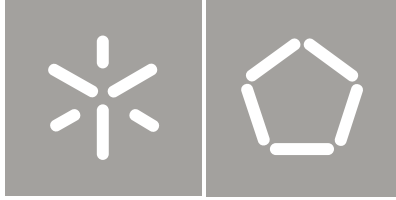


Universidade do Minho
Escola de Engenharia

Ângelo Miguel de Oliveira Conde

Mapas de Mobilidade Pessoal



Universidade do Minho

Escola de Engenharia

Ângelo Miguel de Oliveira Conde

Mapas de Mobilidade Pessoal

Dissertação de Mestrado
Mestrado em Engenharia e Gestão de Sistemas de
Informação

Trabalho efetuado sob a orientação do
Professor Doutor Adriano Moreira

Agradecimentos

A realização deste estudo foi duro e trabalhoso mas foi também o cumprir de um objetivo. Contudo, não teria sido possível sem a ajuda e o apoio de várias pessoas que me acompanharam ao longo desta dissertação.

Em primeiro lugar quero agradecer ao meu orientador, Professor Doutor Adriano Moreira, pela sua constante disponibilidade e inspiração ao longo deste trabalho.

Aos meus pais que sempre me apoiaram para continuar os meus estudos. Tenho também de agradecer à restante família por todo apoio e coragem que me transmitiram para seguir em frente.

Um grande obrigado também à minha namorada Cátia que me acompanhou ao longo de todo o processo com o seu apoio, carinho, amor e amizade.

Aos meus colegas de laboratório um muito obrigado pelo companheirismo, amizade, apoio e contribuição para um bom ambiente de trabalho.

Finalmente, mas não menos importante, quero agradecer a todos os meus Amigos.

A todos, manifesto aqui a minha sincera gratidão,

Obrigado!

Resumo

A computação móvel e ubíqua tem evoluído de forma significativa desde 1991, ano em que Weiser teve uma visão pioneira do que esta seria. O surgimento dos *smartphones*, o desenvolvimento de aplicações para esses dispositivos, novas formas de interação e sobretudo a inclusão de sensores com capacidade para recolher e capturar o contexto do utilizador, permitiu fornecer serviços centrados nas pessoas que melhorem a sua qualidade de vida.

A análise de movimento é, atualmente, uma das atividades mais relevantes no contexto da computação ubíqua e na computação urbana. É uma disciplina bastante abrangente que recorre a vários tipos de sensores para recolha de dados, a tipos de dados de natureza diversa, a estratégias de recolha de dados participativas e/ou colaborativas, a um leque alargado de técnicas de análise desses dados, e a um conjunto de técnicas de visualização de dados espaço-temporais.

Tirando partido da grande proliferação das redes Wi-Fi, no âmbito deste trabalho foi desenvolvido um algoritmo que, usando somente a identificação do AP com maior força de sinal, respetivos APs vizinhos e sem conhecimento prévio da topologia da rede, é capaz de identificar os locais relevantes no quotidiano do utilizador. Esses locais, e respetivas transições entre eles, permitem a construção dos Mapas de Mobilidade Pessoal que representam os hábitos de mobilidade do utilizador.

Abstract

Mobile and ubiquitous computing has evolved significantly since 1991, the year when Weiser had a pioneering vision of what would be. The emergence of new mobile devices, the development of applications for those devices, new forms of interaction and especially the inclusion of sensors with the capacity to collect and capture the context of the user, allowed to provide people-centric services which improve their quality of life.

The movement analysis is currently one of the most relevant activities in the context of ubiquitous and urban computing. It is a broad discipline that uses many types of sensors for data collection, various data types of different nature, various strategies of participatory and/or collaborative data collection, a wide range of techniques for data analysis, and a set of visualization techniques for spatio-temporal data.

As part of this work, and taking advantage of the great Wi-Fi networks proliferation, a new algorithm was developed, using only the identification of the AP with the highest signal strength, as well as the respective neighboring APs, and without prior knowledge of network topology, is able to identify the relevant places in the user's everyday life. These places, and the displacements between them, allow the creation of Personal Mobility Maps which represent the mobility habits of the user.

Conteúdo

1	Introdução	1
1.1	Enquadramento e Motivação	1
1.2	Objetivos e Resultados Esperados	4
1.3	Abordagem Metodológica	5
1.3.1	Identificação do Problema	5
1.3.2	Abordagem e Desafios	9
1.3.3	Avaliação	12
1.4	Estrutura da Dissertação	13
2	Estado da Arte	15
2.1	Contexto	16
2.2	Posicionamento	20
2.2.1	Posicionamento via Satélite	21
2.2.2	Posicionamento Alternativo ao Satélite	22
2.3	Inferência de Informação de Localização	26
2.3.1	Algoritmos Geométricos	27
2.3.2	Algoritmos <i>Fingerprint</i>	30
2.4	Mobilidade Humana	35
3	Descoberta de Locais	37
3.1	Arquitetura da Solução	37
3.1.1	Dados em Bruto	39

3.1.2	Observações	40
3.1.3	Locais Potenciais	41
3.1.4	Estadias	43
3.1.5	Locais Visitados e Locais Relevantes	43
3.2	Algoritmo	44
3.2.1	Extração do Grafo	44
3.2.2	Deteção de Comunidades	45
3.2.3	Extração de Estadias	48
3.2.4	Classificação de Locais	49
3.2.5	Mapa de Mobilidade Pessoal	51
4	Implementação	53
4.1	Dados em Bruto	53
4.1.1	Recolha	53
4.1.2	Tratamento	56
4.1.3	Armazenamento	57
4.2	Carregamento dos Dados	58
4.3	Extração das Observações e do Grafo	62
4.4	Deteção de Comunidades	65
4.5	Extração de Estadias	67
4.6	Classificação de Locais	70
4.7	Mapas de Mobilidade Pessoal	72
5	Resultados e Validação	75
5.1	Recolha de Dados	76
5.2	Descoberta de Locais	78
5.3	Avaliação	80
5.3.1	Confronto com outros Algoritmos	89
5.4	Classificação de Locais	91

5.5	Mapas de Mobilidade Pessoal	96
6	Conclusões e Trabalho Futuro	103
6.1	Conclusões do Projeto	103
6.2	Trabalho Futuro	105
	Referências	107

Lista de Figuras

1.1	Informação dos locais e respetivas trajetórias entre eles relevantes para o utilizador.	8
1.2	Informação detalhada sobre a quantidade de CO ₂ produzida, custos associados e conselhos úteis para uma mobilidade eficiente.	9
1.3	Arquitetura geral do PPS6.	10
1.4	Exemplo da representação de um Mapa de Mobilidade Pessoal.	12
2.1	Arquitetura geral da plataforma LifeMap	34
3.1	Arquitetura de todo o processo de descoberta de locais disposta em camadas	38
3.2	Representação abstrata de um local	42
3.3	Exemplo das duas abordagens que serviram de entrada ao algoritmo de deteção de comunidades	45
3.4	Método de Louvain	47
3.5	Classificação dos locais segundo a sua mobilidade e permanência	51
4.1	Aplicação MOVE	54
4.2	Exemplo dos dados em bruto provenientes da aplicação MOVE	55
4.3	Processo de filtração dos endereços MAC associados a APs móveis	57
4.4	Repositório de suporte a todo o processo de descoberta de locais	58

4.5	Classe de dados JAVA representativa da estrutura dos ficheiros de dados JSON	60
4.6	Classe de dados JAVA representativa da lista WIFI_DATA presente no ficheiro JSON	61
4.7	Conversão de um ficheiro JSON para um objeto Java	62
4.8	Consulta SQL para a extração do grafo (primeira e segunda abordagem)	64
4.9	Funções existentes no algoritmo de deteção de comunidades desenvolvido pelos próprios autores do algoritmo do método de Louvain	65
4.10	Processo de deteção de estadias desenvolvido em Java	69
4.11	Extração dos dois critérios a utilizar no filtro de classificação de locais	71
5.1	Representação das observações em forma de grafo (1 ^a abordagem)	79
5.2	Locais recordados pelo utilizador no diário e locais descobertos pelo algoritmo de deteção	81
5.3	Estadias detetadas pelo algoritmo vs Estadias reais presentes no diário	85
5.4	Função densidade de probabilidade da diferença de tempos de entrada e saída	88
5.5	A forma dos locais mais relevantes	92
5.6	Distribuição dos locais visitados ao longo de 3 meses	95
5.7	Mapas de Mobilidade Pessoal dos locais visitados e dos locais relevantes dos três utilizadores	97
5.8	Perfis temporais do utilizador A nos locais relevantes ao longo de três meses	101

Lista de Tabelas

4.1	Conjunto de observações obtidas a partir dos dados em bruto .	63
4.2	Exemplo da organização de um ficheiro de entrada ao algoritmo de Louvain	64
4.3	Exemplo da organização de um ficheiro de saída gerado pelo algoritmo de Louvain	67
4.4	Exemplo de um conjunto de estadias obtidas a partir das observações e das comunidades detetadas	68
4.5	Exemplo dos critérios obtidos a partir da consulta SQL da Figura 4.11	72
5.1	Exemplo da organização de um diário produzido por um utilizador	77
5.2	Níveis hierárquicos obtidos com uma partição inicial de N nodos e E arestas	78
5.3	Avaliação quantitativa dos locais descobertos a partir de dados recolhidos pelo utilizador A	82
5.4	Avaliação quantitativa dos locais descobertos a partir de dados recolhidos pelo utilizador B	83
5.5	Resultados por estadias em locais	84
5.6	Resultados por estadias em locais	85
5.7	Caraterização temporal dos locais relevantes (Utilizador A) . .	96

5.8	Caraterização temporal das transições mais frequentes (Utilizador A)	98
-----	--	----

Lista de Acrónimos

AP	<i>Access Point</i>
BSSID	<i>Basic Service Set Identification</i>
GPS	<i>Global Positioning System</i>
GSM	<i>Global System for Mobile Communications</i>
IMU	<i>Inertial Measurement Unit</i>
JDBC	<i>Java Database Connectivity</i>
MAC	<i>Media Access Control</i>
MMP	Mapas de Mobilidade Pessoal
PDA	<i>Personal Digital Assistant</i>
POI	<i>Point Of Interest</i>
PPS	Produtos, Processos ou Sistemas
SGBD	Sistema de Gestão de Base de Dados
SSID	<i>Service Set Identifier</i>
RSSI	<i>Received Signal Strength Indication</i>
RTLS	<i>Real Time Location Systems</i>
UI	<i>User Interface</i>
WEP	<i>Wired Equivalent Privacy</i>
WPA	<i>Wi-Fi Protected Access</i>
WPS	<i>Wi-Fi Positioning System</i>

Capítulo 1

Introdução

Este capítulo contém uma introdução à construção dos Mapas de Mobilidade Pessoal. Em primeiro lugar, apresenta-se uma visão geral do trabalho através do respetivo enquadramento, seguindo-se os objetivos e principais resultados esperados com a realização deste trabalho. Na parte final do capítulo é apresentada, de forma sucinta, a organização da presente dissertação.

1.1 Enquadramento e Motivação

Nos últimos anos tem-se assistido à proliferação das redes Wi-Fi, não só em ambientes públicos, mas também em ambientes residenciais. Estas redes garantem uma enorme mobilidade aos dispositivos móveis, podendo o acesso à informação ser realizado em qualquer local e a qualquer momento. Esta evolução deveu-se sobretudo à necessidade que as pessoas têm em aceder a serviços e dados existentes nas redes.

Os *smartphones* são cada vez mais tratados como um objeto intrínseco à vida quotidiana dos seres humanos. Tem-se assistido a uma ampla utilização destes dispositivos na vida diária dos indivíduos. Espera-se que em breve estes pequenos dispositivos atinjam níveis elevados de penetração, se é que já

não atingiram. Estes possuem a capacidade de se ligarem a várias redes de comunicações, uma elevada capacidade de processamento e armazenamento de dados, e estão dotados de um conjunto variado de sensores. Estes últimos geram uma sequência de sinais espaço-temporais que permitem o estudo dos padrões comportamentais das pessoas, como os seus hábitos de mobilidade.

A análise de movimento é, atualmente, uma disciplina bastante abrangente que recorre a vários tipos de sensores para a recolha de dados, a tipos de dados de natureza diversa (geométricos e simbólicos), a estratégias de recolha de dados participativas e/ou colaborativas, a um leque alargado de técnicas de análise desses dados, e a um conjunto de técnicas de visualização de dados espaço-temporais. Genericamente, a análise de movimento aplica-se a vários tipos de contextos que não apenas o movimento humano, mas também o do movimento de veículos (automóveis, barcos, etc.), e ao movimento de animais.

A movimentação de pessoas e bens em espaços urbanos é, atualmente, um dos grandes desafios das grandes cidades. Nas últimas décadas tem-se assistido a uma transferência da população dos meios rurais para as cidades, o que tem conduzido a um crescimento muito rápido do número de habitantes nos espaços urbanos. Este crescimento das cidades contribui para agravar um conjunto de problemas associados aos espaços urbanos, tais como a poluição, o fornecimento de alimentos e as necessidades de mobilidade. No caso das necessidades de mobilidade, o uso de meios de transporte individuais, e em particular dos automóveis, conduz a maiores níveis de poluição, à necessidade de uma rede viária mais alargada, à necessidade de mais espaço de estacionamento e a um maior consumo dos recursos energéticos.

No âmbito do projeto TICE.Mobilidade, de que a Universidade do Minho é parceira, estão a ser estudados mecanismos que contribuam para resolver ou atenuar alguns destes problemas de mobilidade das cidades atuais. Em particular, pretende-se fomentar a utilização de meios de transporte públicos.

A grande motivação para a realização do presente trabalho é contribuir para a resolução dos problemas de mobilidade pessoal, através da modelação do comportamento espaço-temporal dos humanos. No contexto do projeto TICE.Mobilidade, mais propriamente no Produtos, Processos ou Sistemas (PPS) Eco.Circuitos, interessa sobretudo produzir mapas que representem os hábitos de mobilidade de cada indivíduo. Esses mapas, designados por Mapas de Mobilidade Pessoal, deverão refletir os padrões de movimento. Com a construção destes mapas, pretende-se tornar os habitantes das cidades mais conscientes dos seus hábitos de mobilidade e, conseqüentemente, corrigirem algumas das suas práticas que sejam menos eficientes. Pretende-se sobretudo que as pessoas utilizem modos alternativos de transporte de modo a otimizar o seu tempo (por exemplo, evitando congestionamentos), conforto (como o da dificuldade/custo de estacionar) e diminuir os seus consumos mensais.

Para a construção dos Mapas de Mobilidade Pessoal, será necessário adquirir informação de contexto do utilizador, utilizando um ou mais sensores. Estes dados recolhidos representam uma ou mais dimensões do contexto do utilizador. O contexto é qualquer informação que pode ser utilizada para caracterizar a situação de uma entidade, seja esta uma pessoa, local ou um mero objeto (Dey, 2001). Para além de adquirir dados através de sensores, pretende-se estimar ou inferir novas dimensões do contexto através da utilização de algoritmos de inferência. Desta forma, o contexto poderá conter dados que não resultem diretamente de nenhum sensor mas que sejam o resultado de processos de inferência aplicados aos dados recebidos dos sensores. Adquirida e inferida a informação de contexto do utilizador, esta será fundamental para a realização de análises de mobilidade humana.

1.2 Objetivos e Resultados Esperados

O trabalho a realizar no âmbito desta dissertação tem como objetivo principal dar resposta à seguinte questão de investigação: “**Como criar automaticamente, a partir de dados recolhidos por sensores, mapas que representem os hábitos de mobilidade de um indivíduo?**”.

Todo o trabalho focar-se-á na construção dos Mapas de Mobilidade Pessoal (MMP). Estes mapas deverão refletir os padrões de movimento, e serão representados como base no conceito de grafo: os nodos do grafo representam os locais relevantes para cada pessoa, enquanto que as arestas do grafo representam as transições entre locais, ou viagens.

A construção dos MMP recorre a dados sobre as redes Wi-Fi, recolhidos através de *smartphones* utilizados por cada indivíduo e posteriormente processados por um conjunto de algoritmos de inferência. Assim, interessa identificar e/ou desenvolver algoritmos que, a partir dos dados recolhidos pelos sensores presentes nos dispositivos, sejam capazes de atingir os seguintes resultados:

- Identificar e caraterizar os locais relevantes (descoberta dos locais, classificação dos locais, caraterização dos perfis temporais de permanência em locais) e representá-los como nodos do grafo;
- Identificar e caraterizar as viagens mais frequentes entre esses locais (detecção de fluxos e respetiva frequência e tempos de viagem) e representá-las como arestas do grafo.

Apesar de não ser um objetivo desta dissertação, os MMP serão responsáveis por enriquecer um conjunto de relatórios de eficiência pessoais sobre a mobilidade de um utilizador. Estes relatórios pretendem consciencializar e assistir os utilizadores para uma utilização eficiente e sustentável dos sistemas de transportes nas áreas urbanas.

Não constituem objetivos deste trabalho os processos de recolha de dados, nem a avaliação do impacto que os relatórios de eficiência pessoais possam ter na mudança de hábitos de mobilidade.

1.3 Abordagem Metodológica

O trabalho desta dissertação basear-se-à na abordagem *Design Science* ou *Constructive Research* para atingir os resultados esperados. Por norma, todos os trabalhos idênticos são construtivos, pois envolvem o desenvolvimento de um algoritmo e devida validação do mesmo, bem como resultados quantitativos provenientes de uma avaliação ou *benchmarking* com outros algoritmos semelhantes. A abordagem metodológica a seguir é definida como um procedimento de investigação para a produção de construções inovadoras, artefatos que visam servir os propósitos dos humanos, isto é, que tenham utilidade para nós e que essa utilidade seja demonstrada (March & Smith, 1995). O objetivo passa por contribuir para a resolução dos problemas de mobilidade pessoal, através da modelação do comportamento espaço-temporal dos humanos. No contexto do projeto TICE.Mobilidade, mais propriamente no Produtos, Processos ou Sistemas (PPS) Eco.Circuitos, interessa sobretudo produzir mapas que representem os hábitos de mobilidade de cada indivíduo.

