 36, podemos observar que a Equação 26 possui os melhores resultados do GA para as três métricas. Ocorre um empate (diferenças estatisticamente não relevantes) entre as Equações 25 e 27 para as métricas de pGEH5 e R^2 Arcos.

Na Tabela 37, podemos observar que a Equação 26 também possui os melhores resultados do PSO nas três métricas. Ocorre um empate estatístico entre as Equações 25 e 27 para as métricas de pGEH5 e R^2 Arcos.

A partir destes resultados, a função *fitness* a ser utilizada para o estudo de caso

Tabela 37: Resultados do PSO para diferentes funções *fitness*, para Curitiba

Métrica	Função objetivo	Mínimo	Média	Mediana	Máximo	Desvio padrão	p-value ANOVA
GEH	Equação 25	2,11188	2,69738	2,6456	3,39812	0,354	0,000
	Equação 26	2,0017	2,02524	2,01876	2,09212	0,020	
	Equação 27	2,10559	2,4113	2,40728	2,80228	0,187	
pGEH5	Equação 25	0,80282	0,84319	0,84507	0,88732	0,023	0,000
	Equação 26	0,88732	0,89202	0,88732	0,90141	0,006	
	Equação 27	0,80282	0,84178	0,84507	0,90141	0,024	
R^2 Arcos	Equação 25	0,88309	0,92709	0,92924	0,95214	0,017	0,000
	Equação 26	0,96283	0,96339	0,96349	0,96362	$2,2 \cdot 10^{-4}$	
	Equação 27	0,88017	0,92435	0,9282	0,94543	0,013	

de Curitiba é a Equação 26, independente do algoritmo ser o GA ou o PSO. Em especial, o algoritmo GA se mostrou mais sensível a escolha da função *fitness* em relação ao PSO neste estudo de caso. Outro ponto importante é que a Equação 26 é a única com pGEH5 superior a 0,85 na média, o que é uma questão a ser considerada conforme a Seção 4.1.4.

5.2.3 COMPARAÇÃO DOS RESULTADOS DE ESTIMAÇÃO DA MATRIZ OD

Na última coluna da Tabela 38, são demonstrados os valores ANOVA para comparar PSO, GA, LS, MILP e VISUM. Considerando nível de significância de 5%, a H_0 é rejeitada para todas as métricas avaliadas, e portanto é possível concluir que os conjuntos não são todos iguais. Para os resultados da Tabela 38, são realizados testes T entre os casos considerados de melhor desempenho e os demais. Em caso de não existir diferença estatisticamente significativa, as linhas dos conjuntos estaticamente semelhantes tem cor de fundo cinza.

O desempenho das técnicas avaliadas, de acordo com as métricas propostas na Seção 4.1 para arcos, pode ser observado na Tabela 38. A função *fitness* utilizada para GA e PSO foi a (26).

A técnica com pior desempenho em todos os aspectos foi o LS, apresentando um GEH médio acima de 7, e somente 42,3% de arcos com GEH abaixo de 5. A solução do VISUM apresentou GEH médio de 1,861, porém uma porcentagem de somente 73,2% de arcos com GEH abaixo de 5. Assim, as soluções obtidas pelo LS e VISUM não podem ser consideradas adequadas, de acordo com a Seção 4.1.4.

Para a métrica de GEH, MILP obteve o melhor desempenho, seguido de VISUM e PSO. Para R^2 Arcos, o melhor desempenho foi obtido por PSO, seguido por GA, e

Tabela 38: Comparação das técnicas de estimação da matriz OD para Curitiba

Métrica	Técnica	Mínimo	Média	Mediana	Máximo	Desvio padrão	p-value ANOVA
GEH	PSO	2,0017	2,02524	2,01876	2,09212	0,020	0,000
	GA	2,31036	2,40437	2,39622	2,481	0,051	
	LS	7,14832	7,14832	7,14832	7,14832	0	
	MILP	1,83757	1,83757	1,83757	1,83757	0	
	VISUM	1,861	1,861	1,861	1,861	0	
pGEH5	PSO	0,88732	0,89202	0,88732	0,90141	0,006	0,000
	GA	0,85915	0,88357	0,88732	0,90141	0,011	
	LS	0,42254	0,42254	0,42254	0,42254	0	
	MILP	0,87324	0,87324	0,87324	0,87324	0	
	VISUM	0,73239	0,73239	0,73239	0,73239	0	
R^2 Arcos	PSO	0,96283	0,96339	0,96349	0,96362	$2,2 \cdot 10^{-4}$	0,000
	GA	0,9572	0,95914	0,95947	0,96017	$8,0 \cdot 10^{-5}$	
	LS	0,79595	0,79595	0,79595	0,79595	0	
	MILP	0,95345	0,95345	0,95345	0,95345	0	
	VISUM	0,91047	0,91047	0,91047	0,91047	0	

então MILP. Para pGEH5, o melhor desempenho também foi obtido pelo PSO, seguido do GA e então pelo MILP. Pela superioridade em duas das três métricas, o PSO pode ser considerado o vantajoso para o estudo de caso de Curitiba, em relação ao MILP e GA.

