



**José Luís  
Sanches Tavares  
Semedo**

**Sistema de recomendação de trajectos rodoviários  
orientado ao contexto**

**Context-based routing in road networks**



José Luís  
Sanches Tavares  
Semedo

Sistema de recomendação de trajectos rodoviários  
orientado ao contexto

Context-based routing in road networks

“

*Amar o perdido  
deixa confundido  
este coração.*

*Nada pode o olvido  
contra o sem sentido  
apelo do Não.*

*As coisas tangíveis  
tornam-se insensíveis  
à palma da mão.*

*Mas as coisas findas,  
muito mais que lindas,  
essas ficarão.*

”

— *Memória*, Carlos Drummond de Andrade



Universidade de Aveiro  
2015

Departamento de Eletrónica,  
Telecomunicações e Informática

**José Luís  
Sanches Tavares  
Semedo**

**Sistema de recomendação de trajectos rodoviários  
orientado ao contexto**

**Context-based routing in road networks**

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia de Computadores e Telemática, realizada sob a orientação científica do Doutor José Manuel Matos Moreira, Professor auxiliar do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro.

Dissertation presented to University of Aveiro to fulfill the necessary requirements to obtain the Mestre degree in Engenharia de Computadores e Telemática, done under of scientific guidance of Doctor José Manuel Matos Moreira, auxiliar Professor at of University of Aveiro.

Este trabalho é dedicado a todas as pessoas que de alguma forma contribuíram para esta longa caminhada académica, que num passado não muito longínquo era algo quase impossível de se concretizar. Digo quase porque a motivação e o conhecimento recebido de muitas pessoas fez-me acreditar em continuar esta batalha. Nessas pessoas incluem-se professores com quem aprendi muito (um beijo especial para a Professora Rosa Áurea), amigos de infância, que sempre deixaram explícito o seu apoio incondicional, aos amigos que tive o prazer de conhecer durante esta etapa académica, às pessoas maravilhosas que conheci durante a minha passagem pela bela cidade que é Aveiro e aos familiares que sempre viram este esforço como um futuro exemplo daquilo que os nossos filhos poderão aspirar (Xana, tu és a seguir!!!). Em especial, para o Sr Vitor e para a Dona Domingas, a quem sempre terei um amor incondicional e a quem serei eternamente agradecido por tudo.

**o júri / the jury**

presidente / president

Professor Doutor Joaquim João Estrela Ribeiro Silvestre Madeira  
Professor Auxiliar da Universidade de Aveiro

vogais / examiners committee

Prof. Doutor Alexandre Miguel Barbosa Valle de Carvalho  
Professor Auxiliar da Faculdade de Engenharia da Universidade do Porto

Prof. Doutor José Manuel Matos Moreira  
Professor Auxiliar da Universidade de Aveiro

**agradecimentos /  
acknowledgements**

Agradecimentos ao Professor Doutor José Manuel Matos Moreira, por ter prestado um grande auxílio no desenvolvimento deste trabalho. À Professora Doutora Alda Sofia Pires de Dias Marques e ao Professor Doutor João Manuel de Oliveira e Silva Rodrigues por me terem dado a oportunidade de desenvolver este trabalho e fazer parte do projecto "Formar Profissionais de Saúde para o Uso de Auscultação Pulmonar Computorizada". Um bem haja a todos os colegas que me ajudaram a avançar com trabalho sempre que senti dificuldades. Sabem que estarei disponível sempre que precisarem. Mais uma vez deixo o meu grande agradecimento aos meus amigos e familiares, que cá estiveram de alguma forma presentes nos bons e maus momentos.

## Palavras Chave

data warehouse espacial, sistemas de informação orientados ao contexto, sistemas de recomendação, *map matching*

## Resumo

A utilização de sistemas de geo-localização já faz parte do quotidiano das pessoas, através dos mais variados equipamentos de recepção de dados posicionais. Com o aumento de utilização deste tipo de equipamentos, a estruturas de armazenamento associadas a dados geográficos foram aperfeiçoadas de forma a permitir uma gestão mais eficiente. Neste trabalho é apresentado uma Data Warehouse espacial orientado ao contexto para o armazenamento de dados relativos ao tráfego automóvel. Esses dados são obtidos através de dados posicionais e representados num mapa rodoviário digital. O mapeamento é analisado através da comparação das similaridades de trajectos criados a partir de dados posicionais de veículos em Aalborg (Dinamarca), usando diferentes períodos de amostragem. Após o armazenamento dos trajectos, dos dados posicionais e dos troços percorridos em três Data Marts distintas, foi feita uma análise de rotas em zonas residenciais, estradas secundárias e em pontos turísticos em Pequim (China). Através da análise faz-se a relação entre as rotas mais rápidas, as mais curtas, o número de paragens e o contexto em que esses dados foram registados. A análise mostra que, nas zonas de estradas secundárias as rotas utilizadas com mais frequência são na maioria dos casos a rota mais rápida, o que não acontece nas zonas residenciais e nas zonas turísticas. A análise mostrou que a relevância temporal influenciou o nível de tráfego e que o factor meteorológico não teve influência na escolha das rotas.

**Keywords**

spatial data warehousing, context aware information systems, recommendation systems, map matching

**Abstract**

The use of Geographic Information Systems is part of today's people's lives. As the use of these systems keeps growing, data storage and management structures are improving in order to allow a better use of spatio-temporal data. In this dissertation a Data Warehouse for traffic and contextual storage is presented. The main data sources are GPS datasets and digital road maps. A map matching algorithm was implemented and evaluated using data from vehicles in Aalborg (Denmark). The data stored in three Data Marts, a route analysis in several zones of Beijing city was presented. With this analysis, we show the relationship between the shortest routes, the fastest routes, the number of stoppages and the contextual data. The analysis shows that in secondary road zones, the routes used more frequently are usually the fastest routes. In residential zones and in turistic locations this conclusion was not found valid. It was concluded that time is a dimension that influences traffic levels, but no pattern was found between route selection and weather condition influenced the traffic level.



# CONTEÚDO

---

CONTEÚDO . . . . .	i
LISTA DE FIGURAS . . . . .	iii
LISTA DE TABELAS . . . . .	v
GLOSSÁRIO . . . . .	vii
1 INTRODUÇÃO . . . . .	1
2 ESTADO DE ARTE . . . . .	5
2.1 Map Matching . . . . .	5
2.1.1 Técnicas de Map Matching . . . . .	6
2.1.2 Sistemas de geo-informação online . . . . .	12
2.1.3 Trabalhos Relacionados . . . . .	13
2.2 Similaridade de Trajectos . . . . .	15
2.2.1 Coeficiente de similaridade de Jaccard . . . . .	16
2.2.2 Métrica de deslocamento de terra (Earth Mover's Distance) . . . . .	16
2.2.3 Distância de Hausdorff . . . . .	16
2.2.4 Trabalhos Relacionados . . . . .	16
2.3 Integração e Armazenamento de Dados . . . . .	18
2.3.1 Data Warehousing . . . . .	19
2.3.2 Data Warehouse Espacial . . . . .	21
2.4 Agrupamento de Trajectórias . . . . .	26
2.5 Sumário . . . . .	27
3 ESPECIFICAÇÃO E MODELAÇÃO DO SISTEMA . . . . .	29
3.1 Análise de requisitos . . . . .	30
3.2 Desenho Conceptual . . . . .	31
3.3 Desenho Lógico . . . . .	35
3.4 Desenho Físico . . . . .	35
3.5 Processo de Extracção, Transformação e Carregamento de Dados . . . . .	37
4 RESULTADOS . . . . .	47
4.1 Casos de estudo . . . . .	47
4.1.1 Caso de estudo 1: Aalborg, Dinamarca . . . . .	47
4.1.2 Caso de estudo 2: Pequim, China . . . . .	48

4.2	Map Matching . . . . .	49
4.3	Análise dos Trajectos . . . . .	52
4.4	Discussão . . . . .	58
4.4.1	Testes de similaridade . . . . .	58
4.4.2	Análise das rotas . . . . .	59
5	CONCLUSÃO E TRABALHO FUTURO . . . . .	61
	REFERÊNCIAS . . . . .	63

# LISTA DE FIGURAS

---

2.1	Estratégia ponto a ponto [5]	7
2.2	Problemas da estratégia ponto a curva [5]	7
2.3	Mapeamento (CTC)	8
2.4	Mapeamento topológico	9
3.1	Data Mart das Viagens	33
3.2	Data Mart dos dados posicionais	33
3.3	Data Mart dos troços	34
3.4	Esquema em constelação	35
3.5	Conversão Inicial de um desenho lógico para uma estrutura física	36
3.6	Processo de Extracção, Transformação e Carregamento	38
3.7	Exemplo de troços e respectivas intersecções antes da fase de segmentação	40
3.8	Exemplo de troços e respectivas intersecções após a fase de segmentação	41
3.9	Exemplo dos buffers criados à volta dos troços, após a fase de segmentação	42
3.10	Exemplo da execução de Map Matching	44
3.11	Problema de ciclos no Map Matching	45
4.1	Trajectos registados na Dinamarca, maioritariamente em Aalborg	48
4.2	Trajectos registados da cidade de Pequim, China	49
4.3	Nível similaridade dos trajectos resultantes - 1 a 16	51
4.4	Nível similaridade dos trajectos resultantes - 18 a 37	52
4.5	Par de intersecções A e B da análise 1	53
4.6	Rotas utilizadas com maior frequência	54
4.7	Par de intersecções A e B da análise 2	54
4.8	Rotas utilizadas com maior frequência	55
4.9	Par de intersecções A e B da análise 3	56
4.10	Rotas utilizadas com maior frequência	56
4.11	Duas rotas atípicas de dois pontos de intersecção	57
4.12	Dados posicionais de taxis em Pequim	59

# LISTA DE TABELAS

---

4.1	Resultados dos testes de Map Matching . . . . .	50
4.2	Média dos níveis de similaridade . . . . .	51
4.3	Distância máxima, mínima, média, mediana e desvio padrão da análise 1 . . . .	54
4.4	Distância máxima, mínima, média, mediana e desvio padrão da análise 2 . . . .	55
4.5	Distância máxima, mínima, média, mediana e desvio padrão da análise 3 . . . .	56

# GLOSSÁRIO

---

**DFM** Dimentional Fact Model

**DW** Data Warehouse

**ETC** Extração, Transformação e Carregamento

**GPS** Global Positioning System

**GPX** GPS eXchange

**MM** Map Matching

**MOLAP** Multi-dimensional Online Analytical Processing

**OLAP** Online Analytical Processing

**OLTP** Online Transaction Processing

**OSM** OpenStreetMaps

**plpgsql** Linguagem Procedural SQL

**OWM** Open Weather Maps

**PBF** Protocol Binary Format

**QGIS** Quantum GIS

**SIG** Sistemas de Informação Geográfica

**SOLAP** Spatial Online Analytical Processing

**SQL** Structured Query Language

**UML** Unified Modeling Language

**XML** eXtensible Markup Language

**WS** Web Service

# INTRODUÇÃO

---

Os sistemas modernos de geo-localização têm permitido avanços tecnológicos significativos na área dos Sistemas de Informação Geográfica (SIG). O aumento da recolha de dados foi uma das grandes consequências dessa evolução tecnológica e a gestão eficiente de dados espaciais tem sido um problema analisado pelos investigadores. Alguns Sistemas de Gestão de Base de Dados suportam dados espaciais (por exemplo, PostgreSQL ou Oracle), disponibilizando métodos de indexação espacial e um conjunto de operações para a análise e processamento de dados em SIG.

Tal como os SIG, os sistemas de armazenamento de dados de grandes dimensões foram aperfeiçoados através de novos paradigmas que se focam na concepção de Base de Dados multi-dimensionais, orientados para a análise de dados em múltiplas perspectivas, facilitando a criação de relatórios baseados em dados passados e a tomada de decisão.

O objectivo desta dissertação consiste:

- No desenho e implementação de um modelo multidimensional de dados associados ao tráfego de veículos e aos dados que representam o seu contexto;
- Na representação de dados posicionais num mapa de estradas digital;
- Na obtenção e utilização de dados externos para a valorização contextual dos dados posicionais;
- Na implementação de métodos de análise dos dados capazes de obter informação relativa ao comportamento das trajectórias dos veículos.

A partir do modelo multidimensional, deve ser criada uma Data Warehouse para o armazenamento de dados relativos a trajectórias registadas por um conjunto de veículos. Essas trajectórias devem ser criadas sob um mapa de estradas digital, onde os dados posicionais devem ser associados ao mapa através de um algoritmo de mapeamento de dados.

A análise será focada na verificação de múltiplas trajectórias com origens e destinos comuns, ou seja, verificar a existência de diferentes rotas a partir de uma determinada origem e de um destino e caso existam, verificar se as rotas mais utilizadas são as mais rápidas, as mais curtas ou outras, procurando padrões nas selecções das rotas.

A análise das trajectórias deverá ser feita através de um cubo de dados. Esse cubo de dados deverá ter a capacidade de obter informação necessária da Data Warehouse (DW) para fazer as análises

mencionadas anteriormente, tais como a apresentação das rotas mais rápidas entre uma origem e um destino, as rotas alternativas e a taxa de utilização de cada rota.

Para fazer o armazenamento das trajectórias, foram criados um conjunto de ferramentas para a extracção e limpeza de dados Global Positioning System (GPS). Através do PostgreSQL foi criada a Data Warehouse e o Cubo de Dados necessários para a análise das trajectórias. Essas ferramentas serviram para a criação de trajectos e inserção de dados nas DWs.

Todas as ferramentas foram cruciais para o desenvolvimento da Data Warehouse e do Cubo de Dados e para a análise das trajectórias.

Esta Dissertação está organizada em 6 capítulos:

**1 - Introdução** Na introdução é feita uma contextualização dos assuntos abordados nesta dissertação através da descrição actual do estado das tecnologias de geo-localização, bem como do armazenamento de dados espaciais em modelos multidimensionais.

De seguida são apresentadas as tarefas principais que foram delineadas no início desta dissertação e posteriormente, uma breve explicação dos passos utilizados para o desempenho das tarefas.

**2 - Estado de Arte** Neste capítulo são apresentados trabalhos que estão relacionados com os assuntos abordados nesta dissertação.

Inicialmente, é apresentado o conceito de Map Matching e as várias técnicas de mapeamento utilizadas ao longo do tempo.

De seguida são apresentados os Sistemas de geo-informação mais utilizados actualmente. Os sistemas analisados foram o Google Maps, Microsoft Bing Maps, Yandex Maps e OpenStreetMap. Todos os sistemas foram analisados usando o mesmo conjunto de critérios, de forma a facilitar a comparação entre os sistemas. Os critérios foram o nível de cobertura do sistema, a qualidade dos serviços de criação de rotas, a qualidade de serviços de Geocoding e restrições ou limitações encontradas.

De seguida apresenta-se um conjunto de métricas de similaridade utilizadas para medir a semelhança entre formas geométricas. Estas métricas foram analisadas com o objectivo de, após a associação dos dados posicionais ao mapa de estradas digitais, possibilitar a análise e validação dos resultados mapeados. As métricas analisadas foram o coeficiente de similaridade de Jaccard, a métrica de deslocamento de terra e a Distância de Hausdorff.

De seguida é feita a apresentação do conceito de Data Warehousing. Após a apresentação do conceito, são apresentados os tipos de abordagens existentes para a modelação do sistema, e as fases de desenvolvimento inerentes à criação e gestão dos dados neste tipo de paradigma. Após a apresentação de trabalhos relacionados com Data Warehousing, é apresentado o conceito de Data Warehousing espacial e os modelos dimensionais abordados até à data. Depois são apresentados trabalhos relacionados com este paradigma, havendo uma secção específica para os trabalhos relacionados com o armazenamento de trajectórias.

A secção seguinte apresenta trabalhos relacionados com o agrupamento de trajectórias, utilizando modelação multidimensional.

No final deste capítulo, é apresentado uma breve análise aos trabalhos feitos previamente e a contribuição que este trabalho acrescenta aos trabalhos apresentados.

**3 - Problema** Neste capítulo apresentam-se as soluções propostas durante a dissertação.

Inicialmente é feita uma análise de requisitos para uma posterior organização eficiente dos dados no sistema e conseqüentemente, para uma melhor análise.

Após a análise de requisitos, procedeu-se à criação do modelo conceptual do sistema. Este

modelo permite fazer a apresentação e a descrição geral do sistema. Através deste modelo, é possível perceber o funcionamento do sistema e de como os requisitos definidos anteriormente são abordados. O modelo foi representado usando a notação Dimensional Fact Model (DFM) [1]. De seguida, foi implementado o desenho lógico do sistema. Este desenho é obtido a partir do modelo conceptual criado anteriormente. O modelo lógico permite determinar o tipo de armazenamento de dados mais apropriado para este sistema.

A partir do desenho lógico, foi implementado o modelo físico do sistema. O modelo físico descreve a forma como os dados são armazenados, as estruturas de armazenamento e as funcionalidades de que permitem um acesso eficiente aos dados.

Depois foi delineado o processo de Extração, Transformação e Carregamento (ETC) dos dados relativos às posições dos veículos e ao contexto. O processo é descrito em detalhe com o auxílio de um diagrama, falando da limpeza dos dados, passando pela associação dos dados posicionais ao mapa digital e o carregamento dos dados.

Durante o carregamento dos dados, são realçadas as fontes externas utilizadas para definir o contexto dos dados.

#### **4 - Resultados** Neste capítulo são apresentados os resultados obtidos através das resoluções propostas.

Inicialmente, são apresentados os resultados relativos ao mapeamento dos dados posicionais no mapa digital usando diferentes períodos de amostragem. Esses trajectos são comparados com o trajecto original, através de uma medida de similaridade.

De seguida é apresentado uma análise de trajectos em diferentes tipos de zonas numa cidade. Essa análise é feita através de um conjunto de consultas capazes de obter informação tal como a rota mais rápida, a rota mais curta, rotas alternativas, durações, médias, medianas das distâncias feitas e o desvio padrão.

#### **5 - Discussão** Neste capítulo faz-se uma análise aos resultados obtidos neste trabalho. São discutidos os resultados obtidos no mapeamento de 2 conjuntos de dados posicionais com características distintas. É realçado a taxa de resultados válidos e a influência das características dos dados nos resultados. Em relação à análise de trajectos, é feita uma análise aos dados estatísticos das rotas utilizadas pelos veículos, tirando ilações das rotas seleccionadas tendo em conta a zona analisada e dados relativos ao contexto.

#### **6 - Conclusão** Neste capítulo apresentam-se as conclusões da dissertação, apresentando as contribuições dadas e faz-se referência ao trabalho futuro que poderá ser feito tendo como base o trabalho apresentado.



## ESTADO DE ARTE

---

Este capítulo foca 5 tópicos principais relacionados com as tarefas propostas para este trabalho.

O primeiro tópico, Map Matching, apresenta o conceito e diferentes técnicas de mapeamento conhecidas. De seguida são apresentados Sistemas de geo-informação. Esta pesquisa foi feita com o objectivo de analisar e seleccionar o sistema com melhores mapas rodoviários digitais a serem utilizados para o mapeamento dos dados posicionais. De seguida são apresentados o resultado de uma pesquisa de trabalhos relacionados com os sistemas mencionados anteriormente.

No segundo tópico, Similaridade de Trajectos, são abordadas métricas de similaridade da associação tendo em vista avaliação dos trajectos resultantes dos dados posicionais com as estradas. Após a apresentação das métricas, são apresentados trabalhos relacionados com as métricas propostas.

No terceiro tópico, Integração e Armazenamento de Dados, é apresentado o conceito de Data Warehousing. De seguida, são apresentados vários trabalhos que utilizam este tipo de armazenamento. Esses trabalhos realçam técnicas de modelação e criação de Data Warehouses, bem como um conjunto de técnicas, estruturas e operações utilizados para o acesso aos dados (por exemplo, operações de agregação, utilização de vistas materializadas, entre outros). No mesmo capítulo é apresentado o conceito de Data Warehousing espacial e vários modelos apresentados até à data. De seguida são apresentados vários trabalhos que utilizam este paradigma.

No capítulo seguinte, Agrupamento de Trajectórias, são apresentados trabalhos relacionados com o agrupamento de trajectórias. Nestes trabalhos são analisadas as ferramentas utilizadas para o agrupamento de trajectórias, bem como a performance dos algoritmos de agrupamento mais utilizados. No último tópico é feito um pequeno sumário da análise feita aos trabalhos pesquisados nos tópicos mencionados anteriormente.

### 2.1 MAP MATCHING

As técnicas de *Map Matching* permitem o mapeamento de posições geográficas, isto é, de dados GPS relativos à posição de veículos ao longo do tempo em mapas de redes rodoviárias digitais.

Estas técnicas permitem não só estimar o segmento onde uma localização GPS foi registado, mas também o ponto dentro do segmento que melhor corresponde a essa localização. Porém, a precisão do

registo dos pontos GPS difere entre diferentes receptores, bem como nas suas características inerentes (frequência de amostragem das localizações, capacidade diferencial [2], entre outras características). Estes factores fazem com que a margem de erro do registo de uma localização GPS possa ser de 30 metros. O mesmo problema depende também da qualidade do mapa rodoviário digital [3]. No processo de digitalização, é possível que hajam estradas que não tenham sido registadas, características incompletas, por exemplo, que não tenham informação relativa ao sentido, tipo (estrada, ponte, túnel), entre outras limitações.

Estando ciente destes factores, as técnicas de *Map Matching* procuram fazer o melhor mapeamento entre estes dados, de forma a conseguir criar trajectórias rodoviárias com a maior fiabilidade possível.

As técnicas de *Map Matching* podem ser executadas em tempo real (*online*), isto é, quando um ponto GPS (ou um conjunto de pontos, dependendo do algoritmo) é registado, o algoritmo de mapeamento é executado, produzindo resultados imediatos. Quando o mapeamento é feito com dados que não estão a ser processados em tempo real, ou seja, dados obtidos antecipadamente, o processamento pode ser feito *offline*.

Enquanto a primeira abordagem dá mais ênfase à *performance* e ao tempo de execução, já que é necessário que os resultados sejam obtidos quase em tempo real, a segunda abordagem dá mais ênfase na qualidade dos resultados.

Podemos então verificar que a qualidade dos resultados de um algoritmo de *Map Matching* depende muito do tempo de execução que é necessário para obter esses dados.

### 2.1.1 TÉCNICAS DE MAP MATCHING

Os algoritmos de *Map Matching* podem seguir várias técnicas, tais como a geométrica, topológica, probabilística ou avançadas. De seguida são apresentadas algumas técnicas de *Map Matching*.

#### MAPEAMENTO GEOMÉTRICO

Mapeamento Geométrico é uma técnica de *Map Matching* que tem apenas em conta a informação geométrica das redes rodoviárias [4]. Nesta técnica, os dados relativos à forma como os segmentos estão ligados entre eles não é considerada. Esta técnica pode ser dividida em três categorias:

**Ponto-a-ponto (*Point-to-point* [5], [6])** Esta estratégia é a mais simples de implementar. Como pode-se constatar a partir da Figura 2.1, o algoritmo utilizado nesta estratégia faz corresponder o ponto  $pt$ , com o nó mais próximo do mapa rodoviário. Apesar de ser uma técnica eficaz relativamente ao seu tempo de execução, esta estratégia tem desvantagens de relevância consideráveis para a criação de uma trajectória. Uma delas é o facto de existir uma grande probabilidade dos pontos GPS corresponderem a segmentos com um número elevado de nós associado. Na Figura 2.1, podemos verificar que, se seguirmos uma correspondência relativa aos nós, o segmento mais próximo seria o  $[B^0 - B^2]$ , com o nó  $B^1$  mais próximo da posição  $pt$ , quando na realidade, é visível que o segmento mais próximo é o segmento  $[A^0 - A^1]$ .

Outra desvantagem é o facto das correspondências não terem em conta correspondências prévias, ou seja, não é mantido um historial. Isto possibilita a produção de trajectos irreais. Um caso típico em que este problema pode acontecer é no cenário de estradas paralelas, onde devido à (falta de) precisão dos registos dos pontos de localização, a correspondência desses pontos pode

ser feita em diferentes estradas paralelas. A Figura 2.2a ilustra esse caso, em que os pontos  $p0$  e  $p2$  são correspondidos a *Arc A*, enquanto que, devido a uma ligeira deslocação, o ponto  $p1$  é correspondido a *Arc B*.

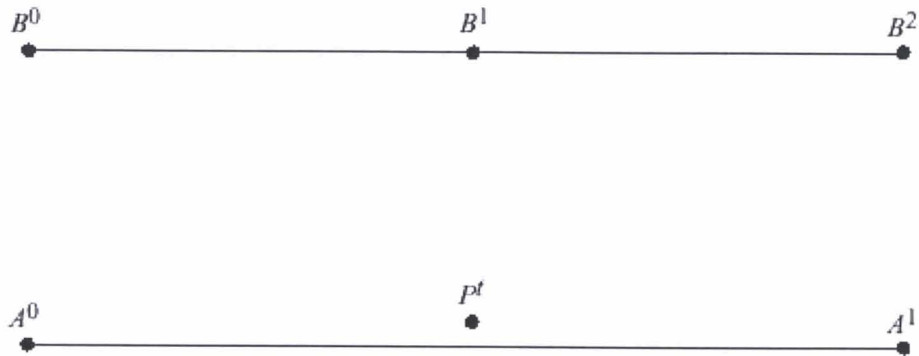
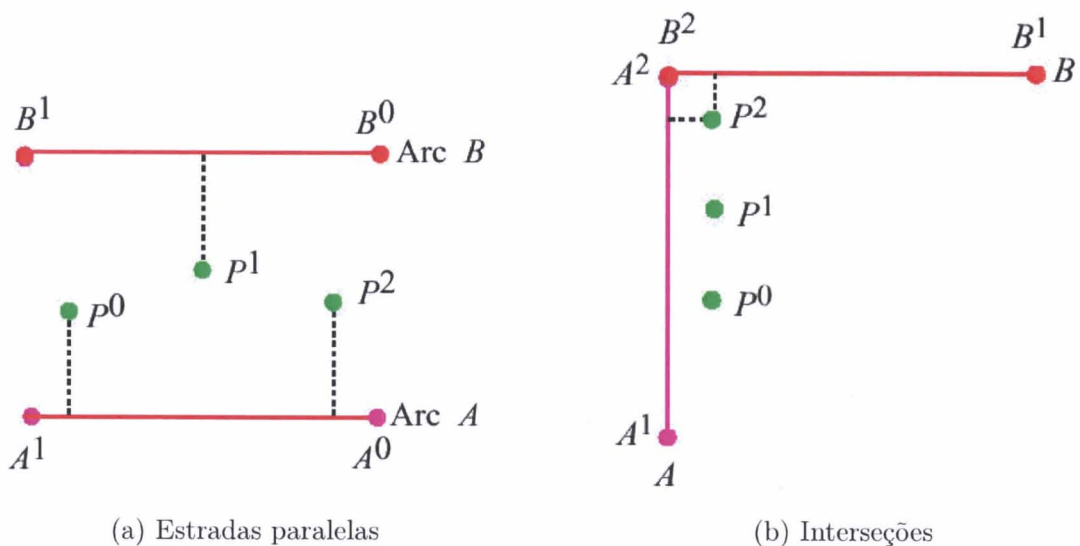


Figura 2.1: Estratégia ponto a ponto [5]

**Ponto-a-curva (Point-to-curve)** [5], [7] Esta estratégia faz a correspondência dos dados de cada ponto com o segmento mais próximo num mapa rodoviário digital. Tal como a estratégia ponto-a-ponto, o facto de não armazenar um historial de correspondências pode levar à produção de trajectos irrealis, com pontos consecutivos a serem correspondidos, a segmentos paralelos distintos (Figura 2.2a). Na estratégia ponto-a-curva também podem haver problemas de decisão em pontos que estejam próximos de nós que representem intersecções. Na imagem 2.2b, podemos ver que sem um historial de correspondências precedentes, é impossível definir qual é o melhor segmento (*A* ou *B*) para fazer a correspondência com o ponto  $p2$ .