1.3.1 Identificação do Problema

A mobilidade é um dos grandes problemas da sociedade moderna. Devido à concentração de mais de 60% da população nas zonas urbanas, onde o PIB é criado, emerge um conjunto de problemas. O aumento do tráfego nas cidades tem consequências cada vez mais nefastas em termos sociais, económicos e ambientais: o número de acidentes aumenta; a economia europeia perde 1% do PIB devido aos problemas de Mobilidade; a movimentação nas zonas urbanas

é responsável por cerca de 40% das emissões de CO_2 ; 70% das emissões de outros poluentes provem de transportes rodoviários.

O projeto mobilizador “TICE.Mobilidade – Sistema de mobilidade Centrado no utilizador” conta com a presença de várias entidades que reúnem esforços colaborativos para a concepção e comercialização de produtos inovadores no domínio da mobilidade e transportes utilizando para tal a infraestrutura da internet e os dispositivos móveis. O projeto prevê o desenvolvimento de uma plataforma de comercialização de serviços de mobilidade designada “One.Stop.Transport” e o desenvolvimento de diversos PPSs de mobilidade. Pretende-se que os PPSs delineados correspondam às necessidades de mercado bem como às soluções tratadas nas políticas da União Europeia (UE), isto é, que sejam capazes de tornar mais eficiente a rede de mobilidade urbana de modo a combater muitos dos problemas que resultam daí.

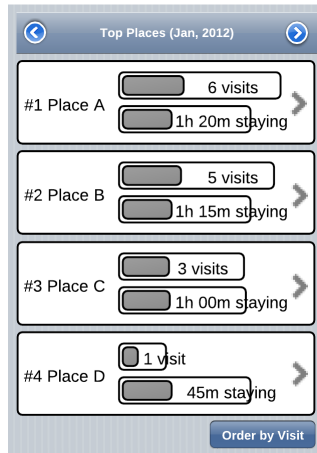
Dos vários PPSs existentes, esta dissertação insere-se mais precisamente no PPS6 – ECO.Circuitos. Uma série de entidades encontram-se reunidas de modo a atingir os objetivos estabelecidos: SMARTMOVE, S.A, Faculdade de Ciências e Tecnologia da Universidade de Coimbra (FCT/UC), Universidade do Minho (UM), Intelligent Sensing Anywhere (ISA), S.A e Monitar.

O objetivo principal do PPS6 consiste no desenvolvimento de um sistema personalizado que permita conhecer e traçar os padrões individuais de mobilidade urbana e analisar esses padrões de modo a poder disponibilizar quer informação relativa a esses padrões, nomeadamente informação ambiental, quer fornecer alternativas dirigidas ao aumento sustentável da eficiência energética de cada indivíduo e ao bem-estar geral. O PPS6 pretende sobretudo consciencializar o cidadão da sua mobilidade, em particular o seu custo ambiental através da disponibilização de relatórios de eficiência de mobilidade. Esta consciencialização só será possível se os relatórios apresentados forem eficazes no modo como transmitem a mensagem ao utilizador.

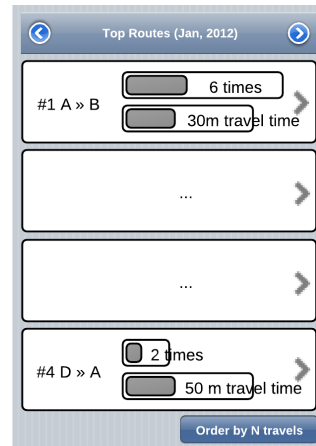
De modo a tornar as pessoas conscientes dos seus comportamentos, pretende-se fornecer informação acerca de como os comportamentos do utilizador afetam o ambiente envolvente. Existe bastante literatura na área da psicologia e sociologia que aborda e promove comportamentos responsáveis para com o meio ambiente (Maibach et al., 2003; Ampt & Gleave, 2004; Abrahamse, Steg, Vlek, & Rothengatter, 2005). O uso de motivadores como o compromisso pessoal, estabelecer objetivos, oferecer incentivos financeiros, fornecer frequentemente *feedback* na forma adequada, bem como fornecer comparações históricas (comparar com dados passados) ou normativas (comparar com outros utilizadores) podem ter um impacto positivo no comportamento dos utilizadores (Fischer, 2008).

Atualmente os *smartphones* com capacidades de recolher informações sobre o ambiente que rodeia o utilizador atingiram uma alta penetração na vida quotidiana das pessoas. Aproveitando as capacidades destes pequenos dispositivos (elevada capacidade de processamento e armazenamento dos dados, e a possibilidade de recolher dados a partir dos sensores embutidos) estas serão exploradas de modo a recolher informação acerca da sua mobilidade diária. Na rotina diária de um individuo existe uma quantidade de locais importantes e significativos. Pretende-se que o utilizador tenha uma visão geral dos locais mais relevantes na sua vida diária. Através do desenvolvimento de uma aplicação para *smartphones*, esta terá como objetivo fornecer informação sobre os padrões de mobilidade do utilizador. A seguir são apresentados alguns dos possíveis exemplos dos relatórios de eficiência a apresentar aos utilizadores.

A Figura 1.1a ilustra uma lista ordenada dos locais frequentados pelo utilizador, seja por número vezes que visita ou tempo médio despendido. Para cada um dos locais da lista, o utilizador será capaz de visualizar os gráficos temporais associados, saber a frequência com que os visita, bem como o tempo despendido em cada um deles.



(a) Ranking dos locais com significado para o utilizador

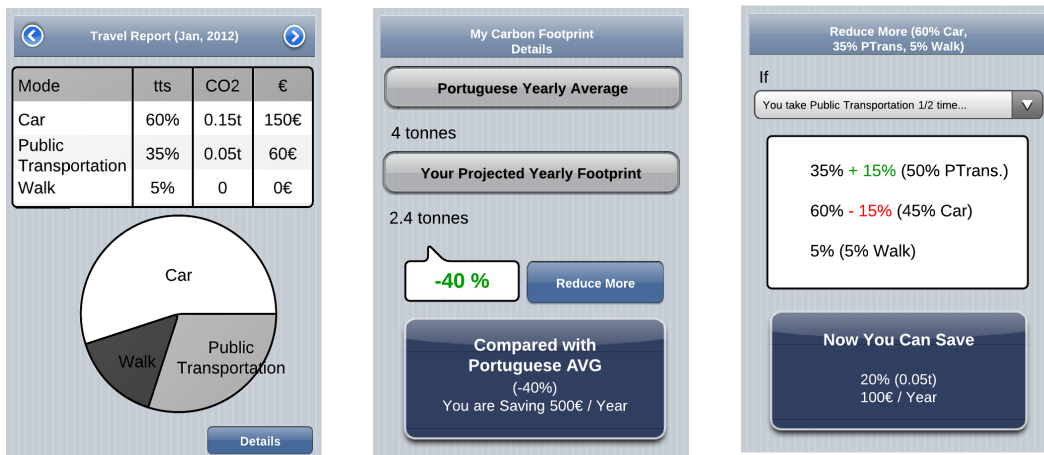


(b) Ranking das trajetórias mais frequentes para o utilizador.

Figura 1.1: Informação dos locais e respetivas trajetórias entre eles relevantes para o utilizador.

Similar aos locais, a Figura 1.1b lista de forma ordenada as trajetórias mais frequentes entre dois locais, ordenada pelo número de viagens ou pelo tempo médio de viagem entre locais. Para cada trajetória entre locais, será possível visualizar um gráfico temporal relativo às horas e frequência com que o utilizador viaja entre esses locais.

Para cada mês do ano, será possível visualizar com detalhe o modo de transporte mais utilizado, quantidade de CO₂ produzida e custos associados, como mostra a Figura 1.2a. Ainda com maior detalhe, o utilizador poderá ter uma noção das suas emissões produzidas comparada com a média dos portugueses, sabendo quando está a poupar ou a gastar em demasia, como é observável na Figura 1.2b. Por fim, a Figura 1.2c retrata uma opção onde serão facultados conselhos para o utilizador conseguir poupar ainda mais através do ajuste dos seus hábitos de transporte. Por exemplo, se passar a fazer viagens de transporte público na metade das viagens que realiza, o utilizador será informado da quantidade de emissões que produz a menos e os ganhos monetários associados com essa redução.



(a) Modo de transporte e custos associados

(b) Detalhes sobre a quantidade de CO₂ produzida

(c) Conselhos para melhorar os hábitos de mobilidade

Figura 1.2: Informação detalhada sobre a quantidade de CO₂ produzida e custos associados (a, b). Conselhos úteis para uma mobilidade eficiente (c).

1.3.2 Abordagem e Desafios

De modo a facilitar a distribuição de tarefas a cada entidade, foi elaborada uma arquitetura que representa todas as camadas de software e todas as principais componentes de cada camada (Figura 1.3).

As camadas de *software* especificadas foram as seguintes: (i) *Mobile*; (ii) *Data Collection*; (iii) *Data Processing*; (iv) *Integration*; (v) *Mobility Portal*.

A camada *Mobile* diz respeito à aplicação móvel responsável pela recolha dos dados provenientes dos sensores, bem como pela apresentação de alternativas à mobilidade do utilizador e informação sobre a mobilidade individual. A *Data Collection* é a camada responsável pela recolha e armazenamento dos dados provenientes dos *smartphones* e pela gestão das comunicações entre eles. A camada *Data Processing* é responsável pelo tratamento dos dados armazenados pela camada *Data Collection* e pela inferência dos padrões de mobilidade com base nestes dados e em informação proveniente de sistemas externos (por exemplo, informação das operadoras de transporte). Em rela-

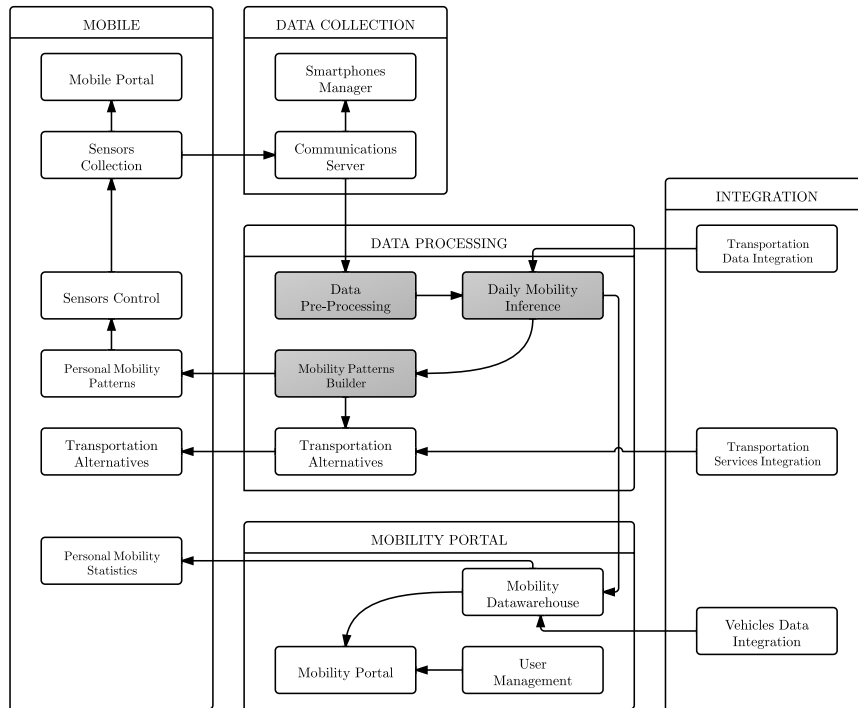


Figura 1.3: Arquitetura geral do PPS6.

ção à camada *Integration*, está é responsável pela integração com os restantes PPSs e com os sistemas externos. Por fim, a camada *Mobility Portal* diz respeito ao portal de acesso aos mapas de mobilidade, relatórios de eficiência ambiental e outras informações disponibilizadas pelo sistema.

Das cinco camadas existentes, o trabalho desta dissertação focar-se-á na camada *Data Processing*. Dentro desta camada serão focados esforços somente em três das quatro componentes (a componente *Transportation Alternatives* encontra-se fora do âmbito desta dissertação):

- *Data Pre-Processing* – Nesta componente é realizado o primeiro processamento dos dados recolhidos pelos sensores presentes nos *smartphones*, corrigindo e/ou filtrando os dados relevantes, aplicando um pré-processamento antes da aplicação dos algoritmos de inferência;
- *Daily Mobility Inference* – Componente responsável pelos algoritmos de

inferência, construindo a informação necessária aos mapas de mobilidade pessoal;

- *Mobility Patterns Builder* – Componente responsável pela construção dos padrões de mobilidade individuais.

A construção dos Mapas de Mobilidade Pessoal (MMP) envolve estas três componentes. Pretende-se que estes representem os hábitos de mobilidade de cada indivíduo, isto é, os padrões de movimento (locais frequentados, transições entre locais, tempos médios de estadia em locais, tempos médios de transição entre locais, entre outra informação importante para caracterizar a mobilidade). Como ilustrado na Figura 1.4, um MMP é representado em forma de um grafo ponderado $P = (L, T)$ onde os nodos L são os locais visitados pelo utilizador e as suas arestas T são as transições entre eles. Cada aresta tem um peso associado que representa o número de transições entre os locais unidos por essa aresta.

A construção dos MMP levanta alguns desafios, quer no pré-processamento dos dados provenientes dos sensores, quer no desenvolvimento dos algoritmos de inferência e respetiva validação:

- A deteção de locais e respetivas estadias, transições entre locais, tempos de entrada e saída de um local, serão os desafios principais a ultrapassar no âmbito do algoritmo de descoberta de locais;
- A existência de cada vez mais redes móveis Wi-Fi, devido à facilidade com que podem ser criadas quer nos computadores portáteis bem como nos *smartphones*, poderá interferir no algoritmo de inferência de locais, devido à deteção de falsos locais;
- A construção de *groundtruth* será o desafio mais complicado da fase de validação do algoritmo. A validação requer a realização de diários de

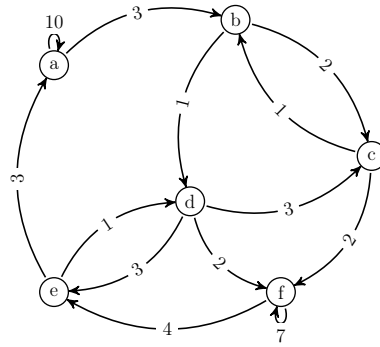


Figura 1.4: Exemplo da representação de um Mapa de Mobilidade Pessoal.

mobilidade realizados pelos utilizadores. Conseguir que os voluntários consigam anotar todas as suas entradas e saídas nos locais visitados, durante um período suficientemente longo, e que o grau de erro seja mínimo, será um desafio extremamente difícil de ultrapassar.

1.3.3 Avaliação

Após percebido o **problema**, **sugeridas** novas funcionalidades, sejam elas criativas ou baseadas noutros trabalhos de investigação, e **desenvolvido** o processo para a construção automática de mapas de mobilidade pessoal, o próximo passo será **avaliar** o algoritmo em termos de desempenho. Nesta fase da metodologia, pretende-se realizar alguns testes com o algoritmo desenvolvido de modo a medir o seu desempenho em relação a outros concorrentes, utilizando métricas desenvolvidas por outros autores, para avaliar o número de locais corretamente detetados bem como a precisão nos tempos de entrada e saída de um local. Assim que atingir um resultado satisfatório, dá-se por concluído todo o esforço de investigação que resultará nos MMP.

A construção dos MMP permitirá dar resposta a duas questões fundamentais nesta metodologia de investigação: (i) Qual a utilidade dos mapas de mobilidade? e (ii) Como é demonstrada essa utilidade?

Os MMP permitirão enriquecer os relatórios de eficiência com informação que permitirá, ou não, consciencializar o utilizador dos seus hábitos de mobilidade. Os MMP serão úteis na perspetiva de fornecer contributo para o módulo onde estão inseridos dentro da arquitetura definida para o PPS6.

A utilidade só poderá ser demonstrada quando os relatórios de eficiência forem desenvolvidos com base na informação alimentada pelos MMP. Como o desenvolvimento dos relatórios se encontra fora do âmbito desta dissertação, coloca-se a hipótese que estes poderão consciencializar os utilizadores. Esta consciencialização dependerá de muitos fatores, como o *design* da aplicação móvel, e como a informação é apresentada ao utilizador, entre outros. Caso seja aceite pelos utilizadores, pretende-se que estes corrijam algumas práticas menos eficientes, tornando-os mais conscientes dos seus hábitos de mobilidade. Deste modo seria possível resolver muitos dos problemas das cidades atuais, como por exemplo a fluidez do tráfego, através da utilização de transportes públicos.

1.4 Estrutura da Dissertação

Este documento encontra-se estruturado em vários capítulos. Neste primeiro capítulo é realizado o enquadramento do trabalho e definidos os objetivos e resultados esperados. É ainda apresentada a abordagem metodológica seguida para atingir os resultados esperados.

No capítulo 2 são apresentadas algumas das formas de aquisição e inferência de informação de contexto existentes, fundamentais para o processo de análise de mobilidade humana. São ainda apresentados alguns estudos quanto à mobilidade humana que têm vindo a ser realizados por outras equipas de investigação.

O capítulo 3 diz respeito à solução proposta, onde é descrito todo o pro-

cesso de descoberta de locais, desde a definição da arquitetura, passando por cada uma das camadas da arquitetura e respetivos processos de transformação entre camadas.

No capítulo 4 é descrita a implementação do algoritmo de descoberta de locais especificado no capítulo anterior. No capítulo 5 é feita a validação desse algoritmo, apresentando-se os resultados obtidos com o processamento dos dados recolhidos por vários utilizadores em ambiente real.

Finalmente, no sexto e último capítulo, são apresentadas as conclusões obtidas com o trabalho desenvolvido e são apontadas algumas propostas para trabalho futuro de investigação.

Capítulo 2

Estado da Arte

A área da computação móvel e ubíqua evoluiu significativamente na última década. Em particular, o surgimento de serviços e aplicações dependentes do contexto e da localização ganhou visibilidade em resultado da evolução tecnológica que permitiu a criação de novos dispositivos e novas tecnologias de comunicação que permitem o acesso à Internet nos mais variados locais. Esta evolução deveu-se à necessidade que as pessoas têm em aceder a serviços e informação presente nas redes.

A alta penetração dos *smartphones*, a sua capacidade de se ligarem a várias redes de comunicações, bem como a existência de diversos sensores, torna possível a recolha dos mais variados tipos de dados ao longo da rotina diária do utilizador. Tirando partido do modo como os *smartphones* estão intrinsecamente ligados à vida das pessoas, será possível estudar a forma como as pessoas se movimentam.