Na Fig. 31, tem-se um gráfico de dispersão dos fluxos de veículos estimados e observados nos arcos com sensores para GA, LS, MILP, PSO e VISUM. O ideal é ter os fluxos estimados iguais aos observados em cada arco e, portanto, uma reta diagonal. O LS, no geral, apresenta valores estimados superiores aos medidos, independente da faixa de valores medidos nos sensores. Este comportamento que pode ser explicado pelo fato de que sem a compensação de alguns pares OD negativos, que são transformados em zero para reparar a solução, a tendência é que os erros sejam para mais. A solução VISUM apresenta erros dispersos em todas as faixas de valores medidos nos arcos, para mais e para menos.

As técnicas de GA, PSO e MILP apresentam desvios maiores principalmente para arcos com fluxo medido inferior a 1000 veículos por hora, aparentemente nos mesmos arcos, o que indica uma dificuldade inerente ao problema de otimização por si. No entanto, a maioria dos arcos é próxima da diagonal para estas três técnicas, o que reflete o fato de que somente estas técnicas obtiveram valor $GEH < 5$ em 85% dos casos, conforme a Tabela 38.

Na Figura 32, são mostrados os fluxos nos arcos da rede viária do estudo de caso de Curitiba, a partir da matriz OD obtida pela execução do PSO. Considerando o fluxo

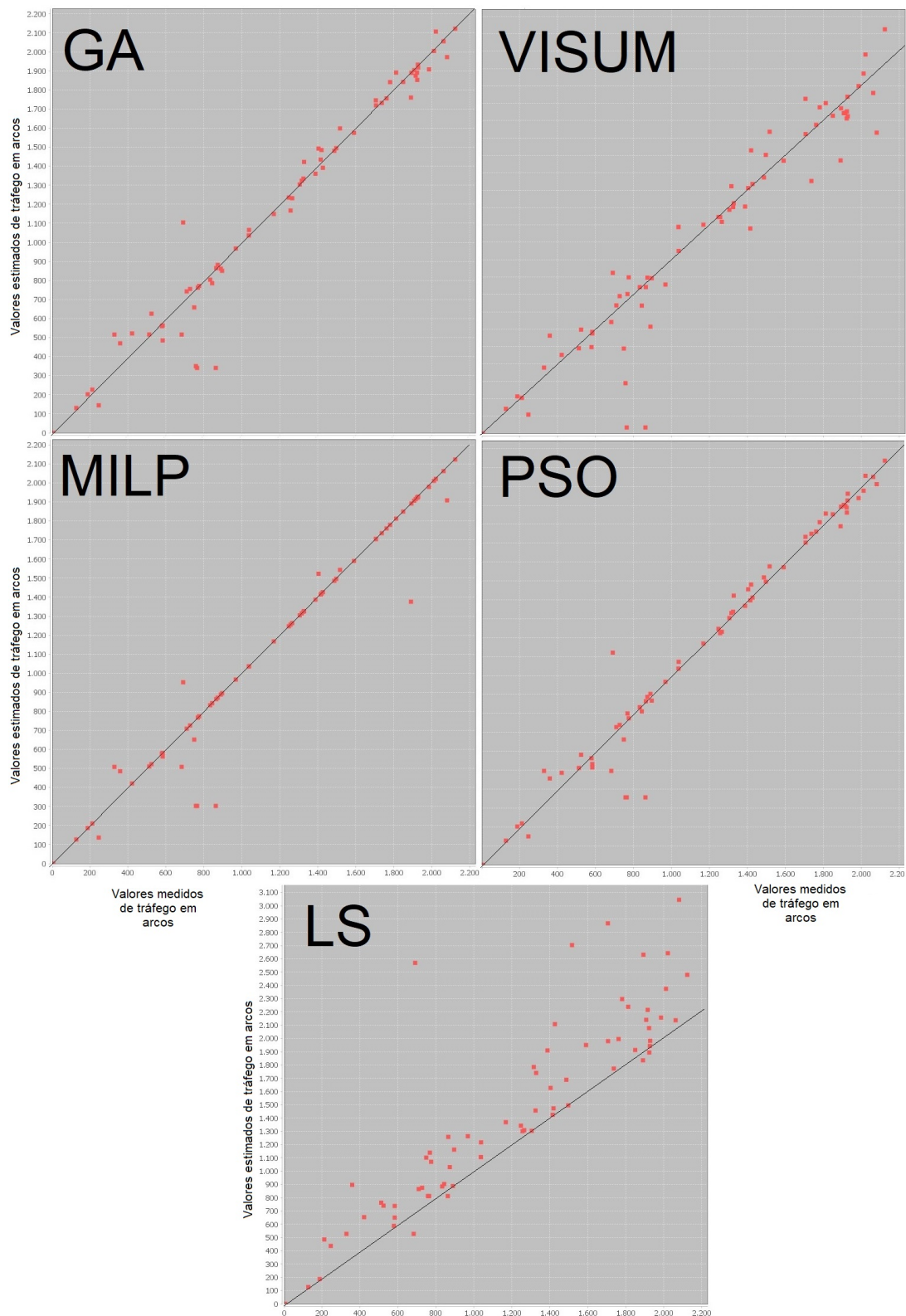


Figura 31: Gráficos de dispersão para o fluxo de veículos estimados e observados nos arcos da rede viária de Curitiba, usando GA, VISUM, MILP, PSO e LS.

de veículos por hora, as cores simbolizam a intensidade do fluxo observado. A partir da alocação da matriz OD na rede em conjunto com a *assignment matrix*, é possível obter