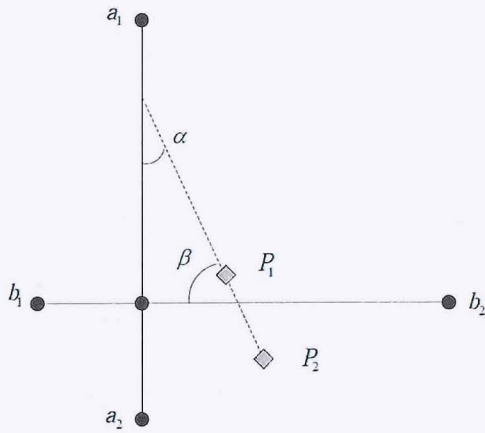
**Curva a curva (Curve-to-curve)** [5], [7], [8]) Esta estratégia faz o cálculo do ângulo mais pequeno entre a uma trajectória constituída por um conjunto de pontos com segmentos do mapa rodoviário



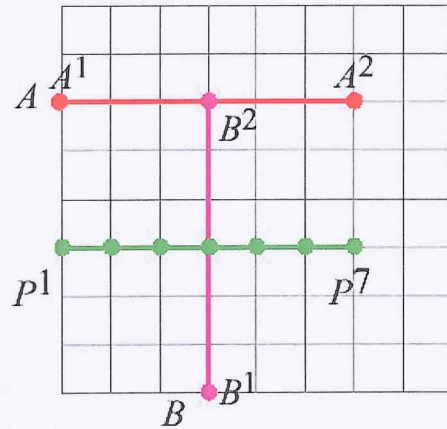
(a) Estradas paralelas

(b) Intersecções

Figura 2.2: Problemas da estratégia ponto a curva [5]



(a) Estrat\u00e9gia CTC



(b) Estrat\u00e9gia CTC relativa \u00e0 dist\u00e2ncia

Figura 2.3: Mapeamento (CTC)

digital (Figura 2.3a). Na Figura 2.3b, o conjunto de pontos  $[p1-p7]$  s\u00e3o correspondidos ao segmento A, j\u00e1 que o \u00e2ngulo entre o segmento rodovi\u00e1rio e o segmento resultante do conjunto de pontos ( $0^\circ$ ) \u00e9 menor do que o \u00e2ngulo resultante com o segmento B ( $90^\circ$ ). Apesar de utilizarem um historial de pontos precedentes, esta estrat\u00e9gia \u00e9 sens\u00edvel a valores at\u00edpicos, j\u00e1 que a na constru\u00e7\u00e3o inicial da traject\u00f3ria \u00e9 utilizada a estrat\u00e9gia ponto-a-ponto. Por isso, caso a correspond\u00eancia do primeiro ponto seja errada, \u00e9 muito prov\u00e1vel que as correspond\u00eancias seguintes tamb\u00e9m o sejam [9].

**Filtro de redu\u00e7\u00e3o de estradas (Road Reduction filter [10])** Esta estrat\u00e9gia utiliza dados relativos \u00e0 posi\u00e7\u00e3o, velocidade e sentido (*height-aiding*), obtidos atrav\u00e9s de dados espaciais do mapa de estradas e da correc\u00e7\u00e3o de pontos GPS diferenciais virtuais [11]. Com a utiliza\u00e7\u00e3o dos dados relativos \u00e0 posi\u00e7\u00e3o, velocidade e sentido, esta t\u00e9cnica melhora significativamente a precis\u00e3o do registo dos pontos GPS. Tal como na estrat\u00e9gia Curva-a-curva, esta estrat\u00e9gia \u00e9 sens\u00edvel a valores at\u00edpicos j\u00e1 que utiliza a estrat\u00e9gia ponto-a-ponto para fazer a correspond\u00eancia do ponto inicial [7].

## MAPEAMENTO TOPOL\u00d3GICO

S\u00e3o considerados algoritmos de mapeamento topol\u00f3gicos (Figuras 2.4) as t\u00e9cnicas que utilizam informa\u00e7\u00e3o relativa \u00e0s rela\u00e7\u00f5es entre formas geom\u00e9tricas [12]–[14], como a adjac\u00eancia (com pol\u00edgonos), a conectividade (com linhas) e a conten\u00e7\u00e3o (com pontos). Desta forma, \u00e9 poss\u00edvel evitar a cria\u00e7\u00e3o de traject\u00f3rias imposs\u00edveis num cen\u00e1rio real, evitando segmentos que n\u00e3o sejam alcan\u00e7\u00e1veis a partir de um determinado segmento. Isto \u00e9 evitado n\u00e3o s\u00f3 devido \u00e0 verifica\u00e7\u00e3o das adjac\u00eancias e conectividade das estradas, mas tamb\u00e9m devido \u00e0 informa\u00e7\u00e3o das caracter\u00edsticas das estradas como o sentido, a velocidade o \u00e2ngulo das curvas, entre outras caracter\u00edsticas. Na Figura 2.4a, temos um conjunto de pontos  $p1$ ,  $p2$  e  $p3$  que n\u00e3o est\u00e3o representados no mapa rodovi\u00e1rio digital. Na figura 2.4b, s\u00e3o verificados os segmentos v\u00e1lidos para o mapeamento dos pontos. Ap\u00f3s essa selec\u00e7\u00e3o, s\u00e3o analisadas as rela\u00e7\u00f5es geom\u00e9tricas e as caracter\u00edsticas dos segmentos rodovi\u00e1rios para a selec\u00e7\u00e3o do segmento onde os pontos devem ser mapeados.

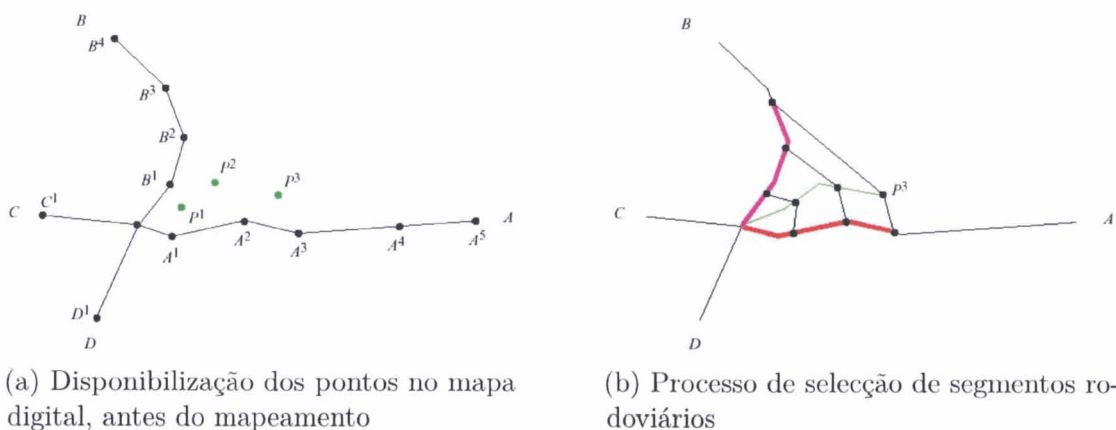


Figura 2.4: Mapeamento topológico

Em 2001 foi proposto um método de mapeamento que utiliza a técnica de múltiplas hipóteses [15], de forma a determinar a estrada de uma forma probabilística. A técnica de múltiplas hipóteses obtém múltiplos trajectos possíveis numa determinada área sem uma ordem aparente, através de uma função de similaridade. Para a execução da técnica de múltiplas hipóteses, é necessário fazer o cálculo de medidas (*pseudo-measurements*), medidas essas geradas através de informação topológica das estradas, tais como as ligações entre segmentos, direcção e informação inerente aos segmentos. Para aumentar a qualidade dos resultados de mapeamento provenientes dos erros dos mapas digitais, é aplicado um filtro de *Kalman* para estimar esses erros. Segundo os autores, os resultados experimentais de campo mostraram que a técnica de mapeamento proposto tem performances consistentes não só em áreas simples, mas também em áreas complexas como em áreas urbanas centrais, áreas com estradas cruzadas e áreas com estradas adjacentes paralelas.

Em 2002, foi proposto um algoritmo baseado em pesos para uma rede rodoviária topológica que faz apenas a utilização de dados referentes às coordenadas das posições GPS [4], descartando informação relacionada com sentido e velocidade, sendo estes valores calculados. Como é sensível a valores atípicos, a determinação do sentido dos pontos pode ser incorrecta.

Em 2003, foi publicada uma versão melhorada [16] do algoritmo mencionado anteriormente [4], utilizando critérios semelhantes uma vez que o tempo de execução é linear relativamente à dimensão do conjunto de dados, é recomendado a utilização do algoritmo com conjuntos de dados de pequenas dimensões, de forma a que o tempo de processamento não seja elevado.

Em 2004, foi publicado um algoritmo [17] de mapeamento que se baseia apenas nas coordenadas GPS e na topologia das estradas. Segundo os autores, o algoritmo foca-se principalmente no mapeamento *offline* de conjuntos de dados de grandes dimensões. Através de vários testes, os autores demonstraram a eficiência do algoritmo em termos da exactidão dos resultados e da velocidade de processamento.

Em 2006, foi apresentado um algoritmo [18] baseado na análise topológica de uma rede rodoviária. O algoritmo faz a correlação entre a trajectória dos veículos e as características da estrada (curvatura, ligações, pontos de viragem). O algoritmo foi implementado utilizando dados de navegação adquiridos de GPS, técnicas de navegação estimada (*Dead Reckoning*), e dados espaciais da rede rodoviária, incluindo informação referente a pontos de saída e entroncamentos, de forma a melhorar a performance. Este algoritmo não foi aconselhado para o processamento de dados em tempo real, devido ao facto de, nos casos em que hajam intersecções, o algoritmo necessita de fazer um pós-processamento de forma a identificar a ligação correcta.

Em 2007, foi apresentado um algoritmo [19] de mapeamento que procura uma sequência de segmentos rodoviários que correspondam aos pontos de veículos a serem mapeados e que sejam transversais em intervalos de tempo associados aos pontos GPS. Inicialmente, o algoritmo utiliza uma variação do mapeamento topológico para obter rotas rodoviárias candidatas ao mapeamento de um conjunto de pontos. Na selecção de segmentos rodoviários, é utilizado o modelo oculto de Markov para eliminar transições entre segmentos que não são adjacentes. Ao obter as rotas candidatas, são comparados os dados temporais relativos pontos com dados temporais relativos aos segmentos rodoviários, ou seja, os tempos em que os segmentos são passíveis de serem percorridos por um veículo. Essas medidas temporais são obtidas através de um software de *routing*. O algoritmo foi treinado e testado com dados obtidos de um conjunto grande de condutores reais e segundo os autores, o nível de exactidão dos resultados dos mapeamentos são positivos.

Em 2010, foi proposto um algoritmo [20] iterativo de mapeamento baseado em votos (*Voting-based Map Matching*). Na produção de resultados, este algoritmo tem em conta três pontos:

- i O contexto posicional dos pontos GPS, bem como a informação topológica da rede de estradas;
- ii A influência mútua entre pontos GPS. O resultado do mapeamento de um ponto é uma referência para as posições dos pontos vizinhos. Então, ao mapear esses pontos, o ponto mapeado previamente também será referenciado.
- iii O peso da influência mútua entre os pontos GPS calculado através da distância entre os pontos GPS. Quanto maior for a distância, menor será o peso da influência mútua.

Segundo os autores, o algoritmo não só tem em conta a informação espaço-temporal das trajectórias mas também uma estratégia de voto para a modelação das influências mútuas entre pontos GPS. Para a avaliação do algoritmo, foi utilizado um conjunto de trajectórias reais feitas por utilizadores.

## MAPEAMENTO PROBABILÍSTICO

Este tipo de abordagem foi proposto pela primeira vez em 1989 [21] e utiliza uma área de confiança, elíptica ou rectangular, fixa à volta dos pontos fixos obtidos pelo GPS, fazendo a correspondência entre os esses pontos e os segmentos contidos na área de confiança, através de um sensor de navegação estimada (*Dead reckoning*). Um sensor de navegação estimada calcula a posição actual através da posição anterior, a direcção do movimento e pela distância percorrida.

Em 1997 foi proposto que o tamanho da área de confiança seja determinado tendo em conta os erros provenientes dos dispositivos GPS [22]. Essa área é depois sobreposta na rede rodoviária para fazer a identificação do segmento onde o ponto foi registado. Caso a área de confiança contenha múltiplos segmentos, são então tidos em conta dados referentes ao sentido da viagem, ligação entre segmentos, proximidade, velocidade do veículo e a distância para o junção mais próxima.

Em 2004, é apresentada uma versão melhorada dos algoritmos apresentados anteriormente [3]. Este algoritmo cria a área de confiança só quando o ponto fixo estiver perto de uma junção, ao contrario dos anteriores que criam para todos os pontos fixos, independentemente da sua localização em relação aos segmentos. Este detalhe faz com que o processamento do algoritmo seja mais rápido, já que essa área não é criada quando os pontos estão numa ligação, o que é o cenário mais comum. Este algoritmo também leva em conta critérios baseados em estudos empíricos para a detecção de viragens em junções, de forma a identificar de uma forma mais precisa a mudança de um veículo entre segmentos. Este algoritmo também tem em conta o erro relativo ao sentido da viagem dos veículos proveniente dos

equipamentos de captura quando a viagem é feita em velocidade lenta. Nestes casos, é feita uma estimativa óptima da localização do veículo. Estes casos podem ser aplicados em zonas urbanas onde o tráfego é lento com paragens frequentes.

## MAPEAMENTO AVANÇADO

Este tipo de abordagem utiliza conceitos aperfeiçoados para o processamento das correspondências, tais como os filtros de Kalman [23]–[28], a teoria de Dempster–Shafer [29], [30], modelos de espaço de estados com filtros de partículas [31], modelos de interacção múltipla [32], modelos de lógica difusa [22], [28], [33]–[36] ou a aplicação da inferência de Bayes [15], [31]. De seguida são mencionados trabalhos relativos a esses conceitos.

Em 2002 é apresentado um sistema de navegação integrado composto por um dispositivo GPS, um sensor de navegação estimada e um algoritmo de correspondência aplicáveis em sistemas de transporte inteligente. Inicialmente, é aplicada uma estratégia ponto a curva, de forma a encontrar o segmento correto. De seguida, através de uma projecção ortogonal, é determinada a posição da localização do veículo no segmento.

Para corrigir o erro relativo ao desvio lateral, é aplicado um filtro de Kalman para determinar a posição do veículo na estrada. A precisão desse cálculo depende da qualidade da rede rodoviária e da representação das curvaturas.

Note-se que, como é inicialmente aplicado a estratégia ponto a curva, não é certo que o segmento escolhido seja o correto devido à sua sensibilidade a *outliers*, especialmente em cenários urbanos. Por isso, a precisão dos resultados do filtro irá sempre depender da precisão dessa selecção.

Em 2002, Gustafsson et al. criaram uma solução para o posicionamento, navegação e rastreamento de partículas através de um filtro baseado no método recursivo de *Bayes* [31]. O argumento necessário para o processamentos do algoritmo é a velocidade. O artigo afirma que com a utilização deste filtro, uma eventual correspondência incorrecta na fase inicial é corrigida com uma margem de erro de um metro com a utilização deste filtro.

Inicialmente, a primeira posição fixa é registada pelo utilizador através de um dispositivo GPS ou outro sistema. A área de correspondência do mapa rodoviário deve cobrir cerca de dois quilómetros, para que a performance do algoritmo seja boa. Este algoritmo mostrou bons resultados em áreas abertas, dependendo sempre da qualidade dos dados obtidos do dispositivo GPS. No que toca a cenários urbanos não foram publicados quaisquer resultados.

Em 2003, Gui e Ge criaram um algoritmo que visa resolver a correspondência das posições fixas em cenários urbanos [32], onde existem elementos como árvores ou prédios que podem bloquear os sinais para os dispositivos GPS. Assumindo que ao entrar numa área urbana o segmento de entrada é conhecido, o algoritmo determina a partir daí as posições do veículo. Em junções, algoritmo utiliza um método de correspondência probabilístico integrado com um filtro de *Kalman* para estimar a posição do veículo e o segmento correto para qual o veículo circula. A performance deste algoritmo não foi avaliada.

Dos quatro tipos de mapeamento apresentados, para este trabalho foi utilizado o mapeamento topológico. A escolha teve em conta o facto dos dados posicionais utilizados não terem os atributos necessários para serem usados por algoritmos de mapeamento probabilístico ou avançado (por exemplo, dados relativos à elevação e sentido do veículo). Em comparação com o mapeamento geométrico, o mapeamento topológico tem menor probabilidade de criar trajectos irrealis, utilizando informação rodoviária para o mapeamento dos dados.

## 2.1.2 SISTEMAS DE GEO-INFORMAÇÃO ONLINE

Para fazer o mapeamento das localizações, foi necessário seleccionar um Sistema de geo-informação adequado às necessidades deste trabalho [37]. De seguida são analisados um conjunto de serviços de geo-informação em termos de cobertura (área onde o serviço pode ser disponibilizado, em termos de nível de detalhe e precisão), *routing* (funções e API fornecidos a o auxílio no *routing* e na navegação), *Geocoding* (processo de conversão das de dados geográficos em coordenadas geográficas, por exemplo, latitude e longitude) e restrições/limitações.

### GOOGLE MAPS [38]

**Cobertura** Fornece cobertura mundial, com zonas do globo mais detalhadas do que outras. O nível de detalhe pode ser consultado através de uma tabela de cobertura disponibilizada pelo Google [39]. A zona com maior nível de detalhe são os Estados Unidos da América.

**Routing** Para o encaminhamento, o *Google Directions API* [40] é um serviço com capacidade de fornecer direcções consoante o tipo de navegação, através de pedidos HTTP. As direcções podem ser delineadas através de pontos iniciais e finais, bem como pela representação textual das coordenadas. Como as direcções são calculadas a partir de pontos conhecidos, este serviço não deve ser utilizado em tempo real. Para isso, são disponibilizados os serviços *JavaScript Directions* [41] e *Distance Matrix* [42].

**Geocoding** Para o *Geocoding*, existe o serviço Google Geocoding API [43]. Este serviço deve ser utilizado através de pedidos HTTP. Tal como no Google Directions API, este serviço não foi concebido para responder em tempo real.

**Restrições/Limitações** Todos os serviços mencionados anteriormente têm restrições no número de pedidos possíveis num determinado período de tempo. havendo uma diferenciação no número de pedidos entre utilizadores registados e não registados.

### MICROSOFT BING MAPS [44]

**Cobertura** Fornece cobertura mundial, com zonas do globo mais detalhadas do que outras. O nível de detalhe de cobertura pode ser consultado através de uma tabela de cobertura disponibilizada na página do Bing MSDN [45]. Existe ainda uma ferramenta de análise de cobertura que possibilita o registo de novas áreas [46]. Criada pela comunidade do *OpenStreetMaps*, esta ferramenta possibilita a comparação entre mapas Bing e mapas OSM. As áreas mais detalhadas são os Estados Unidos e a Europa Ocidental.

**Routing** A Microsoft disponibiliza o *Bing Map Routes API* [47] onde é possível criar rotas através de duas ou mais localizações através de pedidos HTTP. Os utilizadores podem escolher o tipo de transporte, bem com adicionar informação relativa a tráfego.

**Geocoding** O *Microsoft Bing Maps* fornece um API de localização capaz de fazer *geocoding* [48], bem como a sua inversão em todas as regiões disponíveis, mas com precisão variável.

**Restrições/Limitações** Para além do número de pedidos limitados, para cada rota é possível especificar apenas 25 pontos de localização.



## YANDEX MAPS [49]

**Cobertura** Fornece cobertura mundial, com grande detalhe na Rússia e na Turquia;

**Routing** As operações de *routing* são disponibilizadas através do *JavaScript Yandex Map API* [50].

O número de pontos entre o início e o fim de uma rota são ilimitados.

**Geocoding** O *Javascript API* [50] permite o acesso às operações de *geocoding* através de pedidos HTTP. A resposta dos pedidos pode ser enviada em formato XML, YMapsML ou em JSON.

**Restrições/Limitações** Para além do número de pedidos limitados, este serviço não pode ser utilizado por serviços que tenham vertentes comerciais.

## OPENSTREETMAP [51]

**Cobertura** Fornece cobertura mundial, com o nível de detalhe a variar dependendo do nível de actividade da comunidade de utilizadores numa determinada região.

**Routing** Este serviço fornece apenas os dados, sendo a comunidade responsável pela criação e manutenção de aplicações e funções de *routing*. As aplicações mais conhecidas para a criação de rotas são o *pgRouting* [52], *GraphHopper* [53], *pyRoute* [54], entre outras.

**Geocoding** As funções de *geocoding* também são criadas e mantidas pela comunidade de utilizadores.

**Restrições/Limitações** Os dados do OpenStreetMaps são de livre utilização, podendo ser modificados por qualquer utilizador. O facto destes dados serem de livre acesso, possibilita aos utilizadores a criação dos seus próprios serviços com características feitas de acordo com as suas necessidades. No entanto, a livre modificação dos mapas digitais compromete a fiabilidade dos segmentos rodoviários.

### 2.1.3 TRABALHOS RELACIONADOS

Vários trabalhos que necessitam de mapas rodoviários têm sido desenvolvidos ao longo dos anos, muitos deles ligados ao Map Matching, à apresentação de rotas, entre outros. Porém, nem todos os mapas são os mais adequados para o trabalho que se pretende fazer.

Nesta secção apresenta-se o resultado de pesquisas feitas aos trabalhos desenvolvidos baseados em diferentes mapas de estradas, de forma a ser posteriormente possível fazer uma decisão ponderada do mapa de estradas a ser utilizado para este trabalho.

Em 2014, Mattheis et. al. apresentam um sistema escalável de código aberto com a capacidade de fazer Map Matching online baseado, com um algoritmo baseado no modelo oculto de Markov [55]. Este sistema foi desenvolvido com o objectivo de fornecer serviços baseados na localização. Neste artigo, os autores mencionam os factores necessários para obter um trajectória com qualidade, tais como a determinação dos segmentos candidatos a rotas, das emissões e das probabilidades de transição. Os autores realçam a capacidade do sistema suportar dados espaço-temporais e a utilidade da integração com outros projectos de código aberto, tais como o OpenStreetMap e *Apache Foundation Software*.

Teslya apresentou um serviço de Map Matching web para uma aplicação turística móvel [37]. Este serviço é orientado ao contexto, tendo a capacidade de fornecer informação aos utilizadores antes, durante e depois de uma viagem. Este serviço permite ao utilizador usar múltiplos sistemas de

*mapping*, tais como a Google (Maps), Microsoft e a Yandex. Este serviço também fornece serviços de projectos *mapping* de código livre tais como o OpenStreetMap, Leaflet, PostGIS, pgRouting e Nominatim. Depois de uma análise aos serviços de *mapping*, o autor concluiu que a maioria dos serviços fornecem um conjunto de funcionalidades de visualização de mapas, encaminhamento, *geocoding* entre outras. Porém, muitas dessas funcionalidades são restringidas de forma livre. O autor realça que OpenStreetMap fornece os dados e funcionalidades de manipulação dos dados OpenStreetMap de forma livre, havendo a possibilidade do desenvolvimento de serviços baseados nos dados desse projecto. Depois de executado um caso de estudo, o autor verificou que os testes à performance mostraram tempos de execução baixos na utilização das funções do sistema, podendo o sistema dessa forma ser utilizado fornecer serviço de alta qualidade aos utilizadores.

Hann desenvolveu uma tese baseada na análise do efeito das auto-estradas na performance das firmas na Hungria [56], utilizando dados de cerca de 20.000 empresas entre 1992 e 2003 com várias fontes de dados SIG. Segundo o autor, o mapa de estradas utilizado foi o OpenStreetMap, Overpass Turbo e Overpass API. Com estas ferramentas, o autor obteve os mapas de estradas entre 1992 e 2003, podendo desta forma verificar os desenvolvimentos feitos nas estradas nesse período. O utilizador põe em causa a confiabilidade dos mapas de estradas, que podem ser actualizados por qualquer pessoa e essas mudanças são armazenadas sem qualquer verificação. Por essa razão, o autor fez uma comparação dos mapas de estradas fornecidas pelo OpenStreetMap com o *Google Maps*. Após a comparação, o autor detectou pouquíssimos erros nas estradas e algumas áreas pouco detalhadas, especificamente em algumas estradas de menor importância.

Hart, Sim e Urquhart apresentaram uma aplicação hiper-heurística criada para encontrar soluções rápidas e aceitáveis para o *Workforce Scheduling* (regras pré-definidas para a optimização automática de agendamentos e da utilização de recursos tais como pessoas ou veículos [57]), e problemas de encaminhamento [58]. Os autores realçam uma função interactiva utilizada através da aplicação, com a capacidade de activar 5 objectivos diferentes utilizados como pesos e de acordo com as preferências do utilizador. Segundo os autores, a aplicação utiliza uma rede de estradas, o OpenStreetMap, para calcular as distâncias entre localizações. A aplicação também utiliza o *GraphHopper* para tratar do encaminhamento. Os resultados de testes não foram publicados devido a questões de confidencialidade. Os mesmos autores, este ano, apresentaram um modelo para o problema de encaminhamento de veículos que captura várias restrições da vida real [59]. O modelo utiliza o *GraphHopper* para calcular distâncias e tempos de percurso e segue uma taxonomia proposta que se foca nesse problema e também no problema das características físicas. Segundo os autores, o modelo apresentado gera 4.800 instâncias de encaminhamento de veículos, pelo que foram todas disponibilizadas ao público. Os autores realçam que este modelo é um recurso para problemas de encaminhamento de veículos que pode ser usado como plataforma para os investigadores fazerem análise e comparação com novos métodos e soluções, aproximando assim no futuro a investigação académica à prática.

Khachay criou um navegador pedestre baseado em dados do OpenStreetMap[60]. A criação da aplicação deveu-se ao facto do autor querer investigar aplicações de navegação de código livre. O autor analisa várias aplicações livres que se baseiam em dados OpenStreetMap, tais como a *Open Source Route Machine*, *CycleStreets* e *GraphHopper*. Após a análise, o autor apresenta uma aplicação de encaminhamento desenvolvida em Java para criar trajectórias pedestres com custo mínimo e com as passagens mais seguras. Após feitos os testes, o autor menciona que o tempo de execução da aplicação pode ser reduzido significativamente caso se utilizem técnicas de *caching* em grafos previamente construídos.

Kulakov e Shabaev apresentam uma estratégia para a criação de serviço de planeamento de viagens

baseado num espaço inteligente, através da plataforma *Smart-M3* [61]. A tecnologia *Smart-M3* permite criar serviços pro-activos baseados no contexto inserido e com utilizações para múltiplos serviços. O serviço de navegação utilizado é o *GraphHopper*, juntamente com o OpenStreetMap. Os autores descrevem o problema de planeamento de viagens através de uma lista de tarefas e apresenta modelos matemáticos para as tarefas comuns. Após mostrarem cenários de utilização, possíveis fontes de dados e algoritmos utilizados para a transformação de dados, os autores afirmam que os resultados devem ainda ser trabalhados, mencionando que no futuro irão ser apresentados o modelo de desenho dos dados e a ontologia do planeamento de viagens.

Mora e Squillero analisaram a entrega de produtos lacticínios a famílias de três áreas urbanas [62]. O requisito necessário deste trabalho é que a optimização devia ser feita tendo em conta o processo de negócio utilizado pela empresa. O desafio da optimização prende-se em reduzir a distância feita pelos distribuidores e balancear a carga pelo número de viagens de forma a minimizar o número de mudanças na rede de entregas. A estratégia utiliza um algoritmo evolucionário para a ordenação das entregas e uma estratégia de múltiplos agentes para re-atribuir entregas entre rondas. Para o calculo de percursos mais curtos e mais rápidos, foi utilizado o *GraphHopper*. Um caso de estudo revelou que a distância percorrida pode ser reduzida a 19%, o desvio entre o comprimento das rondas pode ser reduzido consideravelmente com apenas 10% das entregas dos clientes terem de ser re-atribuídas entre rondas.

Em 2015, Damiani et. al. apresentam um sistema livre de planeamento de percursos inteligente, denominado por SMART-GH [63]. Este sistema utiliza dados livres, onde os utilizadores participam na colecção de dados obtidos no seu quotidiano, tais como o nível de ruído, de poluição, entre outros dados. Em relação aos percursos, o *GraphHopper* foi escolhido pelos autores devido à rapidez à eficiência na criação de percursos, tendo a capacidade de calcular o percursos mais rápido e/ou o mais curto entre duas localizações GPS. O mapa de estradas utilizado é o OpenStreetMap. Posteriormente, SMART-GH utiliza esses dados para responder a consultas feitas pelo utilizador. O sistema permite que os utilizadores façam decisões inteligentes no que toca aos percursos que estes fazem, melhorando assim a sua qualidade de vida. Os autores realçam que SMART-GH é o primeiro SIG que implementa características ecológicas à utilização comum dos SIGs.

## 2.2 SIMILARIDADE DE TRAJECTOS

Após a correspondência dos dados posicionais ao mapa digital rodoviário, é necessário avaliar a precisão desses resultados com os trajectos feitos pelo veículos. Os resultados obtidos após o *Map Matching*, foram resultantes de vários conjuntos de pontos com diferentes frequências de amostragem, de forma a que fosse verificado em que frequência de amostragem os trajectos obtidos eram diferentes dos trajectos reais. Para isso, foi necessário pesquisar por uma métrica de similaridade adequada para comparar os trajectos resultantes dos dados posicionais com diferentes períodos de amostragem, de forma a verificar os níveis de similaridade.