Este capítulo apresenta, resumidamente, alguns dos trabalhos mais relevantes na área de abrangência desta dissertação, apresentando uma perspetiva geral do estado da arte.

2.1 Contexto

Em 1991, Weiser (1991) teve uma visão pioneira do que seria a computação móvel e ubíqua. O autor anteviu uma realidade onde os sistemas computacionais seriam *pervasive*, isto é, onde o processamento da informação seria integrado em objetos e atividades do dia a dia. O grande objetivo do autor era alcançar o tipo mais eficaz de tecnologia, que estivesse disponível em todo o ambiente físico, tornando-a invisível para o utilizador.

A alta popularidade e penetração dos dispositivos móveis, como os *notebooks*, *netbooks*, *Personal Digital Assistant* (PDA) e *smartphones*, contribuiu significativamente para a evolução na área das tecnologias de computação móvel e ubíqua na última década. Esta evolução levou ao surgimento de aplicações inovadoras e novas utilizações para os dispositivos e tecnologias existentes.

Estes factos vão de encontro ao que Weiser previu, permitindo assim constatar a validade dos seus princípios e verificar que os sistemas computacionais estão cada vez mais integrados no ambiente e com capacidade de oferecer funcionalidades em função do contexto em que são utilizados.

Um dos grandes avanços prende-se com a capacidade de recolha e captura de contexto do utilizador. Perceber o contexto do utilizador é um pré-requisito para fornecer serviços centrados nas pessoas e que melhorem a sua qualidade de vida (Chon & Cha, 2011).

Definir a palavra contexto não é uma tarefa fácil. Vários investigadores tentaram definir por si próprios o que realmente inclui o contexto. Desde a primeira vez que foi introduzido o termo *context-aware* por Schilit e Theimer (1994), um grande número de definições do termo contexto foram propostos na área das ciências da computação.

Schilit e Theimer (1994) descrevem o contexto como localização, identidades de pessoas próximas, objetos e mudanças nesses objetos. Brown, Bovey, e

Chen (1997) numa definição similar à anterior, definem contexto como localização, identidades das pessoas à volta do utilizador, hora do dia e temperatura. Ryan, Pascoe, e Morse (1998) referem-se ao contexto como a localização do utilizador, ambiente, identidade e tempo. Este tipo de definições por exemplos, e por utilização de sinónimos, são extremamente difíceis de aplicar na prática (Dey, 2001). Era necessário encontrar uma definição operacional para o termo contexto. O trabalho de Schilit, Adams, e Want (1994) contribuiu com uma melhor definição de contexto. Os autores afirmam que os aspectos importantes do contexto são: onde é que o utilizador está, com quem está e que recursos estão próximos. Apesar de ser uma boa definição, esta pecava por ser muito específica. Dey (2001) tinha outra perceção do contexto, afirmando que este refere-se a toda a situação relevante para uma aplicação e respetivos utilizadores. Abordando as noções bastante limitadas, como as descritas anteriormente, Dey chegou à seguinte definição de contexto, designada por muitos como uma das mais aceites:

Contexto é qualquer informação que pode ser utilizada para caracterizar a situação de uma entidade. Uma entidade é uma pessoa, local, ou objeto que é considerado relevante para a interação entre o utilizador e a aplicação, incluindo o próprio utilizador e a aplicação.

Esta definição facilita a tarefa de quem desenvolve aplicações quanto ao modo de enumerar o contexto para um determinado cenário de uma aplicação. O autor acredita que se um pedaço de informação pode ser utilizado para caracterizar a situação de um participante numa interação, então essa informação é contexto.

Dey, Abowd, e Salber (2001) classificam a informação de contexto com base nas entidades cujo contexto é avaliado. Como a própria definição de contexto refere, as entidades identificadas pelos autores são os locais (escri-

tórios, edifícios e ruas), pessoas (individuais ou em grupos) e objetos (físicos ou componentes de software como aplicações ou ficheiros). Cada uma destas entidades pode ser descrita por vários atributos que podem ser classificadas em quatro categorias: identidade, localização, estado e tempo.

A categoria identidade refere-se à capacidade de atribuir um identificador único a uma entidade (Dey et al., 2001). Por exemplo, utilizar hardware ou software para identificar uma pessoa. No caso do hardware, utilizando o endereço MAC (*Media Access Control*) como identificador, este identificaria o dispositivo mas não a pessoa. Em relação ao software como identificador, este resolve os problemas do hardware. A utilização de uma norma aberta de autenticação como o OpenID (Recordon & Reed, 2006), já utilizada pela rede social Facebook e Google, permite atribuir uma identificação única a uma pessoa, mesmo que seja de forma digital.

Os objetos físicos e dispositivos estão organizados espacialmente e cada vez mais as pessoas se movem em ambientes móveis e ubíquos. A localização é muito mais que a informação sobre uma posição no espaço bidimensional. Esta inclui toda a informação que pode ser utilizada para inferir relações espaciais entre entidades, como a proximidade entre elas (Dey et al., 2001). Zimmermann, Lorenz, e Oppermann (2007) descrevem esta categoria como modelos de localização que classificam a posição física ou virtual de uma entidade, bem como outra informação relacionada como a velocidade e orientação. Os modelos de localização física podem ser divididos em quantitativos ou geométricos, e qualitativos ou simbólicos. Os quantitativos referem-se às coordenadas geográficas, como a informação fornecida pelo *Global Positioning System* (GPS). A informação de localização qualitativa refere-se por exemplo a edifícios, pisos e divisões dentro desses edifícios. A sobreposição dos dois modelos permite a transformação da informação espacial quantitativa em informação espacial qualitativa. Em relação aos modelos de localização virtual,

é algo que é tido como posição de referência dentro de um determinado ambiente. Zimmermann et al. (2007) dá como exemplo o uso do endereço IP como posição dentro de uma rede de computadores. De modo geral, uma entidade possui sempre uma localização física qualitativa, que pode ser representada por diferentes localizações quantitativas, mas também várias localizações virtuais ao mesmo tempo.

Em relação à categoria estado, ou atividade, esta identifica as características intrínsecas de uma entidade que podem ser adquiridas (Dey et al., 2001). No caso dos locais, podem ser adquiridas informações sobre a temperatura, luz ambiente ou nível de ruído. É essencial diferenciar o estado de um dispositivo móvel e de uma pessoa. Quanto aos dispositivos móveis, pode ser qualquer informação referente ao estado de uma aplicação, se está a executar ou receber uma chamada, nível da bateria, ou a existência de um ficheiro no sistema. Para uma pessoa pode referir-se a fatores fisiológicos, ou atividades em que a pessoa está envolvida. Muitas das atividades, como o modo de transporte das pessoas (a pé, corrida, bicicleta, parada ou num transporte público), refletem-se nos sensores dos dispositivos. Sensores embebidos nos dispositivos, como os acelerómetros, são capazes de descrever os movimentos físicos do utilizador de um dispositivo (Gil, Jesus, & Lopez, 2011).

A categoria tempo é um aspecto fundamental para a perceção e classificação do contexto de uma entidade. A utilização da informação temporal em conjunto com outros pedaços de informação, permite conhecer o instante ou período em que ocorreu, ou é relevante, para outra informação contextual. Isto é, permite tirar partido da riqueza e valor da informação histórica (Dey et al., 2001). O contínuo armazenamento da informação temporal cria um conjunto de dados contendo a história da informação de contexto obtida. O acesso a esta história, ordenada numa linha temporal, permite aceder a informação de contexto passada da entidade, analisar o histórico de interação,

inferir hábitos de utilização e prever contextos futuros (Zimmermann et al., 2007).

Esta classificação da informação de contexto pode ser utilizada para inferir contexto adicional. A inferência de contexto acontece quando esta é deduzida a partir de uma simples informação de contexto conhecida, ou então a partir da junção de vários pedaços de informação de contexto que descrevem uma situação.

2.2 Posicionamento

Recorrendo à variedade de sensores existentes nos *smartphones*, estes serão responsáveis por fornecer os dados de entrada para os algoritmos *place learning*. Estes algoritmos têm a capacidade de descobrir os locais visitados pelos dados recolhidos através de um ou vários tipos de dados provenientes dos sensores. Estes sensores incluem as tradicionais tecnologias como as interfaces de rede Wi-Fi, *Global System for Mobile Communications* (GSM), bem como o GPS. Estes estão presentes em praticamente todos os dispositivos móveis atuais, até aos mais exóticos como os acelerómetros, bússolas digitais e barómetros, que estão cada vez mais a ser integradas nos *smartphones*. Esta integração nos dispositivos móveis deve-se sobretudo à obrigação das operadoras de redes móveis em localizar dispositivos em situações de chamadas de emergência, para irem de encontro à diretiva E112 implementada na Europa (D. Kim, Hightower, Govindan, & Estrin, 2009; LaMarca et al., 2005).

Os dispositivos móveis estão cada vez mais a tornar-se ubíquos por toda a parte do mundo, especialmente em áreas urbanas e países industrializados (Candia et al., 2008). A ampla disseminação e a penetração quase a 100%, bem como a sua ligação ao quotidiano dos seres humanos (Chon, Talipov, Shin, & Cha, 2010), tornaram os *smartphones* no principal agente de captura

e recolha de informação que poderá ajudar a caracterizar a atividade ou estado de um utilizador numa determinada situação. Os dados relativos às frequentes interações entre os utilizadores e os seus dispositivos podem ser facilmente recolhidos recorrendo aos sensores embutidos nos *smartphones*. A exploração destes sensores será fundamental para gerar informação sobre a localização de um indivíduo, um dos principais focos da computação ubíqua.

2.2.1 Posicionamento via Satélite

A solução mais frequente de posicionamento em ambientes exteriores é o Sistema de Posicionamento Global (*Global Positioning System* (GPS)) (PNT & Committee, 2012). Baseado em satélites, o GPS foi desenvolvido pelo departamento de defesa dos Estados Unidos da América para ser utilizado na obtenção exata da posição de modo a satisfazer necessidades militares. Contudo, a partir de 1980, foi tornado livre para utilização civil e empresarial.

O GPS foi concebido para proporcionar uma cobertura mundial, podendo ser acedido em qualquer parte do mundo e utilizado por um número ilimitado de pessoas. O sistema é composto por uma constelação de 24 satélites operacionais em órbita da Terra, estações terrestres e receptores GPS possuídos pelos utilizadores.

Cada satélite possui um relógio atómico de elevada precisão, onde os sinais por eles emitidos são diferenciados em uso civil ou para uso militar (mais preciso). Os receptores descodificam esses sinais, de modo a calcular a posição e determinar a hora local. A posição calculada é expressa na forma de coordenadas Cartesianas XYZ ou em latitude, longitude e altitude de acordo com o sistema de coordenadas WGS84 (Raghunath, Malleswari, Sridhar, et al., 2011), fornecendo uma acuidade de cerca de 10 metros (LaMarca et al., 2005).

Devido à contínua descida do preço da tecnologia GPS, e também com

influência da diretiva E112, praticamente todos os dispositivos móveis têm receptores GPS integrados. Contudo, com o passar dos anos, tem vindo a ser identificados vários inconvenientes nesta solução. LaMarca et al. (2005) afirmam que apesar da alta disponibilidade, medida segundo a percentagem da superfície da Terra abrangida, este tem fraca cobertura segundo a percentagem de tempo que o sinal está disponível nos locais onde passamos a maior parte do nosso tempo, os ambientes interiores. Da mesma forma, em ambientes urbanos com grandes edifícios e ruas estreitas o sistema não estará sempre disponível (Hightower, Consolvo, LaMarca, Smith, & Hughes, 2005) devido às dificuldades em estabelecer linha de vista entre o recetor e um número mínimo de satélites.

Devido aos problemas de disponibilidades mencionados, ao elevado consumo de energia da bateria nos dispositivos móveis (Chon, Talipov, & Cha, 2012), e o facto de os seres humanos passarem a maior parte dos tempos em ambientes interiores será importante encontrar alternativas eficazes ao GPS para posicionamento no interior de edifícios.

2.2.2 Posicionamento Alternativo ao Satélite

O sistema de posicionamento Wi-Fi (*Wi-Fi Positioning System* (WPS)) é uma solução eficaz para resolver os problemas da localização em ambientes fechados (LaMarca et al., 2005; Bahl & Padmanabhan, 2000). O WPS fornece um grau razoável de acuidade, contudo implica a disponibilidade de um mapa rádio, de *Access Points* (APs) Wi-Fi, que deve ser construído para se poder obter informação precisa sobre a localização (Chon, Talipov, & Cha, 2012). Após construído o mapa, mediante um pedido de um utilizador sobre a sua localização, o dispositivo móvel faz uma comparação entre os APs Wi-Fi do mapa com os detetados no ambiente que o rodeia (Chon, Talipov, & Cha, 2012). Além dos custos excessivos (Constandache, Choudhury, & Rhee, 2010;

Chon, Talipov, & Cha, 2012), a localização que produz não é precisa visto que a construção do mapa é realizada via *war-driving*. *War-driving* é o ato de conduzir aleatoriamente com um dispositivo móvel equipado com GPS e dispositivos rádio, tais como adaptadores de rede 802.11, GSM ou *Bluetooth*, de modo a recolher sinais das redes disponíveis (LaMarca et al., 2005; Constandache et al., 2010).

Existem já aplicações comerciais que constroem os mapas através dos dados submetidos pelos utilizadores, tais como o Skyhook¹ e Navizon². Outros métodos para localização em ambientes interiores incluem os sistemas de localização em tempo real (*Real Time Location Systems* (RTLS)) e os sistemas de navegação baseados em unidades de medida inercial (*Inertial Measurement Unit* (IMU)) (Chon & Cha, 2011). Um dos grandes problemas da utilização de IMU é a constante acumulação de erro, visto que o processo adiciona, continuamente, posições relativas às suas localizações anteriores (Chon, Talipov, & Cha, 2012). A aplicação LifeMap (Chon & Cha, 2011) é um exemplo da utilização de IMU. Utilizando sensores como o acelerómetro e bússola digital, permite obter informação sobre a localização de um dispositivo. Alguns trabalhos têm sido desenvolvidos utilizando sensores menos comuns: câmaras e microfones para recolha de som ambiente, luz e cor em detrimento dos habituais sinais de rádio (Lu, Pan, Lane, Choudhury, & Campbell, 2009; Azizyan, Constandache, & Roy Choudhury, 2009).

Será importante não descurar a localização em ambientes interiores, pois a qualidade dos serviços de localização depende da precisão da informação acerca da localização dos utilizadores. Como as pessoas passam a maior parte do tempo em ambientes interiores, esses serviços terão que fornecer, sobretudo, informações sobre a nossa localização em ambientes fechados.

Descrevem-se, a seguir, alguns dos sistemas mais conhecidos para posicio-

¹<http://www.skyhookwireless.com/>

²<http://www.navizon.com/>

namento em ambientes interiores.

RADAR

Bahl e Padmanabhan (2000) desenvolveram o RADAR, um sistema de localização e registo de movimentação de utilizadores em ambientes interiores. O sistema baseia-se em *fingerprints*³ Wi-Fi, registando e processando a informação do nível do sinal obtido dos APs presentes no meio, sendo capaz de obter alta precisão em ambientes interiores. O RADAR consegue estimar a localização de utilizadores com uma precisão média de 2 a 3 metros.

Para estimar a localização dos utilizadores, o sistema necessita de duas fases: *offline* e *realtime*. Na fase *offline* são recolhidos os dados relativos ao nível do sinal nos diversos pontos assinalados na planta. Em cada ponto é registada informação relativa a 4 direções d (norte, sul, este, oeste) na forma $\{t, x, y, d\}$, onde t representa o instante de tempo e x, y as coordenadas do utilizador. Para cada AP (bs) regista a força de sinal medida ss conjuntamente com uma marca temporal sincronizada t , sendo esta informação preservada na forma $\{t, bs, ss\}$. Relativamente à fase *realtime*, também são recolhidos os dados relativos à força de sinal em tempo real. Os dados são comparados com os dados recolhidos na fase *offline*. Por fim, é aplicado um algoritmo de triangulação, de modo a estimar as coordenadas de um utilizador.

Mais tarde, os autores procederam a diversos melhoramentos ao sistema RADAR, tais como (Bahl, Padmanabhan, & Balachandran, 2000): (i) ter em conta que as mudanças no ambiente têm impacto significativo nos sinais de rádio, como por exemplo, alteração do número de pessoas e das obstruções no edifício e mudanças de temperatura; (ii) desenvolvimento de um novo algoritmo baseado no algoritmo Viterbi para uma localização contínua dos utilizadores, tendo em conta o tempo real e o padrão de mobilidade dos utili-

³Adota-se a terminologia inglesa *fingerprint* para designar a caracterização de algo que é feito num dado instante de tempo.

zadores; (iii) ter em consideração os edifícios com vários andares.

PlaceLab

LaMarca et al. (2005) propuseram o PlaceLab, um sistema com a finalidade de fornecer informação sobre a localização de computadores portáteis, PDAs e telefones móveis, através da escuta de *beacons*⁴ rádio provenientes de redes Wi-Fi, estações base de transmissão GSM ou dispositivos fixos de *Bluetooth* já existentes no meio. Por via desta escuta, obtém-se a informação relativa a identificadores únicos ou semi-únicos, como por exemplo os endereços MAC dos APs Wi-Fi, e pode efetuar-se uma estimativa da sua localização tendo por base a localização dos emissores de *beacons* instalados no meio. Um mapa rádio é construído através da conhecida atividade de *war-driving* e do mapeamento das estações base de transmissão Wi-Fi/GSM com as coordenadas GPS. Posteriormente, o mapa é distribuído pelos dispositivos móveis dos clientes que se localizam a si próprios comparando as estações base de transmissão Wi-Fi/GSM visíveis no meio com as presentes no mapa previamente construído.

A ideia básica por detrás do sistema PlaceLab é a localização de *beacons*. Segundo Hightower, LaMarca, e Smith (2006), é uma abordagem viável para atingir uma capacidade de localização ubíqua. O principal objetivo do sistema PlaceLab é fornecer o máximo de cobertura, inclusivamente em ambientes interiores, onde as pessoas passam a maior parte do seu tempo. A principal razão da escolha de sinais 802.11 e GSM prende-se com o fato de estes estarem a ser implementados rápida e ubíquamente pelas áreas urbanas. É deixado para segundo plano a precisão das estimativas de localização, preferindo perder alguma precisão mas ganhar uma maior ubiquidade. Quanto a

⁴Adota-se a terminologia inglesa *beacon* para designar os dados transmitidos por um determinado AP Wi-Fi, contendo o seu *Basic Service Set Identification* (BSSID), *Service Set Identifier* (SSID), número de canal e protocolos de segurança como o *Wired Equivalent Privacy* (WEP) e *Wi-Fi Protected Access* (WPA).

questões de privacidade, os dispositivos estimam a sua localização localmente, sem necessidade de acesso à Internet ou a outra qualquer infraestrutura.