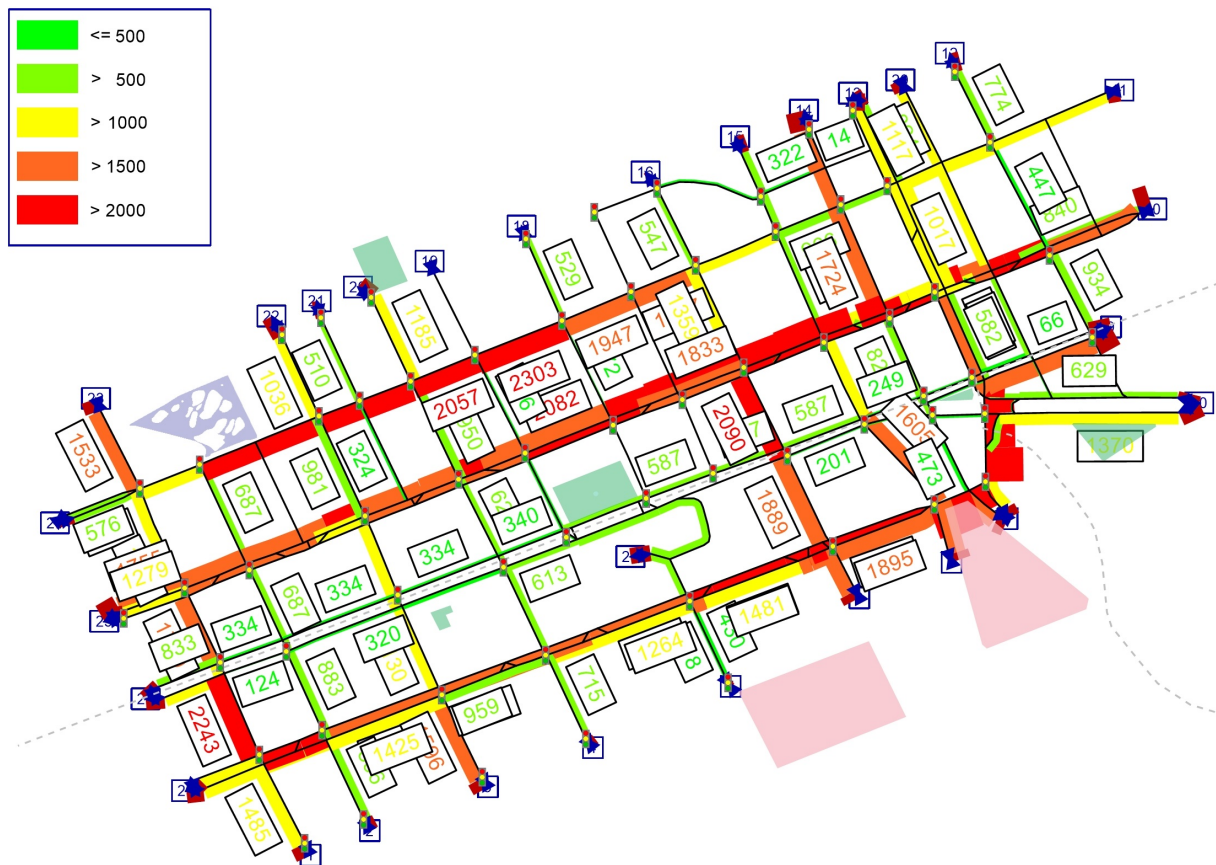


Figura 32: Representação no VISUM da matriz OD encontrada pelo PSO para Curitiba.

estimações do fluxo de veículos em arcos que não possuem sensores.

5.3 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo, a fim de obter os resultados dos experimentos inicialmente propostos, foi necessário realizar vários passos auxiliares, tais como o teste de normalidade de Shapiro-Wilk para identificar a normalidade dos dados de saída e então aplicar os testes estatísticos adequados para comparação de resultados, e procedimentos de agrupamento espacial e *map matching* para a definição do estudo de caso da cidade do Porto.

Na Seção 5.1.1, o agrupamento espacial realizado permitiu definir os pontos de origem e destino para o estudo de caso da cidade do Porto. Foi também utilizado um processo iterativo para descobrir o número ideal de *clusters*. Por outro lado, outros critérios podem ser analisados para definição das áreas em aplicações futuras, como por exemplo, utilizar dados demográficos.

A reconstrução das trajetórias de táxis apresentou algumas falhas na identificação

dos trajetos. Porém, os resultados de *map matching* da Seção 5.1.2 permitiram, em conjunto com o agrupamento espacial, a definição do estudo de caso da cidade do Porto.

Na Seção 5.1.3 para a cidade do Porto, e na Seção 5.2.1 para Curitiba, foi realizada a seleção de parâmetros dos algoritmos GA e PSO através do iRace. Os parâmetros indicados foram diferentes para cada um dos estudos de caso. Tal procedimento foi necessário para obter um melhor desempenho do GA e PSO, uma vez que o comportamento dos algoritmos é influenciado por sua configuração.

A Seção 5.1.4 avaliou a influência da escolha da função *fitness* para Porto, enquanto que a Seção 5.2.2 para Curitiba. A melhor função *fitness* variou entre os estudos de caso de Curitiba e da cidade do Porto, mas dentro de um mesmo estudo de caso, foi a mesma para GA e PSO. A conclusão é de que a melhor função *fitness* varia mais de acordo com o problema do que com o algoritmo escolhido.