### 2.2.1 COEFICIENTE DE SIMILARIDADE DE JACCARD

O Coeficiente de similaridade de Jaccard [64], [65], ou índice de Jaccard, é um método estatístico apresentado por Paul Jaccard utilizado para fazer a comparação da similaridade entre dois objectos que podem conter um conjunto de amostras finito. O coeficiente é calculado através da seguinte equação:

$$d_J(A, B) = \frac{\sum_i \min(a_i, b_i)}{\sum_i \max(a_i, b_i)}, a_i, b_i \geq 0.$$

Em que A e B representam vectores constituídos por pontos reais.

O seu complemento é conhecido por Distância de Jaccard, representando a dissimilaridade entre dois objectos.

### 2.2.2 MÉTRICA DE DESLOCAMENTO DE TERRA (EARTH MOVER'S DISTANCE)

A distância de deslocamento de terra, matematicamente conhecida por métrica de Wasserstein, é um método utilizado para calcular a similaridade entre duas distribuições. Dadas duas distribuições, em que uma representa terra correctamente espalhada num dado espaço e a outra representa um conjunto de “buracos” nesse espaço, este método mede o trabalho mínimo necessário para tapar todos os “buracos” com terra. O custo do trabalho é o produto da quantidade de terra movida e a distância pela qual é feita o movimento.

Esta medida de similaridade é muito utilizada na área da recuperação de informação [66], bem como no reconhecimento de padrões [67].

### 2.2.3 DISTÂNCIA DE HAUSDORFF

A Distância de Hausdorff [68] é uma métrica utilizada para a medição da semelhança entre dois subconjuntos. Apresentada por Felix Hausdorff em 1914, este método calcula a maior distância possível entre um ponto de um dos conjuntos ao ponto mais próximo do outro conjunto.

A definição da Distância de Hausdorff,

$$d_H(A, B)$$

é dada pela seguinte fórmula:

$$d_H(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b) \right\}$$

Onde *sup* representa o supremo, *inf* representa o ínfimo, A e B representam dois subconjuntos não vazios pertencentes a um espaço métrico  $(M, d)$ .

### 2.2.4 TRABALHOS RELACIONADOS

Em 1993, Huttenlocher, Klanderman e Rucklidge fizeram um estudo relativamente à comparação de imagens utilizando a Distância de Hausdorff [69]. Os autores começaram por discutir a computação da Distância de Hausdorff sob uma tradução eficiente em imagens binárias, comparando um modelo *bitmap*

de tamanho 32x32 com uma imagem bitmap de tamanho 256x256 numa fracção de segundo, através do computador *SPARCstation 2*. Os autores verificaram que a Distância de Hausdorff sob tradução é similar em vários casos com a correlação binária. Os autores também verificaram que o método é mais tolerante a perturbações nas localizações dos pontos do que a correlação binária devido ao facto do primeiro medir a proximidade ao invés de medir a super-posição exacta. Verificou-se também que o cálculo da Distância de Hausdorff parcial produz bons resultados em casos que a correlação falha.

Em 1994 Dubuisson e Jain apresentaram 24 versões de cálculos de similaridade baseados na Distância de Hausdorff para o mapeamento de objectos [70]. Os autores realçam que os cálculos podem ser utilizados para o cálculo parcial de dois objectos. Através testes feitos em imagens sintéticas com vários níveis de ruído, os autores concluíram que a versão 22, denominada de Distância Modificada de Hausdorff, obteve as melhores performances no mapeamento de objectos. A equação da Distância Modificada de Hausdorff é a seguinte:

$$f(d(A, B), d(B, A)) = \frac{N_a d(A, B) + N_b d(B, A)}{N_a + N_b}$$

Em que N representa o número de pontos num determinado conjunto,  $d(A, B)$  e  $d(B, A)$  representam um cálculo generalizado de Hausdorff:

$$d(A, B) = \frac{1}{N_a} \sum_{a \in A} d_H(a, B)$$

Várias vantagens deste cálculo em relação aos restantes apresentados são demonstrados pelos autores.

Em 1996, Helmut e Guibas utilizaram várias técnicas (incluindo a Distância de Hausdorff) para a medição de similaridade, da distância entre formas geométricas e para o cálculo de aproximações e interpolações entre elas [71]. Os autores focaram-se em técnicas baseadas em geometria computacional que foram criados para mapeamento de formas, simplificação e *morphing*.

Em 2004, Baudrier et. al. apresentam uma nova forma de cálculo da similaridade entre imagens [72]. Tipicamente, o processo de medição é feito através de uma análise de cada imagem que resulta numa assinatura. A similaridade das imagens é feita através da comparação dessas assinaturas. A proposta feita pelos autores não necessita de um conhecimento prévio da imagem. Segundo os autores, o processo de comparação proposto é feito da seguinte forma:

- i É feita uma análise morfológica com multi-resoluções a cada imagem;
- ii É feito um mapa distância em cada escala através cálculo da Distância de Hausdorff, restringida através de uma janela deslizante;
- iii Uma assinatura é então extraída do mapa distância e é utilizada para fazer a decisão de similaridade.

Segundo os autores, o algoritmo foi testado através de uma aplicação, utilizando uma base de dados de ilustrações antigas.

Em 2005, Hwang, Kang e Li analisam as propriedades de trajectórias similares num mapa de estradas [73]. Os autores propuseram um método que produz trajectos similares baseando-se na sua observação em medidas similares entre trajectos e o mapa de estradas. Através de resultados experimentais, os autores verificaram que este método não só é prático para a procura de trajectórias similares mas também é um bom método para fazer o agrupamento de trajectórias.

Em 2010, Kitagawa et. al. analisaram o problema de agrupamento para dados relativos a trajectórias em redes de estradas [74]. Os autores realçam o facto de vários algoritmos propostos até à

altura não terem em conta a proximidade espacial dos dados com redes de estradas. Para dar conta destes casos, os autores propõem uma nova medida de distância denominada de *NNCluster* que reflecte a proximidade das trajectórias com as estradas e também propõem um método de agrupamento eficiente que reduz o número cálculos de distâncias no processo de agrupamento. Os resultados experimentais demonstraram que o método proposto identifica correctamente agrupamentos sob a rede de estradas, reduzindo em 80 por cento o custo de distâncias.

Em 2011, Nutanong, Jacox e Samet exploraram a utilização da Distância de Hausdorff [75]. Os autores começam por explicar o funcionamento da função generalizada e de seguida mostram a complexidade da execução da função, comparando-a com o problema do vizinho mais próximo/longínquo. Os autores de seguida apresentam uma versão da Distância de Hausdorff que se faz o exame linear de um conjunto de pontos  $X$  e utiliza um índice para ajudar na computação do vizinho mais próximo no conjunto de dados  $Y$  em cada  $x$ . Para essa versão, os autores apresentam uma solução que permite evitar o exame linear do conjunto  $X$  aplicando o conceito de procura agregada do vizinho mais próximo. Também é proposto pelos autores um método que permite a análise incremental dos índices de  $X$  e  $Y$  de forma simultânea. Para a aplicação das técnicas propostas, os autores utilizam a função da Distância de Hausdorff para a medição de similaridade entre duas trajectórias representadas por um conjunto de pontos e para a comparação da performance relativamente a abordagens tradicionais. Os resultados experimentais mostraram que o método proposto mostra melhores performances do que os métodos tradicionais em uma ordem de magnitude em termos de custo transversal e em tempo de resposta total.

Roh e Hwang apresentaram o *Trajectory Pattern Mining* (TPM) [76], um software criado para a consulta de padrões em trajectórias em redes de estradas. Na criação do software foram feitos esforços para que seja possível fazer consultas em janelas espaço-temporais para serviços baseados na localização e para suportar o espaço euclidiano sem qualquer tipo de restrições. O software suporta consultas com padrões totais, sub-padrões e padrões inversos, todos eles utilizados para a correspon [74]dências de padrões para trajectórias em redes de estradas. O software foi testado com trajectórias reais de larga escala nos três tipos de padrões.

Em 2013, Enayatifar e Salam apresentam um sistema de reconhecimento de formas bidimensionais [77]. No método proposto, é feita numa primeira fase uma detecção de arestas baseado no método *fuzzy celular automata*. Numa segunda fase as formas são agrupadas através das diferenças do grau de cada ângulo. Por fim é aplicada a Distância de Hausdorff para determinar a taxa de similaridade. Segundo os autores, os resultados de várias simulações demonstram a invariância do sistema nas operações de rotação, translação e de mudança de escala.

## 2.3 INTEGRAÇÃO E ARMAZENAMENTO DE DADOS

Os dados relativos às posições, aos trajectos feitos pelos veículos, aos segmentos rodoviários utilizados nesses percursos e toda a informação contextual devem ser armazenados de forma a que a análise seja feita de forma eficiente. Desta forma, foi feita uma pesquisa a técnicas de armazenamento de dados espaciais em *Data Warehouses*.

### 2.3.1 DATA WAREHOUSING

Uma Data Warehouse é um armazém de dados tipicamente utilizado para armazenar informação relativa a uma determinada organização de uma forma consolidada. As Data Warehouses mantêm a informação armazenada em forma de históricos. Este facto contribui no processo de tomada de decisões, através de relatórios em que é possível obter o desempenho real de um determinado negócio. Dependendo da dimensão dos dados a serem armazenados, bem como a dimensão do negócio, uma Data Warehouse pode estar subdividida em Data Marts, que armazenam subconjunto de dados. Alguns autores como Ralph Kimball defendem que a modelação de uma Data Warehouse deve seguir abordagem de "dividir para conquistar" [78], ou seja, a construção de uma Data Warehouse deve passar inicialmente pela criação de vários Data Marts para que posteriormente esses sejam integrados numa Data Warehouse. Esta abordagem é conhecida como a estratégia "*Bottom-up*". Bill Inmon defende que inicialmente deve construir-se uma Data Warehouse de forma a que toda a organização siga um modelo comum [79]. Posteriormente, deve-se construir as Data Marts baseando-se em assuntos ou em secção departamentais. Esta abordagem é conhecida por estratégia "*Top-Down*". Independentemente da estratégia utilizada, a criação de uma Data Warehouse ou Data Mart passam pelas seguintes fases:

- i Análise de Requisitos;
- ii Design Conceptual do Sistema;
- iii Design Lógico do Sistema;
- iv Processo de ETC de dados;
- v Design Físico do Sistema.

A ferramenta mais utilizada para a exploração de dados numa Data Warehouse é a ferramenta Online Analytical Processing (OLAP). Esta tecnologia permite fazer a análise de grandes quantidades de dados a partir de várias perspectivas diferentes.

Em 1992, William H. Inmon publicou "Building da Data Warehouse" [79], um livro que expõe os desenvolvimentos e estratégias utilizadas na altura em que este paradigma era pouco utilizado.

Franklin aborda o impacto da informatização de mapas em relação ao acesso a informação de negócio e governamental[80]. O autor sugere a criação de um novo campo, denominando-a de Gestão de informação geográfica. No mesmo artigo são analisadas aplicações Geographic Information Systems (GIS) que visam postos de trabalho e microcomputadores.

Em 1995, vários artigos foram publicados tendo em conta a gestão e otimização de consultas através de vistas materializadas. As vistas materializadas têm como objectivo a salvaguarda (local ou remota) de resultados de uma consulta a uma Base de Dados proporcionando um acesso mais eficiente aos dados. Tipicamente, as vistas materializadas são criadas para consultas feitas com mais frequência pelos utilizadores.

Kimball e Strehlo apresentaram num artigo um conjunto de razões nas quais Online Transaction Processing (OLTP) e os modelos de base de dados relacionais não são os mais adequados, apresentando alguns conceitos do paradigma multidimensional como soluções para o suporte à decisão [78].

Gupta, Harinarayan e Quass abordaram o processamento de consultas agregadas em ambientes de Data Warehouse "Aggregate-Query Processing in Data Warehouse Environments" [81], propondo projecções generalistas capazes de aglomerar projecções (agregações, *group by's*, entre outros) numa

plataforma comum, sendo capaz de reescrever regras para operações de agregação mais eficientes que as usadas na altura. Essas propostas foram testadas em tabelas através de vistas materializadas.

Zhuge et al. publicaram "View Maintenance in a Warehousing Environment" [82], onde abordam o problema da atualização das vistas materializadas de uma Data Warehouse nos cenários em que estas atualizam as suas fontes.

Nesse ano, Chaudhuri et al. procuraram otimizar as consultas através de vistas materializadas [83], propondo uma abordagem para o caso.

Em 2001, Goldstein e Larson apresentam um algoritmo rápido e escalável de otimização de consultas através de vistas materializadas [84], onde parte de uma consulta pode ser processada através de vistas materializadas, não havendo necessidade de ser processada na íntegra caso a eficiência não seja a melhor. Os testes foram baseados em implementações feitas em Structured Query Language (SQL).

Levy et al. abordaram em algoritmos escaláveis para a consulta através de vistas [85], tais como o algoritmo *bucket* e o algoritmo *Minicom*. Para esse trabalho foi feita uma revisão, apresentando várias aplicações do problema em questão.

Em 1996, Ralph Kimball e Margy Ross publicaram "*The Data Warehouse Toolkit*" [86]. Estando na terceira edição, este livro tem servido como ferramenta base para quem se inicia no paradigma das Data Warehouse.

Nesse mesmo ano, Harinarayan, Rajaraman e Ullman publicaram "Implementing Data Cubes Efficiently" [87], onde analisam quais os melhores conjuntos de dados que devem ser materializados em vistas de forma a aumentar a performance das consultas num cubo de dados. Vários algoritmos *greedy*, foram testados para a procura das melhores materializações.

Agarwal et al. apresentam algoritmos a serem aplicados em operações de agregação em dados OLAP e multidimensional [88].

Em 1997, Surajit Chaudhuri e Umeshwar Dayal fazem uma análise à Data Warehouse em conjunção com a tecnologia OLAP [89].

No mesmo ano, Gray et al. apresenta o cubo de dados (ou cubo simples) [90] como uma função de agregação de dados, generalizada para múltiplas dimensões, capaz de construir histogramas, tabulações cruzadas e sub-totais, tais como vemos em relatórios.

Wu et al. apresentam alguns problemas inerentes ao processo de ETC [91], especificamente na modelação e desenho do sistema, limpeza e carregamento dos dados, estruturas de indexação especiais, entre outras. Ao apresentar os problemas, o autor também propõe algumas sugestões de melhoramento.

Em 2004, Ralph Kimball e Joe Caserta publicaram "*The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*" [92], onde são abordadas várias estratégias para o processo de ETC de dados.

Song et al apresentaram HaoLap (*Hadoop based oLap*) [93], um sistema OLAP que visa a gestão de megadados. Este sistema foi concebido num modelo Multi-dimensional Online Analytical Processing (MOLAP) de forma a suportar o mapeamento de dimensões e medidas. Neste artigo, os autores apresentam vários algoritmos utilizados pelo sistema, tais como:

- i um algoritmo transversal que permite fazer operações *roll-up* na hierarquia dimensional;
- ii um algoritmo de partição e linearização que permite o armazenamento eficiente de dimensões e medidas;
- iii um algoritmo de selecção de dados que optimiza a performance das operações OLAP.
- iv a utilização do modelo de programação *MapReduce* para a execução das operações OLAP.



O sistema foi avaliado numa aplicação real e os resultados obtidos foram comparados com sistemas semelhantes (Hive, HadoopDB, HBaseLattice e OLAP4Cloud). Os autores concluíram que o sistema HaoLap apresenta melhores performances no carregamento de dados, e tem grandes vantagens na performance das operações OLAP em consultas complexas e na definição do tamanho dos dados. Os autores também realçam o facto deste sistema suportar operações dimensionais totais.

### 2.3.2 DATA WAREHOUSE ESPACIAL

Uma Data Warehouse espacial junta todas as características de um modelo multi-dimensional adoptado em Data Warehouses tradicionais, adicionando a capacidade de gestão de dados geográficos. Desta forma, verifica-se que uma Data Warehouse espacial acaba por ser uma extensão de uma Data Warehouse convencional.

Ao longo do tempo foram propostos vários modelos dimensionais para Data Warehouses espaciais. Em 1998 foi proposto um modelo [94] em que é feita uma distinção clara entre dimensões/medidas convencionais e dimensões/medidas espaciais. Nesta publicação, uma dimensão espacial contém dados capazes de serem representados geograficamente num mapa. As dimensões que contém dados que não se enquadram dentro dessas propriedades são consideradas não espaciais. Enquanto que em Data Warehouses convencionais as medidas contém dados numéricos, os autores propõem uma medida espacial capaz de conter um ponto ou uma colecção de pontos geométricos de objectos espaciais. Segundo os autores, este modelo não mantém a consistência dos dados quando a Data Warehouse é actualizada.

Com base no modelo anterior foi proposto em 2001 um modelo semelhante [95], diferenciando-se na caracterização das medidas. Neste artigo, os autores propõem três tipos de medidas:

- i A primeira medida deve conter uma ou um conjunto de formas geométricas obtidas a partir de múltiplas dimensões espaciais;
- ii A segunda medida deve conter dados resultantes de processamentos das medidas espaciais ou operadores topológicos, sendo estes guardados nas células de um cubo;
- iii A terceira medida deve conter um conjunto de referências a formas geométricas guardadas numa estrutura externa.

Em 2003, um grupo de investigadores propuseram um modelo de criação de Data Marts [96]. Após uma clarificação dos conceitos relacionados com as características das Data Warehouses, os autores fazem a descrição conceptual de um Data Mart através da ferramenta UML. No artigo, constata-se que cada Data Mart é descrito por um esquema multidimensional constituído por dimensões e medidas, que podem ser espaciais ou não-espaciais. Em relação à tabela de factos, um Data Mart pode ter uma ou várias tabelas de factos, e cada uma delas é caracterizada pelo nome, o qual nunca poderá ser modificado. Cada tabela de factos pode ser relacionada com múltiplas dimensões que podem ser espaciais ou não. Cada dimensão é caracterizada por uma classe Hierárquica, utilizada para manter um historial dos níveis da hierarquia. Os autores propõem 4 tipos de dimensões:

**Temporal** para a caracterização dos dados a serem manipulados;

**Geométrica Não Espacial** para dados não geométricos, mas passíveis de serem localizados no espaço;

**Espacial Geométrica para Não Geométrica** constituída por uma dimensão geométrica generalizada para uma dimensão não geométrica;

**Totalmente Geométrica** onde os níveis hierárquicos de uma dimensão contém dados geométricos, mantendo-se dessa forma após generalizações.

Tal como em [94], existem dois tipos de medidas: espacial e não-espacial.

Em 2008, foi proposto um modelo [97] que utiliza conceitos baseados no modelo Entidade-Relacionamento, denominado por *MultiDimER*. Os autores começam por apresentar três conceitos relativos a relacionamentos de dimensões/tabelas de facto.

Inicialmente é apresentado o conceito de níveis espaciais, onde a aplicação mantém as características espaciais a serem armazenadas, podendo ser linhas, pontos, áreas ou um conjunto dessas representações. De seguida, os autores apresentam o conceito de hierarquia espacial. Uma hierarquia deve conter pelo menos um nível espacial e pode conter diferentes estruturas, sendo a criação destas influenciadas pelo relacionamento entre as dimensões (um-para-um, um-para-muitos, entre outras). Por fim, os autores apresentam um conceito de relacionamento entre factos, que representa o relacionamento entre uma tabela de facto e as dimensões directamente ligadas a ela. Este tipo de relacionamento é considerado espacial caso pelo menos duas dimensões numa hierarquia espacial sejam espaciais.

Em relação às medidas, os autores utilizam os mesmo conceitos apresentados em [94] e [96], onde existem medidas não-espaciais e medidas espaciais.

Na representação conceptual, é utilizada a notação MADs para a descrição destes conceitos.

## MÉTODOS DE PROCESSAMENTO DE DADOS ESPACIAIS

Em 1990, Beckmann et al. apresentam o método de indexação  $R^*$ -tree [98]. Este método é baseado na optimização heurística de uma área rectangular fechada em cada nó interior e foi criado para o acesso rápido e eficiente a pontos e rectângulos. No artigo, os autores realçam dois pontos característicos deste método: o facto de suportar pontos e dados espaciais em simultâneo e o facto dos custos de implementação serem ligeiramente maiores do que outros métodos  $R$ -Trees existentes, sendo que o rácio de custo-eficiência é positivo.

Em 1993, Brinkhoff, Kriegel e Seeger analisam a performance de várias abordagens utilizadas para o processo de junções espaciais nas consultas às bases de dados, utilizando  $R$ -trees, e  $R^*$ -trees [99]. Os autores apresentam várias técnicas que visam reduzir o tempo de execução na utilização dos dois métodos, tanto em relação a operações feitas no CPU como em operações I/O. Posteriormente, eles apresentam um algoritmo personalizado em que, segundo os autores, com um *buffer* de tamanho razoável, o ganho de performance nas operações I/O é óptimo, quase correspondendo com o tempo de leitura de apenas uma das páginas de memória.

Brinkhoff, Kriegel e Seeger mostraram o quão conveniente e vantajoso é fazer o processamento de junções espaciais em plataformas de *hardware* paralelo [100], tirando partido do facto destes sistemas estarem equipados com memória virtual partilhada. Os autores apresentam um algoritmo e várias variantes, em que todas elas são executadas em três fases: (i) criação, (ii) atribuição e (iii) execução paralela de tarefas. Estas fases são executadas de forma a permitir a redução dos custos das operações de CPU e I/O. Os algoritmos são analisados, sendo mencionados suas vantagens e desvantagens. Os autores concluíram que um dos algoritmos atinge uma velocidade óptima, assumindo sempre que o tamanho de discos rígidos são suficientemente grandes.

Em 1998, Ester et. al. apresentam o problema da execução de operações OLAP em Data Warehouses. Inicialmente, os autores mostram que as técnicas de agregação utilizadas em dados não espaciais não podem ser aplicadas a dados espaciais, já que as possíveis hierarquias e agrupamentos que podem existir entre esses dados não são conhecidos na fase de modelação. Este problema foi exemplificado através de um sistema de monitorização de tráfego. Consequentemente, os autores propõem uma estrutura de dados denominada de aR-Tree, que acaba por ser uma combinação de um índice espacial com uma técnica de materialização. O índice espacial é usado para fazer a agregação dos dados usando uma estrutura hierárquica de dados baseada em rectângulos englobantes mínimos. Isto faz com que as funções de operação não necessitem de aceder directamente aos objectos, mas apenas aceder aos nós intermediários. A estrutura aR-Tree foi aplicada em simulações da vida real para testar a sua performance. No final, os autores realçam que trabalhos estão a ser feitos para que esta estrutura seja aplicada em dados espaço-temporais.

Em 2000, Wang, Zhou e Lu abordam a gestão de dados em modelos multidimensionais (espaciais e espaço-temporais) [101]. Os autores apresentam os conceitos de objetos espaço-temporais e suas aplicações, técnicas para manipulação de dados espaço-temporais em modelos multi-dimensionais tais como a indexação multidimensional, estruturas de dados e avaliação de consultas.

Os autores referem que caso os sistemas de base de dados mais recentes (orientados a objeto / relacionais) suportem eficientemente dados espaço-temporais, será um passo fulcral para a criação de extensões para operações multi-dimensionais com capacidade de explorar essas características, de forma devolver dados espaço-temporais através de consultas realizadas de forma eficiente.

Tao analisa várias tecnologias de gestão de dados espaciais aplicados a ambientes de dados urbanos [102]. Após a apresentação dos conceitos e princípios na área de Data Warehousing bem como as suas características e arquitectura, o autor propõe uma arquitectura de três camadas para a criação de uma Data Warehouse espacial apresentando potenciais problemas que podem aparecer no design e na implementação desta.

Em 2001, Merret e Han mostram uma visão geral de dos conceitos relacionados com Data Warehouse espacial, para a descoberta de conhecimento geográfico [103]. Nessa publicação, os autores realçam a importância das Data Warehouse na tomada de decisões estratégicas e na descoberta de conhecimento. Os autores afirmam que a integração eficiente de dados espaciais permite obter uma visão uniforme, com um armazenamento de dados limpos e transformados, o que facilitaria a análise multidimensional desses dados. A tecnologia OLAP também é referida como uma ferramenta de análise rápida e flexível para gerir dados espaciais multi-dimensionais. Porém, é referida a implementação complexa que essa tecnologia necessita. Alguns campos a serem investigados são no fim realçados pelos autores.

Em 2002, Yvan et al. apresentam uma ferramenta denominada de *Perceptory*, capaz de fazer uma representação de forma eficiente de múltiplos objectos geométricos e representações cartográficas [104].

Em 2003, Zghal et al. apresentam uma ferramenta denominada de CASME (*Computer Aided Spatial Mart Engineering*) [105] capaz de criar uma Data Mart espacial. Esta ferramenta permite construir um modelo multidimensional, criado utilizando os conceitos Unified Modeling Language (UML). Após criado o modelo lógico, é gerado automaticamente o modelo físico. A base de dados é criada em *Oracle Spatial*.

Nadi e Delavar analisam os principais conceitos de espaço e tempo, associados aos parâmetros mais relevantes do SIG temporal [106]. Após analisarem várias abordagens para a modelação de dados SIG temporais, é apresentado e discutido um protótipo de um SIG temporal para a simulação de tráfego. Os autores concluem que no futuro, a utilização de GIS temporais em detrimento de GIS convencionais será inevitável devido à aplicação e utilidade que eles representam.

Em 2005, Gorawski e Malczok apresentam uma lista de materialização que visa o processamento de listas de agregação grandes e um armazenamento eficiente [107]. A lista de materialização contém agregações calculadas através dos dados armazenados na base de dados e após criados, as agregações são materializadas para uso futuro. Estando a lista estruturada através de páginas, esta foi analisada de forma a obter a melhor performance possível através da configuração do número de páginas em disco, o tamanho de cada página e o número de ligações à base de dados. Os autores dizem que esta lista pode ser conjugada com estruturas de agregação tais com a aR-Tree.

No ano seguinte, os autores apresentam sistemas de Data Warehouse espaciais centralizados e distribuídos [108], utilizados para a análise e agregação de grandes quantidades de dados, propondo tipos de distribuição de dados e de carga de trabalho, bem como técnicas de indexação com a estrutura aR-Tree. Os autores justificam a utilização desta estrutura de indexação pelo facto desta armazenar as agregações em disco ao invés de utilizar intensivamente a memória, o que seria um problema na indexação de grandes quantidades de dados. Neste artigo, os autores mostraram que os sistemas de Data Warehouse distribuídos obtêm melhores performances do que os sistemas de Data Warehousing centralizados. Os autores realçam também que a materialização selectiva de partes da estrutura de índices aR-Trees aumenta significativamente a eficiência do sistema.

Em 2006, Malinowski e Zimany propõem três abordagens para a análise e captura de requisitos para a criação de uma Data Warehouse espacial. Os autores referem o facto de uma fraca análise de requisitos poder levar à criação de sistemas que são passíveis a falhas. Antes de abordarem os métodos, os autores apresentam um modelo *MultiDimER*. Este modelo permite a representação de conceptual de dados multidimensionais com apoio a dados espaciais. Os métodos são *Demand-driven*, *Supply-driven* e misto.

Em 2007, Escribano et. al. apresentaram *Piet*, um sistema que permite a integração entre o sistema GIS e operações OLAP [109]. Este sistema faz a decomposição de cada camada temática no GIS em polígonos convexos, e o processamento e armazenamento dessas camadas numa base de dados para serem mais tarde utilizados por um processador de consultas. Após o sistema ser descrito e analisado, os autores concluem que o pré-processamento das camadas GIS pode obter melhores performances em comparação com sistemas GIS que utilizem indexação baseada em R-Trees.

Em 2008, Glorio e Trujillo apresentaram uma Data Warehouse para dados espaciais [110]. O modelo apresentado define um conjunto de regras através de consultas, vistas e transformações. Essas regras permitem obter uma representação lógica do sistema de uma forma automática. O modelo proposto é implementado em ferramentas eclipse.

Huibing apresenta a criação de uma base de dados espaço-temporal desde o desenho até à implementação [111], através de uma base de dados relacional-objecto. Esta apresentação deveu-se ao facto do autor constatar que, apesar de na altura haver investigação considerável relacionada com modelação de base de dados espaço-temporais, pouco havia sido publicado no que toca às implementações das mesmas em bases de dados objecto-relacionais. Durante a demonstração, o autor apresenta a definição de um objecto espaço-temporal, um modelo generalizado de dados espaço-temporal, e a implementação de um sistema de informação espaço-temporal. Para concluir, o autor apresenta um caso de estudo onde dados espaço-temporais são utilizados no sistema proposto.