A arquitetura do sistema PlaceLab consiste em três elementos chave: (i) os *beacons* de rádio detetados no meio, (ii) bases de dados que contêm informação crucial sobre a localização dos emissores de *beacons* e (iii) os clientes PlaceLab que utilizam esta informação para estimar a sua localização. A precisão média de localização é de 15 a 20 metros, desde que esteja disponível, somente, uma densidade razoável de emissores de *beacons* 802.11. No caso de estarem disponíveis apenas *beacons* GSM, a precisão decresce para 100 a 200 metros.

2.3 Inferência de Informação de Localização

Segundo um estudo descrito em Klepeis et al. (2001), os seres humanos passam aproximadamente 87% do seu tempo em ambientes interiores, 6% em veículos e os restantes 7% em ambientes exteriores. É nos ambientes interiores, onde passamos a maior parte do nosso tempo, que se localizam os principais pontos de interesse (*Point Of Interest* (POI)) da sociedade moderna (Chon, Talipov, & Cha, 2012). Um POI, também designado por local significativo, é caracterizado como um local onde um indivíduo passa uma quantidade substancial do seu tempo, ou permanece por mais de um determinado tempo (por exemplo 10 minutos) e/ou visita com regularidade esse local (D. Kim et al., 2009; Jiang, Li, Yang, Lv, & Liu, 2011). Devido a estes factos, a monitorização dos POI será o principal foco dos serviços móveis centrados no utilizador (Chon et al., 2010).

No entanto, a localização geométrica - par de coordenadas, latitude e longitude - é insuficiente para distinguir os POI. Primeiro porque o GPS que fornece localização precisa de céu aberto e, por isso, só será útil em aproximadamente 13% do tempo. Segundo Chon, Talipov, e Cha (2012), a localização física

de um POI não é necessariamente descrita pelo mesmo ponto. Por exemplo, tendo em conta a localização física, dois ou mais pontos geográficos diferentes referem-se a dois sítios diferentes, mas, no entanto, podem referir-se ao mesmo POI. Sendo assim, será necessário recorrer à informação lógica dos locais significativos para o utilizador, isto é, atribuir nomes aos POIs, como por exemplo “local de trabalho”, “local de estudo”, “casa”, para agregar diferentes pontos geográficos como um único POI. O LifeMap (Chon & Cha, 2011), PlaceSense (D. Kim et al., 2009) e o SensLoc (D. Kim, Kim, Estrin, & Srivastava, 2010) são alguns exemplos de aplicações que tiram partido dos algoritmos *place learning* para determinar locais que sejam importantes e significativos para os utilizadores.

A principal função dos algoritmos *place learning* é tentar encontrar locais que sejam importantes e com significado para uma pessoa (D. Kim et al., 2009). Estes algoritmos podem ser divididos em duas classes baseadas nas características da fonte dos dados de localização: geométricos e *fingerprint*.

2.3.1 Algoritmos Geométricos

Os algoritmos geométricos produzem pontos, círculos, ou polígonos para descrever lugares que o algoritmo acredita serem significativos para uma pessoa (D. Kim et al., 2009; Hightower et al., 2005). Estes algoritmos reconhecem lugares verificando se a localização geográfica atual do dispositivo está contida na forma geográfica de um determinado local (Jiang et al., 2011). Os trabalhos de Hariharan e Toyama (2004), Kang, Welbourne, Stewart, e Borriello (2005), Marmasse e Schmandt (2000), Ashbrook e Starner (2003) desenvolveram exemplos de algoritmos geométricos em que as coordenadas geográficas referentes a um determinado POI são agregadas.

comMotion

O sistema comMotion (Marmasse & Schmandt, 2000) é composto por um computador portátil equipado com um sensor de GPS. Tem como objetivo relacionar a informação pessoal com a localização, apresentando uma lista com tarefas a fazer consoante a localização do utilizador.

O sistema é capaz de identificar um local através da perda de sinais GPS. Quando o recetor GPS deixa de receber sinais dentro de um determinado raio, o sistema considera que o utilizador se encontra dentro de um edifício. Após três visitas a um local este é considerado relevante e o sistema pede ao utilizador que classifique o local e lhe atribua um nome. Assim que identificado, esse local fica associado a uma lista de tarefas a realizar.

Esta abordagem é suficiente para identificar locais interiores de pequena dimensão, como por exemplo uma casa. Em locais de grande dimensão, como um parque de exposições, o sistema tende a falhar. Outro problema na identificação de locais reside nos centros urbanos, onde a grande concentração de edifícios leva a atrasos na aquisição de sinal, gerando falsos positivos.

Ashbrook e Starner

No trabalho de Ashbrook e Starner (2003) os locais são determinados com base nos movimentos detetados a partir das coordenadas geográficas recolhidas de um recetor GPS. Quando existe uma falha no sinal GPS, e este reaparece após 10 minutos, um conjunto de coordenadas são agregadas como locais utilizando uma variante do algoritmo de *clustering*⁵ *k-means*. Para além de determinar as localizações onde o utilizador esteve, o sistema tem a capacidade de prever os movimentos dos utilizadores considerando as várias localizações e os movimentos anteriormente efetuados entre as várias localizações para

⁵Adota-se a utilização do termo *clustering* para o processo de decisão de atribuir um dado elemento a um determinado *cluster* (aglomerado).

construir um modelo de Markov.

O algoritmo é capaz de resolver as limitações do comMotion (Marmasse & Schmandt, 2000), como as dimensões dos locais interiores e os falsos positivos. Contudo, a utilização da perda do sinal como mecanismo para inferência de locais não permite inferir alguns locais interiores importantes, bem como múltiplos locais dentro de um mesmo edifício.

Kang et al.

Em Kang et al. (2005) é apresentado um algoritmo de *clustering* para extrair locais significativos a partir de uma sequência de coordenadas com marcas temporais associadas. Estas coordenadas geográficas são obtidas a partir da plataforma PlaceLab que determina a localização do utilizador detetando os pontos de acesso Wi-Fi e torres GSM existentes.

O sistema considera os locais mais significativos aqueles em que o utilizador passa mais tempo. Contudo, existem locais onde o utilizador passa pouco tempo mas é considerado significativo devido à elevada frequência de visitas.

Apesar de extrair os locais sem grandes recursos computacionais, esta abordagem ainda não resolve os problemas do GPS. Como utiliza um sistema de localização, o algoritmo necessita de executar um passo intermediário de modo a obter as coordenadas geográficas com base nos sinais Wi-Fi ou GSM detetados (D. Kim et al., 2009).

Moreira e Santos

Moreira e Santos (2005) baseado-se apenas em coordenadas geográficas e na “hora de pico”, desenvolveram um algoritmo que permite determinar os locais correspondentes à casa e ao local de trabalho do utilizador. Esta abordagem assenta no pressuposto de que o utilizador passa mais horas por dia em casa. Este é um método que não carece de qualquer conhecimento prévio ou da

intervenção dos utilizadores.

2.3.2 Algoritmos *Fingerprint*

As abordagens baseadas em *fingerprint* produzem uma lista de pontos sem dados geográficos associados. Estas listas contêm as “assinaturas” de cada local, permitindo ao dispositivo detetar quando este retorna a um local (Hightower et al., 2005), utilizando o *fingerprint* de um local como um vetor de células GSM e/ou de Wi-Fi visíveis (D. Kim et al., 2009). Os algoritmos *fingerprint* detetam sinais de rádio estáveis que indicam um local de permanência, mas não fornecem informação sobre a localização absoluta desse local (D. Kim et al., 2009). Comparando a similaridade entre dois *fingerprints*, é possível afirmar se estão, ou não, perto um do outro (Jiang et al., 2011). Incluem-se nos algoritmos de *fingerprint* os trabalhos de Hightower et al. (2005), Laasonen, Raento, e Toivonen (2004), Krumm e Horvitz (2004), Ahmad, d’Auriol, Lee, e Lee (2007) e D. Kim et al. (2009).

Laasonen et al.

Em Laasonen et al. (2004) é proposta uma solução baseada somente na célula GSM à qual o dispositivo móvel se encontra ligado. O sistema utiliza a identificação da célula ativa e das outras células detetáveis num dado instante para reconhecer localizações pessoais importantes e caminhos entre esses locais, sem o conhecimento prévio da topologia física da rede.

Apesar de apresentar resultados adequados e tratar da questão da privacidade, esta é uma solução que utiliza apenas a informação sobre as células GSM disponíveis no local. Como os próprios autores admitem, se utilizassem o mesmo algoritmo mas em redes com células de menor dimensão, seria possível delinear os locais de forma mais precisa e adicionar detalhes aos caminhos entre locais. Outro problema deve-se à necessidade do utilizador ter que atri-

buir um nome a cada local, implicando que haja interação com o utilizador tornando o processo aborrecido para quem o utiliza.

BeaconPrint

O algoritmo BeaconPrint (Hightower et al., 2005) tem a finalidade de, automaticamente, descobrir POIs e detetar quando um utilizador retorna a esses POIs. Contrariamente a muitos outros algoritmos, o BeaconPrint não utiliza os sinais GPS para localizar dispositivos. São utilizados apenas ops *fingerprints* dos sinais de rádio Wi-Fi e GSM recolhidos pelo dispositivo móvel do utilizador. A utilização dos sinais rádio Wi-Fi e GSM permite a utilização do BeaconPrint na maioria dos dispositivos móveis.

A localização é realizada recolhendo os sinais provenientes de APs Wi-Fi e estações base GSM, denominados por *beacons*. Os sinais recebidos pelo dispositivo serão os dados de entrada para o algoritmo. Após verificados os dados, se o dispositivo permanecer no mesmo local, pelo menos durante uma janela de tempo w , esse local é marcado como um POI e um *fingerprint* desse POI é gerado. Um *fingerprint* corresponde a uma lista de todos os identificadores *beacons*, com uma marca temporal associada, recebidos num determinado POI. Quando o dispositivo retorna a esse local e outro *fingerprint* é gerado, o algoritmo trata de os fundir num só.

Para validar a solução, os resultados provenientes do algoritmo são complementados com inquéritos sobre os locais frequentados pelos indivíduos. O algoritmo tem uma precisão superior a 90% na aprendizagem e reconhecimento de locais. É demonstrada uma precisão de 63% para locais a que um indivíduo retorna apenas uma vez ou que visita por um tempo inferior a 10 minutos, e 80% de para locais visitados duas vezes.

PlaceSense e SensLoc

D. Kim et al. (2009) aperfeiçoaram o já existente algoritmo BeaconPrint (Hightower et al., 2005). O PlaceSense para além de reconhecer se um utilizar está num POI, deteta também a partida e chegada a esse POI através da exploração dos *beacons* de radio Wi-Fi ou GSM. A grande contribuição do algoritmo reside na descoberta de locais visitados em curtos instantes de tempo ou em locais onde o dispositivo estava em constante movimento.

Nos resultados obtidos, o algoritmo teve uma precisão de 92% na descoberta de POIs, e deteta com precisão os seus tempos de entrada e saída. A utilização das taxas de resposta para selecionar quais os *beacons* representativos e suprimir os *beacons* com frequência intermitente, aumenta a precisão do PlaceSense principalmente em ambientes rádio inconsistentes.

A chegada a um local é identificada através dos *scans* estáveis realizados aos *beacons* existentes no meio. Um *scan* é estável se não possuir novos *beacons* comparado com o *scan* anterior. A variável s_{max} indica quantos *scans* estáveis serão necessários para o algoritmo concluir que se trata de uma chegada a um local. Sempre que é realizado um *scan* estável é incrementado um valor ao contador s . Caso seja realizado um *scan* onde sejam identificados novos *beacons*, o contador s volta a 0, o histórico de *scans* estáveis é eliminado e recomeça um novo ciclo de *scans* até o contador s atingir o valor definido em s_{max} . Assim que s atinge o valor definido em s_{max} é declarada uma entrada num local.

A saída de locais é declarada quando um *scan* retorna apenas novos *beacons*, isto é, nenhum *beacon* representativo do local é retornado. Um *beacon* é considerado representativo, se a sua taxa de resposta for maior que o *threshold* r_{rep} definido. Quando nenhum *beacon* representativo do local é retornado, indica que o ambiente rádio que rodeia o dispositivo está a mudar e que o dispositivo está a abandonar o local.

Um ano mais tarde, os autores com base no trabalho desenvolvido no PlaceSense, desenvolveram o SensLoc (D. Kim et al., 2010). A grande diferença entre eles, reside na utilização de uma função de similaridade, mais propriamente o coeficiente de Tanimoto, que permite medir a similaridade entre dois *fingreprints*. A inclusão de um acelerómetro para conhecer se um utilizador está ou não em movimento, permite ao SensLoc reduzir significativamente o consumo de energia.

Em termos de desempenho, o SensLoc é capaz de distinguir muitos dos locais que estão próximos uns dos outros, enquanto o PlaceSense os considerava como um único local. O SensLoc tira partido das variações do nível de sinal (RSSI) para detetar corretamente os locais e da utilização de robustas funções de similaridade conseguiu reduzir assim muitas das falsas deteções presentes no PlaceSense. Numa avaliação entre os dois algoritmos, utilizando as mesmas condições (dados, frequência de recolha, entre outras), os autores conseguiram provar uma precisão de 94% contra os 89% do PlaceSense.

LifeMap

Chon e Cha (2011) desenvolveram o LifeMap, um fornecedor de contexto em *smartphones* para serviços de localização. O sistema é prático e eficaz, fornecendo informação precisa da localização dos utilizadores em ambientes interiores e exteriores. Para ambientes interiores, o LifeMap utiliza os sensores inerciais presentes nos *smartphones* para fornecer informação sobre a localização. Utilizando o método *dead reckoning*, baseado em sensores inerciais (IMU), é estimada a localização atual baseada na localização anterior, utilizando somente o acelerómetro e a bússola digital. Essa informação lógica é combinada com dados geográficos do GPS e dos WPS, de modo a gerar contexto do utilizador no seu dia a dia.

Como mostra a Figura 2.1, a arquitetura do sistema é separada em qua-

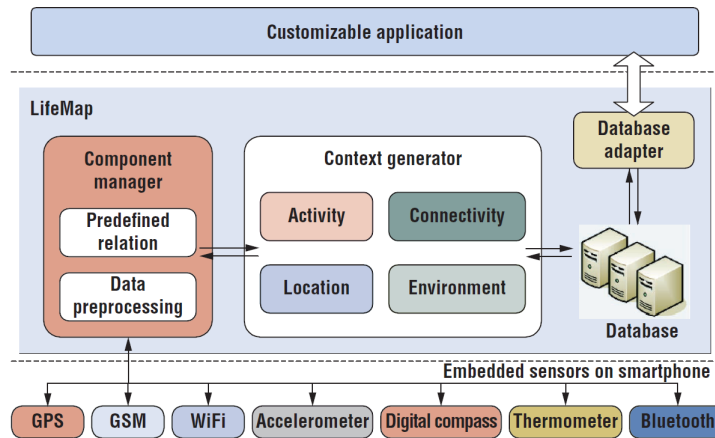


Figura 2.1: Arquitetura geral da plataforma LifeMap (Chon & Cha, 2011).

tro componentes: (i) no nível mais baixo estão os sensores (embebidos nos smartphones) que enviam a informação recolhida (ii) para o gestor de componentes onde a informação é processada e posteriormente fornece informação de alto nível. Utilizando essa informação de alto nível, (iii) o gerador de contexto cria nodos (POIs, que contêm o contexto do utilizador) e arestas (trajetórias sobre o movimento do utilizador). Com o conjunto de nodos e arestas é construído um mapa de contexto na forma de um grafo, que é guardado numa base de dados para comparar e agregar contextos dos utilizadores. Finalmente, (iv) o adaptador de base de dados é responsável por fornecer contexto de utilizador à própria *User Interface* (UI) do LifeMap, bem como a outras aplicações.

A principal preocupação era desenvolver um sistema que tivesse o mínimo de impacto no tempo útil da bateria do *smartphone*. Sendo assim, foram criadas regras de decisão baseadas nos movimentos realizados pelos utilizadores. Se fosse detetada uma alteração no movimento do utilizador, um número mínimo de sensores seria ativado de acordo com as regras de decisão.

O LifeMap é muito mais que um simples sistema de posicionamento e/ou localização com base em *fingerprints*. Para além de fornecer contexto, o LifeMap foca-se na construção de mapas de mobilidade pessoal, possibilitando aos utilizadores visualizar quais os locais mais frequentados e tempo médio de

estadia nesses locais.

2.4 Mobilidade Humana

Nos últimos anos tem-se assistido a um aumento do êxodo rural. A busca por melhores condições de vida tem levado a população a transferir-se das zonas rurais para os grandes centros urbanos. Este crescimento das cidades contribui para agravar um conjunto de problemas, tais como a propagações de vírus (humanos e eletrônicos), e para dificultar tarefas como o planeamento urbano ou as previsões de tráfego (Gonzalez, Hidalgo, & Barabasi, 2008; C. Song, Qu, Blumm, & Barabasi, 2010). Perceber e prever a mobilidade dos indivíduos, e os locais por eles frequentados, será fundamental para combater estes problemas. Outras das contribuições da percepção e previsão da trajetória dos indivíduos será o melhoramento do desempenho e fiabilidade dos protocolos de redes Wi-Fi e da própria infra-estrutura (L. Song, Kotz, Jain, & He, 2006).

Hoje em dia o telemóvel tem uma penetração de quase 100%. Enormes quantidades de dados são recolhidos através dos telemóveis de pessoas individuais, durante longos períodos de tempo.

Os estudos realizados por Farrahi e Gatica-Perez (2009, 2008), recorrendo a um *dataset* com dados, da vida real dos utilizadores, recolhidos pelos telemóveis, permitiram descobrir rotinas de locais diários frequentados pelos utilizadores. O principal objetivo dos autores é caracterizar, em termos dos seus padrões de localização, o comportamento individual e em grupo dos seres humanos.