Quanto aos critérios de escolha dos sensores no estudo de caso da cidade do Porto (Seção 5.1.5), a conclusão é de que o critério de escolha dos sensores influencia as técnicas de forma similar. Ye e Wen (2017) realizaram testes utilizando R2, e também R4 em conjunto com R7 (Ver Seção 2.2.1). Nesta dissertação, estes dois casos foram testados, mas também um caso utilizando apenas R7. Neste último caso, o desempenho obtido foi superior.

A partir dos experimentos da Seção 5.1.7, considerando a variação no número de sensores e consequentemente, quantidade de informação envolvida, notou-se que as técnicas são afetadas num mesmo sentido (com mais informação, geralmente melhor), porém não na mesma intensidade. Outro ponto importante foi que mesmo técnicas determinísticas de MILP e LS, apresentam dificuldades em algumas configurações do problema, sendo que acrescentar mais sensores piora o resultado, ao invés de melhorá-lo. Com mais informação envolvida (maior número de sensores) o PSO se beneficia mais rapidamente que o MILP.

Considerando as Seções 5.1.6 para a cidade do Porto e 5.2.3 para Curitiba, com relação à comparação geral de diferentes técnicas de estimação da matriz OD, o melhor desempenho foi observado em MILP (para a cidade do Porto) e PSO (para Curitiba). Porém, os resultados dependem das métricas utilizadas. Para as métricas relativas a arcos para a cidade do Porto, há um empate entre MILP e PSO, com MILP sendo considerado superior neste estudo de caso.

Apesar das diferentes metodologias para formulação dos dois estudos de caso, as

técnicas demonstraram um ordenamento de desempenho semelhante em ambos os casos, com MILP e PSO obtendo os melhores resultados, seguidos de GA e, por fim, do LS. É interessante notar que o desempenho do PSO e GA foi mais regular em Curitiba, sem grandes variações nos piores valores, como aconteceu para a cidade do Porto. Outra questão importante é relativa à comparação da mediana em relação à média das métricas utilizadas. Para o PSO e GA no estudo de caso da cidade do Porto, foi possível observar, principalmente nas métricas relativas à matriz OD, que os piores casos eram bem mais distantes da mediana do que os melhores casos, indicando situações em que as métricas de arcos estavam bem ajustadas, mas a matriz OD não estava. Tal situação ocorre mesmo quando a função de *fitness* utilizada para PSO e GA é similar a utilizada para o MILP.

Quanto à seleção das métricas de desempenho para o problema de estimação da matriz OD, algumas métricas (como RMSE e R^2 ODM, por exemplo) demonstraram comportamentos semelhantes, enquanto outras demonstraram características totalmente diferentes (como MAE e TDD, por exemplo). Usar poucas métricas pode levar a resultados enviesados. As métricas são complementares, identificando diferentes características de erros da estimação da matriz OD.

6 CONCLUSÃO

Nesta dissertação, foi realizada a comparação de quatro diferentes técnicas de estimação da matriz OD (PSO, GA, mínimos quadrados e MILP), em diferentes configurações do problema. Além disso, a técnica *T-Flow Fuzzy* foi também considerada para o caso de utilização do software VISUM. O melhor desempenho foi observado com MILP para a cidade do Porto e PSO para Curitiba. Porém, os resultados dependem das métricas utilizadas. As métricas são complementares, identificando diferentes características de erros de estimação da matriz OD. Para as métricas relativas a arcos na cidade do Porto, há um empate entre MILP e PSO, com MILP sendo considerado superior neste estudo de caso. O VISUM se mostrou uma solução com bom desempenho nas métricas GEH e R^2 Arcos, porém não apresentando 85% dos arcos com $GEH < 5$, o que não é aceitável de acordo com a Seção 4.1.4. O LS apresentou desempenho inferior em relação às outras técnicas na maioria dos casos. Outra questão importante é relativa à comparação da mediana com a média das métricas utilizadas. Para o PSO e GA no estudo de caso da cidade do Porto, foi possível observar que, principalmente nas métricas relativas à matriz OD, os valores médios são de pior desempenho em relação aos valores medianos, indicando uma distribuição assimétrica dos resultados, com valores elevados de erros que poderiam ser considerados como *outliers*. Ocorreram situações em que as métricas de arcos estavam bem ajustadas, mas as métricas de matriz OD não estavam. Esta situação ocorreu mesmo quando a função de *fitness* utilizada para PSO e GA foram similares à utilizada para o MILP.

O critério de escolha dos sensores influencia todas as técnicas de forma similar. Considerando a variação no número de sensores e consequentemente, da quantidade de informação envolvida, as técnicas são beneficiadas com mais informação, porém não na mesma intensidade. Com mais informação envolvida (maior número de sensores) o PSO se beneficia mais rapidamente que o MILP. O agrupamento espacial definiu as áreas no estudo de caso da cidade do Porto. Porém, outros critérios podem ser analisados para definição das áreas em aplicações futuras, como por exemplo, utilizar dados demográficos.

Em relação a trabalhos futuros, nota-se atualmente a geração de um grande volume de dados de mobilidade, porém com acesso ainda muito restrito. Entretanto, novas fontes de dados podem surgir com o passar do tempo, com políticas *open source* e *open data*. Uma possibilidade é o aplicativo *Arcade City*, que, em princípio, tem o potencial de democratizar o acesso a dados de uma grande quantidade de corridas do próprio aplicativo. Maior disponibilidade de dados de veículos sonda beneficiaria as pesquisas na área de mobilidade urbana, permitindo novas abordagens. Por exemplo, no estudo de caso de Curitiba, poderiam ser utilizados dados de veículos sonda para obter uma matriz OD semente e as probabilidades de uso de arco por par OD (*assignment matrix*), potencialmente melhorando a qualidade da estimação da matriz OD. Outra possibilidade de trabalho futuro é a utilização de veículos sonda e *map matching* como um auxílio para resolução do problema de escolha de sensores em uma rede.