Em 2012, Aissi e Gouider apresentam um trabalho em que são analisados modelos multidimensionais espaciais e espaço-temporais [112]. Os autores, mencionam que 80 por cento dos dados utilizados para tomadas de decisões são dados espaço-temporais, concluindo assim que este tipo de dados devem ser integrados nos modelos OLAP, bem como nos sistemas de Data Warehousing. Através de um conjunto de critérios e de estudos de *benchmarking*, vários modelos foram avaliados de

forma a encontrar possíveis tendências e problemas que possam necessitar de investigação adicional. Os autores concluíram que entre os modelos multi-dimensionais avaliados [96], [104], [113]–[117], a maioria integra dados espaciais baseando em dimensões espaciais não geométricas, sendo as dimensões espaciais geométricas e mistas negligenciadas. A análise da selecção de dados que devem ser incluídos nas Data Warehouses e a sua forma de inclusão é algo que deve continuar a ser explorado, de acordo com os autores.

Kyung, Yom e Kim apresentaram uma Data Warehouse espacial, baseada num modelo multi-dimensional [118], criada para ajudar o processo de decisão na actualização de dados espaciais. O modelo criado inclui um esquema em estrela e a implementação da tecnologia Spatial Online Analytical Processing (SOLAP). Na criação da Data Warehouse espacial, um conjunto de considerações teve de ser tomado conta, tais como as consequências da adição de níveis hierárquicos nas tabelas de dimensões durante a fase de implementação, a disponibilidade dos valores de medida nas tabelas de factos, tendo em conta que a nulidade desses valores resultará em consultas nulas, descredibilizando o sistema devido retorno de resultados inúteis para o utilizador. Outra consideração que os autores tiveram em conta foi a disponibilidade dos dados espaciais.

Em 2013, Aji et al. apresentam Hadoop GIS, um sistema de Data Warehouse escalável e de alta performance capaz de executar consultas de dados espaciais em grande escala na plataforma Hadoop [119]. As consultas podem ser feitas a vários tipos de dados espaciais no MapReduce através de técnicas de particionamento espacial e de um motor de consultas espaciais denominado de RESQUE. Hadoop GIS está integrado com Apache Hive, havendo a possibilidade de executar consultas espaciais declarativas com uma arquitectura integrada. De acordo com os autores, os testes mostraram grande eficiência nas respostas e grande escalabilidade ao ser executado. A performance demonstrada está a par com outros sistemas de base de dados espaciais, sendo melhor em cenários de consultas intensivas.

## ARMAZENAMENTO DE TRAJECTÓRIAS

Para o trabalho proposto, o armazenamento de trajectórias criadas através dos dados de localização é um requisito necessário para uma posterior análise. Por isso, foi feita uma pesquisa a trabalhos que envolvem este tipo de armazenamento.

Marketos et. al. apresentaram soluções para a criação de uma Data Warehouse de trajectórias [120]. Os autores inicialmente analisam modelos de Data Warehousing tradicionais e investigam a forma de adaptação desses modelos para o armazenamento de trajectórias. Neste artigo, o trabalho foi focado em três problemas críticos para a construção da Data Warehouse. Os tópicos foram :

- i a reconstrução da trajectória a partir do carregamento dos dados do dispositivo que a originou;
- ii O processo de extracção, transformação e carregamento que adiciona as trajectórias na Data Warehouse;
- iii as agregações das medidas do cubo para a utilização da tecnologia OLAP.

A solução foi testada num conjunto de dados constituídos por cerca de 6 milhões de registos posicionais relativos a movimentos feitos por carteiros em Londres num período de um mês, com uma frequência de amostragem de 10 segundos. Os dados foram utilizados para a reconstrução e armazenamento de trajectórias. Depois da solução ter sido devidamente testada, os autores consideraram esta abordagem eficiente.

Para trabalhos futuros, os autores prometem analisar medidas úteis para Data Warehouses de trajectórias, tais como a *trajectória típica* [121], [122]. Os autores também prometem analisar as capacidades analíticas deste sistema para a aplicação de técnicas de mineração nos dados agregados na Data Warehouse de trajectórias.

Em 2014, Andersen et. al. propõem uma Data Warehouse criada em PostgreSQL para a gestão de dados relacionados com níveis de combustível e com condições meteorológicas [123], realçando o facto de que as abordagens apresentadas até a altura não consideram a combinação dos dados GPS com dados externos. Com uma tabela de facto de 3.4 biliões de registos de 16 diferentes fontes, os autores mostraram que este sistema pode ser aplicado para a análise de tráfego relacionado com o consumo dos veículos e o congestionamento das estradas.

Xie, et. al propuseram um uma estrutura de dados baseado em árvore [124] para a contagem de trajectórias através do seu mapeamento num histograma espacial com diferentes granularidades. Os autores apresentam também uma abordagem para o processamento de consultas espaciais por abrangência, consistindo na agregação de histogramas em consultas por abrangência rectangular. Segundo os autores, este método permite preservar a privacidade dos veículos através da manutenção das trajectórias agregadas, e pode ser utilizado para resolver o problema da contagem distinta, da mesma forma utilizada por Leonardi et. al [125]. Os estudos experimentais feitos pelos autores mostraram que a estrutura de dados proposta atinge grandes níveis de exactidão nos resultados das consultas e tem melhores performances do que outras abordagens baseadas em histogramas.

## 2.4 AGRUPAMENTO DE TRAJECTÓRIAS

O agrupamento (ou *clustering*) de trajectórias possibilitam análises úteis para a detecção de padrões das rotas feitas pelos veículos. Este tipo de agrupamento é tipicamente feito no espaço euclidiano, ou seja, não estão condicionadas pela existência de obstáculos ou de mapas rodoviários. Com a possibilidade de uma futura utilização deste tipo de análise, foi feita uma pesquisa de trabalhos relacionados com este tópico.

Em 1999, Gaffney e Smyth abordam o problema relacionado com o agrupamento de trajectórias semelhantes, apresentando um algoritmo de agrupamento (algoritmo *EM*) baseado em princípios metodológicos para a modelação probabilística de um conjunto de trajectórias, tratando-as como uma sequência de pontos individuais geradas por um modelo de mistura finita constituído por componentes de modelos de regressão [126]. A aprendizagem é feita sem supervisão, através do método de máxima verosimilhança. Segundo os autores, o algoritmo *EM* tem a capacidade de lidar com o problema de dados ocultos encontrado noutros algoritmos de agrupamento. O algoritmo foi feito genericamente de forma a ter a capacidade de lidar com componentes de regressão não paramétricas e com saídas multi-dimensionais. Os resultados provenientes tanto de simulações como de dados reais foram comparados com os resultados de outros algoritmos de agrupamento (*Naive K-means*, Mistura Gaussiana). Segundo os autores, o algoritmo apresentado mostrou melhores performances, comparando com os outros algoritmos.

Em 2004, Li, Han e Yang propõem abordagem para o agrupamento de dados relativos a objectos móveis, através de micro-agrupamentos [127]. Segundo os autores, a técnica de micro-agrupamento foi aplicada pelo facto a detectar padrões dos objectos espaço-temporais e de gerir grandes quantidades de dados. Os autores afirmam que através de técnicas eficientes para manter os micro-agrupamentos

em pequenas dimensões e através da identificação de colisões entre micro-agrupamentos móveis, é possível fazer a gestão dinâmica de micro-agrupamentos, sendo possível obter agrupamentos de forma rápida em qualquer instante temporal. Os resultados experimentais, mostram melhorias no tempo de execução em várias ordens de magnitude, comparando com o método *K-means*.

Em 2006, Nanni e Pedreschi propõem um algoritmo de agrupamento de dados móveis, orientado ao tempo [128]. Numa primeira fase, é aplicado um algoritmo de agrupamento baseado na densidade dos dados, focando-se apenas na distância entre as trajectórias. Depois de serem executados testes ao algoritmo, os resultados são comparados com a resultados obtidos com algoritmos *standard*. Após serem feitos os testes, é apresentada e aplicada a abordagem, denominada de focagem temporal, onde se explora as semânticas da dimensão temporal (intervalos de tempo) para melhorar a qualidade do agrupamento de trajectórias. Segundo os autores, a qualidade de agrupamentos obtidos é ótima.

Em 2009, Leonardi et. al. propuseram uma abordagem para o armazenamento e agregação de padrões provenientes de objectos móveis numa Data Warehouse de trajectórias [125], com o objectivo de permitir a avaliação rápida de padrões que ocorram numa determinada zona espacial ou num determinado intervalo de tempo. Com um sistema modelado para a exploração de um cubo de dados, os autores adicionaram à Data Warehouse de trajectórias uma medida que contém os padrões frequentes obtidos no processo de mineração de dados das trajectórias, permitindo assim a análise dos padrões em diferentes níveis e granularidade através de operações OLAP.

Os autores discutiram o processo ETC de armazenamento de padrões em cada registo da Data Warehouse tendo em conta o acesso aos dados em ordens diversas e em grandes quantidades de objectos móveis. A agregação espaço-temporal desses padrões foi realçada, sendo mencionada a necessidade de resposta a consultas multidimensionais analíticas.

Em 2014, os mesmos autores apresentaram um sistema que permite fazer a modelação de uma Data Warehouse de trajectórias [129], ou seja, dados agregados de objectos móveis. Este sistema também fornece operações OLAP visuais para a análise de dados. Os autores realçam o facto deste sistema suportar dados espaciais e espaço-temporais, sendo suficientemente flexível para lidar com objectos livres ou com movimentos limitados (por exemplo, movimentos em relação ao processamento de medidas agregadas), os autores provaram que a medida representativa do número total de visitas numa determinada área espacial com um conjunto de trajectórias pode ser processada de forma eficiente e independente da discretização do domínio espaço-temporal e da estrutura hierárquica que a Data Warehouse adopte. Esta medida também é uma boa aproximação à medida representativa do número de trajectórias numa determinada área espacial, medida esta que representa problemas no seu processamento, já que é uma medida holística. Note-se que este problema insere-se no problema da contagem distinta, muito conhecida na consulta de dados nas Data Warehouses. O sistema foi aplicado em dois cenários: na navegação de barcos no Mar Adriático e no tráfego de estradas numa área urbana em Itália. Os autores comprometeram-se no futuro a adicionar ao sistema medidas mais complexas, realçando medidas como os padrões frequentes e trajectórias representativas. Resultados desses trabalhos podem ser vistos no artigo publicado em 2009.

## 2.5 SUMÁRIO

Após a conclusão do estado de arte, foi possível constatar que existem alguns trabalhos que procuram fazer a gestão de trajectórias através de Data Warehouses [120]. Porém, com o conhecimento adquirido

através das pesquisas, trabalhos relacionados com contextualização de dados espaço-temporais usando dados externos ainda não é algo que seja consideravelmente investigado. Uma possível razão para a escassez que investigação pode ter haver com o facto de, para além de existirem poucos conjuntos de dados posicionais disponíveis, muitas fontes de dados contextuais úteis para este tipo de trabalho não estão disponíveis de forma gratuita.



# ESPECIFICAÇÃO E MODELAÇÃO DO SISTEMA

---

Neste capítulo é descrito um sistema de apoio à decisão orientado para a análise de tráfego. Os trajectos, serão obtidos a partir de históricos de dados GPS.

Para além dos dados posicionais, também serão utilizadas fontes de informação contextuais, atribuindo mais valor aos dados posicionais. Os trajectos serão analisados através de um mapa rodoviário.

## CONCEITOS

Os pontos de localização dos veículos encontram-se armazenadas em ficheiros texto. Os pontos foram obtidos através de um dispositivo GPS. Cada ponto GPS é caracterizado por um conjunto de atributos, nomeadamente as coordenadas e o instante temporal em que foi registado. Os trajectos são um conjunto de posições registadas com um intervalo de tempo mínimo de 5 minutos entre cada amostra. Cada trajeto irá ser associado a uma estrada no mapa rodoviário digital a partir de um processo de Map Matching.

O mapa rodoviário digital utilizado neste trabalho é fornecido pelo OpenStreetMaps. Neste mapa a identificação das estradas pode ser feita de várias formas, tanto com um critério individual como num critério colectivo. Por isso, foi necessário uniformizar essas identificações para que a análise das estradas percorridas seja feita de uma forma mais eficiente.

A uniformização passou por processar todas as estradas onde foram registados trajectos. Após a uniformização, foram definidos os seguintes conceitos:

- Um segmento é um troço rodoviário que começa e acaba numa interseção;
- Uma interseção é um ponto representa a adjacência entre três ou mais segmentos (cruzamento, entroncamento);

Desta forma, cada segmento pode estar adjacente a outro segmento através de uma interseção e cada ponto de um trajeto irá ser associado a um segmento do mapa rodoviário.

## 3.1 ANÁLISE DE REQUISITOS

O sistema a ser concebido tem como objectivo a análise de trajectos. Este sistema também poderá ser utilizado como base para a recomendação de percursos rodoviários baseando no contexto.

Tipicamente, os sistemas de recomendação baseados em técnicas de filtragem colaborativa recorrem a avaliações feitas por vários utilizadores a determinados itens. Essa avaliação é utilizada como uma componente preditiva, comparando a avaliação de um determinado utilizador com a de outros utilizadores com avaliações semelhantes.

Outro tipo de técnica de recomendação muito utilizado é a filtragem baseada em conteúdo, onde a predição é feita através da descrição de um conjunto de itens e as preferências históricas do utilizador. Quando um item é seleccionado, a sua descrição (tipicamente metadados) são utilizados para a recomendação de outros itens com descrições semelhantes.

Para o sistema proposto, pretende-se criar um sistema de recomendação baseado na informação contextual em que os trajectos foram registados. Este sistema também poderá vir a suportar um sistema de recomendação colaborativa, bem como um sistema de recomendação baseado em conteúdos. Consequentemente, para além de ser necessário armazenar informação relacionada com a localização dos utilizadores (posições GPS, bem como o segmento rodoviário a que as posições pertencem), será necessário armazenar informação contextual em que esses dados foram registados. O contexto associado aos trajectos estará relacionado com:

- i As localizações geográficas no contexto rodoviário em que o trajecto é iniciado e concluído;
- ii O instante temporal (data, e hora/minutos);
- iii A informação relacionada com o veículo utilizado para fazer o percurso;
- iv As condições meteorológicas e
- v A relevância da data do trajecto.

As localizações GPS, que são dados geográficos tal como os segmentos rodoviários, deverão ser armazenados juntamente com a sua componente espacial, podendo assim ser representados geograficamente através de coordenadas. Os dados geográficos são úteis para a análise do espaço geográfico, permitindo determinar relações topológicas tais como adjacências, intersecções, e cruzamentos.

Para a recomendação de trajectos, é necessário registar na Data Warehouse informação relacionada com trajectos realizados pelos utilizadores, nomeadamente o trajecto percorrido na íntegra e os pontos/segmentos iniciais e finais desse trajecto. Das viagens registadas, deverá ser necessário que sejam armazenados todos os pontos GPS, bem como o segmento rodoviário em que esses pontos foram registados.

Para criar uma solução que forneça ao utilizador o melhor trajecto a ser percorrido entre dois pontos, serão usados pesos para a informação contextual em que a viagem pode ser feita, visto que o melhor percurso vai ser determinado pelo contexto em que ele poderá vir a ser feito. Essas condicionantes devem ser representadas e armazenadas juntamente com os dados espaciais.

Os instantes temporais em que o trajecto pode ser percorrido é um factor importante na recomendação. Podemos assumir que em muitos cenários, os trajectos que contenham vias rápidas sejam os mais rápidos a serem feitos pelos utilizadores em horas mortas; o mesmo poderá não acontecer em horas de ponta em que o tráfego muitas vezes está congestionado.

Uma das condicionantes que também pode influenciar a recomendação de trajectos é o veículo utilizado e as várias componentes relacionadas, como a marca, o modelo, o tipo de veículo utilizado (Motociclo,

Automóvel, etc), o tipo de combustível utilizado (Gasolina, Gasóleo, Gás, Eléctrico, etc), o consumo de combustível por cada quilómetro e a capacidade de ocupação.

Outra condicionante que pode influenciar a selecção de trajectos são as condições meteorológicas [67], [130]–[135] e neste trabalho pretende-se fazer uma estimativa dos melhores trajectos a realizar tendo em conta o estado do tempo.

O nível de utilização de uma estrada pode ser influenciado por critérios temporais, nomeadamente o tráfego nas estradas normalmente difere entre dias úteis, fins de semana e feriados.

Numa perspectiva mais ampla, podemos até considerar que essa afluência difere em diferentes fases do ano. Por isso, os dados relativos à data das viagens devem estar agrupados por diferentes fases do ano.

## 3.2 DESENHO CONCEPTUAL

Nesta fase, procedeu-se à criação conceptual do esquema das Data Warehouses. Na Figura 3.1, apresenta-se o esquema da Data Warehouse que contém dados relativos às viagens registadas pelos utilizadores. Na Figura 3.2, apresenta-se o esquema da Data Warehouse que contém dados relativos a segmentos rodoviários onde foram registadas passagens de utilizadores. Na Figura 3.3, apresenta-se o esquema da Data Warehouse que contém dados relativos a localizações GPS registadas pelos utilizadores. Todos os esquemas foram criados usando a notação DFM[1]. Os factos são as seguintes:

**Travel** Armazenam dados relativos a viagens. Note-se que as viagens correspondem a trajectos criados a partir das posições GPS dos utilizadores, numa fase prévia ao processo de Map Matching. Esta tabela de factos irá ter associada a velocidade média, a velocidade máxima, a duração das viagens e a distância percorrida como medidas.

**Segments** Armazena todos os troços rodoviários utilizados pelos condutores através das viagens. Esta irá ter como medidas a velocidade média feita em cada segmento.

**GPS sample** Armazena todas as amostras GPS dos utilizadores. Esta irá ter como medidas a velocidade, a elevação bem como a direcção e a Diluição horizontal de Precisão (hdop).

As dimensões serão as seguintes:

**Date** Armazena dados temporais relativos à data. Esta dimensão está hierarquicamente organizada em quatro níveis: Ano, Trimestre, Mês e Dia (dia do ano, dia de semana, fim de semana, feriado).

**Time** Armazena dados temporais. Esta dimensão está hierarquicamente organizada em dois níveis: Hora e Minutos.

**Driver** Armazena dados relativos ao condutor.

**Location** Armazena dados relativos à localização do evento. Esta dimensão está hierarquicamente organizada em dois níveis: País e Cidade.

**Weather** Armazena dados relativos às condições meteorológicas (temperatura, nebulosidade, pressão, humidade, nascer do sol, pôr do sol, temperatura mínima e máxima). Esta está organizada num único nível.

**Vehicle** Armazena dados relativos ao veículo utilizado no registo do ponto GPS, num segmento ou numa viagem. Esta está organizada num único nível e terá dados relacionados com o tipo de veículo, o tipo de combustível utilizado, marca, modelo e o número de lugares.

**First Point** Armazena dados relativos a um ponto GPS registado num segmento. Esta está organizada em um único nível e é utilizada quando existirem múltiplos pontos GPS num segmento de um determinado condutor. O facto de se saber qual é o primeiro ponto, auxilia o cálculo da velocidade média do condutor num segmento.

**Point** Armazena dados relativos a um ponto GPS. Esta dimensão está organizada em um único nível e é uma tabela espacial.

**Segment** Armazena dados relativos a um segmento do mapa rodoviário. Esta dimensão está organizada em um único nível e é uma tabela espacial.

**Start Path** Armazena dados relativos a um segmento relativo ao início de uma viagem. Esta dimensão está organizada em um único nível e é uma tabela espacial.

**End Path** Armazena dados relativos a um segmento relativo ao fim de uma viagem. Esta dimensão está organizada em um único nível e é uma tabela espacial..

**Path** Armazena dados relativos a um trajecto. Esta dimensão está organizada em um único nível e é uma tabela espacial.

**Snapped Points** Armazena os pontos mapeados aos segmentos seleccionados pelo algoritmo de mapeamento.

**Day Event** Armazena um acontecimento relevante que tenha decorrido no dia em que o trajecto foi registado.

Enquanto que na Data Warehouse que regista as viagens (Figura 3.1) é possível obter informação relativa a viagens na íntegra, as restantes Data Warehouses (Figuras 3.2 e 3.3) podemos obter dados estatísticos relativos a cada segmento rodoviário e a cada localização GPS registada de forma a que, no processo de recomendação de trajectos, a selecção dos trajectos seja mais fiável. Esta Data Mart também possibilita determinar o nível de utilização de cada segmento, bem como a velocidade média em cada segmento rodoviário.

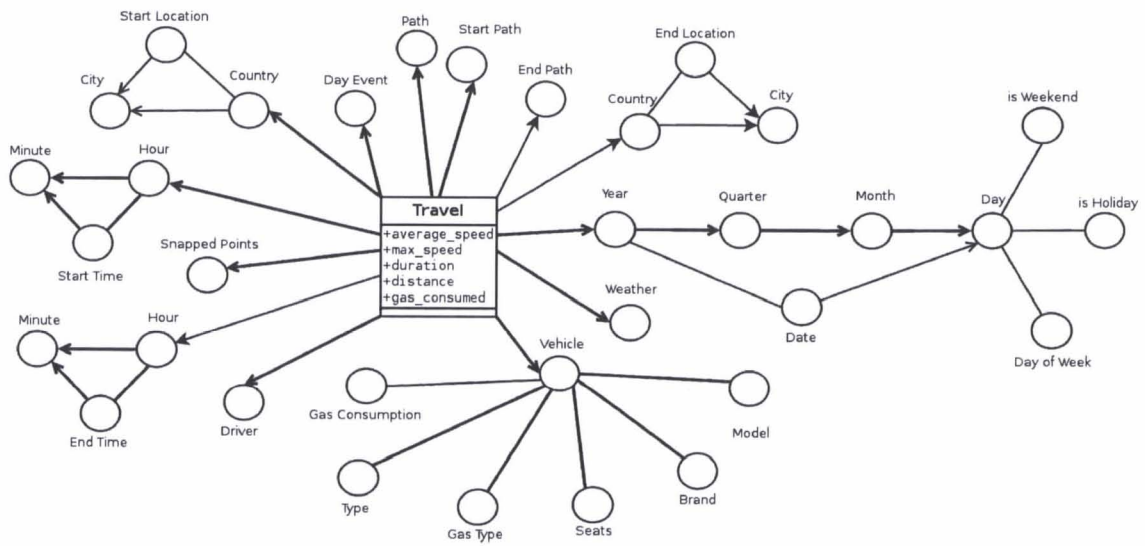


Figura 3.1: Data Mart das Viagens

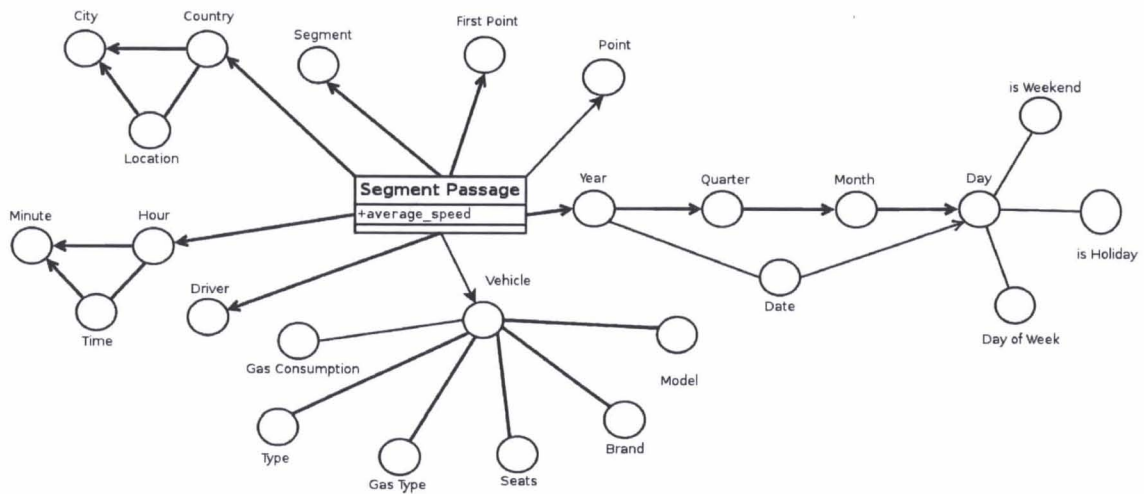


Figura 3.2: Data Mart dos dados posicionais

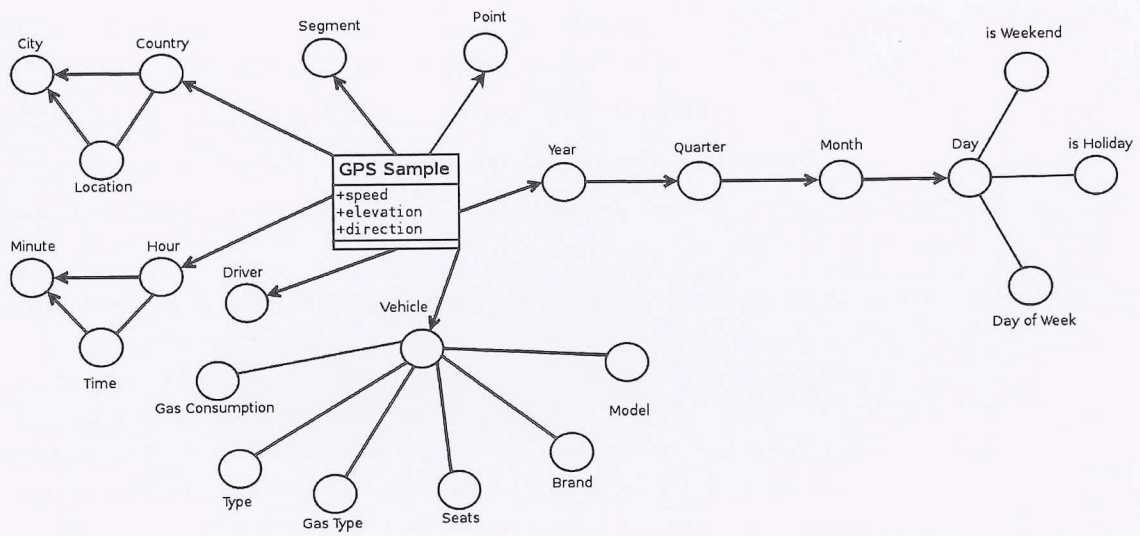


Figura 3.3: Data Mart dos troços

### 3.3 DESENHO LÓGICO

O modelo lógico apresentado na secção anterior foi implementado usando um esquema em constelação. Relativamente a outros métodos analisados (Estrela, Flocos de neve), concluiu-se que este esquema é o mais apropriado pelo facto de possibilitar a introdução de múltiplas tabelas de facto que partilham várias tabelas dimensionais (Figura 3.4).

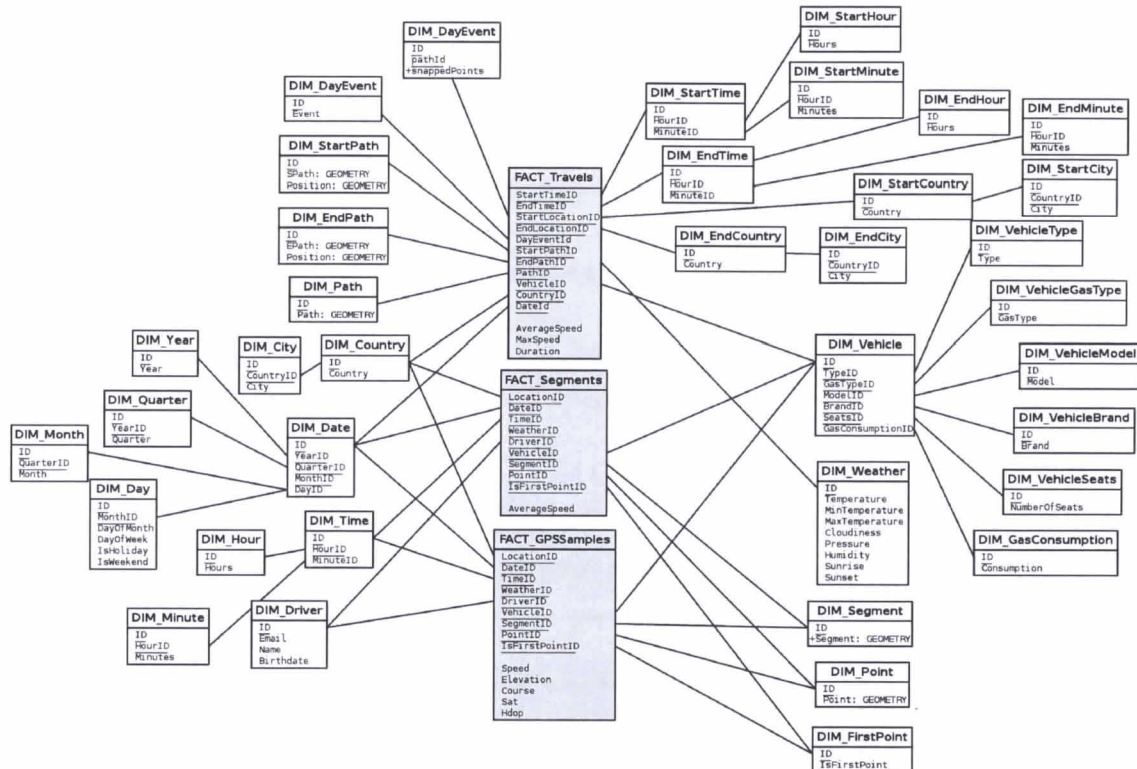


Figura 3.4: Esquema em constelação

### 3.4 DESENHO FÍSICO

Depois de concebido o desenho lógico, iniciou-se a criação do desenho físico, onde são considerados os dados recolhidos nessas fases para a construção estrutural da base de dados. A base de dados escolhida foi o *PostgreSQL*.