Outros estudos importantes sobre a mobilidade dos indivíduos têm sido realizados. Com acesso a dados provenientes das estações base das redes móveis, Gonzalez et al. (2008) têm alcançado algumas conclusões sobre os padrões de mobilidade humana. Os autores mostram que as trajetórias humanas apresen-

tam um alto grau de regularidade temporal e espacial, isto é, os seres humanos seguem padrões de mobilidade simples e reproduzíveis.

C. Song et al. (2010) concluem que apesar do desejo profundamente enraizado de mudança e espontaneidade, a mobilidade diária é caracterizada por uma regularidade profundamente enraizada.

Eagle e Pentland (2009), recorrendo ao *dataset Reality Mining* do MIT⁶ como fonte de dados de mobilidade humana, utilizaram o princípio da análise de componentes principais (Principal Component Analysis (PCA)) de modo a identificar as principais componentes que estruturam o comportamento diário dos seres humanos. Estas componentes são um conjunto de vectores, denominados *eigenbehaviors*, que mostram as principais rotinas diárias de um grupo de utilizadores, como por exemplo estar em casa durante a noite.

Para além dos dados recolhidos de estações base das redes móveis, existem estudos da mobilidade humana recorrendo a dados GPS (Rhee et al., 2011), Wi-Fi (M. Kim, Kotz, & Kim, 2006), bem como no rastreamento de notas bancárias (Brockmann, Hufnagel, & Geisel, 2006). Estes dados relatam os comportamentos da mobilidade humana a várias escalas de tempo e espaço.

⁶Massachusetts Institute of Technology.

Capítulo 3

Descoberta de Locais

Os locais visitados por uma pessoa no seu dia a dia, bem como os deslocamentos efetuados entre eles, são a informação nuclear dos MMP. Deste modo, foi necessário desenvolver um algoritmo com a capacidade de inferir locais e transições entre eles com base em dados recolhidos pelos *smartphones*.

Os recentes avanços tecnológicos nos *smartphones*, tais como o seu pequeno tamanho e a capacidade de recolher grandes quantidades de informação através dos sensores embutidos, possibilita o estudo dos hábitos de mobilidade dos indivíduos sem se tornar um obstáculo durante o seu dia a dia.

Neste capítulo é realizada uma descrição de todo o processo de descoberta de locais e respetivas transições, começando pela definição da arquitetura, percorrendo cada uma das camadas da arquitetura e respetivos processos de transformação ocorridos entre as camadas.

3.1 Arquitetura da Solução

Tratar grandes quantidades de dados não é uma tarefa fácil. Para o sucesso de todo este projeto de investigação, é necessária uma estrutura capaz de organizar e armazenar todos os dados. O primeiro passo foi adaptar uma

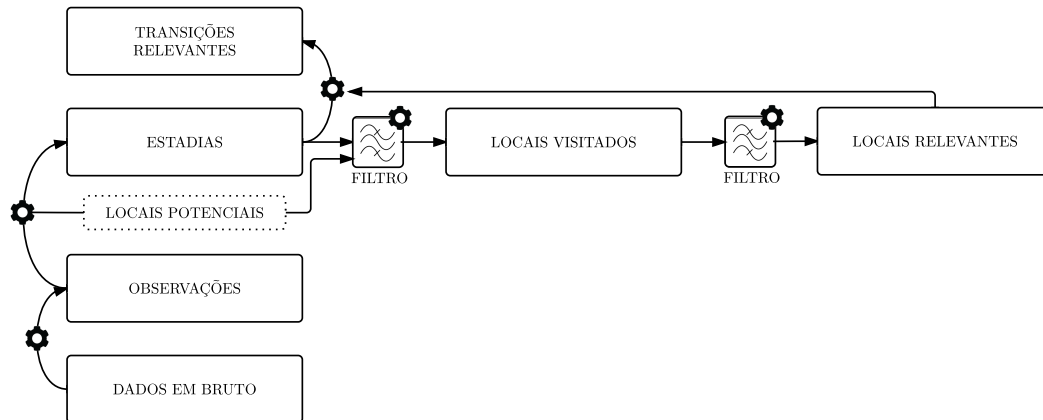


Figura 3.1: Arquitetura de todo o processo de descoberta de locais disposta em camadas. Cada camada é responsável por um determinado tipo de dados fundamental para a representação da mobilidade de um indivíduo.

arquitetura já existente, proposta por Peixoto e Moreira (2012) (Figura 3.1). No seu trabalho, os autores foram capazes de verificar que os conceitos propostos são adequados para representar a mobilidade dos indivíduos utilizando três tipos de dados (GPS, GSM e Wi-Fi). Os autores definem oito conceitos, concebidos para ajustar os dados desde a sua aquisição até à representação do movimento de um único indivíduo, ou de um grupo de pessoas.

Pela observação da Figura 3.1 cada conceito, ou camada, representa o modelo de dados responsável por armazenar, organizar e estruturar a informação resultante dos processos de transformação. Estes processos são representados pelas setas que contêm uma roldana. O objetivo destes é realizar simples transformações nos dados ou inferir nova informação a partir de informação já existente.

As grandes diferenças entre a arquitetura original, e a arquitetura adaptada a este trabalho dizem respeito aos conceitos utilizados, ao tipo de dados, aos processos de transformação e processos de inferência de locais.

Apesar da disponibilidade de vários tipos de dados (como por exemplo GPS, GSM, acelerómetro, entre outros), no contexto deste trabalho, foram utilizados apenas os dados do tipo Wi-Fi. O rápido crescimento das redes

Wi-Fi tornou-as numa das tecnologias de comunicação sem fios que está cada vez mais ubíqua. Isto é, em praticamente todos os locais que visitamos, há uma grande probabilidade de existir pelo menos um ponto de acesso. Como o GPS não está disponível em maior parte das horas do dia a dia dos humanos, e como o objetivo deste trabalho não era estimar a posição exata dos locais visitados por uma pessoa mas sim a sua representação simbólica, a inclusão de dados GPS no processo de deteção de locais foi excluída.

Relativamente aos processos de transformação, existem pequenas diferenças. Uma delas é a não inclusão de certas variáveis nas estrutura de dados. A existência de uma variável identificadora do indivíduo não é necessária, pois os MMP são pessoais, são característicos de uma determinada pessoa e não se encontram reunidos com dados de outros utilizadores.

No que diz respeito aos processos de inferência de locais, descartou-se o uso de complexas funções de similaridade. Em seu lugar, foi utilizado um novo algoritmo baseado em deteção de comunidades. Estes algoritmos tiram partido da topologia da rede, isto é, da forma como os APs se distribuem no espaço, da proximidade e visibilidade entre eles.

Em geral, foram adaptados apenas quatro conceitos dos oito existentes na arquitetura original: dados em bruto, observação, local e estadias. O objetivo da utilização desta arquitetura, passou por perceber como é que os dados, depois de recolhidos pelos sensores, se iam organizar após sofrerem os vários processos de transformação e inferência baseada em dados existentes.

3.1.1 Dados em Bruto

A disponibilidade das redes Wi-Fi aumentou significativamente, não só em locais públicos, como universidades, aeroportos e centros comerciais, mas também em locais privados, como zonas habitacionais. Sendo assim, estamos perante um excelente ambiente para perceber o movimento de cada indiví-

duo, bem como os locais de interesse por eles frequentados. Com a grande instalação de redes Wi-Fi, será possível obter dados relativos a APs Wi-Fi, visíveis a um indivíduo, em particular através do seu dispositivo móvel.

De modo a tornar possível todo o processo de *place learning* são obrigatórios um conjunto de dados provenientes dos sensores Wi-Fi. Esses dados devem incluir o instante de tempo em que foram recolhidos e a informação acerca de todos os AP visíveis no momento da recolha, tais como o endereço MAC (BSSID) e nível do sinal recebido (RSSI). De notar que o dispositivo móvel não necessita de estar ligado a um determinado AP Wi-Fi. Precisa apenas de detetar os sinais Wi-Fi presentes no ambiente, mesmo que estes sejam privados.

3.1.2 Observações

O conceito de observação tem como objetivo definir um conjunto de dados necessários para os próximos passos do processo de *place learning*. Esta informação diz respeito à observação do ambiente Wi-Fi que rodeia o dispositivo num determinado instante de tempo. Sendo assim, o conceito observação vai de encontro à definição de contexto utilizada nesta dissertação. Uma observação é a informação que pode ser utilizada para caracterizar a situação de uma entidade. Neste caso a utilização do ambiente Wi-Fi foi utilizada para determinar um contexto de localização do utilizador do dispositivo que recolheu esses dados. Por si só, os *beacons* Wi-Fi enviados pela rede, indicam o contexto.

Para além do contexto de localização, a informação de contexto temporal permitiu conhecer o instante ou período em que ocorreu uma determinada observação, bem como a possibilidade de ordenar temporalmente todas as observações. Esta linha temporal ordenada permite tirar partido da riqueza e valor da informação histórica.

Uma observação é representada pelos seguintes atributos:

(Id_Observation, Timestamp, StrongestAP, SAP_Neighbors)

O primeiro atributo é um identificador único de cada observação. O atributo *Timestamp* contém o instante de tempo em que essa observação foi obtida. O atributo *StrongestAP* representa o AP Wi-Fi com maior nível de sinal presente no ambiente que rodeia o utilizador no momento em que foi observado. Os *SAP_Neighbors* identificam os restantes APs vizinhos que foram vistos em simultâneo.

A transformação dos dados brutos em observação é um processo simples. Cada elemento da lista dos dados Wi-Fi recolhidos pelos sensores é transformado numa observação. O aspecto chave reside na identificação do AP com maior força de sinal e nos seus APs vizinhos. Um exemplo de uma lista de observações ordenadas temporalmente é representado na tabela 4.1. O atributo *TIME* dos dados brutos é correlacionado com o *Timestamp* das Observações.

3.1.3 Locais Potenciais

Entre a camada Observação e Locais ocorre o processo de transformação mais importante de todo o processo de descoberta de locais. É nessa transformação que ocorre a deteção dos locais potenciais a partir das observações (*StrongestAP* e *SAP_Neighbors*) presentes na camada Observações.

Um local potencial é definido pela agregação de um ou mais APs (Figura 3.2). Atualmente, é possível observar mais que um AP associado a um local. Nas grandes zonas urbanas este fenómeno é verificado com maior intensidade. Por exemplo, numa zona residencial praticamente todos os apartamentos possuem um AP. Se um individuo morar num desses apartamentos, é certo que

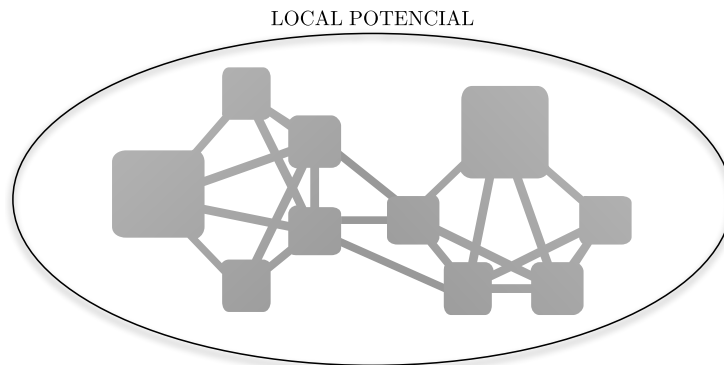


Figura 3.2: Representação abstrata de um local.

deteta mais que um único AP. Esta realidade obrigou à necessidade de representar um local como um conjunto de vários APs. Contudo, existem casos isolados onde existe um único AP associado a um local. Nestes casos o local potencial resume-se apenas a um AP.

Este processo de agregação de APs em locais foi realizado recorrendo a um algoritmo de deteção de comunidades. O objetivo destes algoritmos é identificar comunidades em grafos baseando-se somente na topologia da rede. De acordo com Fortunato (Fortunato & Castellano, 2007), as comunidades são grupos de nodos que possuem características em comum, ou que assumem papéis semelhantes dentro da rede. No contexto desta dissertação, uma comunidade será um conjunto de um ou mais APs e a característica em comum é a proximidade entre eles. Neste trabalho, assume-se que dois APs são próximos se forem visíveis em simultâneos na mesma observação.

A deteção das comunidades e suas fronteiras permitiu a descoberta dos locais potenciais. Um local potencial é aquele que ainda não foi classificado como sendo realmente relevante para a caracterização da mobilidade de um indivíduo. Nesta fase do processo, ainda não é possível inferir se um local é relevante, ou não, na vida do utilizador. A sua classificação só é possível com base nas estadias, o próximo passo no processo de descoberta de locais.

3.1.4 Estadias

Com base em observações consecutivas $\{o_i, o_{i+1}, o_{i+2}, \dots\}$, uma estadia pode ser definida pelo intervalo de tempo entre a primeira e última observação num mesmo local, e é representada pelos seguintes atributos:

$$(Id_Stay, Place, Timestamp0, Timestamp1)$$

O primeiro atributo é um identificador único de cada estadia. O *Place* representa o local em que ocorreu essa estadia. Os atributos *Timestamp0* e *Timestamp1* representam o tempo de entrada e saída da estadia num determinado local.

Existem contudo algumas restrições na extração das estadias. Uma estadia só ocorre quando, para um conjunto de observações consecutivas, o local é o mesmo. Para ser considerada uma estadia, o intervalo de tempo entre a observação inicial e final tem que ser, pelo menos, superior a zero. Por fim, o intervalo entre quaisquer duas observações deve ser o mais curto possível (por exemplo, 2 minutos) para ser possível de assumir que o utilizador não deixou o local entre essas duas observações. Um exemplo desta última restrição ocorre quando existe uma falha na recolha de dados. Se porventura falhar a recolha e o intervalo de tempo entre duas observações consecutivas for demasiado grande, não é possível assumir que o indivíduo permaneceu sempre no mesmo local.

3.1.5 Locais Visitados e Locais Relevantes

A partir da obtenção da informação das estadias estão reunidas então as condições necessárias para conhecer se um local potencial é afinal um local que foi realmente visitado e/ou é relevante na vida do indivíduo.

Um local potencial só é classificado como visitado quando ocorre pelo

menos uma estadia com um tempo de permanência superior a um determinado tempo mínimo.

Um local relevante deve ser um local importante e significativo no cotidiano do indivíduo (i.e., local visitado frequentemente e/ou onde o indivíduo permanece por longos períodos de tempo). Isto é, relevante no contexto da mobilidade do indivíduo.

3.2 Algoritmo

Nesta secção serão descritos todos os processos de transformação que ocorrem entre as camadas da arquitetura. Estes processos são essenciais para executar simples transformações nos dados bem como inferir nova informação a partir da informação já existente.

3.2.1 Extração do Grafo

Entre a camada observação e local existem dois processos de transformação intrínsecos à seta representada na Figura 3.1. O primeiro processo envolve a extração de um grafo ponderado a partir das observações. Um grafo ponderado é representado como $G = (N, E)$ onde N é um conjunto de nodos e E um conjunto de arestas. Cada aresta e_i representa a ligação entre dois nodos n_a e n_b com um peso w_i e é definida por $e_i = \{n_a, n_b, w_i \mid n_a, n_b \in N, w_i \in \mathbb{N}\}$.

Utilizando a informação dos atributos *StrongestAP* S , *SAP_Neighbors* SN de cada observação e do número de vezes que foram observados em simultâneo, foram adotadas duas abordagens na criação dos grafos (Figura 3.3).

Na primeira abordagem, ilustrada na Figura 3.3a, o grafo G_α é gerado com base apenas na proximidade entre S e SN , para cada observação. O conjunto de nodos N é composto pela intersecção de S e SN para todas as observações, e as arestas E representam a proximidade entre os nodos. O número de vezes

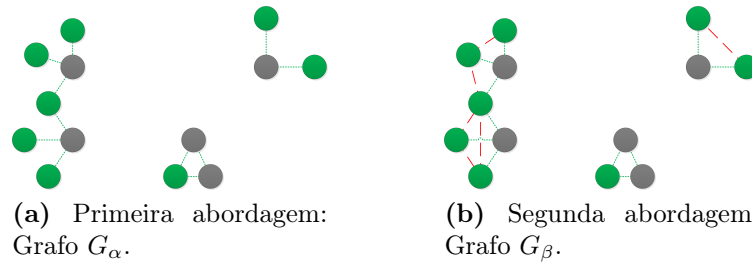


Figura 3.3: Exemplo das duas abordagens calculadas a partir de uma lista de observações. Os nodos a cinza representam os nodos S e os nodos SN aparecem a verde. As arestas a verde representam as ligações entre os nodos S e SN . A ligação entre os nodos SN estão representadas a cor vermelha.

em que S e SN foram observados simultaneamente é utilizado como peso w_i de cada aresta e_i . Os nodos S estão coloridos a cor cinza e os SN a verde. As arestas estão a cor verde.

Na segunda abordagem, representada na Figura 3.3b, o grafo G_β para além da proximidade de S e SN , utilizada na abordagem anterior, considera-se também na proximidade entre os SN . A grande diferença entre as duas abordagens reside na inclusão da visibilidade entre os APs vizinhos como arestas do grafo. Este novo tipo de arestas é representado a cor vermelha.

As duas abordagens resultantes serão utilizadas como dados de entrada no outro processo de transformação intrínseco à seta representada entre a camada observação e local: o processo de deteção de comunidades.

3.2.2 Deteção de Comunidades

O grafo obtido durante o processo de transformação apresentado na subsecção anterior, possibilita a extração de informação da sua estrutura (Albert & Barabási, 2002). Utilizando essa rede como partição inicial do algoritmo de comunidades, o objetivo foi descobrir quais os nodos (APs) que pertencem à mesma comunidade através da frequência com que são visíveis em simultâneo.

O algoritmo escolhido para a deteção de comunidades foi o método de Louvain (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Existem muitos

outros algoritmos que podiam ter sido utilizados (Clauset, Newman, & Moore, 2004; Pons & Latapy, 2005; Wakita & Tsurumi, 2007). Contudo, Blondel et al. (2008) conseguiram provar uma ótima precisão do algoritmo que fornece uma boa decomposição em comunidades. Em relação ao tempo de computação, este algoritmo ultrapassa todos os outros. À medida que o número de observações for aumentando, o grafo poderá conter um número considerável de nodos e arestas. Desta forma, o tempo de computação foi um aspecto chave na decisão de escolher este algoritmo para o processo de descoberta de comunidades.