Além disso, existem outras técnicas na literatura de estimação da matriz OD que não foram consideradas neste trabalho, como as técnicas de inferência bayesiana e máxima verossimilhança, que poderiam ser exploradas.

REFERÊNCIAS

- AG, P. P. T. V. **PTV Visum. Traffic in cities - simulation model**. 2018. Acesso em Maio de 2018. Disponível em: <<http://vision-traffic.ptvgroup.com/en-us/products/ptv-visum/>>.
- ALEXANDER, L. et al. Origin–destination trips by purpose and time of day inferred from mobile phone data. **Transportation Research Part C: Emerging Technologies**, v. 58, p. 240 – 250, 2015. ISSN 0968-090X. Big Data in Transportation and Traffic Engineering. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0968090X1500073X>>.
- BAEK, S.; KIM, H.; LIM, Y. Multiple vehicle origin destination matrix estimation from traffic counts using genetic algorithm. **Journal of Transportation Engineering**, v. 130, n. 3, p. 339–347, 2004.
- BARANN, B.; BEVERUNGEN, D.; MÜLLER, O. An open-data approach for quantifying the potential of taxi ridesharing. **Decision Support Systems**, Elsevier, v. 9, p. 86–95, 2017. ISSN 0167-9236.
- BELL, M. G. H. The estimation of origin-destination matrices by constrained generalised least squares. **Transportation Research Part B: Methodological**, v. 25B, n. 1, p. 13–22, 1991.
- BERA, S.; RAO, K. V. K. Estimation of origin-destination matrix from traffic counts : the state of the art. **European Transport**, v. 49, p. 3–23, 2011.
- BIEM, A. et al. Ibm infosphere streams for scalable, real-time, intelligent transportation services. In: **ACM sigmon international Conference on management of data**. New York, USA: ACM, 2010. p. 1093–1104. ISBN 978-1-4503-0032-2. Disponível em: <<http://doi.acm.org/10.1145/1807167.1807291>>.
- BIERLAIRE, M. The total demand scale : a new measure of quality for static and dynamic origin – destination trip tables. **Transportation Research Part B**, v. 36, p. 837–850, 2002.
- CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics**, v. 3, n. 1, p. 1–27, 1974. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/03610927408827101>>.
- CASCETTA, E. Estimation of trip matrices from traffic counts and survey data. A generalized least squares estimator. **Transportation Research Part B**, I, n. 415, p. 289–299, 1984.
- CASCETTA, E.; NGUYEN, S. A unified framework for estimating or updating origin/destination matrices from traffic counts. **Transportation Research Part B: Methodological**, v. 22, n. 6, p. 437–455, 1988. ISSN 0191-2615. Disponível em: <<http://www.sciencedirect.com/science/article/pii/0191261588900240>>.

CAVALCANTE, E. et al. Thinking smart cities as systems of systems: A perspective study. In: **Proceedings of the 2Nd International workshop on smart cities**. New York, NY, USA: ACM, 2016. (16), p. 9:1–9:4. ISBN 978-1-4503-4667-2. Disponível em: <<http://doi.acm.org/10.1145/3009912.3009918>>.

CHEN, A.; CHOOTINAN, P.; RECKER, W. W. Examining the quality of synthetic Origin Destination Trip table estimated by path flow estimator. **Journal of Transportation Engineering**, v. 131, n. July, p. 506–513, 2005.

CIUFFO, B.; PUNZO, V.; TORRIERI, V. Comparison of simulation-based and model-based calibrations of traffic-flow microsimulation models. **Transportation Research**, v. 2088, p. 36–44, 2008.

CONOVER, W. **Practical Nonparametric Statistics**. John Wiley & Sons, 1971. ISBN 9780471168515. Disponível em: <<https://books.google.com.br/books?id=Nv4YAAAAIAAJ>>.

CORMEN, T. H. et al. Introduction to algorithms. In: _____. 2nd. ed. Cambridge, USA and London, England: MIT Press and McGraw-Hill, 2001. cap. 29.3. The simplex algorithm, p. 790–804. ISBN 0-262-03293-7.

CURITIBA, P. de. **Pesquisa origem destino começa nesta sexta feira em Curitiba**. 2016. Acesso em Maio de 2018. Disponível em: <<http://www.curitiba.pr.gov.br/noticias/pesquisa-origem-destino-comeca-nesta-sexta-feira-em-curitiba/39381>>.

DU, J.; WEI, L. Quantum behaved particle swarm optimization for origin destination matrix prediction. In: **2nd International Conference on Power Electronics and Intelligent Transportation System**. Shenzhen, China: IEEE, 2009. v. 1, p. 133–136. ISBN 978-1-4244-4544-8.

ELWOOD, S. Geographic information science: emerging research on the societal implications of the geospatial web. **Progress in Human Geography**, v. 34, n. 3, p. 349–357, 2010. Disponível em: <<https://doi.org/10.1177/0309132509340711>>.