#### CONVERSÃO DO DESENHO LÓGICO PARA O FÍSICO

Inicialmente, foi necessário fazer um mapeamento das componentes utilizadas no desenho lógico para as respectivas componentes utilizadas num desenho físico.

Com a ajuda da Figura 3.5, mostra-se que enquanto que um desenho lógico consiste num modelo com um conjunto de Entidades, Atributos, Identificadores Únicos e Relações, num desenho físico teremos um estrutura com um conjunto de Tabelas, Colunas, Chaves Primárias e Chaves Estrangeiras.

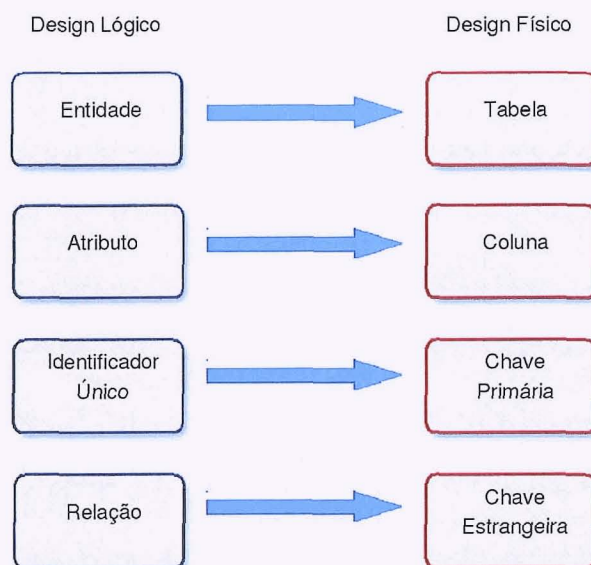


Figura 3.5: Conversão Inicial de um desenho lógico para uma estrutura física

## OUTRAS COMPONENTES ESTRUTURAIS

Após a execução mapeamento, foi necessário adicionar componentes adicionais, tais como Tabelas Particionadas, Vistas, Restrições de Integridade, Dimensões e Factos.

As Tabelas Particionadas foram utilizadas com o objectivo de dividir tabelas de grandes volumes de dados em tabelas mais pequenas.

Esta divisão aumentou a performance na execução de consultas de acesso e actualização de tabelas, já que estas são assim mais facilmente geridas relativamente à gestão de uma única tabela com uma quantidade considerável de dados. Se tivermos em conta que os dados mais consultados irão provavelmente estar numa partição ou num pequeno grupo de partições, o acesso a elas é feito com mais rapidez, já que o tamanho dos índices relativos a cada partição são mais pequenos e mais adequados para caberem em memória.

A estratégia de partição das tabelas será baseada por data de registo, com uma granuralidade mensal.

As vistas serão criadas de forma a que o acesso aos dados seja feito de uma forma menos complexa, criando tabelas virtuais através de consultas pré-definidas.

Para além de diminuir a complexidade dos dados ao criar uma visão mais lógica e compreensível para quem os acede, cria uma abstracção entre a arquitectura do sistema e os seus utilizadores, protegendo a estrutura da Base de Dados e também facilitando uma eventual actualização da mesma.

As vistas mais utilizadas foram criadas com os seguintes objectivos:

- Listar os trajectos detectados entre dois pontos de intersecção. Cada trajecto terá associada a identificação do trajecto e da viagem, a duração e a distância do trajecto, e os dados posicionais deslocados nos segmentos do mapa digital da parte do trajecto que integra a rota identificada;
- Listar o número de ocorrências em determinadas rotas, bem como a distância dessas rotas e o número de paragens médio;
- Listar rotas em função da média do tempo decorrido na realização dos trajectos;



- Listar rotas em função do número de paragens de cada trajecto;
- Listar rotas considerando critérios meteorológicos como a média e mediana da temperatura em função do nível de nebulosidade (*clear-[day/night]*, *cloudy*, *party-cloudy[day/night]*) e do tipo de precipitação (*rain/snow*).
- Listar as rotas, média e mediana da temperatura em função do tipo de precipitação. A descrição da precipitação foi encontrada nas análises feitas foram duas: chuva e neve.
- Listar o número de ocorrências de trajectos em determinadas rotas, a média e mediana da temperatura em função da nebulosidade e do tipo de precipitação;

Em relação às Restrições de Integridade, estas foram verificadas após a fase de Mapeamento do desenho lógico para o físico, através da certificação das Chaves Primárias, das Chaves Estrangeiras e do tipo de dados inerente a eles.

### 3.5 PROCESSO DE EXTRACÇÃO, TRANSFORMAÇÃO E CARREGAMENTO DE DADOS

Como podemos verificar na Figura 3.6 e como o próprio nome indica, o processo está dividido em três fases.

Essas fases são utilizadas para a execução de duas tarefas semi-paralelas: A transformação dos segmentos rodoviários e a transformação das localizações GPS em trajectos. De seguida explica-se de forma mais detalhada o processamento de cada tarefa.

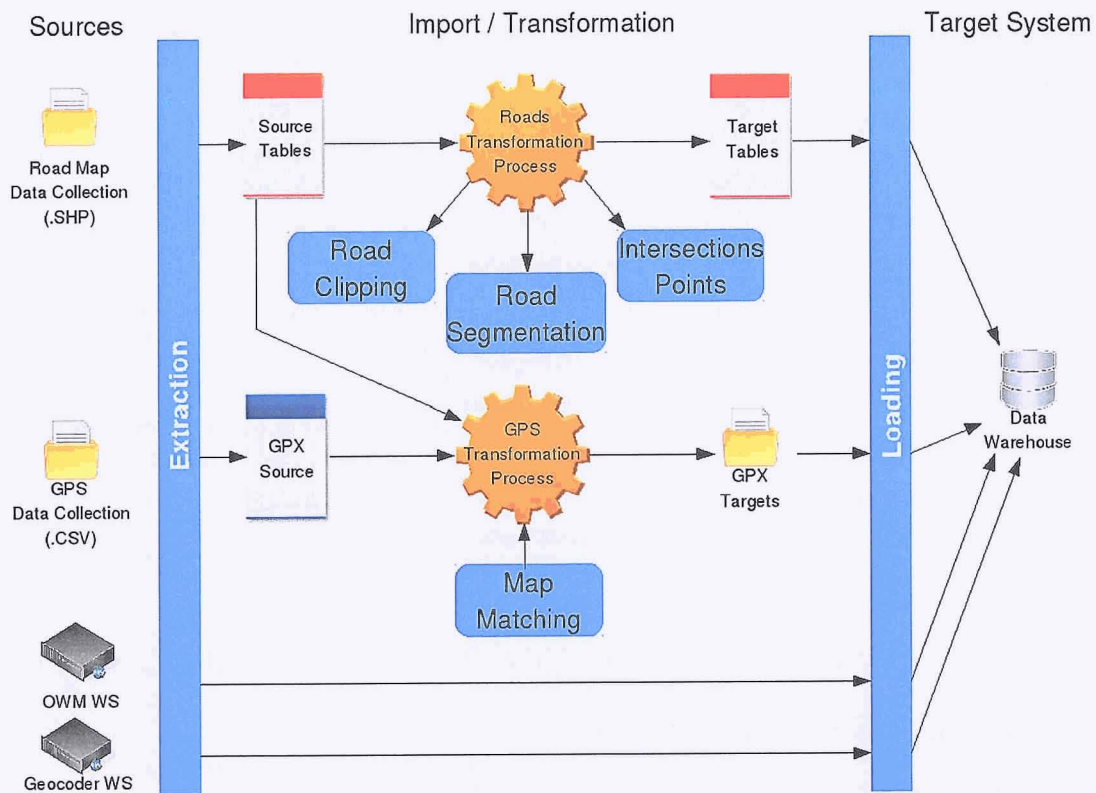


Figura 3.6: Processo de Extração, Transformação e Carregamento

## MAPAS RODOVIÁRIOS

Após o estudo realizado para a selecção do mapa de rodoviário a ser utilizado, foram escolhidos os mapas rodoviários disponibilizados pelo projecto *OpenStreetMap*.

O facto do projecto ser livre foi um motivo muito importante para a sua selecção e, entre os mapas rodoviários *open-source* analisados, este projecto tem os mapas mais usados pelos utilizadores, com padrões bem definidos e com a informação mais atualizada. Os mapas foram obtidos no portal da GEOFABRIK através de formatos *Shapefile*. Os mapas das estradas foram importados para uma base de dados na plataforma *PostgreSQL*. O processo de importação foi feito pela aplicação *pgAdmin* utilizada para a gestão de base de dados PostgreSQL. Esta aplicação tem disponível uma extensão denominada por *PostGIS Shapefile Import/Export Manager*, utilizada para fazer a extração das estradas para a base de dados, para posteriormente serem devidamente processados.

Na base de dados, cada estrada tem os seguintes conjunto de atributos associados:

### **gid**

Identificação única do dado geométrico;

### **osm\_id**

A identificação única no projecto OpenStreetMaps (OSM). Este atributo pode ser utilizado para obter mais atributos relativos a uma estrada, armazenados em ficheiros OSM;

### **name**

Localização da estrada. Tipicamente este atributo é utilizado para armazenar a morada onde a estrada está localizada;

**ref**

Descrição da estrada. Este atributo é utilizado para armazenar a identificação de uma estrada (e.g. IC19, EN122, ...);

**type**

Este atributo armazena a classificação das estradas, nomeadamente estradas residenciais, primárias, secundárias, terciárias, de serviço, caminhos pedestres, entre outras classificações;

**oneway**

Este atributo indica se a estrada tem apenas um sentido de circulação;

**bridge**

Este atributo indica se a estrada é uma ponte;

**tunnel**

Este atributo indica se a estrada é um túnel;

**maxspeed**

Este atributo indica a velocidade máxima a circular numa determinada estrada;

**geom**

Este atributo armazena a representação geométrica de uma estrada. Esta representação pode ser feita através de uma linha contínua (*Line String*), um conjunto de linhas não contínuas (*Multiline Strings*), de polígonos *Polygons*, entre outros objectos passíveis de serem utilizados através do *PostGIS*.

Depois dos mapas rodoviários serem analisados, foi necessário fazer uma segmentação das estradas de forma a possibilitar o seu mapeamento com os dados GPS e a melhorar a sua performance na agregação de dados das estradas.

Tendo em conta que no OSM, as estradas podem estar representadas de várias formas, o processo de segmentação foi necessário para que a representação das estradas ficassem uniformes.

Inicialmente, foi executado um *clipping* ao mapa rodoviário para que fossem usadas apenas áreas onde os dados GPS a serem testados foram registados, para acelerar o processo de segmentação. O *clipping* foi feito através da aplicação Quantum GIS (QGIS).

Para o processamento da segmentação das estradas, foram consideradas as seguintes definições:

**(Ponto de) Intersecção:** Ponto onde 3 ou mais segmentos rodoviários se interceptam. Num cenário rodoviário, podemos considerar que uma intersecção é um entroncamento (Figura 3.8);

**Troço:** Conjunto de segmentos que, num cenário rodoviário, constituem uma estrada. O início e o fim de um troço devem estar entre pontos de intersecção (Figura 3.8).

O processo de segmentação é composto por várias tarefas:

- i A partição das estradas em segmentos; a detecção de intersecções deve ser feita antes e após a criação de troços;
- ii A criação de troços, onde os segmentos que obedecerem a um conjunto de critérios, que serão apresentados posteriormente, são transformados em troços;
- iii A criação de raios para os troços e intersecções.

## SEGMENTAÇÃO

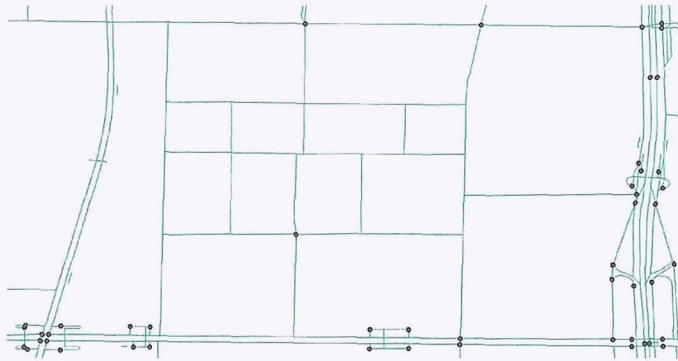


Figura 3.7: Exemplo de troços e respectivas intersecções antes da fase de segmentação

A segmentação consiste na partição das estradas do mapa em segmentos retos. As estradas que no projecto OSM estão representadas através de um conjunto variado de objectos (*Line Strings*, *Multiline Strings* ou *Polygons*) estarão, após a partição, representadas por linhas contínuas (*Line Strings*) (Figura 3.7). O resultado dessa partição é armazenada numa tabela temporária na base de dados, para ser posteriormente utilizada para a criação de troços.

A segmentação é executada em todas as estradas do mapa que não sejam túneis e pontes visto que estes não terão possíveis intersecções ao longo da sua rota. Note-se que as estradas dos mapas têm um conjunto de atributos associados, sendo dois deles a possibilidade destes serem estradas serem um túnel ou uma ponte. A rejeição de túneis e pontes evitará situações de falsas intersecções nos mapas rodoviários.

## DETECÇÃO DE INTERSECÇÕES

A detecção de pontos de intersecção consiste na procura de cruzamentos entre segmentos, em que os pontos considerados num segmento são apenas o inicial e final.

Para haver condições suficientes para a criação de uma intersecção, é necessário que este tenha pelo menos três segmentos a intersectá-lo. A detecção é feita através de uma busca exaustiva, ou seja, o ponto inicial e final de cada segmento é comparado com todos os pontos iniciais e finais de todos os segmentos produzidos na fase de segmentação.

É possível que existam abordagens que registem melhores performances em relação aos tempos de execução. Esta abordagem foi a escolhida não só pela sua facilidade de implementação mas também porque era necessário assegurar que todos os segmentos são analisados, garantindo assim que a criação de troços a ser feita posteriormente seja a melhor possível, já que a eficácia depende directamente do número de intersecções encontradas. Note-se que a organização interna dos dados dos mapas não asseguram qualquer tipo de ordem entre o registo das estradas, pelo que os resultados não seriam fiáveis ao diminuir a área de procura de intersecções. Resumindo, a remoção de segmentos na procura de possíveis intersecções poderia resultar em falhas na detecção de intersecções.

Tal como os segmentos, os pontos de intersecção são armazenados numa tabela temporária na base de dados para que, após a criação dos troços, haja uma refinação desses pontos.

## TROÇOS

Após a fase de detecção de intersecções, o processo de criação de troços é executado. Nesta fase é feita a análise de todos os segmentos obtidos na fase de segmentação, com o auxílio dos pontos de intersecção criados na fase anterior.

Essa análise consiste em verificar se um ponto (inicial e final) de um segmento intercepta um ponto (inicial e final) de outro segmento. Caso haja uma intersecção entre esses pontos, é feita uma procura dessa intersecção no conjunto de pontos obtidos na fase anterior. Caso não pertença, os dois segmentos são fundidos num único segmento.

Após a transformação, os segmentos utilizados para fazer a transformação são descartados.

Para que haja garantia de que todos os troços são criados, é necessário executar este processo pelo menos duas vezes, para que o segundo processamento verifique casos que não foram verificados na primeira execução. Um exemplo que põe em evidência a necessidade dessa verificação são os casos em que, após dois segmentos serem transformados num só segmento, não são comparados com os segmentos adjacentes. Uma segunda execução foi o suficiente para verificar todos os casos não analisados na primeira execução, nos testes executados para este trabalho. Note-se que, como já foi dito anteriormente, não existe uma ordem no que toca à organização nos dados no OSM. Isto pode causar falhas de verificação na transformação de segmentos.

Após a criação dos troços, a lista de pontos de intersecção é atualizada. Esta atualização é necessária devido ao facto da primeira detecção de pontos de intersecção ter sido executada baseando-se nos segmentos criados na fase de segmentação. Esta nova detecção é feita baseando-se nos troços já criados.

Esta lista de pontos de intersecção será a lista final sendo armazenada na base de dados de forma permanente para utilização futura.

Após a criação dos troços e a actualização dos pontos de intersecção, segue-se a criação de raios nesses objectos. O motivo para a criação dos raios foi a prever uma possível utilização para a aplicação no processo de Map Matching. Como se pode verificar na Figura 3.9, o raio criado foi um raio circular de 10 metros em torno dos troços/intersecções, valor escolhido tendo em conta a margem de erro passível de acontecer na cálculo da posição em equipamentos GPS diferenciais. O tamanho do raio foi escolhido tendo em conta o desvio de erro do registo dos dados posicionais.

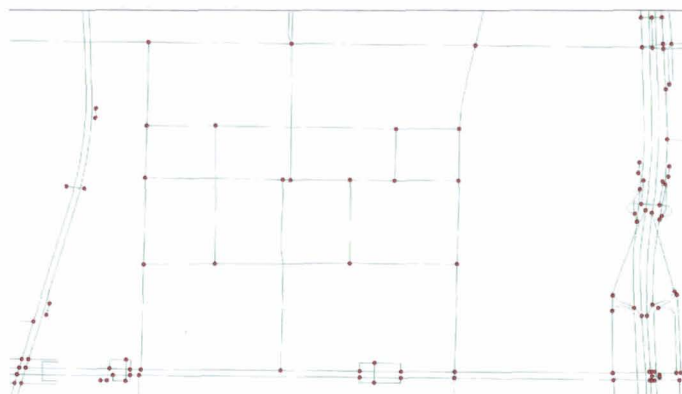


Figura 3.8: Exemplo de troços e respectivas intersecções após a fase de segmentação

O processo de segmentação, detecção de pontos de intersecção e de criação de troços e respetivos *buffers* foram executados através de um script criado na Linguagem Procedural SQL (plpgsql).

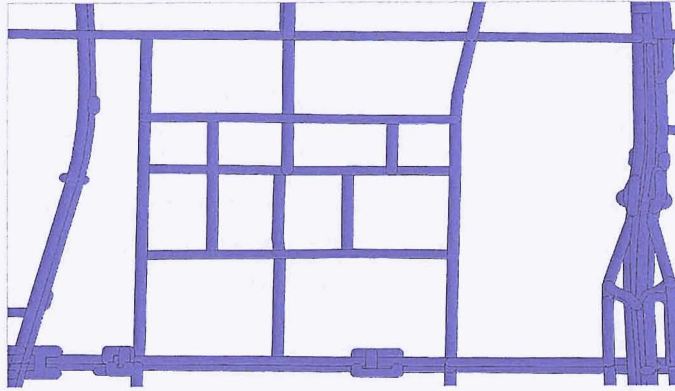


Figura 3.9: Exemplo dos buffers criados à volta dos troços, após a fase de segmentação

## NOVA ABORDAGEM PARA A SEGMENTAÇÃO RODOVIÁRIA

Após uma pesquisa posterior relativa à utilização do *pgrouting*, esse conhecimento foi utilizado para a criação de uma função que, dados a tabela fonte, um conjunto de atributos e restrições, a função cria uma tabela que é preenchida com os segmentos uniformizados. A segmentação é executada através do apoio de três operações fornecidas pelo *PostGIS*:

**ST\_Dump(*geom*):** Esta função devolve um conjunto de objetos geométricos resultantes de um objeto *geom*. Por exemplo, caso *geom* seja uma *MultiLineString*, a função retorna o conjunto de *LineStrings* que compõem o objeto dado como argumento.

Esta função foi utilizada na abordagem anterior.

**ST\_Union(*geom*):** Esta função permite agregar um objeto geométrico *Multi\** (i.e., *MultiLineString*, *MultiPolygon*, ...) num objeto geométrico singular, removendo as regiões de intersecção. *ST\_Union* é muito similar a *ST\_Node*, diferenciando-se no facto deste em alguns casos não aplicar a agregação na íntegra, necessitando de funções adicionais (*ST\_LineMerge*) para o fazer.

**ST\_Node(*geom*):** Esta função faz a junção de geometrias lineares numa só, utilizando o número mínimo de nós possíveis.

Tal como a função anterior, esta função foi utilizada na abordagem anterior.

Esta abordagem para além de obter melhores performances tendo em conta o tempo de execução, também não necessita que os pontos de intersecção sejam previamente determinados.

Essa melhoria pode ser verificada na segmentação das estradas da Dinamarca. Enquanto que na primeira abordagem o processo não foi possível ser feito em tempo útil, nesta abordagem o processo demorou sensivelmente 48 horas.

## MAP MATCHING

Para o mapeamento dos dados GPS com mapa rodoviário, foi escolhido o projecto GraphHopper em conjunção com mapas do OSM. GraphHopper é um projecto livre desenvolvido na linguagem Java, que contém um motor de *routing* utilizado no GraphHopper Maps.

O projecto tem um sub-projecto que visa o mapeamento de dados GPS com o mapa das estradas. Esse sub-projecto foi em algumas instâncias modificado para obter da forma mais útil os dados relativos

à execução do Map Matching. O processo de Map Matching implementado no GraphHopper, pode ser dividido em duas fases: a fase de carregamento do mapa rodoviário e fase de carregamento e mapeamento dos dados GPS.

Para fazer o carregamento do mapa rodoviário devem ser utilizados mapas no formato Protocol Binary Format (PBF). Visto como uma alternativa ao formato eXtensible Markup Language (XML) para o armazenamento de dados relativos a mapas rodoviários (especificamente o *.osm.bz2*, utilizado no OSM baseado em XML), o formato PBF permite uma maior capacidade de compressão (cerca de 30%) e melhores performances em operações de leitura (6x mais rápido) e escrita (5x mais rápido) em relação a outros formatos utilizados pelo OSM. Este modo pode ser executado da seguinte forma:

---

```
java -jar gh action=import datasource=./some-dir/osm-file.pbf vehicle=car
```

---

Após o carregamento do mapa rodoviário, os dados contidos nele são utilizados para a construção de um grafo do mapa de estradas que depois é armazenado para ser posteriormente utilizado para o processo de Map Matching com os dados GPS. No processo de carregamento do mapa, através do atributo *vehicle* é possível escolher o perfil de encaminhamento, ou seja, o cenário em que os dados GPS devem ser tidos em conta. As opções passíveis de selecção podem ser: carro (*car*), motociclo (*motorcycle*), bicicleta (*bike*) ou a pedestre (*foot*).

Para este trabalho, foram utilizados mapas zonas da área da China e da Dinamarca. Os mapas dos respectivos países foram obtidos através da GEOFABRIK.

No processo de carregamento dos dados GPS para a posterior execução do Map Matching, os dados GPS devem ser carregados através de ficheiros do formato GPS eXchange (GPX). Como inicialmente os dados GPS estão armazenados numa base de dados, foi necessário criar um programa capaz de obter os dados GPS da base de dados para ficheiros GPX. Nesta conversão, foi necessário estabelecer um conjunto de critérios para que cada um dos ficheiros represente um trajecto feito pelo utilizador. Para evitar que o programa crie um trajecto com todos os pontos na íntegra, pontos esses registados durante um período de tempo considerável (semanas, até meses de registos), é necessário estabelecer um conjunto de trajectos nesse conjunto de pontos de forma a ter uma percepção realista dos percursos feitos pelos utilizadores.

É considerado um trajecto um conjunto de localizações GPS em que:

- i A marca temporal entre duas posições GPS consecutivas devem ter um tempo máximo de 5 minutos;
- ii A distância geográfica entre duas posições GPS consecutivas não excedam os 2,5 Quilómetros.

O primeiro critério é o ponto fulcral para a criação de trajectos, enquanto que o segundo critério é importante para a remoção de posições atípicas nos trajectos GPS.

Após a criação de trajectos, estes foram visualizados através do programa QGIS. QGIS é uma plataforma que permite a criação, edição visualização e análise de dados geoespacial.

Ao serem analisados, foram encontrados posições atípicas em alguns trajectos, ou seja, pontos que foram registados com coordenadas fora do contexto da trajectória feita pelo utilizador. Estes valores atípicos tiveram de ser removidos, visto que são uma grande influência no resultado do mapeamento dos pontos aos mapas das estradas.

Para a remoção de valores atípicos, foi criada uma função que permite fazer a verificação do tempo e espaço percorrido por cada duas posições consecutivas. Entre cada par de localizações GPS, é verificado



Figura 3.10: Exemplo da execução de Map Matching

a distância percorrida entre eles em função do período de amostragem. Caso a distância percorrida seja considerada impossível de ser feita no período de amostragem utilizado, essa localização é removida.

Após a remoção de posições atípicas, o processo de Map Matching é executado com a seguinte instrução:

---

```
java -jar gh action=match gpx=/path/to/gpx/directories/
```

---

A aplicação executa o algoritmo de Map Matching sobre todos os ficheiros GPX recursivamente, a partir do directório dado como argumento. Os resultados de cada trajecto são devolvidos no mesmo tipo de ficheiros (GPX).

O processo de Map Matching é dividido em quatro fases:

1. **Fase de Procura:** Fase onde é feita a busca de quatro segmentos mais próximos a cada localização GPS;
2. **Fase de Ponderação:** Fase em que, para cada segmento encontrado anteriormente é associado um peso.  
O peso corresponde à distância entre segmento e à localização GPS; O valor do peso é inversamente proporcional à probabilidade do segmento pertencer ao trajecto;
3. **Fase de Pesquisa:** É feita a pesquisa do melhor trajecto a partir da primeira até à última localização GPS, tendo em consideração os pesos estabelecidos anteriormente. Esta pesquisa é feita utilizando um algoritmo de Dijkstra personalizado;
4. **Fase de Correspondência:** Nesta fase, cada localização GPS é mapeada no mapa rodoviário.

O trajecto a ser processado é dividido num conjunto de sub-trajectos, aos quais os passos anteriores são posteriormente aplicados. No final é feita a união de cada sub-trajecto resultante de forma a obter o trajecto final, como podemos ver na Figura 3.10.

O motivo pelo qual o algoritmo é aplicado aos sub-trajectos ao invés de ser aplicado ao trajecto na íntegra prende-se com o facto de haver a necessidade de evitar trajectos com *loops*. Ao serem encontrados, o passo 3 não é executado corretamente, visto que o algoritmo de Dijkstra descarta os pontos que constituem o *loop*, como podemos ver na Figura 3.11.

Após a análise dos resultados, foi necessário reposicionar as localizações GPS no mapa rodoviário, de forma a melhorar a performance do algoritmo de Map Matching. Note-se que com o erro no registo



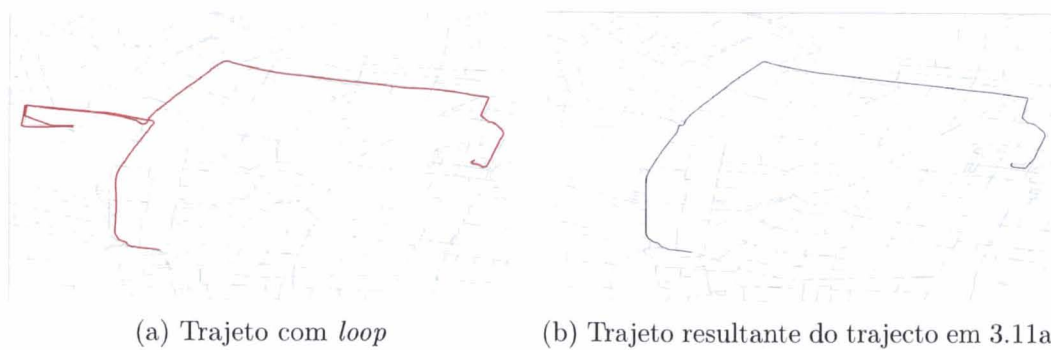


Figura 3.11: Problema de ciclos no Map Matching

das localizações GPS, bem como a característica dos mapas rodoviários (sistema de referência das coordenadas) faz com que as localizações estejam tipicamente com uma margem de erro relativamente ao mapa rodoviário. Esse erro foi colmatado com o reposicionamento das localizações GPS.