Método de Louvain

O algoritmo de Louvain é dividido em duas fases, (i) otimização da modularidade e (ii) agregação em comunidades, que se repetem iterativamente até a modularidade não conseguir ser melhorada (Figura 3.4).

A partição inicial tem como base um grafo ponderado¹ com N nodos. A primeira fase começa por associar uma comunidade a cada nodo. Depois o nodo i é comparado com os seus vizinhos j até ser encontrado o máximo ganho positivo de modularidade. O ganho de modularidade é calculado a partir de uma equação que envolve a soma do peso das arestas pertencentes à comunidade a ser testada, a soma do peso das arestas incidentes ao nodo i , a soma do peso das arestas do nodo i para os nodos da comunidade a ser testada e a soma dos pesos de todas as arestas do grafo. A comparação é realizada colocando o nodo i na comunidade do nodo j . Caso haja um ganho positivo, o i é colocado na comunidade que lhe permite ter um maior ganho positivo. Se não existir um ganho positivo então o i permanece na comunidade de origem. Este processo é aplicado repetida e sequencialmente para todos os nodos até que não seja possível melhorar a modularidade com os deslocamentos dos nodos i para as comunidades vizinhas, e assim termina a primeira fase.

¹Um grafo é ponderado quando suas arestas possuem um peso.

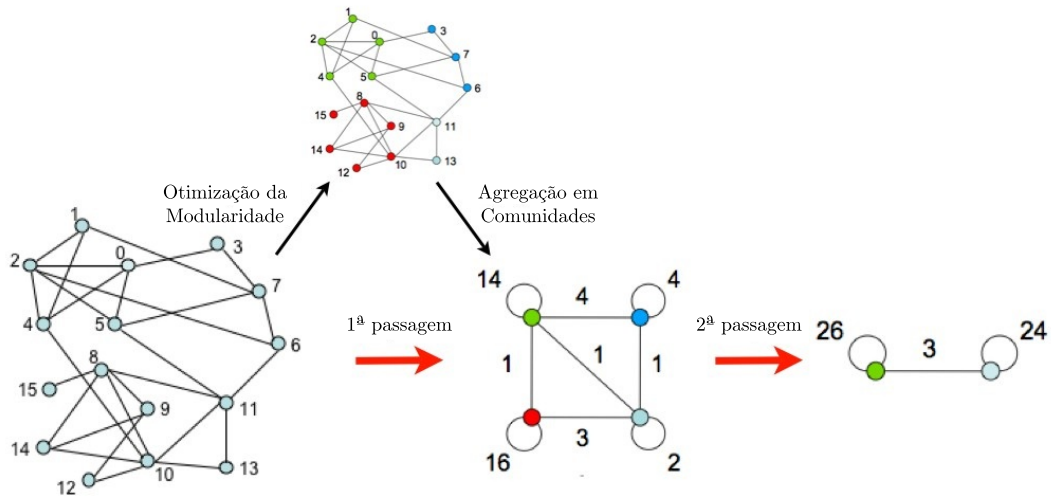


Figura 3.4: Método de Louvain composto por duas fases (adaptado de (Blondel et al., 2008)): otimização da modularidade e agregação em comunidades. Neste exemplo estão representadas apenas duas passagens. Contudo, o número de passagens depende da quantidade de iterações realizadas até a modularidade não conseguir ser melhorada.

Na segunda fase, uma nova rede é criada onde os nodos c são agora as comunidades descobertas na primeira fase. É criada uma ligação entre dois nodos c caso tenha existido pelo menos uma aresta entre os nodos das comunidades correspondentes na fase anterior. O peso da aresta entre dois c é dado pela soma dos pesos das arestas entre os nodos das comunidades correspondentes na fase anterior.

Assim que concluída a segunda fase, as duas fases são novamente repetidas utilizando a rede resultante da passagem anterior. Uma passagem é a combinação das duas fases. A cada passagem o número de comunidades vai diminuindo, e estas passagens são iteradas até que não haja mais nenhuma alteração e um máximo de modularidade seja atingido.

A modularidade é uma medida utilizada para avaliar a qualidade das partições obtidas e assume valores entre -1 e 1, sendo que quanto mais próximo está de 1 melhor é a qualidade da partição realizada (Blondel et al., 2008; Fortunato, 2010).

Fortunato (2010) mostra que a partição com maior modularidade, mais próxima de 1, nem sempre é a melhor. Isto acontece porque muitas das comunidades de tamanho relativamente pequeno podem passar despercebidas ao algoritmo de comunidade. Blondel et al. (2008) também reconhecem este problema. Contudo, este não é um problema para o algoritmo escolhido, pois este fornece uma decomposição da rede em vários níveis hierárquicos, onde cada um dos níveis revelados é determinado pelas partições encontradas em cada passagem. Isto permite percorrer os vários níveis da estrutura hierárquica originada pelo algoritmo e escolher o nível hierárquico com melhor partição em termos de comunidades para cada aplicação concreta.

Como a modularidade não é um critério de escolha fiável, o nível da hierarquia a escolher terá que ser aquele que fornecer a partição onde as comunidades não estejam nem muito divididas (baixo nível na hierarquia) nem agregadas em demasia (alto nível na hierarquia), de forma reduzir a probabilidade de descoberta de locais divididos ou agregados.

3.2.3 Extração de Estadias

A extração de uma estadia num determinado local implica a agregação de observações consecutivas que pertencem ao mesmo local.

Dado um conjunto de observações no mesmo local $O = \{o_1, o_2, \dots, o_n\}$, de modo a assumir a deteção de uma estadia, o tempo passado entre quaisquer duas observações consecutivas $\{o_i, o_{i+1}\}$ não deve ser maior que um determinado tempo máximo (T_{max}). O tempo de estadia é dado por $S = t_n - t_1$ e deve ser superior a zero.

Assim que o local entre duas observações consecutivas se alterar, assume-se que o utilizador movimentou-se de um local para outro e a estadia no local anterior é interrompida.

A informação presente nas estadias é das mais importantes no que diz

respeito ao enriquecimento de informação dos locais e transições entre eles. As estadias permitem a caracterização de cada local, como os tempos de estadia, número de visitas, tempos médio de entrada e saída. Esta camada possui a informação necessária para a deteção das transições entre dois locais.

3.2.4 Classificação de Locais

A análise das estadias permite a transformação dos locais potenciais em locais visitados através da aplicação de um filtro. Este passo tem como objetivo classificar os locais como visitados se realmente o utilizador permaneceu nesse local por um tempo considerável.

Os locais visitados podem agora ser classificados como relevantes ou não. A análise das estadias permite a classificação dos locais visitados em locais relevantes através da aplicação de um filtro. Este passo tem como objetivo classificar os locais como relevantes se estes forem realmente importantes e significativos no quotidiano dos indivíduos (i.e., locais visitados frequentemente e/ou onde o indivíduo permanece por longos períodos de tempo).

A classificação de locais não é uma técnica nova. Alguns autores de algoritmos de *place learning* utilizaram a classificação com o mesmo propósito da classificação de locais nesta dissertação. Marmasse e Schmandt (2000) classificam um local relevante para o utilizador quando este o visita pelo menos três vezes. Kang et al. (2005) consideram como um local significativo aquele local onde o utilizador passa mais tempo e/ou apresenta elevada frequência de visitas. Os autores do LifeMap (Chon, Shin, Talipov, & Cha, 2012), definem como critério diferenciador, um *threshold* de mais de três visitas a um local para este ser considerado como um local frequente.

Aproveitando as ideias dos autores acima, pretendeu-se estudar um conjunto de critérios para a classificação de locais, com base na análise dos dados referentes às estadias em locais.

Para o primeiro filtro, classificação dos locais potenciais em locais visitados, este tem em consideração apenas um critério: estadias com tempo de permanência superior a um determinado tempo mínimo.

O segundo filtro de decisão não é tão simples como o anterior. Este contém os dois critérios, que são responsáveis por decidir se um local deve ser considerado relevante ou não: (i) frequência de visitas a um local, por dia, e (ii) tempo total de permanência num local sobre o tempo total de observações. Por exemplo, se durante as observações foram recolhidos dados durante um período de 15 dias, o tempo total de observações serão esses 15 dias.

Ainda em relação ao segundo filtro, em termos de análise de mobilidade, o primeiro critério é muito mais relevante do que o segundo. Se um local apresenta uma alta frequência de visitas, significa que o indivíduo visita regularmente esse local. Isso implica a existência de custos de transporte e ambientais. Uma visita a um local é considerada quando o indivíduo dispense mais do que um determinado tempo no mesmo local. A existência deste limite de tempo permite diferenciar uma visita de uma simples passagem num local.

O tempo de permanência num local não fornece nenhuma informação acerca da mobilidade do indivíduo. Mas se ao longo das observações permanece uma grande parte do seu tempo nesse local, então é porque esse local é significativo no dia a dia do indivíduo.

A combinação destes dois critérios permite diferenciar os locais que realmente são relevantes dos que não o são (Figura 3.5). Os locais relevantes devem ser aqueles com alta mobilidade ou com alta permanência. Em contraste, os locais visitados são aqueles que ainda não atingiram os requisitos para serem locais relevantes (baixa mobilidade e baixa permanência).

A questão que emerge agora é: “O que é considerada uma alta mobilidade? E uma alta permanência?” O desafio nesta etapa do processo é descobrir as fronteiras que separam os dois tipos, nas duas dimensões.

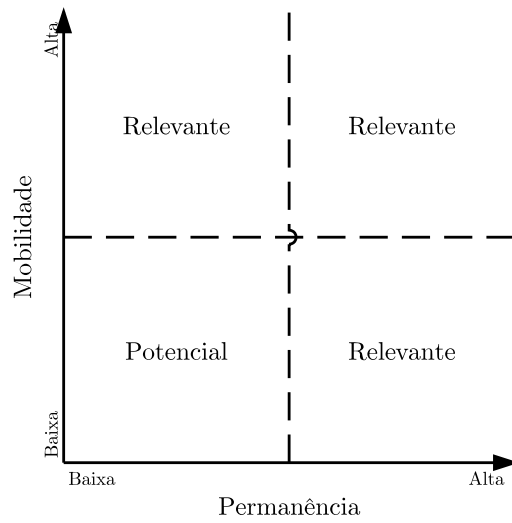


Figura 3.5: Classificação dos locais segundo a sua mobilidade e permanência.

3.2.5 Mapa de Mobilidade Pessoal

A construção dos MMP implica, ou não, a classificação dos locais como relevantes. Isto é, os mapas não necessitam obrigatoriamente dos locais relevantes para serem construídos.

Os dados realmente necessários provêm da camada das estadias e locais visitados. Como referido na secção 3.2.3, as estadias permitem enriquecer os MMP com informação caracterizadora dos locais e respetivas trajetórias entre eles.

Os locais, representados pelos nodos do grafo, são caracterizados pelas visitas (total de visitas e total de visitas por dia) e pela permanência/estadia (total de permanência, total de permanência sobre o tempo total observado, tempo médio de permanência). Quanto às trajetórias entre locais, representadas pelas arestas do grafo, são caracterizadas pelo número de ocorrências (total de ocorrências e total de ocorrências por dia) e pelo tempo despendido a percorrer desde o local origem até ao local destino (tempo total despendido e tempo médio despendido).

A informação caracterizadora, para além de enriquecer os MMP, permite

também a personalização da sua visualização de acordo com o atributo desejado. Por exemplo, com a informação do número de visitas, é possível destacar os nodos com mais visitas dos restantes, através da cor dos nodo ou pelo seu diâmetro. Em relação às arestas, estas podem também ser distintas umas das outras pela cor. Por exemplo, as arestas com maior número de ocorrências, possuem uma cor mais preenchida do que as restantes.

Para além da representação dos locais e respetivas trajetórias, pretende-se sobretudo que os MMP causem impacto logo na primeira visualização. Isto é, para quem os visualiza, seja fácil de perceber quais os locais e trajetórias que se destacam das restantes, de acordo com os atributos desejados.

Capítulo 4

Implementação

Neste capítulo é realizada uma descrição da implementação do algoritmo de descoberta de locais, desde os dados em bruto até aos MMP. Inicialmente é explicado o cenário da implementação, seguido da escolha e justificação das opções tomadas para um bom funcionamento do algoritmo.

4.1 Dados em Bruto

Apesar do processo de recolha não estar inserido no âmbito da dissertação, existiram contudo preocupações em estudar o funcionamento da aplicação utilizada, formato de saída dos dados e se era necessário algum tratamento.

4.1.1 Recolha

O trabalho desenvolvido no contexto desta dissertação não incluiu atividades específicas de recolha de dados, uma vez que já existiam várias atividades neste domínio em curso no grupo de investigação, bem como no projeto TICE.Mobilidade. Deste modo, aproveitando a utilização de uma aplicação já desenvolvida pela Universidade de Coimbra, foi possível a recolha de grandes quantidades de dados ao longo de vários períodos de tempo, e evitou-se a

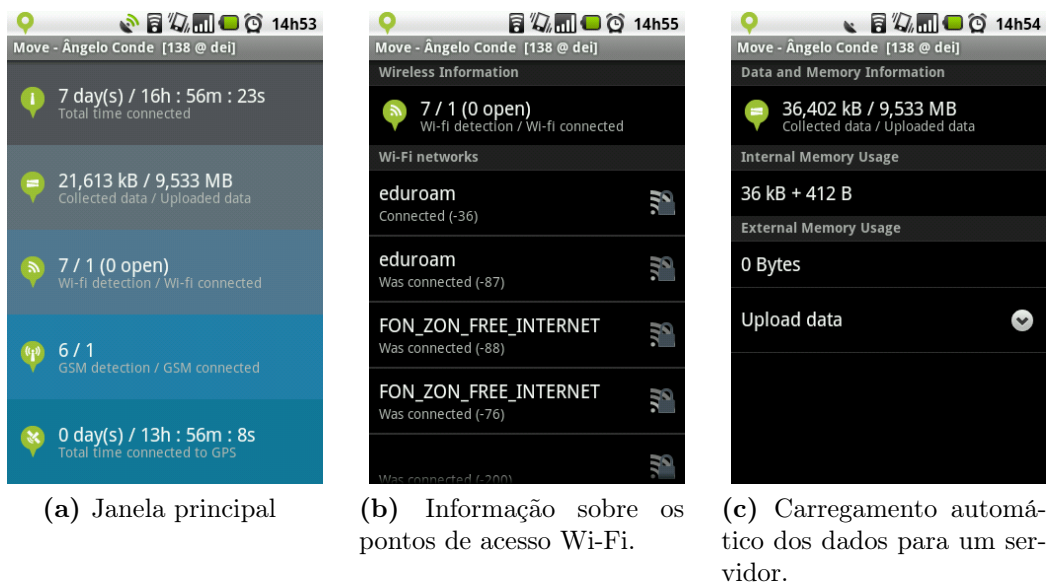


Figura 4.1: Aplicação MOVE.

perda de tempo de desenvolvimento e revisões de estado de arte sobre recolha de dados em *smartphones* (como por exemplo, otimização do desempenho da bateria e a frequência de recolha).

A MOVE é uma simples aplicação que apenas recolhe os dados provenientes dos sensores de forma automática. Existe também uma plataforma web que dá suporte à MOVE de forma a validar todos os dados recolhidos por esta. A validação dos dados inclui a associação de atividades e modos de transporte às trajetórias identificadas pela plataforma.

De modo a recolher os dados necessários para o processo de descoberta de locais, o utilizador só necessita de instalar a aplicação, e esta é capaz de se iniciar automaticamente sempre que o *smartphone* reinicia ou ocorre algum erro inesperado. Sempre que detetar uma ligação à internet, a MOVE encarrega-se de fazer os carregamentos para o servidor de forma periódica (por exemplo, de 10 em 10 minutos).

A UI é muito simples, sendo constituída por uma série de janelas. A Figura 4.1 mostra três das janelas mais importantes no MOVE. Na primeira

```
{ WIFI_DATA : [  
  {  
    BSSID: "88:43:e1:13:6f:20",  
    EXTRA: "00:0e:d7:cd:e7:90,eduroam,-61,2427|  
            48:5b:39:3b:2a:ee,LID3,-61,2432|  
            00:0e:a6:b7:e1:08,DSI WiFi,-76,2412",  
    SSID: "eduroam",  
    TIME: 1330603506237,  
    RSSI: -47  
  }  
]  
}
```

Figura 4.2: Exemplo dos dados em bruto provenientes da aplicação MOVE.

janela (4.1a) é apresentada a informação básica, como os dias consecutivos em execução, quantidade de dados recolhidos e enviados para o servidor, número de APs Wi-Fi detetados (e se está ligado a algum ou não), bem como outras informações que não são relevantes para este projeto. Na janela seguinte (Figura 4.1b), é mostrada a informação relativa aos APs que estão visíveis e aos quais o utilizador está ou já esteve ligado. Na última janela (Figura 4.1c) é possível ver a quantidade de dados recolhida e enviada, bem como a possibilidade de forçar o envio para o servidor.

Os dados enviados para o servidor encontram-se em formato JSON¹ (Figura 4.2). A simplicidade de como os dados estão organizados, em listas possíveis de iterar pelos elementos que as constituem recorrendo a diversas bibliotecas existentes (como por exemplo a GSON desenvolvida pela Google), tornou simples todas as atividades que implicavam o tratamento dos dados em bruto.

Os dados incluem o atributo *TIME* que identifica o instante de tempo em que os dados foram recolhidos. Em relação ao AP a que o dispositivo móvel se

¹*JavaScript Object Notation* (Crockford, 2006).

encontra ligado, o atributo BSSID identifica o endereço MAC do AP, o *SSID* é um conjunto de caracteres que identifica a rede e o *Received Signal Strength Indication* (RSSI) é o nível de sinal de rádio recebido. Caso o dispositivo móvel não se encontre ligado a nenhum AP, os campos BSSID, SSID e RSSI encontram-se vazios. O atributo *EXTRA* representa todos os APs vizinhos no formato [BSSID, SSID, RSSI, FREQUENCY].

4.1.2 Tratamento

Um dos problemas que obrigou à necessidade de tratar os dados em bruto foram os cada vez mais usuais pontos de acesso móveis. A possibilidade de criar redes *ad-hoc* nos computadores portáteis, bem como nos *smartphones*, é um problema para qualquer algoritmo de inferência de locais que tenha por base somente dados do tipo Wi-Fi. Como estas redes podem ser criadas em qualquer altura e em qualquer lugar, não podem ser associadas a um determinado local.