GEROLIMINIS, N.; DAGANZO, C. F. Existence of urban-scale macroscopic fundamental diagrams : Some experimental findings. **Transportation Research Part B**, v. 42, p. 759–770, 2008.

GREENFELD, J. Matching GPS Observations to Locations on a Digital Map. In: . Washington, USA: 81th Annual Meeting of the Transportation Research Board, 2002.

HAZELTON, M. L. Statistical inference for time varying origin – destination matrices. **Transportation Research Part B**, v. 42, p. 542–552, 2008.

HERRERA, J. C. et al. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. **Transportation Research Part C: Emerging Technologies**, Elsevier Ltd, v. 18, n. 4, p. 568–583, 2010. ISSN 0968090X. Disponível em: <<http://dx.doi.org/10.1016/j.trc.2009.10.006>>.

HOURDAKIS, J.; MICHALOPOULOS, P. G. A practical procedure for calibrating microscopic traffic simulation models. **Transportation Research**, n. 4167, p. 130–139, 2003.

HU, S.-r.; LIOU, H.-t. A generalized sensor location model for the estimation of network origin – destination matrices. **Transportation Research Part C**, Elsevier Ltd, v. 40, p. 93–110, 2014. ISSN 0968-090X. Disponível em: <<http://dx.doi.org/10.1016/j.trc.2014.01.004>>.

JIANG, X. et al. Large-scale taxi O / D visual analytics for understanding metropolitan human movement patterns. **Journal of Visualisation**, n. 288, p. 185–200, 2015.

KENNEDY, J.; EBERHART, R. Particle swarm optimization. In: . Perth, Australia: IEEE, 1995. v. 4, p. 1942–1948. ISSN 19353812.

LIU, Y. et al. Understanding intra-urban trip patterns from taxi trajectory data. **Journal of Geographical Systems**, Springer, v. 14, n. 4, p. 463–483, 2012. ISSN 1435-5949.

LLOYD, S. P. Least squares quantization in pulse code modulation. **IEEE Transactions on Information Theory**, v. 28, n. 2, p. 129–137, 1982. ISSN 15579654.

LO, H.-p.; CHAN, C.-p. Simultaneous estimation of an origin – destination matrix and link choice proportions using traffic counts. **Transportation Research Part A**, v. 37, p. 771–788, 2003.

LOPEZ-IBANEZ, M. et al. The irace package: Iterated racing for automatic algorithm configuration. **Operations Research Perspectives**, v. 3, n. Supplement C, p. 43 – 58, 2016. ISSN 2214-7160. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S2214716015300270>>.

MA, T.; ABDULHAI, B. Genetic algorithm-based optimization approach and generic tool for calibrating traffic microscopic simulation parameters. **Transportation Research Record: Journal of the Transportation Research Board**, Transportation Research Board of the National Academies, n. 1800, p. 6–15, 2002.

MAHER, M. J. Inferences on trip matrices from observations on link volumes: A Bayesian statistical approach. **Transportation Research Part B**, v. 17, n. 6, 1983.

MATIAS, L. M. et al. Time-evolving od matrix estimation using high-speed gps data streams. **Expert Systems with Applications**, vol. 44, p. 275–288, 2016.

NIHAN, N. L. Recursive estimation of o-d matrices from input output counts. **Transportation Research Part B**, n. 2, p. 149–163, 1987.

OpenStreetMap contributors. **Planet dump retrieved from <https://planet.osm.org>**. 2018. Disponível em <https://www.openstreetmap.org>. Acesso em janeiro de 2018.

PANDO, L. U.; LUDERS, R. Estimation of origin-destination matrix from traffic counts in the city of porto with pso and taxi trip data. In: . Uberlandia, Brasil: Encontro Nacional de Inteligência Artificial, 2017. v. 1, p. 901–911.

PARRY, K.; HAZELTON, M. L. Estimation of origin – destination matrices from link counts and sporadic routing data. **Transportation Research Part B**, Elsevier Ltd, v. 46, n. 1, p. 175–188, 2012. ISSN 0191-2615. Disponível em: <<http://dx.doi.org/10.1016/j.trb.2011.09.009>>.

PMC. **Portal de Dados Abertos**. 2018. Acesso em maio de 2016. Disponível em: <<http://www.curitiba.pr.gov.br/dadosabertos/>>.

POLI, R.; KENNEDY, J.; BLACKWELL, T. Particle swarm optimization. **Swarm Intelligence**, v. 1, n. 1, p. 33–57, Jun 2007. ISSN 1935-3820. Disponível em: <<https://doi.org/10.1007/s11721-007-0002-0>>.

ROSINOWSKI, J. **Entwicklung und implementierung eines ÖPNV-Matrixkorrektur verfahrens mithilfe von methoden der theorieunscharfer Mengen (Fuzzy Sets Theorie)**. 1994.

SHERALI, H. D.; HOBEIKA, A. G. A linear programming approach for synthesizing origin–destination trip tables from link traffic volumes. **Transportation Research Part B**, v. 28, n. 3, p. 213–233, 1994.

SHERALI, H. D.; PARK, T. Estimation of dynamic origin destination trip tables for a general network. **Transportation Research Part B**, v. 35, p. 217–235, 2001.

SHI, Y.; EBERHART, R. A modified particle swarm optimizer. **IEEE International Conference on Evolutionary Computation**, IEEE Press, Piscataway, NJ, p. 69–73, 1998.

SPIESS, H. A maximum likelihood model for estimating origin-destination matrices. **Transportation Research Part B**, v. 21, n. 5, p. 395–412, 1987.