## MIGRAÇÃO DE DADOS

Após a fase de transformação de dados estar concluída, iniciou-se o processo de carregamento de dados para as respectivas Data Warehouses. Esse carregamento foi feito através de uma aplicação Java desenvolvida para preencher as Data Warehouses. Para complementar os dados já transformados, foram utilizados dados externos úteis para o funcionamento do sistema.

Em relação à localização dos pontos, foi utilizado um Web Service (WS) para obter os dados relativos à cidade, bem como ao país relativo às localizações GPS e aos mapas rodoviários. O WS utilizado provém do GeoCode da *Google Maps*. Para obter a cidade e respectivo país de um ponto basta enviar para o WS as coordenadas (latitude e longitude) do ponto.

E relação às condições meteorológicas, o WS utilizado inicialmente foi o Open Weather Maps (OWM) para obter a informação relativa ao estado do tempo. A data e hora são os dados necessários para obter essa informação. Durante a elaboração do trabalho, os dados provenientes do OWM deixaram de ser fornecidos gratuitamente. Por isso, foi utilizado o ForecastIO para obter a informação relativa às condições meteorológicas.

## RESULTADOS

---

Neste capítulo pretende-se mostrar a fiabilidade dos resultados dos mapeamentos das trajectórias executados através do algoritmo de Map Matching e mostrar os resultados obtidos através das consultas feitas à Data Warehouse. Esses resultados são relativos à análise das rotas entre dois pontos de intersecção. Para tal, são utilizados dois casos de estudo. O primeiro caso é relativo a um conjunto de dados proveniente de Aalborg, Dinamarca. O segundo caso é relativo a um conjunto de dados proveniente de Pequim, China. De seguida os casos de estudos são apresentados com mais detalhe. Estes casos de estudo foram escolhidos por terem atributos importantes para o armazenamento dos dados posicionais e dos trajectos como a latitude, longitude e a marca temporal de cada registo. O facto de terem características distintas como o período de amostragem, e a quantidade de dados posicionais e o número de dados atípicos também foi considerado para a sua selecção.

### 4.1 CASOS DE ESTUDO

#### 4.1.1 CASO DE ESTUDO 1: AALBORG, DINAMARCA

Este conjunto de dados é descrito no artigo *The Infati Data* [136]. Os dados posicionais estudados são de 20 veículos distintos registados entre Dezembro de 2000 e Março de 2001 através de um receptor GPS.

Os pontos de localização destes dados GPS têm um período de amostragem de um segundo e foram registadas entre Dezembro de 2000 e Janeiro de 2001.

Devido a razões associadas ao anonimato, os dados dos veículos, bem como dos condutores não foi facultado. Por isso, as dimensões relativas a essa informação não foram utilizadas. Apesar da não utilização dessas dimensões, elas permaneceram na Data Warehouse para futuras utilizações.

Em relação aos trajectos, foram armazenadas 22.847 viagens na Data Warehouse. De relembrar que cada viagem contém associada informação relativa ao trajecto completo, aos pontos de partida e de chegada, às condições meteorológicas no início da viagem, ao instante temporal inicial e final, e às localizações iniciais e finais do trajecto. A duração da viagem foi obtida através do tempo inicial e final de cada trajecto. Os dados facultados pelo Infati Data não foram suficientes para obter a velocidade máxima e o consumo médio de combustível dos veículos.

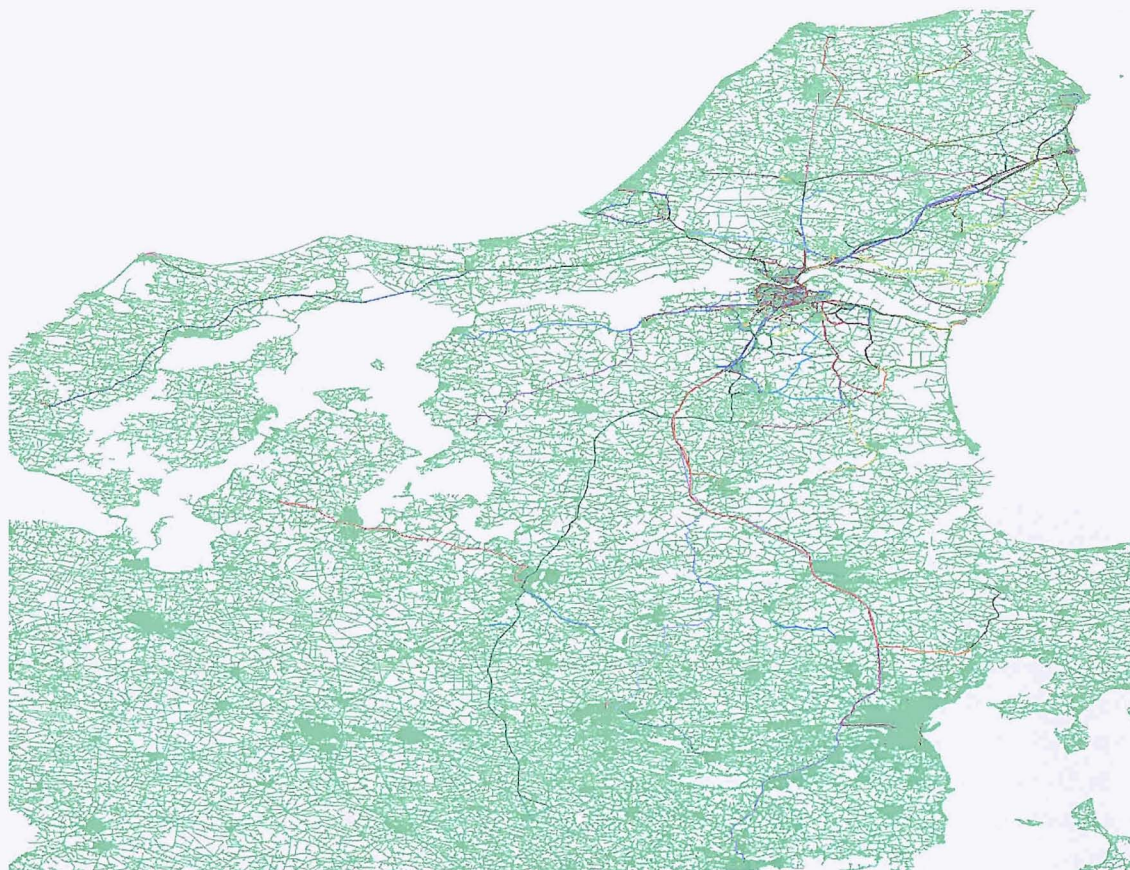


Figura 4.1: Trajectos registados na Dinamarca, maioritariamente em Aalborg

Em relação aos pontos de localização, foram armazenados cerca de 7 milhões de registos que continham informação relativa a quatro milhões de pontos distintos, isto é, dos 7 milhões de pontos cerca de 3 milhões deles são pontos replicados, possivelmente representando casos em que os veículos estão parados. Em relação aos segmentos, foram armazenados 7 milhões de registos de 900 mil troços distintos, ou seja, em 7 milhões de segmentos, cerca de 6 milhões dão replicados. Essas replicações representam segmentos em que os veículos percorreram múltiplas vezes. Por exemplo, se num segmento um veículo registou 6 posições, são armazenados no sistema 6 registos nesse segmento.

#### 4.1.2 CASO DE ESTUDO 2: PEQUIM, CHINA

Este conjunto de dados é descrito no artigo *User Guide of T-Drive Data* [137]–[139]. Os dados posicionais estudados são de 10.357 taxis da cidade de Pequim. Os pontos de localização destes dados GPS têm um período de amostragem de cinco segundos e foram registadas entre 2 e 8 de Fevereiro de ano de 2008. Tal como os dados de Aalborg, os dados dos veículos e dos condutores não foram facultados.

Em relação aos pontos de localização, foram armazenados cerca de 15 milhões de registos que continham informação relativa a nove milhões de pontos distintos.

Em relação aos segmentos, foram armazenados 15 milhões de registos de 1.5 milhões de troços.

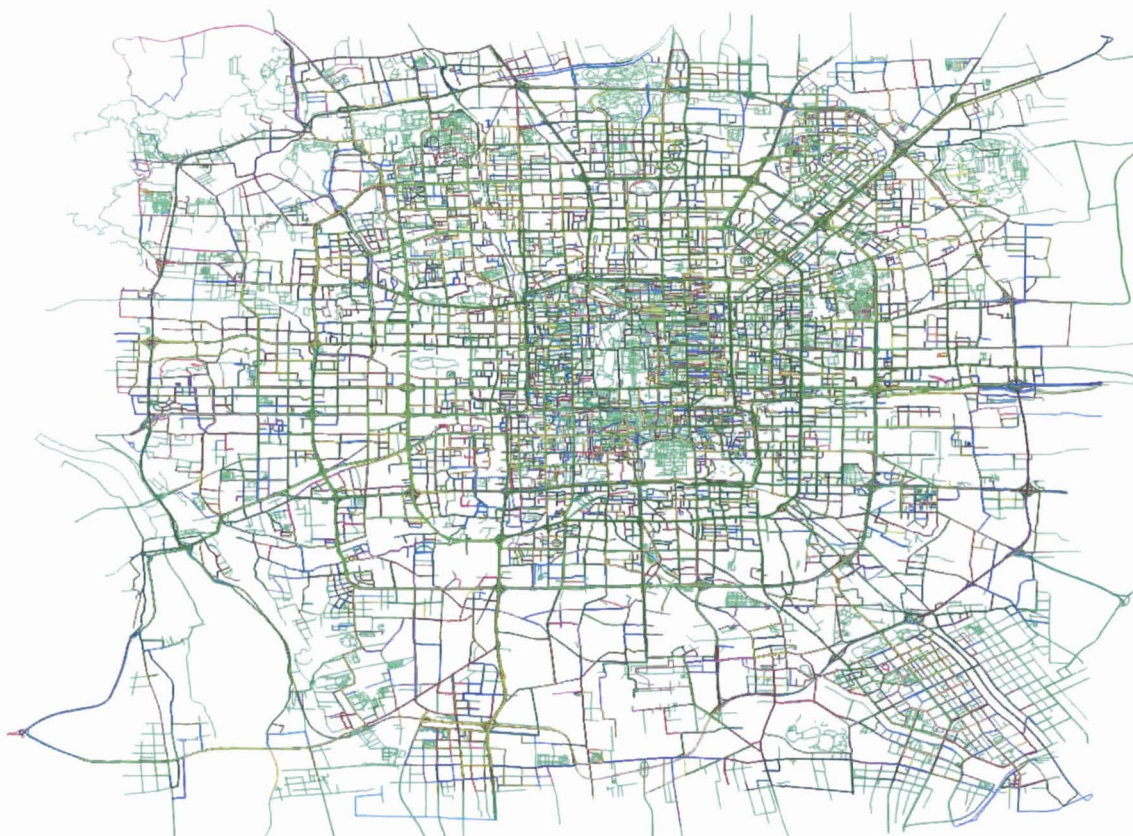


Figura 4.2: Trajectos registados da cidade de Pequim, China

## 4.2 MAP MATCHING

Após terem sido feitas modificações necessárias para a execução do programa de Map Matching (MM), foram feitos um conjunto de testes para a análise dos resultados com diferentes períodos de amostragem.

Estes testes têm como objetivo analisar o grau de similaridade dos mapeamentos feitos pelo *GraphHopper* em diferentes períodos de amostragem. Esse valor pode ser utilizado para classificar a fiabilidade dos resultados.

Para a execução do *GraphHopper*, são definidos valores para várias variáveis, de forma a aumentar a performance dos resultados. O tipo de transporte definido foi um carro; o tamanho mínimo da rede definido foi de 2000 metros, bem como o tamanho mínimo da rede com estradas de um só sentido; A distância de procura foi estabelecida a 5000 metros e a distância de procura máxima para o melhor trajeto foi definida a 3000 metros.

Para os testes que serão apresentados (ver tabela 4.1), foram utilizados um trajecto e um mapa de estradas comum em todos eles. O fator variável será o período de amostragem das posições GPS.

Os períodos de amostragem foram obtidos através da decimação dos dados GPS. Por exemplo, em uma decimação 1:10, foi seleccionada uma amostra entre 9 amostras consecutivas. Como se pode prever, quanto maior for a decimação, menor será a quantidade de dados utilizados para fazer o MM. Numa decimação 1:10, o período de amostragem é de aproximadamente 10 segundos.

Nas imagens resultantes, os dados GPS provenientes dos veículos da Dinamarca são representados por um conjunto de pontos, enquanto que os trajectos resultantes são representados por uma linha.

Foram executados testes em vários períodos de amostragem. Na tabela 4.1 são apresentados apenas 6 nos seguintes períodos: de segundo a segundo, de 10 em 10 segundos, 30 a 30 segundos, 60 a 60 segundos, 90 a 90 segundos e de 120 a 120 segundos.



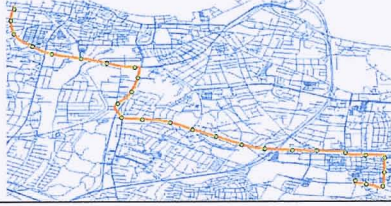

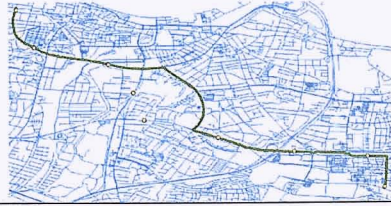

Teste	Decimação	Período (segundos)	Resultado
1	1:1	1	
2	1:10	10	
3	1:30	30	
4	1:60	60	
5	1:90	90	
6	1:120	120	

Tabela 4.1: Resultados dos testes de Map Matching

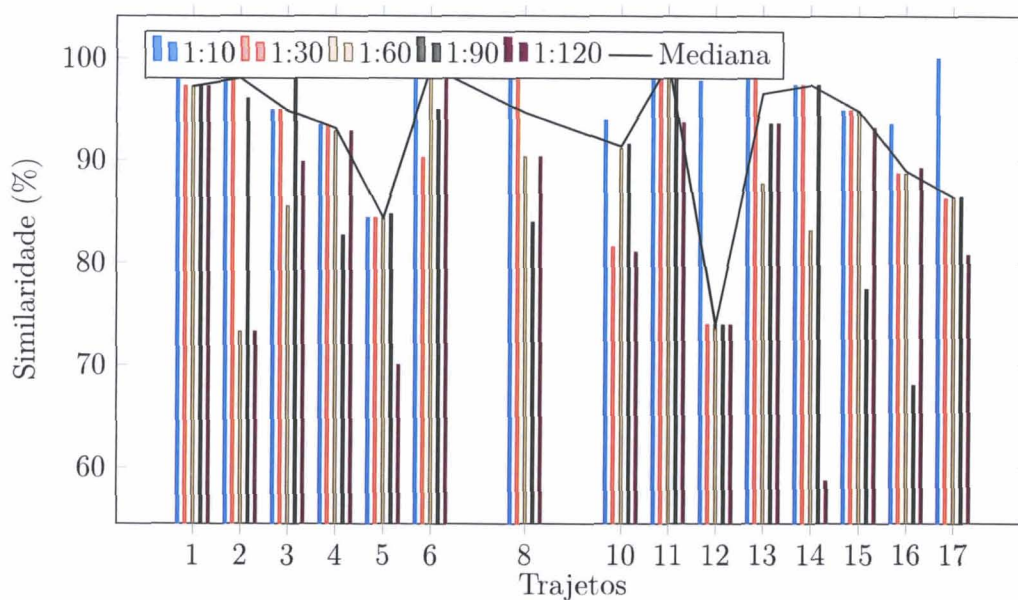


Figura 4.3: Nível similaridade dos trajetos resultantes - 1 a 16

Amostragem	1:10	1:30	1:60	1:90	1:120
Similaridade (%)	96.83	92.65	88.94	85.29	83.30

Tabela 4.2: Média dos níveis de similaridade

Após a execução de MM em todos os trajetos, foi feita uma verificação da qualidade dos resultados nos períodos de amostragem utilizados anteriormente, utilizando a distância de *Hausdiöff*. O nível de similaridade entre os resultados e o trajeto original de um conjunto de 14.292 pontos transformados em 37 viagens (trajetos) feitos por uma família na Dinamarca são apresentados nas figuras 4.3 e 4.4, através do gráfico de barras. Também é possível observar no mesmo gráfico a mediana da percentagem de similaridade dos resultados em cada trajeto.

Na tabela 4.2 são apresentadas as médias dos níveis de similaridade de cada período de amostragem.

Analisando o gráficos 4.3 e 4.4 e tendo em conta a média dos níveis de similaridade (4.2), é possível concluir que o algoritmo de MM apresenta melhores resultados para conjuntos de dados GPS com baixos períodos de amostragem. Porém, a perda de qualidade de similaridade dos trajetos resultantes de baixos períodos de amostragem não é significativa. Podemos ver na Tabela 4.2 que a perda de similaridade de um período de amostragem de segundo a segundo para 30 segundos é de cerca de 8%, demonstrando que, apesar dos mapeamentos não serem iguais, acabam por ser muito semelhantes, diferindo em poucos detalhes.

É de ter em conta que foram utilizados os mesmos argumentos para todas as execuções de MM. Esses argumentos estão relacionados com a distância de procura, a distância de procura máxima, tamanho mínimo de estradas com sentido único e o modo *force repair*, que visa encontrar um caminho mesmo quando não existem rotas candidatas. Foram utilizados os mesmos argumentos em todos visto que estamos a lidar com uma quantidade considerável de dados GPS. Caso esses argumentos fossem

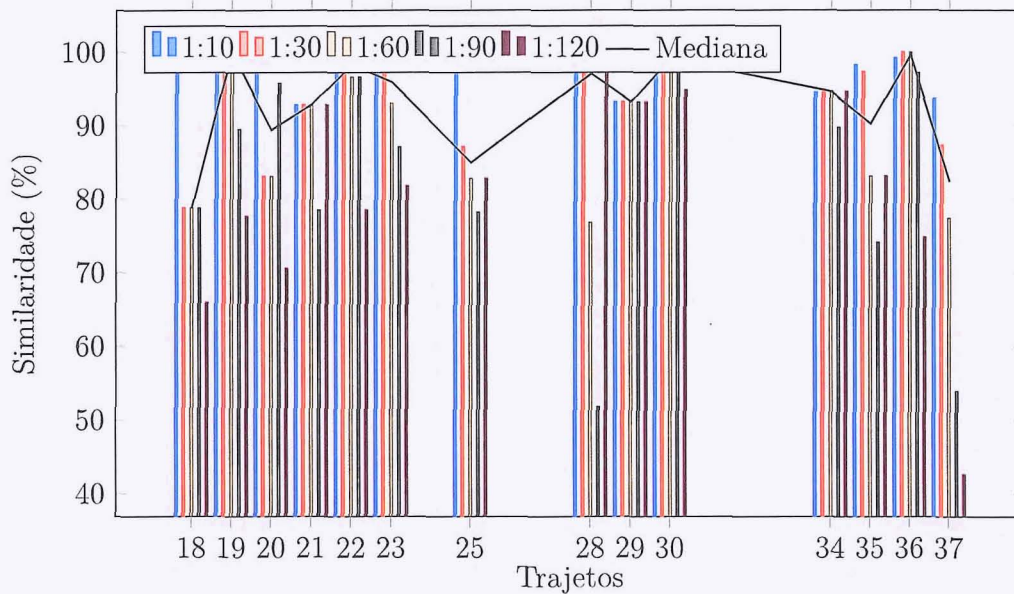


Figura 4.4: Nível similaridade dos trajectos resultantes - 18 a 37

variáveis, é possível que alguns trajectos tivessem melhores mapeamentos e que o número de trajectos inválidos fosse mais baixo.

Resumindo, devemos ter em conta alguns pontos:

- O trajecto resultante depende muito do mapa das estradas e as respectivas características (por exemplo, o sentido);
- Nos casos em que as posições GPS têm um período de amostragem consideravelmente grande, os trajectos mais prováveis são os que têm a distância mais curta, tendo sempre em conta as características das estradas. Note-se que para o cálculo do trajecto é aplicado um algoritmo de Dijkstra modificado;
- Em todos os testes foram utilizadas os mesmos valores para a execução do *GraphHopper*. Esses valores provavelmente deverão ser escolhidos de acordo com as características dos dados GPS e estradas.

### 4.3 ANÁLISE DOS TRAJECTOS

Nesta secção é apresentada uma análise estatística feita aos dados adquiridos de 10.357 taxis que circulam da cidade de Pequim, China.

Foram seleccionadas zonas com múltiplos caminhos alternativos, com o objectivo de verificar o nível de distinção de trajectos nessas zonas, bem como o nível de utilização dos trajectos mais rápidos. Para cada par de pontos de intersecção, foram seleccionadas as rotas entre esses pontos. As rotas são parte do trajecto onde a análise é focada. Esta análise irá basear-se na verificação das rotas utilizadas com mais frequência, através do cálculo do número de rotas registadas entre os pontos de origem e de destino. A partir daí é verificado através da distância da rota se esta é a mais curta, e se esta rota foi

feita com mais rapidez, verificando a duração em que o percurso foi feito. A duração média que os veículos permanecem numa determinada rota é obtida através do tempo que estes permaneceram na rota. Esse tempo é obtido através da diferença temporal entre o último e o primeiro registo posicional dentro da rota.

De seguida, é apresentada uma análise dos casos estudados com este conjunto de dados. O primeiro caso analisa uma zona de estradas secundárias, isto é, estradas utilizadas para ligar zonas residenciais. O segundo caso analisa uma zona residencial e o terceiro caso analisa dois pontos relativamente distantes (em comparação com os dois casos anteriores) que representam pontos turísticos.

## ANÁLISE 1: ROTAS EM ESTRADAS SECUNDÁRIAS

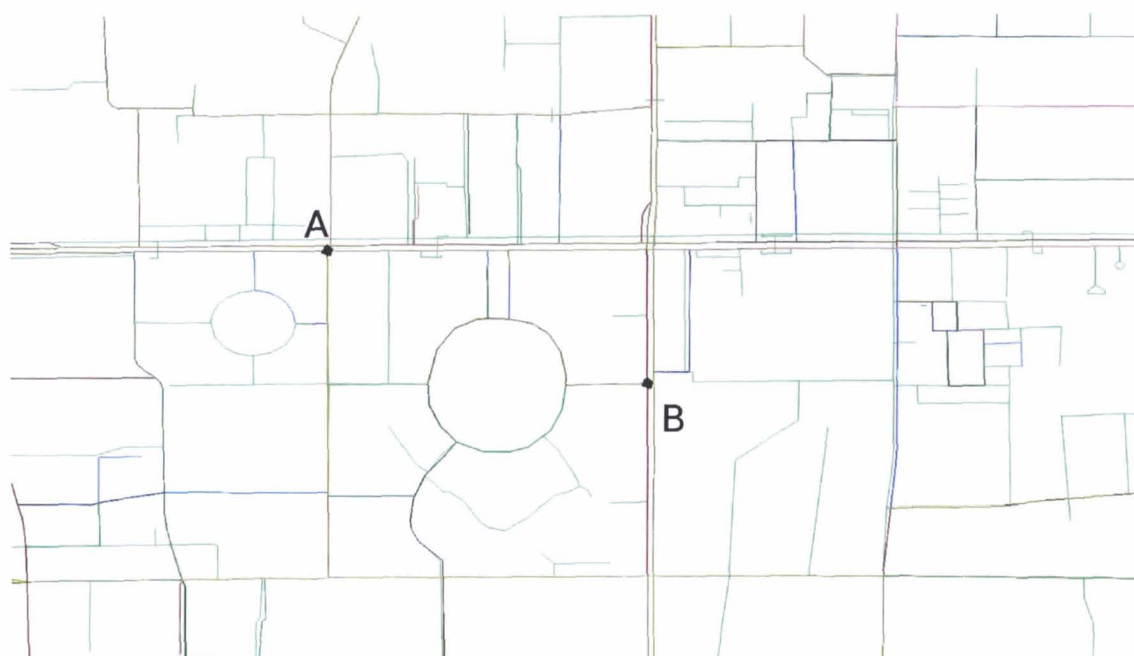


Figura 4.5: Par de intersecções A e B da análise 1

A Figura 4.5, mostra as rotas registadas numa zona de estradas secundárias. Das viagens feitas no sentido de A para B, foram registadas 25 rotas distintas, em que a rota feita com mais frequência regista 81 trajectos (Figura 4.6a). Dos 81 trajectos, 53 precederam do segmento adjacente a norte da intersecção A, 23 precederam do trajecto adjacente a oeste da intersecção A e 5 precederam do trajecto adjacente a sul da intersecção A. As restantes rotas foram percorridas apenas por um trajecto.

No sentido inverso (Figura 4.6b), foram registadas 11 rotas distintas, em que na rota feita com mais frequência registou 5 trajectos; a segunda rota com mais frequência foi feita por três trajectos e as restantes foram registadas apenas por um trajecto.

Através da análise das distâncias, verificou-se que a rota mais curta é a rota utilizada com mais frequência no sentido de A para B, com 866 metros. Esta rota também foi considerada a rota mais rápida, com uma duração média de 2.15 minutos. O número de paragens médio foi de 13.2 por trajecto e quanto ao número de paragens na rota mais rápida, foi registado um número médio de 5.4 paragens por trajecto.



Tabela 4.3: Distância máxima, mínima, média, mediana e desvio padrão da análise 1

Distância máxima	Distância Mínima	Distância Média	Mediana	Desvio Padrão
19105 m	866 m	5654 m	2899 m	5297,98



(a) Rota utilizada por mais trajectos e a mais curta (b) Rota utilizada por mais trajectos, no sentido inverso

Figura 4.6: Rotas utilizadas com maior frequência

## ANÁLISE 2: ROTAS NUMA ZONA RESIDENCIAL

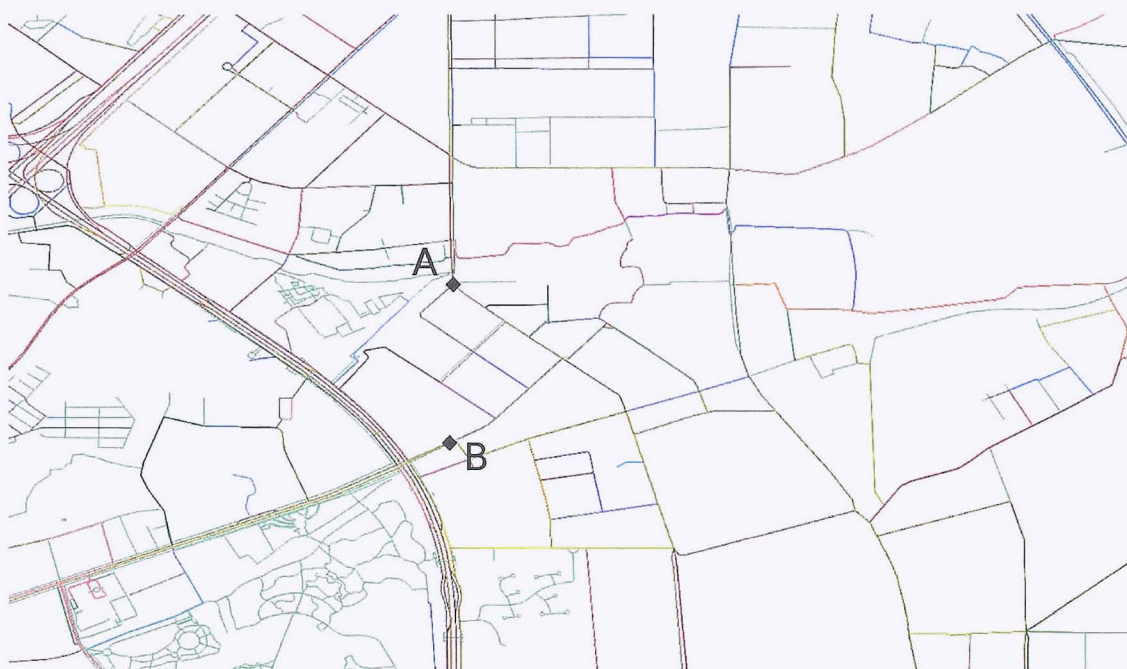


Figura 4.7: Par de intersecções A e B da análise 2

A análise feita na zona demonstrada na Figura 4.7, analisa as rotas registadas numa zona residencial. Das viagens feitas no sentido de A para B, foram registadas 24 rotas distintas, em que a rota feita com maior frequência regista 74 trajectos (Figura 4.6a). Dos 74 trajectos, 69 precederam do trajecto adjacente a noroeste da intersecção A, e 5 precederam do trajecto adjacente a nordeste da intersecção A. As restantes rotas foram percorridas apenas por um trajecto.