De modo a reduzir falsas deteções de locais foi necessário remover todos os APs móveis para ficarem apenas os estáticos. A técnica utilizada consistiu em filtrar todos os APs administrados localmente (por exemplo, redes locais criadas por computadores portáteis) (D. Kim et al., 2010).

De acordo com a norma IEEE 802-2001 (IEEE, 2002, p. 21), o endereço MAC de um AP pode administrado universalmente ou localmente. Estes APs podem ser distinguidos pela verificação do segundo bit menos significativo (lsb²) pertencente ao byte mais significativo (MSB³). O MSB é o que se encontra mais à esquerda na representação hexadecimal, e o lsb o que se encontra mais à direita na representação binária. Sendo assim, o bit que permite distinguir um AP móvel de um estático é o segundo a contar da direita. Se esse bit é 1, então o AP é móvel (administrado localmente). Caso seja 0, o AP é

²Do inglês *least significant bit*.

³Do inglês, *most significant byte*.

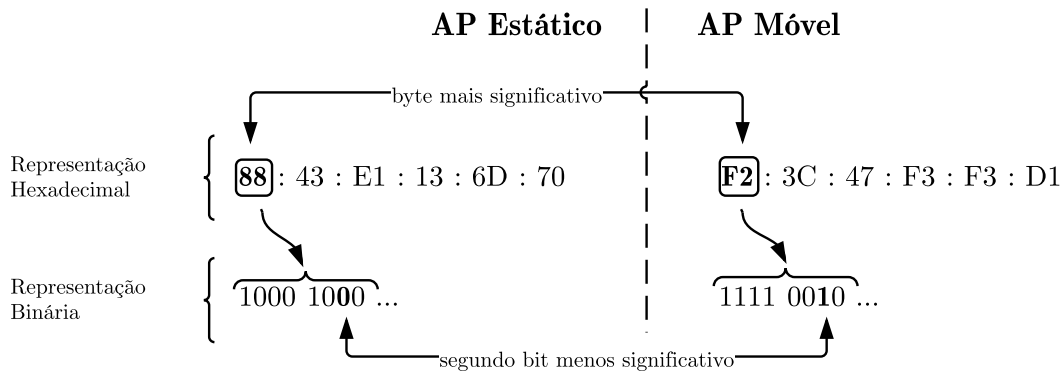


Figura 4.3: Processo de filtragem dos endereços MAC associados a APs móveis.

estático (administrado universalmente). A Figura 4.3 mostra como distinguir se um AP é móvel ou não, através do seu endereço MAC.

4.1.3 Armazenamento

De modo a suportar as camadas de dados definidas na arquitetura do processo de descoberta de locais (Figura 3.1), foi necessário definir e desenvolver um repositório de dados (Figura 4.4).

O modelo de dados foi desenvolvido sobre o Sistema de Gestão de Base de Dados (SGBD) *MySQL*, tendo por base a linguagem *SQL*. O fato de ser um *software* livre e a facilidade com que é possível trocar informação entre o algoritmo e a base de dados através de uma API desenvolvida em linguagem Java (*Java Database Connectivity* (JDBC)) foram os critérios que mais pesaram na escolha do SGBD.

O repositório é constituído por cinco tabelas: *wifi_aps*, *places*, *sap_neighbors*, *observations* e *stays*.

Como o próprio nome o sugere, a tabela *observations* armazena toda a informação da camada *Observação*. Uma observação é representada por um conjunto de APs Wi-Fi. A tabela *wifi_aps* é responsável por armazenar a informação sobre todos os APs. Como existe uma diferenciação entre o AP com maior sinal dos restantes APs, cada observação tem um conjunto de APs

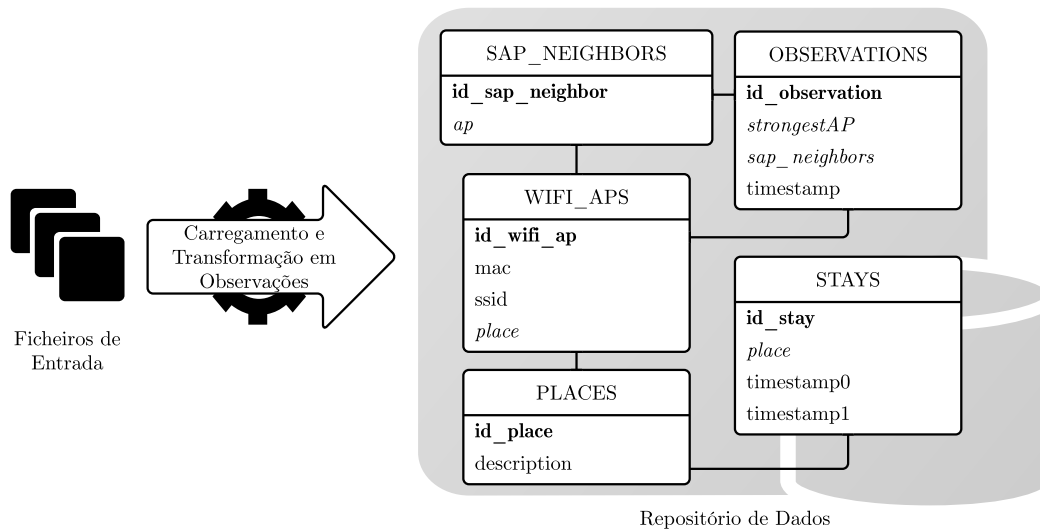


Figura 4.4: Repositório de suporte a todo o processo de descoberta de locais representado na Figura 3.1

vizinhos associados que são armazenados na tabela *sap_neighbors*. Após a detecção de comunidades, cada AP tem uma comunidade associada. A informação sobre essas comunidades descobertas, ou locais, é armazenada na tabela *places*. A informação das observações em conjunto com as comunidades permite a descoberta das estadias. Estas são armazenadas na tabela *stays*.

4.2 Carregamento dos Dados

O primeiro passo do algoritmo desenvolvido foi ler os ficheiros JSON provenientes do servidor da aplicação MOVE. Utilizando uma biblioteca desenvolvida pela Google, a GSON, os ficheiros JSON foram convertidos⁴ em objetos Java.

Apesar de existirem várias listas presentes nos ficheiros JSON (como por exemplo as listas referentes a cada tipo de dados: *GPS_DATA*, *GSM_DATA*,

⁴Este processo de conversão de ficheiros em estruturas de dados ou objetos é denominado de *deserialization*. O processo de *serialization* é o inverso, conversão de estruturas de dados ou objetos em ficheiros de dados.

WIFI_DATA, entre outros) será utilizada apenas uma delas, a *WIFI_DATA* (Figura 4.2). Contudo, é necessário correlacionar todas as listas e respectivos atributos dos ficheiros JSON com as classe em Java. Assim, cada lista presente no ficheiro JSON tem obrigatoriamente uma classe Java associada.

Da mesma forma que um ficheiro JSON contem um conjunto de listas, o mesmo deve ser feito com a criação de uma classe Java. Denominada de *MOVE_DATA*, nesta classe são definidas as variáveis de instância que dizem respeito às listas, e os métodos que permitem definir e obter essas listas (Figura 4.5). A anotação `@Expose` permite identificar se uma variável é exposta ou não ao processo de conversão. A não exposição à conversão significa que não há transferência da informação dessa variável do ficheiro JSON para o objeto Java. Como só necessitamos da lista *WIFI_DATA*, apenas a variável que identifica essa lista se encontra anotada. Mais tarde, se eventualmente for necessário a utilização de outro tipo de dados, basta apenas anotar cada uma das variáveis de instância.

A Figura 4.6 ilustra o aspeto de uma lista em forma de classe de dados. Esta contém uma série de variáveis de instância com o mesmo nome dos atributos presentes na lista, como o *BSSID*, *EXTRA*, *SSID*, *TIME* e *RSSI*, bem como um conjunto de métodos que permitem definir e obter informação dos objetos. Aqui, a anotação funciona da mesma forma. Se não for necessário um certo atributo do ficheiro, basta não anotar a variável de instância correspondente.

Após definidas as classes de dados, a leitura e conversão dos ficheiros JSON é um processo bastante simples através da utilização da biblioteca GSON. A Figura 4.7 retrata esse processo. Primeiro ocorre o carregamento do ficheiro de dados, e através do método `fromJson()` é convertido para um objeto de dados. Este objeto é uma instância da classe *MOVE_DATA* definida na Figura 4.5. Com os métodos definidos nessa classe, bem como nas classes relativas às

```
1 package pt.uminho.ubicomp.mobilitymaps;
2 import java.util.List;
3 import com.google.gson.annotations.Expose;
4
5 public class MOVE_DATA {
6     List<ACCE_DATA> ACCE_DATA;
7     List<BATT_DATA> BATT_DATA;
8     List<CELL_CMDA_DATA> CELL_CMDA_DATA;
9     List<CELL_GSM_DATA> CELL_GSM_DATA;
10    @Expose List<WIFI_DATA> WIFI_DATA;
11    List<LOGOUT_DATA> LOGOUT_DATA;
12    List<GPS_DATA> GPS_DATA;
13
14    public List<WIFI_DATA> getWIFI_DATA() {
15        return WIFI_DATA;
16    }
17    public void setWIFI_DATA(List<WIFI_DATA> wWIFI_DATA)
18        {
19        WIFI_DATA = wWIFI_DATA;
20    }
21    ...
22 }
```

Figura 4.5: Classe de dados JAVA representativa da estrutura dos ficheiros de dados JSON. Esta permite a conversão dos ficheiros de dados JSON em objetos Java.


```
1 package pt.uminho.ubicomp.mobilitymaps;
2 import com.google.gson.annotations.Expose;
3
4 public class WIFI_DATA {
5
6     @Expose String BSSID;
7     @Expose String EXTRA;
8     @Expose String SSID;
9     @Expose long TIME;
10    @Expose int RSSI;
11
12    public String getBSSID() {
13        return BSSID;
14    }
15
16    public void setBSSID(String bSSID) {
17        BSSID = bSSID;
18    }
19    ...
20 }
```

Figura 4.6: Classe de dados JAVA representativa da lista WIFI_DATA presente no ficheiro JSON.

```

1
2 //carregamento do ficheiro JSON
3 BufferedReader br = new BufferedReader(new FileReader(
4     "filename.json"));
5
6 //conversao para um objeto de dados
7 Gson gson = new Gson();
8 MOVE_DATA move = gson.fromJson(br, MOVE_DATA.class);
9
10 //iterar pela lista dos dados wi-fi
11 for (int i = 0; i < move.getWIFI_DATA().size(); i++) {
12     String wifiBSSID = move.getWIFI_DATA().get(w).
13         getBSSID();
14     int wifiRSSI = move.getWIFI_DATA().get(w).getRSSI();
15     ...
16 }
17 ...

```

Figura 4.7: Conversão de um ficheiro JSON para um objeto Java.

listas (Figura 4.6), a leitura da informação é simples e intuitiva para quem lê o código.

4.3 Extração das Observações e do Grafo

Iterando pela informação da lista que contém os dados Wi-Fi, dá-se então início ao processo de transformação dos dados em bruto para observações. O resultado desta transformação é ilustrado na Tabela 4.1.

A partir dos dados das observações é extraída a informação sobre o grafo. Como explicado na secção 3.2.1, existem duas abordagens possíveis para o grafo de entrada no método de Louvain. Através de duas consultas SQL (Figura 4.8), uma para cada abordagem, são obtidos os dados com a estrutura desejada: nodo origem, nodo destino e peso.

A primeira consulta é muito simples, pois é necessário apenas contar o número de vezes que um AP, presente no campo *strongestAP* da tabela *ob-*

Tabela 4.1: Conjunto de observações obtidas a partir dos dados em bruto.

Timestamp	StrongestAP	SAP_Neighbors
01-03-2012 12:05:06	88:43:e1:13:6f:20	[00:0e:d7:cd:e7:90, 48:5b:39:3b:2a:ee]
01-03-2012 12:06:07	88:43:e1:13:6f:20	[00:0e:d7:cd:e7:90, 48:5b:39:3b:2a:ee]
...		
01-03-2012 20:32:32	00:26:5b:2e:30:58	[00:26:5b:2e:30:59, 7c:4f:b5:da:a7:60]

servations, foi visível em simultâneo com outro AP presente no campo *ap* da tabela *SAP_Neighbors*.

Em relação à segunda consulta, esta é muito mais complexa do que a primeira. Esta incluiu a consulta SQL usada na primeira abordagem juntamente (através da utilização do operador UNION) com outra consulta que conta o número de vezes em que os APs da tabela *SAP_Neighbors* são visíveis entre si. Como há a possibilidade de um par de APs origem-destino estar presente nas duas consultas ligadas pelo operador UNION, houve a necessidade de agrupar esses resultados repetidos de modo a somar o número de vezes em que foram visíveis em simultâneo. Essa agregação foi possível combinando os resultados das duas consultas numa só consulta. Com recurso ao operador GROUP BY todos os pares origem-destino repetidos foram agrupados e o valor da sua visibilidade foi somado utilizando a função SUM().

Os resultados dessas consultas são representados por uma matriz de adjacência $A = [a_{ij}]$ onde $a_{ij} = w(n_i, n_j)$ diz respeito aos pesos do grafo ponderado. Esta matriz é então escrita para um ficheiro de texto com a forma origem, destino e peso separados por um espaço entre eles (Tabela 4.2). Este ficheiro serve como partição de entrada no algoritmo de deteção de comunidades.

```

1  /*
2  Consulta SQL do grafo relativo à primeira abordagem.
3  */
4  SELECT o.strongestAP, n.ap, count(*)
5  FROM observations o, sap_neighbors n, wifi_aps a
6  WHERE o.sap_neighbors = n.id_sap_neighbor
7         and o.strongestAP = a.id_wifi_ap
8  GROUP BY o.strongestAP, n.ap
9  ORDER BY o.strongestAP, n.ap
10
11 /*
12 Consulta SQL do grafo relativo à segunda abordagem.
13 */
14 SELECT source, target, sum(weight) FROM (
15     SELECT o.strongestAP as source, n.ap as target, count
16           (*) as weight
17     FROM observations o, sap_neighbors n, wifi_aps a
18     WHERE o.sap_neighbors = n.id_sap_neighbor
19           and o.strongestAP = a.id_wifi_ap
20     GROUP by o.strongestAP, n.ap
21     UNION
22     SELECT n.ap, n1.ap, count(*)
23     FROM sap_neighbors n, sap_neighbors n1
24     WHERE n.id_sap_neighbor = n1.id_sap_neighbor and n.ap
25           != n1.ap
26     GROUP by n.ap, n1.ap)
27 as graph
28 GROUP by source, target
29 ORDER by source, target

```

Figura 4.8: Consulta SQL para a extração do grafo (primeira e segunda abordagem).

Tabela 4.2: Exemplo da organização de um ficheiro de entrada ao algoritmo de Louvain.

Origem	Destino	Peso
1	2	1
3	4	1
4	3	2

```

convert -i graph.txt -o graph.bin -w graph.weights
community graph.bin -w graph.weights -v -l -l > graph.tree
hierarchy graph.tree -l N > graph_n2c_level_N.txt

```

Figura 4.9: Funções existentes no algoritmo de detecção de comunidades desenvolvido pelos próprios autores do algoritmo do método de Louvain.

4.4 Detecção de Comunidades

Foi utilizada uma versão atualizada⁵ do algoritmo de Louvain, desenvolvida pelos próprios autores do método de Louvain na linguagem de programação C++. Com esta estratégia de utilizar o que já está feito, foi possível poupar muito tempo na implementação da solução final.

Desde a submissão do grafo proveniente das observações até à detecção das comunidades foi necessária a execução de três funções: *convert*, *community* e *hierarchy*. Estas funções devem seguir uma sequência ordenada como ilustra a Figura 4.9. Cada uma das funções contém um conjunto de parâmetros que permitem definir certas variáveis (como por exemplo, os ficheiros de entrada e saída) e alterar os comportamentos das funções.

A primeira função a ser executada é a *convert*. Esta tem como objetivo converter um ficheiro de texto para um ficheiro binário. O parâmetro *i* diz respeito ao ficheiro de entrada. O parâmetro *o* permite definir o nome do ficheiro de saída. Como o grafo é ponderado, o parâmetro *w* permite interpretar o grafo como ponderado e escreve os pesos de cada aresta num ficheiro em separado.

A conversão para ficheiro binário é um requisito imposto pelos autores do algoritmo. A razão mais óbvia diz respeito à redução do tamanho do ficheiro

⁵Disponível para descarregar em <https://sites.google.com/site/findcommunities/>.

(por exemplo, um ficheiro em formato texto passa de 1.5 MB para 558 KB). Por outro lado, no formato binário os dados podem vir a perder muita da sua utilidade que possuíam em formato de texto. Em formato binário torna-se muito mais complexo e por vezes impossível realizar a leitura dos dados utilizando folhas de cálculo, processadores de texto e ferramentas de análise de dados.

A seguinte função a ser executada é a *community*. O principal objetivo é realizar o cálculo das comunidades e mostrar os níveis hierárquicos obtidos. O primeiro argumento tem que ser obrigatoriamente o ficheiro binário resultante da função anterior. O parâmetro w diz respeito ao ficheiro com os pesos de cada aresta proveniente da função anterior. Quanto ao v este permite a visualização da informação detalhada sobre o tempo de computação, níveis hierárquicos obtidos e respetiva modularidade. O parâmetro l permite definir qual o nível da hierarquia do grafo a obter. Contudo, o objetivo aqui é obter a estrutura hierárquica para analisar cada uma das partições. Utilizando o valor -1 é obtida a estrutura hierárquica em vez de um grafo num determinado nível. O sinal $>$ permite exportar a hierarquia obtida para um ficheiro.

A última função a ser executada é a *hierarchy*. Esta permite conhecer a comunidade a que cada nodo pertence. O primeiro argumento é obrigatoriamente o ficheiro com as hierarquias resultantes da função anterior. O argumento l permite definir o nível N que queremos analisar. Com a utilização do sinal $>$ é exportada a estrutura do nível hierárquico definido para um ficheiro de texto. O resultado desta função resume-se a um ficheiro de texto contendo todos os nodos e respetivas comunidades associadas.