TEKNOMO, K.; FERNANDEZ, P. A theoretical foundation for the relationship between generalized origin-destination matrix and flow matrix based on ordinal graph trajectories. **Journal of Advanced Transportation**, v. 48, n. 6, p. 608–626, 2014. ISSN 20423195.

UCI. **Dataset de viagens de Táxi de Porto, Portugal**. 2015. Acesso em novembro de 2016. Disponível em: <<https://archive.ics.uci.edu/ml/datasets/Taxi+Service+Trajectory+-+Prediction+Challenge,+ECML+PKDD+2015>>.

VOORHEES, A. M. A general theory of traffic movement. **Transportation**, v. 40, n. 6, p. 1105–1116, 2013. ISSN 0049-4488. Disponível em: <<http://link.springer.com/10.1007/s11116-013-9487-0>>.

WARDROP, J. G.; WHITEHEAD, J. I. Correspondence. some theoretical aspects of road traffic research. **Proceedings of the Institution of Civil Engineers**, v. 1, n. 5, p. 767–768, 1952. Disponível em: <<https://doi.org/10.1680/ipeds.1952.11362>>.

XIE, J.; NIE, Y. M.; LIU, X. Testing the proportionality condition with taxi trajectory data. **Transportation Research Part B**, Elsevier, v. 104, p. 583–601, 2017. ISSN 0191-2615.

XIONG, H. et al. Enhancing security and privacy in traffic-monitoring systems. **IEEE Pervasive Computing**, IEEE Computer Society, Los Alamitos, CA, USA, v. 5, p. 38–46, 2006. ISSN 1536-1268.

YANG, H. A. I. Optimal traffic countering locations for origin–destination matrix estimation. **Transportation Research Part B**, v. 32, n. 2, p. 109–126, 1998.

YANG, H. A. I.; SASAKI, T.; IIDA, Y. Estimation of origin–destination matrices from link traffic counts on congested networks. **Transportation Research Part B**, v. 268, 1992.

YANG, X.; LU, Y.; HAO, W. Origin-destination estimation using probe vehicle trajectory and link counts. **IEEE Transactions on Intelligent Transportation Systems**, v. 2017, n. 18, 2017.

YE, P.; WEN, D. Optimal traffic sensor location for origin – destination estimation using a compressed sensing framework. **Journal of Advanced Transportation**, v. 18, n. 7, p. 1857–1866, 2017.

YOUSEFIKIA, M. et al. A study on the generalized tflowfuzzy o-d estimation. In: . Brisbane, Australia: Australasian Transport Research Forum, 2013.

YUN, I.; AERDE, V.; AERDE, V. Estimation of Dynamic Origin Destination Matrix: A Genetic Algorithm Approach. **IEEE Conference on Intelligent Transportation Systems**, p. 390–395, 2005.

ZHENG, Y. Trajectory data mining: An overview. **ACM Transaction on Intelligent Systems and Technology**, ACM TIST, September 2015. Disponível em: <<https://www.microsoft.com/en-us/research/publication/trajectory-data-mining-an-overview/>>.

ZHENG, Y.-L. et al. On the convergence analysis and parameter selection in particle swarm optimization. In: **Proceedings of the 2003 International Conference on Machine Learning and Cybernetics**. Xian, China: IEEE, 2003. v. 3, p. 1802–1807.

ZWITTER, A. Big data ethics. **Big Data & Society**, v. 1, n. 2, 2014. Disponível em: <<https://doi.org/10.1177/2053951714559253>>.

APÊNDICE A – DETALHES DOS TESTES DE SHAPIRO-WILK E T DE STUDENT

Tabela 42: p-values para o teste de normalidade de Shapiro-Wilk no estudo de caso da cidade do Porto

Experimento	Métrica	GA	PSO
Influência da função <i>fitness</i> Equações (25), (26) e (27), respectivamente	MAE	0,895; 0,945 e 0,948	0,733; 0,791 e 0,866
	RMSE	0,843; 0,915 e 0,874	0,690; 0,662 e 0,797
	GEH	0,860; 0,972 e 0,819	0,946; 0,940 e 0,962
	R^2 Arcos	0,967; 0,981 e 0,898	0,961; 0,967 e 0,973
	R^2 ODM	0,836; 0,914 e 0,865	0,760; 0,651 e 0,784
Seção 5.1.4, Tabelas 27 e 28	TDD	0,873; 0,919 e 0,910	0,686; 0,912 e 0,823
NSLP Critérios R2, R7, R4 e R4+R7, respectivamente Seção 5.1.5, Tabelas 29 e 30	MAE	0,982; 0,976; 0,971 e 0,978	0,977; 0,913; 0,905 e 0,838
	RMSE	0,920; 0,928; 0,900 e 0,943	0,979; 0,900; 0,865 e 0,767
	R^2 ODM	0,731; 0,916; 0,904 e 0,944	0,952; 0,900; 0,920 e 0,821
	TDD	0,991; 0,916; 0,924 e 0,941	0,979; 0,832; 0,716 e 0,760
Comparativo geral de técnicas Seção 5.1.6, Tabela 33	MAE	0,948	0,866
	RMSE	0,874	0,797
	GEH	0,819	0,962
	R^2 Arcos	0,898	0,973
	R^2 ODM	0,865	0,784
	TDD	0,910	0,823