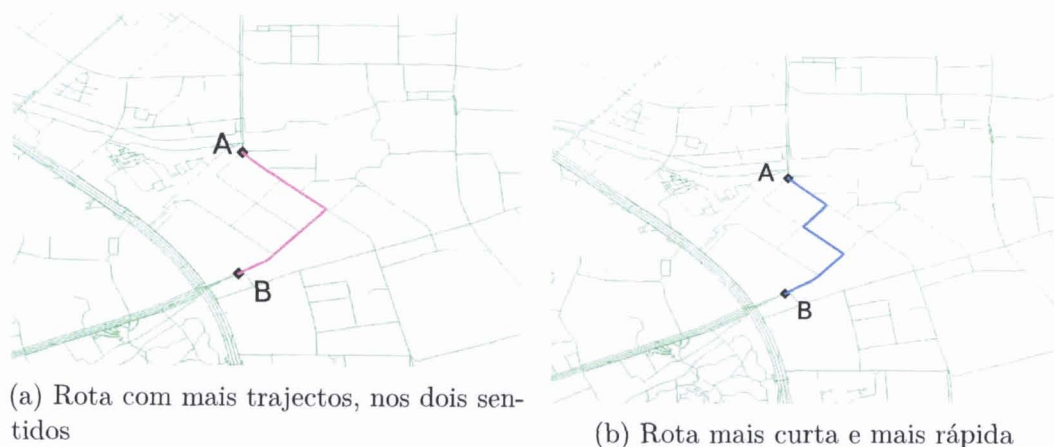


Figura 4.8: Rotas utilizadas com maior frequência

No sentido inverso, foram registadas 20 rotas distintas, em que a rota feita com mais frequência registou 145 trajectos; a segunda rota feita com mais frequência registou 11 trajectos; a terceira rota feita com mais frequência registou 5 trajectos; a quarta e quinta rota feita com mais frequência registou 3 trajectos.

Através da análise das distâncias, verificou-se que a rota mais curta foi percorrida por apenas 11 trajectos, no sentido inverso. Note-se que esta rota é 9 metros mais curta do que a rota utilizada com mais frequência, ou seja, a diferença não é significativa. Em relação à rota utilizada com mais frequência, existem 2 rotas ligeiramente mais curtas, uma utilizada por 11 trajectos (7 metros de diferença), e outra utilizada por 3 trajectos (2 metros de diferença). Verificou-se que existem 5 rotas com distâncias muito semelhantes (aproximadamente 1.2 quilómetros), diferenciando-se em dezenas de metros. Por isso, considerou-se todas essas rotas como rotas óptimas. A rota feita com mais rapidez foi a rota mais curta, (Figura 4.8b), com uma duração média de 1.39 minutos. O número de paragens médio foi de 4.1 paragens por trajecto. Na rota mais rápida, foi registado um número médio de 2.3 paragens por trajecto.

Tabela 4.4: Distância máxima, mínima, média, mediana e desvio padrão da análise 2

Distância máxima	Distância Mínima	Distância Média	Mediana	Desvio Padrão
58963 m	1228 m	10939 m	2604 m	14418,47

### ANÁLISE 3: ROTAS ENTRE LOCAIS TURÍSTICOS

A análise feita na zona demonstrada na Figura 4.9, analisa rotas registadas entre a Cidade Proibida (Ponto A) e o Templo de Céu (Ponto B). Das viagens feitas, foram registadas 19 rotas distintas de A para B, em que a rota feita com mais frequência regista 26 trajectos, em que 24 dos trajecto têm o mesmo segmento precedente em comum, e 2 trajectos têm outro segmento precedente. A segunda rota utilizada com mais frequência foi registou 2 trajectos; as restantes rotas foram percorridas apenas por um trajecto.

No sentido contrário, foram registadas 16 rotas distintas e todas as rotas registaram apenas um trajecto. A rota mais curta utilizada de B para A é cerca de 100 metros maior do que a rota mais

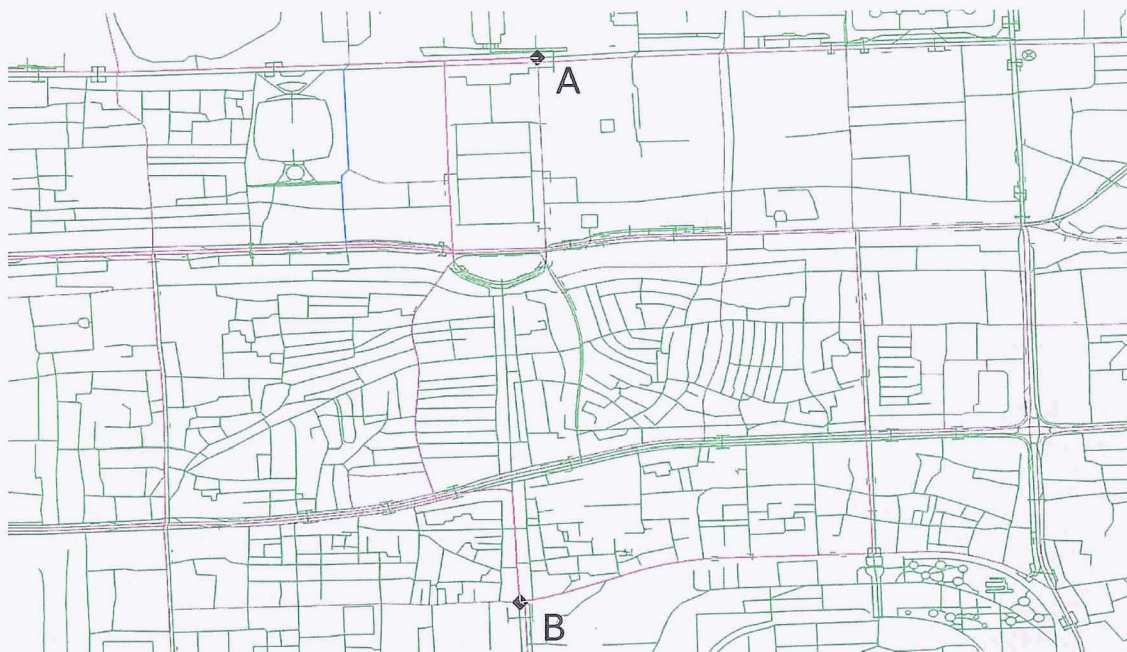


Figura 4.9: Par de intersecções A e B da análise 3



(a) Rota utilizada por mais trajectos e a mais curta, de A para B

(b) Rota mais curta utilizada no sentido de B para A

Figura 4.10: Rotas utilizadas com maior frequência

curta utilizada de A para B.

Através da análise das distâncias, verificou-se que a rota mais curta foi a rota utilizada com mais frequência no sentido de A para B (Figura 4.10a), com 3.01 quilómetros. A segunda rota mais curta foi a rota utilizada com mais frequência no sentido de B para A (Figura 4.10b), 3.07 quilómetros.

A rota feita com mais rapidez foi a rota mais curta, (Figura 4.10a), com uma duração média de 6.36 minutos. O número de paragens médio foi de 34.2 paragens por trajecto. Na rota mais rápida, foi registado um número médio de 19.3 paragens por trajecto.

Tabela 4.5: Distância máxima, mínima, média, mediana e desvio padrão da análise 3

Distância máxima	Distância Mínima	Distância Média	Mediana	Desvio Padrão
80310 m	3010 m	14715 m	10984 m	14960,01

## ANÁLISE GLOBAL

Foram feitos 26 testes em 13 casos distintos. Tal como na análise anterior, cada caso foi caracterizado por um ponto de origem e um ponto de destino.

O número total de rotas distintas registadas nos testes foram de 308. Essas rotas foram percorridas por 936 trajectos. O número de trajectos que percorreram as rotas mais curtas foram de 334 e dessas rotas, 268 são também as rotas mais rápidas. E geral, 303 trajectos percorreram rotas mais rápidas. O número de paragens médio proveniente dos casos de estudo foi de 54.3 paragens por trajecto. Foram ainda registadas 289 rotas, que não foram classificadas como mais curtas ou mais rápidas. Nestas rotas, 142 foram classificadas como rotas atípicas. São consideradas rotas atípicas as rotas que não percorrem zonas rodoviárias entre os pontos de intersecção. Na figura 4.11 são visíveis duas rotas atípicas de um caso de estudo.

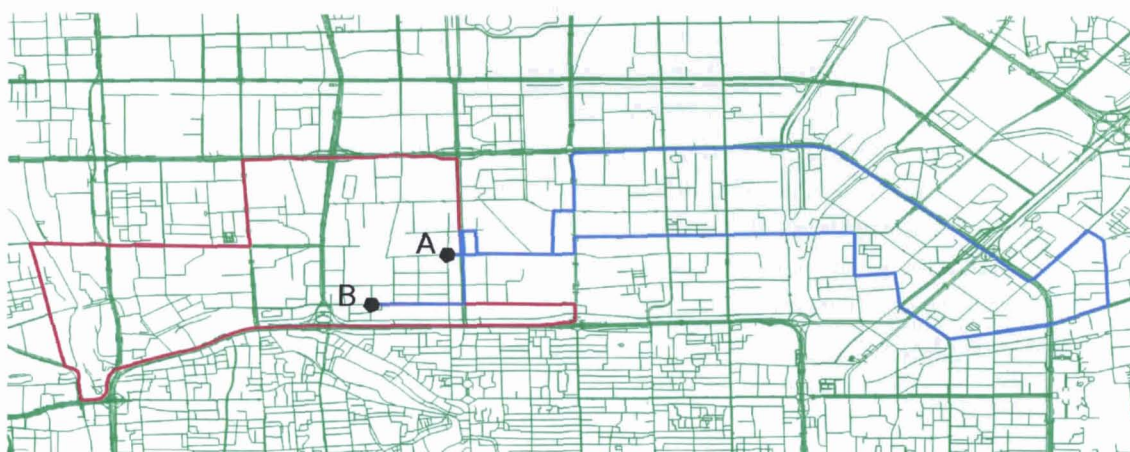


Figura 4.11: Duas rotas atípicas de dois pontos de intersecção

Nos casos focados em zonas com estradas secundárias, foram registados 148 rotas distintas percorridas por 534 trajectos. 478 percorreram a rota mais curta. Desses trajectos, 405 percorreram a rota mais rápida. Em relação à utilização das rotas mais rápidas, estas foram percorridas por 418 trajectos. Foram registadas 131 outras rotas. Nessas rotas, foram classificadas 74 rotas atípicas.

Nos casos focados em zonas residenciais, foram registados 99 rotas distintas percorridas por 285 trajectos. 171 percorreram a rota mais curta. Desses trajectos, 44 percorreram a rota mais rápida. Em relação à utilização das rotas mais rápidas, estas foram percorridas por 85 trajectos. Foram registadas 69 outras rotas. Nessas rotas, foram classificadas 21 rotas atípicas.

Nos casos focados em pontos turísticos, foram registados 61 rotas distintas percorridas por 117 trajectos. 54 percorreram a rota mais curta. Desses trajectos, 26 percorreram a rota mais rápida. Em relação à utilização das rotas mais rápidas, estas foram percorridas por 35 trajectos. Foram registadas 95 outras rotas. Nessas rotas, 71 foram classificadas como rotas atípicas.

## 4.4 DISCUSSÃO

### 4.4.1 TESTES DE SIMILARIDADE

Através dos níveis de similaridade obtidos da comparação do trajectos feitos pelos veículos na Dinamarca com vários períodos de amostragem, é possível verificar que quanto menor for o período de amostragem, maior é a similaridade entre o trajecto real e o trajecto resultante do Map Matching. O facto do período de amostragem ser baixo faz com que o número de trajectos inválidos seja reduzido. Se verificarmos cada trajectória individualmente, verifica-se que o período de amostragem influencia o resultado do mapeamento, já que em poucos casos encontramos trajectos iguais resultantes de diferentes períodos de amostragem (Trajectos 1, 12, 21 e 29 das Figuras 4.3 e 4.4). Apesar dessa influência, podemos verificar pelos resultados da Tabela 4.2 que a perda de qualidade entre diferentes períodos de amostragem não é significativa, isto é, da trajectória original existe em média uma perda de similaridade 4 por cento para a trajectória resultante do mapeamento com período de amostragem de 10 segundos. Isto pode ser útil para aumentar a eficiência do processamento dos dados de localização, bem como da execução do algoritmo de Map Matching. No caso do conjunto de dados da Dinamarca no qual foi necessário processar cerca de 7 milhões de dados posicionais, a utilização de períodos de amostragem de 10 em 10 segundos reduziria os dados para um conjunto de cerca de 700 mil dados. Apesar do facto de que a utilização de um maior período de amostragem ser mais eficiente, esta estratégia de aceleração deve ser utilizada em zonas que não sejam complexas de forma a evitar que entre dados posicionais consecutivos exista um grande conjunto de rotas possíveis de terem sido percorridas num determinado trajecto.

O conjunto de dados correspondente a veículos provenientes da Dinamarca é um conjunto com um período de amostragem ao segundo com poucos dados atípicos. O facto de haver poucas localizações atípicas e de o deslocamento encontrado entre os dados posicionais e o mapa digital ser coerente em todas as posições, fez com que a correcção fosse facilmente executada e que os resultados provenientes do algoritmo de Map Matching tenham sido fiéis aos trajectos originais e com uma taxa de trajectos inválidos muito baixa.

O mesmo algoritmo foi aplicado ao conjunto de dados provenientes de taxis na cidade de Pequim e o que se verificou neste conjunto de dados foi que a taxa de trajectos inválidos subiu consideravelmente visto que a qualidade dos dados é menor, se comparada com os dados da Dinamarca. A dimensão deste conjunto de dados é maior que o conjunto de dados da Dinamarca (15 milhões de pontos vs 7 milhões de pontos), e apesar de ter um período de amostragem baixo (registos posicionais de 5 em 5 segundos), este conjunto apresenta muitas localizações atípicas, o que influencia a qualidade dos resultados do algoritmo de Map Matching, que se mostrou sensível a dados atípicos. Para além disso, o deslocamento das posições aos segmentos rodoviários são muito incoerentes e devido à quantidade de dados existentes, a correcção desse deslocamento foi ineficaz, já que foi impossível determinar visualmente os segmentos rodoviários a que as posições foram registadas (Figura 4.12). Na mesma figura, foram detectadas manchas de dados posicionais dentro de quarteirões. Essas manchas representam parques de estacionamento cobertos. Para a remoção de dados posicionais atípicos, foi implementado um *script* em *pgSQL*. Esse *script* faz a verificação sucessiva das distâncias entre três dados posicionais. Caso o segundo dado posicional esteja registado fora do contexto do primeiro e terceiro ponto, esse ponto é removido. Apesar de ter reduzido consideravelmente o número de dados atípicos, estes não foram removidos na sua totalidade. O algoritmo de Map Matching contém uma opção de *force-repair* utilizada para reduzir o número de resultados inválidos, obrigando o algoritmo a criar trajectórias mesmo que estas não sejam consideradas os melhores mapeamentos. Esta opção não foi utilizada por

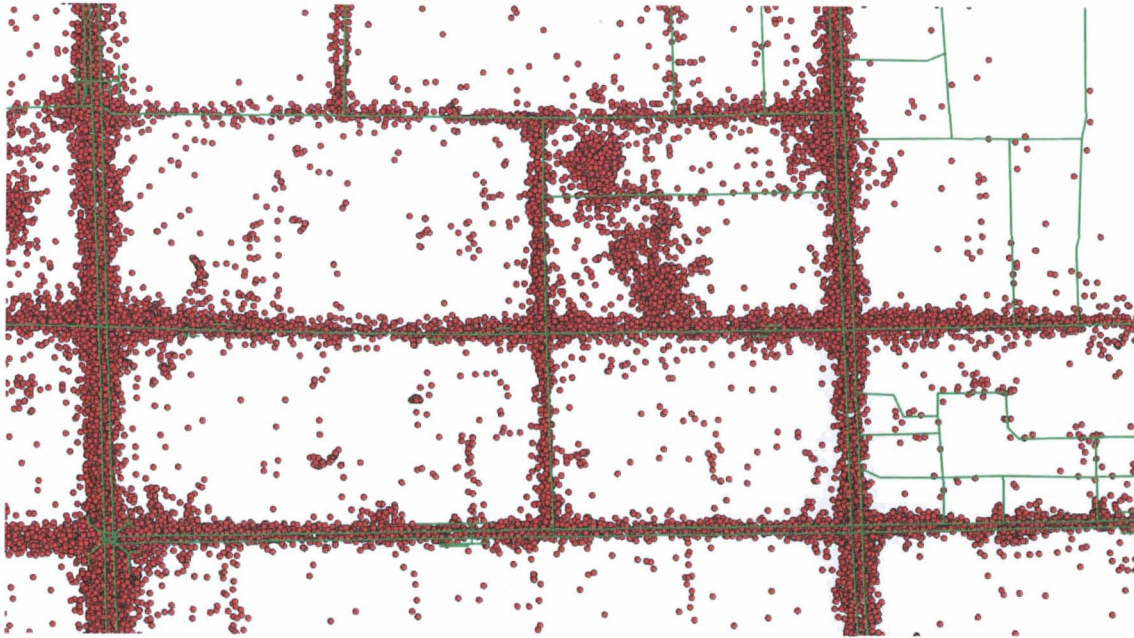


Figura 4.12: Dados posicionais de táxis em Pequim

trajectos irreais devido aos valores atípicos existentes no conjunto de dados. Um exemplo típico é o caso onde se tem um trajecto que contém uma posição atípica, a uma distância considerável (dezenas de quilómetros) do trajecto que foi efectivamente percorrido. Com a opção *force-repair* activa, o algoritmo criaria uma rota entre o trajecto o e dado atípico, resultando num trajecto válido, mas irreal, baixando significativamente o nível de similaridade dos resultados. Relativamente ao nível de trajectos válidos obtidos na execução do algoritmo de Map Matching nos dados da China, em 20.000 trajectos foram mapeados cerca de 50% dos trajectos, ou seja, não foram encontrados rotas possíveis de serem percorridas para cerca de 10.000 trajectos.

#### 4.4.2 ANÁLISE DAS ROTAS

Para a análise de rotas escolhidas pelos taxistas nas estradas da cidade de Pequim, foram escolhidos pares de intersecções tendo em conta zonas com múltiplas rotas alternativas (por exemplo, zonas residenciais, compostas por múltiplos quarteirões), zonas turísticas e zonas de estradas secundárias, ou seja, estradas que tipicamente ligam zonas residenciais e estradas primárias.

Na análise feita a pares de pontos de intersecções localizados em zonas com estradas secundárias, verifica-se que na maior parte dos casos, a rota utilizada com mais frequência é a rota mais curta e a mais rápida. Estes foram os casos que também foram encontrados mais rotas atípicas. Estas rotas atípicas comprovam a utilização dessas zonas como rotas intermediárias para outros destinos.

Na análise dos trajectos em zona residenciais, verifica-se que na maioria dos casos, a rota utilizada com mais frequência é a rota mais curta mas nem sempre é a rota mais rápida. Note-se que a maioria das zonas residenciais apresentaram rotas alternativas com distâncias muito semelhantes devido à organização rectangular das zonas residenciais. Em geral, todos os pares de intersecções em zonas residenciais apresentaram várias rotas distintas.

Na análise feita a pares de intersecções que associam zonas turísticas, verificou-se que em poucos

casos foi utilizada a rota mais rápida. Isso deve-se ao facto de na maioria das rotas feitas entre as duas intersecções, é possível denotar que foram feitos vários transportes entre os pontos de intersecção, ou seja, na maioria das rotas não houve um transporte directo entre as duas zonas turísticas. Tendo em conta o facto de a distância entre os pontos de intersecção ser maior e apresentar mais alternativas, pode ter influenciado o facto de não haver rotas com o número elevado de trajectos.

O valor da rota mais rápida é calculado através da média da diferença temporal entre a saída e a entrada dos veículos entre os pontos de intersecção. O facto de em muitas zonas secundárias as rotas utilizadas com mais frequência serem as mais rápidas, comprova que essas rotas são utilizadas eficientemente, visto que o facto delas registarem mais trajectos com um número elevado de paragens dos veículos poderia implicar uma maior duração do percurso da rota. Nas zonas residenciais, as rotas mais rápidas não eram as rotas utilizadas com mais frequência. De notar que nestes casos, as rotas mais rápidas foram em geral registados poucos trajectos e que o n<sup>o</sup> de paragens nessas rotas foi sempre menor do que o número de paragens médio em todas as rotas.

O facto dos dados terem sido registados num curto espaço de tempo (de 2 a 8 de Fevereiro) fez com que os dados meteorológicos não fossem um factor relevante na análise da selecção das rotas. Nesse espaço de tempo, o estado do tempo variou entre a chuva e a neve, e as temperaturas entre 8 graus negativos a um grau positivo. Na análise das rotas utilizadas com mais frequência, não foram encontrados padrões comportamentais que mostrassem que as condições temporais influenciavam a selecção das rotas.

Em relação ao registo temporal das viagens, foi notório que, nos casos de estudo nas estradas secundárias e nas zonas turísticas, a maioria dos trajectos foram registados em dias referentes ao ano novo chinês, mais concretamente, os dias 6, 7 e 8 de Fevereiro que foram feriados públicos.

# CONCLUSÃO E TRABALHO FUTURO

---

Foi proposto para esta Dissertação o desenho e implementação de um modelo de dados multidimensional representativos do tráfego de veículos, bem como do seu contexto.

O modelo de dados multidimensional implementado é constituído por uma Data Warehouse que pode ser vista como uma composição de 3 Data Marts, em que um faz o registo dos dados orientado ao trajecto feito pelos utilizadores, outro faz o registo dos dados orientado à posição registada pelos veículos e o último faz o registo tendo em conta a passagem de um veículo numa estrada ou troço rodoviário.

Para a contextualização dos dados, foram utilizados dados externos de várias fontes. O OpenStreet-Map foi utilizado para a obtenção do mapa rodoviário digital. Esse mapa sofreu um pós-processamento de forma a que os segmentos rodoviários ficassem devidamente organizados por interseções e segmentos. O ForecastIO foi utilizado para obter dados relativos às condições meteorológicas no momento em que os trajectos foram percorridos.

Também foram utilizados dados com o objectivo de contextualizar as datas de registo dos trajectos. Essas datas podem diferenciar os dias úteis de fins de semana, e podem diferenciar feriados.

Também foi proposto a utilização de um algoritmo de Map Matching com o objectivo de associar os dados posicionais com um mapa de estradas digital. Para isso, foi utilizado o projecto GraphHopper, capaz de receber como argumentos múltiplos trajectos e, para cada um deles, devolver um trajecto mapeado no mapa rodoviário digital. A aplicação foi modificada de forma a que a qualidade dos resultados fosse a melhor possível para o que se pretendeu para este trabalho. Enquanto que a aplicação original devolve apenas os segmentos onde o trajecto foi mapeado sem qualquer informação relativa a instantes temporais, a aplicação modificada devolve também informação relativa aos dados posicionais mapeados aos segmentos rodoviários, mantendo a informação temporal inerente a essa posição. Esta informação foi útil para determinar o número de paragens feitas durante um trajecto.

Os resultados referentes aos testes de Map Matching revelam que os resultados de mapeamento são fiáveis e que, caso o nível de valores atípicos seja baixo e se o nível de erro posicional entre as localizações e o mapa for coerente, a taxa de trajectos inválidos é baixa. A nível de utilização, concluiu-se que a discrepância de similaridade entre períodos de amostragem não são altos (ver Tabela 4.2). No caso demonstrado neste trabalho, a utilização de um período de amostragem de 10 segundos em detrimento



da utilização de um período de amostragem de 1 segundo significa que o volume de dados posicionais a ser processada é 10 vezes menor, sendo o resultado final muito semelhante. Disso podemos concluir que é viável considerar a períodos de amostragem razoáveis para que o processamento dos dados seja eficiente, perdendo um pouco da qualidade dos resultados.

Em relação à análise dos dados, concluiu-se que as rotas mais curtas foram as mais utilizadas nos casos estudados, sendo muitas delas a rota mais rápida. Nas zonas residenciais a distância das rotas eram muito semelhantes devido à organização rectangular das estradas. Nesses casos foi difícil encontrar factores que pudessem influenciar a escolha das rotas, tendo em conta que as condições meteorológicas foram relativamente constantes nesses casos.

O modelo implementado pode ser visto como uma base para uma possível criação de um sistema de recomendação, visto que com dados posicionais e a capacidade de obter o nível de utilização de troços precedentes a um determinado troço seja material suficiente para esses tipos de sistemas. Com dados relativos ao nível de utilização de um troço é possível criar rotas tendo em conta a frequência de utilização, adicionando assim uma capacidade de predição de trajectos. A essa capacidade de predição pode-se associar rotas com troços percorridos com a duração média mais baixa.

A contextualização dos trajectos pode fazer com que a recomendação se baseie em estatísticas de tráfego por estrada ou troço rodoviário de acordo com o contexto, isto é, fazer a recomendação de percursos mediante critérios como as condições meteorológicas ou temporais, por exemplo, considerando o dia da semana ou o horário.

As Data Mart orientadas aos dados posicionais e aos troços são uma fonte de dados útil para a análise de tráfego em cada segmento, bem como a velocidade média e instantânea em que cada segmento é percorrido. Estas Data Marts também poderão servir para fazer uma análise do nível de tráfego em cada segmento, bem com o número de paragens feitos por um veículo num segmento e os intervalos de tempo onde o nível de tráfego e o número de paragens aumenta.

## REFERÊNCIAS

---

- [1] M. GOLFARELLI, D. MAIO e S. RIZZI, «The dimensional fact model: A conceptual model for data warehouses», *International Journal of Cooperative Information Systems*, vol. 07, n° 02n03, pp. 215–247, 1998. DOI: 10.1142/S0218843098000118.
- [2] K. T. Mueller, P. V. Loomis, R. M. Kalafus e L. Sheynblat, *Networked differential gps system*, US Patent 5,323,322, jun. de 1994.
- [3] W. Y. Ochieng, M. Quddus e R. B. Noland, «Map-matching in complex urban road networks», *Revista Brasileira de Cartografia*, vol. 2, n° 55, 2003.
- [4] J. S. Greenfeld, «Matching gps observations to locations on a digital map», em *Transportation Research Board 81st Annual Meeting*, 2002.
- [5] D. Bernstein e A. Kornhauser, «An introduction to map matching for personal navigation assistants», 1998.
- [6] J. Bentley e H. Maurer, «Efficient worst-case data structures for range searching», English, *Acta Informatica*, vol. 13, n° 2, pp. 155–168, 1980, ISSN: 0001-5903. DOI: 10.1007/BF00263991.
- [7] C. E. White, D. Bernstein e A. L. Kornhauser, «Some map matching algorithms for personal navigation assistants», *Transp. Res. Part C Emerg. Technol.*, vol. 8, n° 1-6, pp. 91–108, fev. de 2000, ISSN: 0968090X. DOI: 10.1016/S0968-090X(00)00026-7.
- [8] B. P. Phuyal, «Method and use of aggregated dead reckoning sensor and gps data for map matching», em *Proceedings of the 15th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 2002)*, 2001, pp. 430–437.
- [9] M. A. Quddus, «High integrity map matching algorithms for advanced transport telematics applications», tese de doutoramento, Imperial College London, United Kingdom, 2006.
- [10] G. Taylor e G. Blewitt, «Road reduction filtering using gps map-matching», em *3rd Agil. Conf. Geogr. Inf. Sci.*, vol. 5, Helsinki/Espco, 2000, pp. 114–120.
- [11] C. Kee, B. W. Parkinson e P. Axelrad, «Wide area differential gps», *Navigation*, vol. 38, n° 2, pp. 123–145, 1991, ISSN: 2161-4296. DOI: 10.1002/j.2161-4296.1991.tb01720.x.
- [12] W. Chen, M. Yu, Z. Li e Y. Chen, «Integrated vehicle navigation system for urban applications», em *Proceedings of the 7th International Conference on Global Navigation Satellite Systems (GNSS)*, European Space Agency, Graz, Austria, 2003, pp. 15–22.
- [13] H. Yin e O. Wolfson, «A weight-based map matching method in moving objects databases», em *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*, jun. de 2004, pp. 437–438. DOI: 10.1109/SSDM.2004.1311248.