Tabela 43: p-values para o teste de normalidade de Shapiro-Wilk no estudo de caso de Curitiba

Experimento	Métrica	GA	PSO
Influência da função <i>fitness</i> Equações (25), (26) e (27), respectivamente Seção 5.2.2, Tabelas 36 e 37	GEH	0,949; 0,878 e 0,901	0,962; 0,852 e 0,970
	pGEH5	0,849; 0,961 e 0,942	0,937; 0,597 e 0,955
	R^2 Arcos	0,878; 0,921 e 0,909	0,942; 0,809 e 0,907
Comparativo geral de técnicas Seção 5.2.3, Tabela 38	GEH	0,949	0,852
	pGEH5	0,849	0,597
	R^2 Arcos	0,878	0,809

¹Na Tabela 44, células contendo o símbolo * indicam que os testes t de Student para estes casos não foram realizados, devido aos resultados previamente obtidos de testes ANOVA.

Tabela 44: p-values para os testes t de Student

Experimento	Métrica	Par analisado	p-value GA	p-value PSO
Experimentos com diferentes funções <i>fitness</i> para a cidade de Porto	MAE	Equações (25) e (26)	0,059	*
	MAE	Equações (25) e (27)	0,391	*
	MAE	Equações (26) e (27)	<0,05	*
	RMSE	Equações (25) e (26)	*	*
	RMSE	Equações (25) e (27)	*	*
	RMSE	Equações (26) e (27)	*	*
	GEH	Equações (25) e (26)	<0,05	<0,05
	GEH	Equações (25) e (27)	<0,05	<0,05
	GEH	Equações (26) e (27)	<0,05	<0,05
	R^2 Arcos	Equações (25) e (26)	<0,05	<0,05
	R^2 Arcos	Equações (25) e (27)	<0,05	<0,05
	R^2 Arcos	Equações (26) e (27)	<0,05	<0,05
	R^2 ODM	Equações (25) e (26)	*	*
	R^2 ODM	Equações (25) e (27)	*	*
	R^2 ODM	Equações (26) e (27)	*	*
	TDD	Equações (25) e (26)	*	*
	TDD	Equações (25) e (27)	*	*
	TDD	Equações (26) e (27)	*	*
Experimentos utilizando diferentes critérios de de sensores (NSLP) para a cidade de Porto	MAE	Critérios R2 e R7	<0,05	<0,05
	MAE	Critérios R2 e R4	<0,05	<0,05
	MAE	Critérios R2 e R4+R7	<0,05	<0,05
	MAE	Critérios R7 e R4	<0,05	<0,05
	MAE	Critérios R7 e R4+R7	<0,05	<0,05
	MAE	Critérios R4 e R4+R7	<0,05	<0,05
	RMSE	Critérios R2 e R7	<0,05	<0,05
	RMSE	Critérios R2 e R4	<0,05	<0,05
	RMSE	Critérios R2 e R4+R7	<0,05	<0,05
	RMSE	Critérios R7 e R4+R7	0,154	0,612
	RMSE	Critérios R4 e R4+R7	0,786	<0,05
	R^2 ODM	Critérios R2 e R7	<0,05	<0,05
	R^2 ODM	Critérios R2 e R4	<0,05	<0,05
	R^2 ODM	Critérios R2 e R4+R7	<0,05	<0,05
	R^2 ODM	Critérios R7 e R4	<0,05	<0,05
	R^2 ODM	Critérios R7 e R4+R7	<0,05	0,687
	R^2 ODM	Critérios R4 e R4+R7	0,689	<0,05
	TDD	Critérios R2 e R7	<0,05	<0,05
	TDD	Critérios R2 e R4	<0,05	<0,05
	TDD	Critérios R2 e R4+R7	<0,05	<0,05
Seção 5.1.5 Tabelas 29 e 30	TDD	Critérios R7 e R4	<0,05	0,576
	TDD	Critérios R7 e R4+R7	<0,05	0,222
	TDD	Critérios R4 e R4+R7	0,571	0,133
Comparativo geral de técnicas para a cidade de Porto	MAE	Algoritmos GA e PSO	<0,05	
	RMSE	Algoritmos GA e PSO	0,369	
	GEH	Algoritmos GA e PSO	<0,05	
	R^2 Arcos	Algoritmos GA e PSO	<0,05	
	R^2 ODM	Algoritmos GA e PSO	<0,05	
Seção 5.1.6 Tabela 33	TDD	Algoritmos GA e PSO	0,960	
Experimentos com diferentes funções <i>fitness</i> para Curitiba	GEH	Equações (25) e (26)	<0,05	<0,05
	GEH	Equações (25) e (27)	<0,05	<0,05
	GEH	Equações (26) e (27)	<0,05	<0,05
	R^2 Arcos	Equações (25) e (26)	<0,05	<0,05
	R^2 Arcos	Equações (25) e (27)	0,756	0,503
	R2 Arcos	Equações (26) e (27)	<0,05	<0,05
Seção 5.2.2 Tabelas 36 e 37	pGEH5	Equações (25) e (26)	<0,05	<0,05
	pGEH5	Equações (25) e (27)	0,363	0,820
	pGEH5	Equações (26) e (27)	<0,05	<0,05
Comparativo geral de técnicas para Curitiba	GEH	Algoritmos GA e PSO	<0,05	
	R^2 Arcos	Algoritmos GA e PSO	<0,05	
	pGEH5	Algoritmos GA e PSO	<0,05	