- [14] C. Blazquez e A. Vonderohe, «Simple map-matching algorithm applied to intelligent winter maintenance vehicle data», *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1935, pp. 68–76, 2005. DOI: 10.3141/1935-08.
- [15] J.-S. Pyo, D.-H. Shin e T.-K. Sung, «Development of a map matching method using the multiple hypothesis technique», em *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*, 2001, pp. 23–27. DOI: 10.1109/ITSC.2001.948623.
- [16] M. A. Quddus, W. Y. Ochieng, L. Zhao e R. B. Noland, «A general map matching algorithm for transport telematics applications», *GPS solutions*, vol. 7, n° 3, pp. 157–167, 2003.
- [17] F. Marchal, J. Hackney e K. W. Axhausen, «Efficient map-matching of large gps data sets - tests on a speed monitoring experiment in zurich, volume 244 of arbeitsbericht verkehrs und raumplanung», rel. téc., 2004.
- [18] M. Yu, «Improved positioning of land vehicle in its using digital map and other accessory information», tese de doutoramento, The Hong Kong Polytechnic University, 2006.
- [19] J. Krumm, E. Horvitz e J. Letchner, «Map matching with travel time constraints», SAE Technical Paper, rel. téc., 2007.
- [20] J. Yuan, Y. Zheng, C. Zhang, X. Xie e G.-Z. Sun, «An interactive-voting based map matching algorithm», em *Proceedings of the 2010 Eleventh International Conference on Mobile Data Management*, IEEE Computer Society, 2010, pp. 43–52.
- [21] I. S. K. Honey, A. C. Phillips, L. Altos e M. S. White, *United states patent [191*, 1989.
- [22] Y. Zhao, *Vehicle location and navigation system*. 1997.
- [23] E. Krakiwsky, C. Harris e R. Wong, «A kalman filter for integrating dead reckoning, map matching and gps positioning», em *Position Location and Navigation Symposium, 1988. Record. Navigation into the 21st Century. IEEE PLANS '88., IEEE*, nov. de 1988, pp. 39–46. DOI: 10.1109/PLANS.1988.195464.
- [24] J. Tanaka, K. Hirano, H. Nobuta, T. Itoh e S. Tsunoda, «Navigation system with map-matching method», SAE Technical Paper, rel. téc., 1990.
- [25] J. Takashi, M. Haseyama e H. Kitajima, «A map matching method with the innovation of the kalman filtering», *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 79, n° 11, pp. 1853–1855, 1996.
- [26] W. Kim, G.-I. Jee e J. Lee, «Efficient use of digital road map in various positioning for its», em *Position Location and Navigation Symposium, IEEE 2000*, 2000, pp. 170–176. DOI: 10.1109/PLANS.2000.838299.
- [27] L. Zhihua e W. Chen, «A new approach to map-matching and parameter correcting for vehicle navigation system in the area of shadow of gps signal», em *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, set. de 2005, pp. 449–454. DOI: 10.1109/ITSC.2005.1520086.
- [28] D. Obradovic, H. Lenz e M. Schupfner, «Fusion of map and sensor data in a modern car navigation system», English, *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 45, n° 1-2, pp. 111–122, 2006, ISSN: 0922-5773. DOI: 10.1007/s11265-006-9775-4.
- [29] D. Yang, B. Cai e Y. Yuan, «An improved map-matching algorithm used in vehicle navigation system», em *Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE*, vol. 2, out. de 2003, 1246–1250 vol.2. DOI: 10.1109/ITSC.2003.1252683.
- [30] M. El Najjar e P. Bonnifait, «A road-matching method for precise vehicle localization using belief theory and kalman filtering», English, *Autonomous Robots*, vol. 19, n° 2, pp. 173–191, 2005, ISSN: 0929-5593. DOI: 10.1007/s10514-005-0609-1.

- [31] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson e P.-J. Nordlund, «Particle filters for positioning, navigation, and tracking», *Signal Processing, IEEE Transactions on*, vol. 50, n° 2, pp. 425–437, fev. de 2002, ISSN: 1053-587X. DOI: 10.1109/78.978396.
- [32] Y. Cui e S. S. Ge, «Autonomous vehicle positioning with gps in urban canyon environments», *Robotics and Automation, IEEE Transactions on*, vol. 19, n° 1, pp. 15–25, fev. de 2003, ISSN: 1042-296X. DOI: 10.1109/TRA.2002.807557.
- [33] S. K. KIM, «Development of a map matching algorithm for car navigation system using fuzzy q-factor algorithm», em *TOWARDS THE NEW HORIZON TOGETHER. PROCEEDINGS OF THE 5TH WORLD CONGRESS ON INTELLIGENT TRANSPORT SYSTEMS, HELD 12-16 OCTOBER 1998, SEOUL, KOREA. PAPER NO. 4020*, 1998.
- [34] S. Kim e J.-H. Kim, «Adaptive fuzzy-network-based c-measure map-matching algorithm for car navigation system», *Industrial Electronics, IEEE Transactions on*, vol. 48, n° 2, pp. 432–441, abr. de 2001, ISSN: 0278-0046. DOI: 10.1109/41.915423.
- [35] S. Syed e M. Cannon, «Fuzzy logic-based map matching algorithm for vehicle navigation system in urban canyons», em *ION National Technical Meeting, San Diego, CA*, vol. 1, 2004, pp. 26–28.
- [36] M. A. Quddus, R. B. Noland e W. Y. Ochieng, «A high accuracy fuzzy logic based map matching algorithm for road transport», *Journal of Intelligent Transportation Systems*, vol. 10, n° 3, pp. 103–115, 2006. DOI: 10.1080/15472450600793560.
- [37] N. Teslya, «Web mapping service for mobile tourist guide», em *Open Innovations Association FRUCT, Proceedings of 15th Conference of*, abr. de 2014, pp. 135–143. DOI: 10.1109/FRUCT.2014.6872438.
- [38] *Google maps*, <https://maps.google.pt/>, Google.
- [39] *Maps coverage details | google maps apis | google developers*, <https://developers.google.com/maps/coverage>, Google Developers.
- [40] *The google maps directions api | google maps directions api | google developers*, <https://developers.google.com/maps/documentation/directions/intro>, Google Developers.
- [41] *Directions service | google maps javascript api | google developers*, <https://developers.google.com/maps/documentation/javascript/directions>, Google Developers.
- [42] *The google maps distance matrix api | google maps distance matrix api | google developers*, <https://developers.google.com/maps/documentation/distance-matrix/intro>, Google Developers.
- [43] *The google maps geocoding api | google maps geocoding api | google developers*, <https://developers.google.com/maps/documentation/geocoding/intro>, Google Developers.
- [44] *Bing maps*, <https://www.bing.com/maps/>, Microsoft.
- [45] *Bing maps geographic coverage*, <https://msdn.microsoft.com/en-us/library/dd435699.aspx>.
- [46] *Bing aerial imagery analyzer for openstreetmap*, <http://ant.dev.openstreetmap.org/bingimageanalyzer/>, OpenStreetMap Wiki.
- [47] *Routesapi*, <https://msdn.microsoft.com/en-us/library/ff701705.aspx>, Microsoft.
- [48] *Bing maps rest services*, <https://msdn.microsoft.com/en-us/library/ff701713.aspx>, Microsoft.
- [49] *Yandex maps - a detailed world map*, <https://maps.yandex.com/>, Yandex.

- [50] *Yandex technologies*, <https://tech.yandex.com/maps/doc/jsapi/index-docpage/>, Yandex.
- [51] *Openstreetmap*, <https://www.openstreetmap.org/>, Yandex.
- [52] *Pgrouting project - open source routing library*, <http://pgrouting.org/>, pgRouting Project.
- [53] *Graphhopper directions api with route optimization*, <https://graphhopper.com/>, GraphHopper.
- [54] Ojw, *Pyroute - openstreetmap wiki*, <http://wiki.openstreetmap.org/wiki/Pyroute>.
- [55] S. Mattheis, K. K. Al-Zahid, B. Engelmann, A. Hildisch, S. Holder, O. Lazarevych, D. Mohr, F. Sedlmeier e R. Zinck, «Putting the car on the map: A scalable map matching system for the open source community»,
- [56] A. Hann, «Motorways and firm performance: The case of hungary», tese de doutoramento, Central European University, 2014.
- [57] J. A. Castillo-Salazar, D. Landa-Silva e R. Qu, «A survey on workforce scheduling and routing problems», em *International Conference on the Practice and Theory of Automated Timetabling. Son, Norway*, Citeseer, 2012, pp. 283–302.
- [58] E. Hart, K. Sim e N. Urquhart, «A real-world employee scheduling and routing application», em *Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation*, sér. GECCO Comp '14, Vancouver, BC, Canada: ACM, 2014, pp. 1239–1242, ISBN: 9781450328814. DOI: 10.1145/2598394.2605447.
- [59] K. Sim, E. Hart, N. Urquhart e T. Pigden, «An xml object model for rich vehicle routing problems», 2015.
- [60] D. Khachay, «Gps navigation algorithm based on osm data»,
- [61] K. Kulakov e A. Shabaev, «An approach to creation of smart space-based trip planning service», em *Open Innovations Association (FRUCT16), 2014 16th Conference of*, out. de 2014, pp. 38–44. DOI: 10.1109/FRUCT.2014.7000918.
- [62] N. Urquhart, «Optimising the scheduling and planning of urban milk deliveries». English, em *Applications of Evolutionary Computation*, sér. Lecture Notes in Computer Science, A. M. Mora e G. Squillero, eds., vol. 9028, Springer International Publishing, 2015, pp. 604–615, ISBN: 9783319165486. DOI: 10.1007/978-3-319-16549-3\_49.
- [63] V. Nallur, A. Elgammal e S. Clarke, «Smart route planning using open data and participatory sensing», English, em *Open Source Systems: Adoption and Impact*, sér. IFIP Advances in Information and Communication Technology, E. Damiani, F. Frati, D. Riehle e A. I. Wasserman, eds., vol. 451, Springer International Publishing, 2015, pp. 91–100, ISBN: 9783319178363. DOI: 10.1007/978-3-319-17837-0\_9.
- [64] P. Jaccard, «Étude comparative de la distribution florale dans une portion des alpes et des jura», *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.
- [65] —, «The distribution of the flora in the alpine zone.1», *New Phytologist*, vol. 11, n° 2, pp. 37–50, 1912, ISSN: 1469-8137. DOI: 10.1111/j.1469-8137.1912.tb05611.x.
- [66] K. Grauman e T. Darrell, «Fast contour matching using approximate earth mover’s distance», em *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, IEEE, vol. 1, 2004, pp. I–220.
- [67] H. Rakha, M. Farzaneh, M. Arafah, R. Hranac, E. Sterzin e D. Krechmer, *Empirical studies on traffic flow in inclement weather*. Virginia Tech Transportation Institute, 2007.

- [68] W. Groß, «Grundzüge der mengenlehre», German, *Monatshefte für Mathematik und Physik*, vol. 26, n° 1, A34–A35, 1915, ISSN: 0026-9255. DOI: 10.1007/BF01999507.
- [69] D. Huttenlocher, G. Klanderman e W. Rucklidge, «Comparing images using the hausdorff distance», *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, n° 9, pp. 850–863, set. de 1993, ISSN: 0162-8828. DOI: 10.1109/34.232073.
- [70] M.-P. Dubuisson e A. Jain, «A modified hausdorff distance for object matching», em *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1, out. de 1994, 566–568 vol.1. DOI: 10.1109/ICPR.1994.576361.
- [71] H. Alt e L. J. Guibas, «Discrete geometric shapes: Matching, interpolation, and approximation», *Handbook of computational geometry*, vol. 1, pp. 121–153, 1999.
- [72] E. Baudrier, G. Millon, F. Nicolier e S. Ruan, «A new similarity measure using hausdorff distance map», em *Image Processing, 2004. ICIP '04. 2004 International Conference on*, vol. 1, out. de 2004, 669–672 Vol. 1. DOI: 10.1109/ICIP.2004.1418843.
- [73] J.-R. Hwang, H.-Y. Kang e K.-J. Li, «Spatio-temporal similarity analysis between trajectories on road networks», English, em *Perspectives in Conceptual Modeling*, sér. Lecture Notes in Computer Science, J. Akoka, S. Liddle, I.-Y. Song, M. Bertolotto, I. Comyn-Wattiau, W.-J. van den Heuvel, M. Kolp, J. Trujillo, C. Kop e H. Mayr, eds., vol. 3770, Springer Berlin Heidelberg, 2005, pp. 280–289, ISBN: 9783540293958. DOI: 10.1007/11568346\_30.
- [74] G.-P. Roh e S.-w. Hwang, «Nncluster: An efficient clustering algorithm for road network trajectories», English, em *Database Systems for Advanced Applications*, sér. Lecture Notes in Computer Science, H. Kitagawa, Y. Ishikawa, Q. Li e C. Watanabe, eds., vol. 5982, Springer Berlin Heidelberg, 2010, pp. 47–61, ISBN: 9783642120978. DOI: 10.1007/978-3-642-12098-5\_4.
- [75] S. Nutanong, E. H. Jacox e H. Samet, «An incremental hausdorff distance calculation algorithm», *Proc. VLDB Endow.*, vol. 4, n° 8, pp. 506–517, mai. de 2011, ISSN: 2150-8097. DOI: 10.14778/2002974.2002978.
- [76] G.-P. Roh, J.-W. Roh, S.-W. Hwang e B.-K. Yi, «Supporting pattern-matching queries over trajectories on road networks», *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, n° 11, pp. 1753–1758, nov. de 2011, ISSN: 1041-4347. DOI: 10.1109/TKDE.2010.189.
- [77] R. Enayatifar e R. A. Salam, «Similarity measure using hausdorff distance in 2d shape recognition system», em *2nd International Symposium on Computer, Communication, Control and Automation*, Atlantis Press, 2013.
- [78] R. Kimball e K. Strehlo, «Why decision support fails and how to fix it», *SIGMOD Rec.*, vol. 24, n° 3, pp. 92–97, set. de 1995, ISSN: 0163-5808. DOI: 10.1145/211990.212023.
- [79] W. H. Inmon, *Building the Data Warehouse*. New York, NY, USA: John Wiley & Sons, Inc., 1992, ISBN: 0471569607.
- [80] C. Franklin, «An introduction to geographic information systems: Linking maps to databases», *Database*, vol. 15, n° 2, pp. 12–21, abr. de 1992, ISSN: 0162-4105.
- [81] A. Gupta, V. Harinarayan e D. Quass, «Aggregate-query processing in data warehousing environments», em *Proceedings of the 21th International Conference on Very Large Data Bases*, sér. VLDB '95, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 358–369, ISBN: 1558603794.
- [82] Y. Zhuge, H. Garcia-Molina, J. Hammer e J. Widom, «View maintenance in a warehousing environment», *SIGMOD Rec.*, vol. 24, n° 2, pp. 316–327, mai. de 1995, ISSN: 0163-5808. DOI: 10.1145/568271.223848.

- [83] S. Chaudhuri, R. Krishnamurthy, S. Potamianos e K. Shim, *Optimizing queries with materialized views*, 1995.
- [84] J. Goldstein e P.-bibinitperiod Larson, «Optimizing queries using materialized views: A practical, scalable solution», *SIGMOD Rec.*, vol. 30, n° 2, pp. 331–342, mai. de 2001, ISSN: 0163-5808. DOI: 10.1145/376284.375706.
- [85] A. Y. Levy, A. O. Mendelzon e Y. Sagiv, «Answering queries using views (extended abstract)», em *Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, sér. PODS '95, San Jose, California, USA: ACM, 1995, pp. 95–104, ISBN: 0897917308. DOI: 10.1145/212433.220198.
- [86] R. Kimball, *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. New York, NY, USA: John Wiley & Sons, Inc., 1996, ISBN: 0471153370.
- [87] V. Harinarayan, A. Rajaraman e J. D. Ullman, «Implementing data cubes efficiently», *SIGMOD Rec.*, vol. 25, n° 2, pp. 205–216, jun. de 1996, ISSN: 0163-5808. DOI: 10.1145/235968.233333.
- [88] S. Agarwal, R. Agrawal, P. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan e S. Sarawagi, «On the computation of multidimensional aggregates», em *Proceedings of the 22th International Conference on Very Large Data Bases*, sér. VLDB '96, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996, pp. 506–521, ISBN: 1558603824.
- [89] S. Chaudhuri e U. Dayal, «An overview of data warehousing and olap technology», *SIGMOD Rec.*, vol. 26, n° 1, pp. 65–74, mar. de 1997, ISSN: 0163-5808. DOI: 10.1145/248603.248616.
- [90] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow e H. Pirahesh, «Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals», *Data Min. Knowl. Discov.*, vol. 1, n° 1, pp. 29–53, jan. de 1997, ISSN: 1384-5810. DOI: 10.1023/A:1009726021843.
- [91] M.-c. Wu, R. P. Buchmann, D. F. Informatik e T. H. Darmstadt, «Research issues in data warehousing», em *In Datenbanksysteme in Büro, Technik und Wissenschaft*, 1997, pp. 61–82.
- [92] R. Kimball e J. Caserta, *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data*. John Wiley & Sons, 2004, ISBN: 0764567578.
- [93] J. Song, C. Guo, Z. Wang, Y. Zhang, G. Yu e J.-M. Pierson, «Haolap», *J. Syst. Softw.*, vol. 102, n° C, pp. 167–181, abr. de 2015, ISSN: 0164-1212. DOI: 10.1016/j.jss.2014.09.024.
- [94] J. Han, N. Stefanovic e K. Koperski, «Selective materialization: An efficient method for spatial data cube construction», em *In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'98)*, 1998, pp. 144–158.
- [95] S. Rivest, Y. Bédard e P. March, «Towards better support for spatial decision-making: Defining the characteristics», em *Geomatica*, 2001, pp. 539–555.
- [96] H. B. Zghal, S. Faiz e H. H. B. Ghézala, «Casme: A case tool for spatial data marts design and generation.», em *DMDW*, 2003.
- [97] E. Malinowski e E. Zimányi, «Spatial data warehouses», em *Advanced Data Warehouse Design*, Springer Berlin Heidelberg, 2008, pp. 137–184.
- [98] N. Beckmann, H.-P. Kriegel, R. Schneider e B. Seeger, «The r\*-tree: An efficient and robust access method for points and rectangles», *SIGMOD Rec.*, vol. 19, n° 2, pp. 322–331, mai. de 1990, ISSN: 0163-5808. DOI: 10.1145/93605.98741.
- [99] T. Brinkhoff, H.-P. Kriegel e B. Seeger, «Efficient processing of spatial joins using r-trees», *SIGMOD Rec.*, vol. 22, n° 2, pp. 237–246, jun. de 1993, ISSN: 0163-5808. DOI: 10.1145/170036.170075.

- [100] T. Brinkhoff, H. P. Kriegel e B. Seeger, «Parallel processing of spatial joins using r-trees», em *Data Engineering, 1996. Proceedings of the Twelfth International Conference on*, fev. de 1996, pp. 258–265. DOI: 10.1109/ICDE.1996.492114.
- [101] X. Wang, X. Zhou e S. Lu, «Spatiotemporal data modelling and management: A survey», em *Technology of Object-Oriented Languages and Systems, 2000. TOOLS - Asia 2000. Proceedings. 36th International Conference on*, 2000, pp. 202–211. DOI: 10.1109/TOOLS.2000.885919.
- [102] C. V. Tao, «Spatial data warehousing: A strategy for integrated urban data management in support of decision making», *Geographic Information Sciences*, vol. 6, n° 2, pp. 113–120, 2000.
- [103] T. Merret e J. Han, «Fundamentals of spatial data warehousing for geographic knowledge discovery», *Data Mining and Knowledge Discovery*, 2001.
- [104] B. Yvan, M.-J. Proulx, S. Larrivée e E. Bernier, «Modeling multiple representations into spatial data warehouses: A uml-based approach», em *Symposium on Geospatial Theory, Processing and Applications, Ottawa, Canada*, Citeseer, 2002.
- [105] H. B. Zghal, S. Faïz e H. H. B. Ghézala, «Casmé: A case tool for spatial data marts design and generation.», em *DMDW*, H.-J. Lenz, P. Vassiliadis, M. A. Jeusfeld e M. Staudt, eds., sér. CEUR Workshop Proceedings, vol. 77, CEUR-WS.org, 23 de jan. de 2006.
- [106] S. Nadi e M. R. Delavar, «Spatio-temporal modeling of dynamic phenomena in gis.», em *ScanGIS*, 2003, pp. 215–225.
- [107] M. Gorawski e R. Malczok, «On efficient storing and processing of long aggregate lists», em *Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery*, sér. DaWaK'05, Copenhagen, Denmark: Springer-Verlag, 2005, pp. 190–199, ISBN: 3-540-28558-X, 978-3-540-28558-8. DOI: 10.1007/11546849\_19.
- [108] —, «Materialized ar-tree in distributed spatial data warehouse», *Intell. Data Anal.*, vol. 10, n° 4, pp. 361–377, dez. de 2006, ISSN: 1088-467X.
- [109] A. Escribano, L. Gomez. B. Kuijpers e A. A. Vaisman, «Piet: A gis-olap implementation», em *Proceedings of the ACM Tenth International Workshop on Data Warehousing and OLAP*, sér. DOLAP '07, Lisbon, Portugal: ACM, 2007, pp. 73–80, ISBN: 9781595938275. DOI: 10.1145/1317331.1317345.
- [110] O. Glorio e J. Trujillo, «An mda approach for the development of spatial data warehouses», em *Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery*, sér. DaWaK '08, Turin, Italy: Springer-Verlag, 2008, pp. 23–32, ISBN: 9783540858355. DOI: 10.1007/978-3-540-85836-2\_3.
- [111] W. Huibing, «Extending objectrelational database to support spatiotemporal data», *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 37, pp. 1682–1750, 2008.
- [112] S. Aissi e M. S. Gouider, «Spatial and spatio-temporal multidimensional data modelling: A survey», *CoRR*, vol. abs/1208.0163, 2012.
- [113] N. Stefanovic, J. Han e K. Koperski, «Object-based selective materialization for efficient implementation of spatial data cubes», *Knowledge and Data Engineering, IEEE Transactions on*, vol. 12, n° 6, pp. 938–958, nov. de 2000, ISSN: 1041-4347. DOI: 10.1109/69.895803.
- [114] E. Malinowski e E. Zimányi, «Spatial hierarchies and topological relationships in the spatial multidimer model», em *Proceedings of the 22Nd British National Conference on Databases: Enterprise, Skills and Innovation*, sér. BNCOD'05, Sunderland, UK: Springer-Verlag, 2005, pp. 17–28, ISBN: 3-540-26973-8, 978-3-540-26973-1. DOI: 10.1007/11511854\_2.



- [115] M. Miquel, Y. Bédard e A. Brisebois, «Conception d'entrepôts de données géospatiales à partir de sources hétérogènes exemple d'application en foresterie», *Ingénierie des systèmes d'information*, vol. 12, 2002.
- [116] C. Bauzer-Medeiros, O. Carles, G. Jomier, G. Hébrail, F. De Vuyst, M. Joliveau, B. Hugueney, M. Manouvrier, Y. Naija, G. Scemama et al., «Vers un entrepôt de données pour le trafic routier», 2006.
- [117] O. Glorio e J. Trujillo, «An mda approach for the development of spatial data warehouses», English, em *Data Warehousing and Knowledge Discovery*, sér. Lecture Notes in Computer Science, I.-Y. Song, J. Eder e T. Nguyen, eds., vol. 5182, Springer Berlin Heidelberg, 2008, pp. 23–32, ISBN: 9783540858355. DOI: 10.1007/978-3-540-85836-2\_3.
- [118] M.-J. Kyung, J.-H. Yom e S.-Y. Kim, «Spatial data warehouse design and spatial olap implementation for decision making of geospatial data update», *KSCE Journal of Civil Engineering*, vol. 16, n° 6, pp. 1023–1031, 2012.
- [119] A. Aji, F. Wang, H. Vo, R. Lee, Q. Liu, X. Zhang e J. Saltz, «Hadoop gis: A high performance spatial data warehousing system over mapreduce», *Proc. VLDB Endow.*, vol. 6, n° 11, pp. 1009–1020, ago. de 2013, ISSN: 2150-8097. DOI: 10.14778/2536222.2536227.
- [120] G. Marketos, E. Frentzos, I. Ntoutsis, N. Pelekis, A. Raffaetà e Y. Theodoridis, «Building real-world trajectory warehouses», em *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, sér. MobiDE '08, Vancouver, Canada: ACM, 2008, pp. 8–15, ISBN: 9781605582214. DOI: 10.1145/1626536.1626539.
- [121] F. Giannotti, M. Nanni, F. Pinelli e D. Pedreschi, «Trajectory pattern mining», em *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, sér. KDD '07, San Jose, California, USA: ACM, 2007, pp. 330–339, ISBN: 9781595936097. DOI: 10.1145/1281192.1281230.
- [122] J.-G. Lee, J. Han e K.-Y. Whang, «Trajectory clustering: A partition-and-group framework», em *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, sér. SIGMOD '07, Beijing, China: ACM, 2007, pp. 593–604, ISBN: 9781595936868. DOI: 10.1145/1247480.1247546.
- [123] O. Andersen, B. B. Krogh, C. Thomsen e K. Torp, «An advanced data warehouse for integrating large sets of gps data», em *Proceedings of the 17th International Workshop on Data Warehousing and OLAP*, sér. DOLAP '14, Shanghai, China: ACM, 2014, pp. 13–22, ISBN: 9781450309998. DOI: 10.1145/2666158.2666172.
- [124] H. Xie, E. Tanin, L. Kulik, P. Scheuermann, G. Trajcevski e M. Fanaeepour, «Euler histogram tree: A spatial data structure for aggregate range queries on vehicle trajectories», em *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, sér. IWCTS '14, Dallas/Fort Worth, Texas: ACM, 2014, pp. 18–24, ISBN: 9781450331388. DOI: 10.1145/2674918.2674921.
- [125] L. Leonardi, S. Orlando, A. Raffaetà, A. Roncato e C. Silvestri, «Frequent spatio-temporal patterns in trajectory data warehouses», em *Proceedings of the 2009 ACM Symposium on Applied Computing*, sér. SAC '09, Honolulu, Hawaii: ACM, 2009, pp. 1433–1440, ISBN: 9781605581668. DOI: 10.1145/1529282.1529603.
- [126] S. Gaffney e P. Smyth, «Trajectory clustering with mixtures of regression models», em *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, sér. KDD '99, San Diego, California, USA: ACM, 1999, pp. 63–72, ISBN: 1581131437. DOI: 10.1145/312129.312198.
- [127] Y. Li, J. Han e J. Yang, «Clustering moving objects», em *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, sér. KDD '04,

- Seattle, WA, USA: ACM, 2004, pp. 617–622, ISBN: 1581138881. DOI: 10.1145/1014052.1014129.
- [128] M. Nanni e D. Pedreschi, «Time-focused clustering of trajectories of moving objects», English, *Journal of Intelligent Information Systems*, vol. 27, n° 3, pp. 267–289, 2006, ISSN: 0925-9902. DOI: 10.1007/s10844-006-9953-7.
- [129] L. Leonardi, S. Orlando, A. Raffaeta, A. Roncato, C. Silvestri, G. Andrienko e N. Andrienko, «A general framework for trajectory data warehousing and visual olap», *Geoinformatica*, vol. 18, n° 2, pp. 273–312, abr. de 2014, ISSN: 1384-6175. DOI: 10.1007/s10707-013-0181-3.
- [130] J. C. Tanner, «Effect of weather on traffic flow», *Nature*, vol. 169, n° 4290, pp. 107–107, jan. de 1952. DOI: <http://dx.doi.org/10.1038/169107a0>.
- [131] Y. A. Hassan e D. J. Barker, «The impact of unseasonable or extreme weather on traffic activity within lothian region, scotland», *Journal of Transport Geography*, vol. 7, n° 3, pp. 209–213, 1999, ISSN: 0966-6923. DOI: [http://dx.doi.org/10.1016/S0966-6923\(98\)00047-7](http://dx.doi.org/10.1016/S0966-6923(98)00047-7).
- [132] T. Maze, M. Agarwai e G. Burchett, «Whether weather matters to traffic demand, traffic safety, and traffic operations and flow», *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1948, pp. 170–176, 2006. DOI: 10.3141/1948-19.
- [133] K. Keay e I. Simmonds, «The association of rainfall and other weather variables with road traffic volume in melbourne, australia», *Accident Analysis & Prevention*, vol. 37, n° 1, pp. 109–124, 2005, ISSN: 0001-4575. DOI: <http://dx.doi.org/10.1016/j.aap.2004.07.005>.
- [134] M. Cools, E. Moons e G. Wets, «Assessing the impact of weather on traffic intensity», *Weather, Climate, and Society*, vol. 2, n° 1, pp. 60–68, 2010.
- [135] W. Min e L. Wynter, «Real-time road traffic prediction with spatio-temporal correlations», *Transportation Research Part C: Emerging Technologies*, vol. 19, n° 4, pp. 606–616, 2011, ISSN: 0968-090X. DOI: <http://dx.doi.org/10.1016/j.trc.2010.10.002>.
- [136] C. S. Jensen, H. Lahrman, S. Pakalnis e J. Runge, «The infati data», *ArXiv preprint cs/0410001*, 2004.
- [137] C. S. Jensen, H. Lahrman e J. Runge, «User guide of t-drive data», 2008.
- [138] J. Yuan, Y. Zheng, X. Xie e G. Sun, «Driving with knowledge from the physical world», em *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2011, pp. 316–324.
- [139] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun e Y. Huang, «T-drive: Driving directions based on taxi trajectories», em *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*, ACM, 2010, pp. 99–108.