No fim deste processo, todos os nodos que deram entrada ao algoritmo de Louvain contêm uma comunidade associada. O ficheiro de texto resultante da última função, contém esses nodos e respetivas comunidades sob a forma nodo e comunidade, separados por um espaço (Tabela 4.3).

Tabela 4.3: Exemplo da organização de um ficheiro de saída gerado pelo algoritmo de Louvain.

Nodo	Comunidade
1	1
2	1
3	2
4	2

Posteriormente, esse ficheiro será lido e a informação sobre as comunidades será armazenada na tabela *places* do repositório. A informação sobre a comunidade a que cada AP pertence é definida no atributo *place* da tabela *wifi_aps*. De notar que uma comunidade corresponde a um local potencial, ainda não considerado como local visitado nem local relevante.

4.5 Extração de Estadias

Com os locais descobertos e atribuídos a cada AP, cada observação é agora descrita por esse local e o instante de tempo em que foi recolhida. De modo a descobrir as estadias em locais, é percorrida a lista de Observações por ordem temporal crescente, onde foram agregadas as observações consecutivas em que o local é o mesmo, obedecendo contudo a algumas restrições (a Figura 4.10 ilustra todo o processo de deteção de estadias):

- Se não existir uma observação anterior, são guardados os atributos da observação atual e passa-se imediatamente para a observação seguinte;
- Se o local da observação atual for o mesmo que o da anterior, é verificado o tempo que passou entre essas duas observações. Caso seja menor que o *threshold* T_{max} (foi utilizado um máximo de 5 minutos, que equivale

Tabela 4.4: Exemplo de um conjunto de estadias obtidas a partir das observações e das comunidades detetadas.

Place	Timestamp0	Timestamp1
1	01-03-2012 18:55	02-03-2012 08:31
4	02-03-2012 09:09	02-03-2012 10:05
2	02-03-2012 10:08	02-03-2012 10:25
4	02-03-2012 10:27	02-03-2012 12:46

a 300000 em milissegundos), avança-se imediatamente para a observação seguinte. Caso contrário, dá-se por terminada a estadia. Primeiro verifica-se se o tempo de estadia é superior a zero para confirmar que realmente existiu uma estadia e assim inseri-la na tabela *Stays* através do método *insertStay()*. Por fim, define-se as variáveis com informação da observação atual e dá-se início à descoberta de uma nova estadia;

- Se o local entre duas observações consecutivas alterar, significa que o utilizador abandonou o local e então a estadia é dada por terminada. Novamente, verifica-se se o tempo de estadia é superior a zero para confirmar que realmente existiu uma estadia e assim inseri-la na tabela *Stays*. Depois, são definidas as variáveis com informação da observação atual e retoma-se o processo de descoberta de uma nova estadia.

Resumindo, uma estadia é o conjunto de observações consecutivas onde o local é o mesmo e o tempo total de permanência é superior a zero.

No fim deste processo, isto é, após a leitura de todas as observações e respetiva agregação em estadias, o resultado resume-se a uma tabela denominada *Stays* onde estão presentes todas as estadias detetadas com os atributos relativos ao local, tempo de início e fim, e tempo de estadia calculado com os dois atributos anteriores (Tabela 4.4).


```

1 private static final long T_MAX = 300000;
2 ...
3 ResultSet observations = getConsecutiveObservations();
4 while(observations.next()){
5     timestamp = observations.getLong(3);
6     place = observations.getInt(5);
7     if(!previousObservation){
8         ptimestamp = timestamp;
9         pplace = place;
10        initial_ts = timestamp;
11        previousObservation = true;
12    } else {
13        if(place == pplace){
14            long timebetweenobs = (timestamp-ptimestamp);
15            if(timebetweenobs <= T_MAX){
16                ptimestamp = timestamp;
17            } else {
18                long staytime = (ptimestamp - initial_ts);
19                if(staytime > 0){
20                    insertStay(pplace, initial_ts, ptimestamp);
21                }
22                ptimestamp = timestamp;
23                pplace = place;
24                initial_ts = timestamp;
25            }
26        } else { //local alterou
27            long staytime = (ptimestamp - initial_ts);
28            if(staytime > 0){
29                insertStay(pplace, initial_ts, ptimestamp);
30            }
31            ptimestamp = timestamp;
32            pentity = entity;
33            pplace = place;
34            initial_ts = timestamp;
35        }
36    }
37 }
38 ...

```

Figura 4.10: Processo de detecção de estadias desenvolvido em Java. Aqui ocorre o processo de agregação de observações consecutivas onde o local é o mesmo. Contudo, se o tempo entre duas observações consecutivas for maior que um tempo máximo T_{max} , a observação atual não pode ser agregada com a anterior.

4.6 Classificação de Locais

Cada local é agora caracterizado por um conjunto de estadias. O primeiro filtro de classificação, consiste em classificar os locais potenciais em locais visitados. A partir da informação das estadias, são filtradas todas as estadias com tempo de permanência inferior a 5 minutos. Assim, permanecem apenas os locais em que realmente o utilizador permaneceu.

A partir dos locais visitados e da respetiva informação de estadias são obtidos os dois valores, de frequência e permanência, a utilizar no filtro que permite a classificação de locais visitados em locais relevantes.

Com base na tabela *Stays* é executada uma consulta SQL onde são extraídos esses valores (i) percentagem de tempo total de estadia num local (*tst*) e (ii) frequência de visitas a um local por dia (*fv*). A Figura 4.11 ilustra essa consulta.

Primeiro é obtida a janela temporal *w* a partir da data de início e fim da tabela *Stays*. Os *timestamp* estão por defeito armazenados em milissegundos. Foi realizada então a conversão para minutos (para o cálculo da percentagem de permanência, $w_{minutos}$) e para dias (para o cálculo da frequência de visitas, w_{dias}).

Depois, é executada a consulta SQL que fornece o local *p* e os dois valores pretendidos, *tst* e *fv*. Para cada *p*, o valor *tst* é dado pela soma de todas as estadias superiores a 5 minutos nesse local (aplicação do primeiro filtro), dividida por $w_{minutos}$. A função *FORMAT()* permite o arredondamento para o número de casas decimais desejadas. A função *REPLACE()* permite a remoção da vírgula presente nos resultados numéricos. Para esse mesmo *p* é calculada a frequência de visitas *fv* que é dada pelo número de estadias, onde o tempo de permanência é superior ou igual a 5 minutos (aplicação do primeiro filtro), dividido por w_{dias} .

A Tabela 4.5 é um exemplo do resultado da consulta SQL presente na

```

1  /* data de inicio */
2  SET @t0 = (SELECT timestamp0 FROM stays
3            ORDER BY timestamp0 ASC LIMIT 0,1 );
4  /* data de fim */
5  SET @t1 = (SELECT timestamp1 FROM stays
6            ORDER BY timestamp1 DESC LIMIT 0,1 );
7  /* total de minutos entre as duas datas */
8  SET @totaltime = REPLACE(
9      FORMAT((@t1-@t0)/1000/60,5), ',', ', ');
10 /* total de dias entre as duas datas */
11 SET @totaltime_days = REPLACE(
12     FORMAT((@t1-@t0)/1000/60/60/24,5), ',', ', ');
13
14 SELECT place as p,
15        /* % do tempo total de permanência num local */
16        @tst:=(
17            SELECT REPLACE(FORMAT(sum((timestamp1-timestamp0)
18                                  /1000/60)/@totaltime,5), ',', ', ')
19            FROM stays
20            WHERE place = p and (timestamp1-timestamp0) >= 5
21                   /* filtro visitados */
22            GROUP by place) as totalstaytime,
23        /* frequência de visitas a um local por dia */
24        IFNULL(@fv:=(
25            SELECT REPLACE(FORMAT(count(*)/@totaltime_days,5),
26                              ',', ', ')
27            FROM stays s2
28            WHERE place = p and (s2.timestamp1-s2.timestamp0)
29                   >= 5 /* filtro visitados */
30            GROUP BY place),0) as visits
31 FROM stays s
32 GROUP BY place

```

Figura 4.11: Extração dos dois critérios a utilizar no filtro de classificação de locais.

Tabela 4.5: Exemplo dos critérios obtidos a partir da consulta SQL da Figura 4.11. Estes dados serão utilizados no filtro responsável pela classificação dos locais.

Place	Tempo de Permanência / Tempo Total	Frequência de Visitas / Dia
1	0.54886	1.78832
4	0.20713	2.90326
2	0.03075	3.58767

Figura 4.11. Estes resultados permitem alimentar o filtro responsável pela classificação dos locais visitados em relevantes.

4.7 Mapas de Mobilidade Pessoal

A criação dos MMP requer a transformação da informação presente na tabela *Stays*. A partir dela são extraídos os nodos, as arestas e toda a informação que os caracteriza.

Os nodos e arestas foram extraídos da tabela com recurso a duas consultas SQL. Em relação aos nodos, agrupando os registos por local foram obtidos todos os atributos necessários para enriquecer os MMP. A maneira como foram obtidos é semelhante à consulta SQL presente na Figura 4.11. Quanto às arestas, em vez de se agrupar por local, cada registo foi agrupado segundo o local origem e o local destino que formam uma trajetória.

Os resultados da consulta dos nodos são semelhantes aos presentes na Tabela 4.5. Isto é, cada local tem uma série de informação que o caracteriza. Neste caso, para além dos dois atributos presentes na tabela, estão presentes todos os outros que foram definidos na secção 3.2.5. Os resultados da consulta das arestas difere apenas na existência do local origem e destino.

A informação obtida permitiu a criação de um grafo com um conjunto de nodos e arestas. Este grafo juntamente com a informação caracterizadora

define-se como um MMP. Dois tipos de MMP são então criados: os MMP dos locais visitados e os MMP dos locais relevantes. No primeiro está presente apenas o primeiro filtro. Relativamente aos MMP dos locais relevantes, está presente o primeiro e segundo filtro.

A próxima etapa passa pelo desenho e visualização do MMP. Todo este processo, desde as consultas SQL até à visualização final, poderá ser realizado com recurso à aplicação Gephi⁶.

A Gephi é uma plataforma de exploração e visualização interativa para todos os tipos de redes e sistemas complexos, dinâmicos e grafos hierárquicos (Bastian, Heymann, & Jacomy, 2009). Possui uma extensa lista de aplicações de análise exploratória de dados, análise de ligações entre arestas do grafo, análise de redes sociais e biológicas, bem como um conjunto de métricas implementadas como a *betweenness*, o diâmetro da rede e a modularidade na deteção de comunidades, entre outras.

A plataforma possui integração com vários tipos de base de dados. Isto permitiu a execução de consultas SQL dentro da própria plataforma de modo a obter os nodos e as arestas do grafo e respetiva informação caracterizadora.

Após a execução das consultas com sucesso, define-se o *layout* do mapa. A escolha depende muito do tipo de grafo que se quer obter, bem como da dimensão do mesmo (i.e., do número de nodos). Para a visualização dos MMP, com reduzida dimensão, foi utilizado o *layout* ForceAtlas. No caso dos grafos utilizados na deteção das comunidades, onde estão presentes todos os nodos observados, é utilizado o *layout* ForceAtlas2. Este último tem a capacidade de lidar com grafos de grandes dimensões apresentando um resultado final de boa qualidade.

Desenhado o grafo, o próximo passo será personalizá-lo de acordo com os atributos desejados. Recorrendo à funcionalidade *Ranking* da Gephi, é

⁶Disponível em <http://gephi.org/>.

possível colorir ou ajustar o tamanho dos nodos de acordo com os valores do atributo escolhido. Quanto às arestas, só é possível colorir.

Para o desenho e visualização dos MMP podiam ter sido utilizadas outras aplicações de visualização de grafos, como a Graphviz⁷, Tulip⁸, NetworkX⁹, entre muitas outras¹⁰. Contudo, a escolha recaiu na Gephi por ser a mais completa e permite a obtenção de resultados finais com elevada qualidade. Mais completa no sentido de ser possível trabalhar os dados devido à existência de um “laboratório de dados”, bem como guardar o grafo em vários formatos compatíveis com outras aplicações. Quanto aos resultados finais, os *layouts* de elevada qualidade e precisão obtidos através dos variados algoritmos de *layout* existentes, o baixo tempo de processamento para obter o grafo final (apesar de depender muito do algoritmo escolhido), bem como a possibilidade de exportar os grafos para formatos de imagem vetoriais com grande qualidade, coloca a Gephi em vantagem em relação às outras aplicações de visualização de grafos.

⁷Disponível em <http://www.graphviz.org/>.

⁸Disponível em <http://www.tulip.labri.fr/>.

⁹Disponível em <http://networkx.lanl.gov/>.

¹⁰Lista de várias aplicações para desenho e visualização de grafos disponível em http://www.dmoz.org/Science/Math/Combinatorics/Software/Graph_Drawing/.

Capítulo 5

Resultados e Validação

O processo de descoberta, ou inferência, de locais, como o proposto no capítulo anterior, é um algoritmo que procura, de forma automática e autónoma (sem qualquer conhecimento prévio da topologia da rede ou intervenção dos utilizadores), inferir informação de contexto dos locais relevantes dos indivíduos a partir de um conjunto de dados recolhidos. Contudo, este processo não está isento de erros, uma vez que algumas situações inesperadas (como por exemplo a inexistência de redes Wi-Fi) podem levar o algoritmo a apresentar resultados incorretos.

Na descoberta de locais, é importante verificar se todos os passos do processo, em especial na deteção de comunidades, permitem detetar os locais e respetivas trajetórias realizadas com base somente nos pontos de acesso Wi-Fi.

Ao longo deste capítulo descrevem-se os resultados obtidos para um conjunto de voluntários, provando que o algoritmo desenvolvido apresenta uma boa eficácia na descoberta de locais, bem como na deteção dos momentos de entrada e saída de locais.

5.1 Recolha de Dados

No capítulo 3 é apresentado todo o algoritmo de descoberta de locais, desde a recolha dos dados até à construção dos MMP. No capítulo seguinte é explicado como esse algoritmo foi implementado. Esta proposta apresentada baseia-se na deteção dos sinais, dos cada vez mais ubíquos APs Wi-Fi, permitindo inferir os locais visitados e respetivas trajetórias.

De modo a obter uma boa fonte de dados, três utilizadores (A, B e C) voluntariaram-se para recolher dados ao longo do seu dia-a-dia. A cada utilizador foi disponibilizado um *smartphone* com sistema operativo Android para a recolha de dados, tirando partido dos sensores Wi-Fi embutidos e utilizando uma aplicação própria para o efeito.

Os dados utilizados ao longo deste capítulo foram recolhidos por um período de 91 dias consecutivos. Durante esse tempo, esperava-se que o utilizador não alterasse a sua rotina diária, isto é, que agisse naturalmente como se a aplicação não existisse. Assim, garantia-se uma recolha de dados de mobilidade não forçada e representativa dos hábitos de mobilidade do utilizador.

De modo a validar a qualidade dos resultados obtidos pelo algoritmo desenvolvido com o que realmente acontecia, seria necessário a recolha de *ground truth*. Cada utilizador tinha como tarefa, manter um diário de todos os locais que visitava por tempo igual ou superior a 5 minutos juntamente com os tempos de entrada e saída (Tabela 5.1). Como não é fácil manter um diário atualizado, só foi possível a obtenção de *ground truth* de dois utilizadores (A e B) durante um período de 15 dias.

O utilizador A vive no centro de uma grande cidade e trabalha no campus universitário localizado numa outra cidade, de tamanho médio, localizada a 23 km de distância. O utilizador desloca-se de carro entre os dois locais. Os seus movimentos, para além da deslocação entre estes dois locais, foram realizados em redor destes locais. Neste período de recolha o utilizador alterou

Tabela 5.1: Exemplo da organização de um diário produzido por um utilizador.

Local	Entrada	Saída
Casa	01-03-2012 18:56	02-03-2012 08:30
LID4	02-03-2012 09:10	02-03-2012 10:05
Bar Engenharia	02-03-2012 10:07	02-03-2012 10:25
LID4	02-03-2012 10:27	02-03-2012 12:45
Cantina	02-03-2012 12:57	02-03-2012 13:35
Bar Engenharia	02-03-2012 13:45	02-03-2012 13:55
LID4	02-03-2012 13:58	02-03-2012 16:45
Bar Engenharia	02-03-2012 16:47	02-03-2012 17:05
LID4	02-03-2012 17:08	02-03-2012 18:25
Casa	02-03-2012 18:55	02-03-2012 19:15
Ginásio	02-03-2012 19:25	02-03-2012 20:45
Casa	02-03-2012 21:55	03-03-2012 08:35

de residência durante uma semana, voltando depois à residência habitual.

O utilizador B vive numa cidade de tamanho médio e trabalha no campus universitário situado na mesma cidade. Os seus movimentos são essencialmente dentro da cidade, existindo por vezes visitas em trabalho a outra grande cidade.

Sobre os dados em brutos recolhidos pela aplicação, apesar de serem gravados para um ficheiro em formato JSON e posteriormente processados de acordo com as transformações descritas nos capítulos anteriores, é importante salientar que todos os registos foram processados pela ordem de recolha, não havendo conhecimento dos registos seguintes. Desta forma, apesar do processo não estar implementado para funcionar em tempo-real, os resultados obtidos seriam exatamente os mesmos se o processamento tivesse sido feito em tempo-real.

Tabela 5.2: Níveis hierárquicos obtidos com uma partição inicial de N nodos e E arestas (Utilizador A - 91 dias).

	Dimensão do Grafo		Modularidade
	Nodos	Arestas	
1ª Abordagem			
Nível 0	4382	30334	0.581104
Nível 1	610	1313	0.590774
Nível 2	417	352	0.591056
Nível 3	397	270	0.591059
Nível 4	396	267	0.591059
2ª Abordagem			
Nível 0	4382	30334	0.532429
Nível 1	561	1016	0.538496
Nível 2	417	354	0.538885
Nível 3	398	277	0.538892
Nível 4	396	269	0.538892
Duração Total		0 segundos	

5.2 Descoberta de Locais

Após a transformação dos dados em bruto para observações e da extração do grafo ponderado a partir das observações, ocorre a etapa nuclear de todo o processo: a deteção das comunidades com recurso ao método de Louvain.

Independentemente da abordagem utilizada na extração do grafo, o método de Louvain fornece sempre uma decomposição em vários níveis hierárquicos. A Tabela 5.2 mostra os níveis obtidos e respetivo número de nodos, arestas e modularidade atingida, para as duas abordagens.

O nível 0 representa a partição inicial, isto é, o grafo ponderado que deu

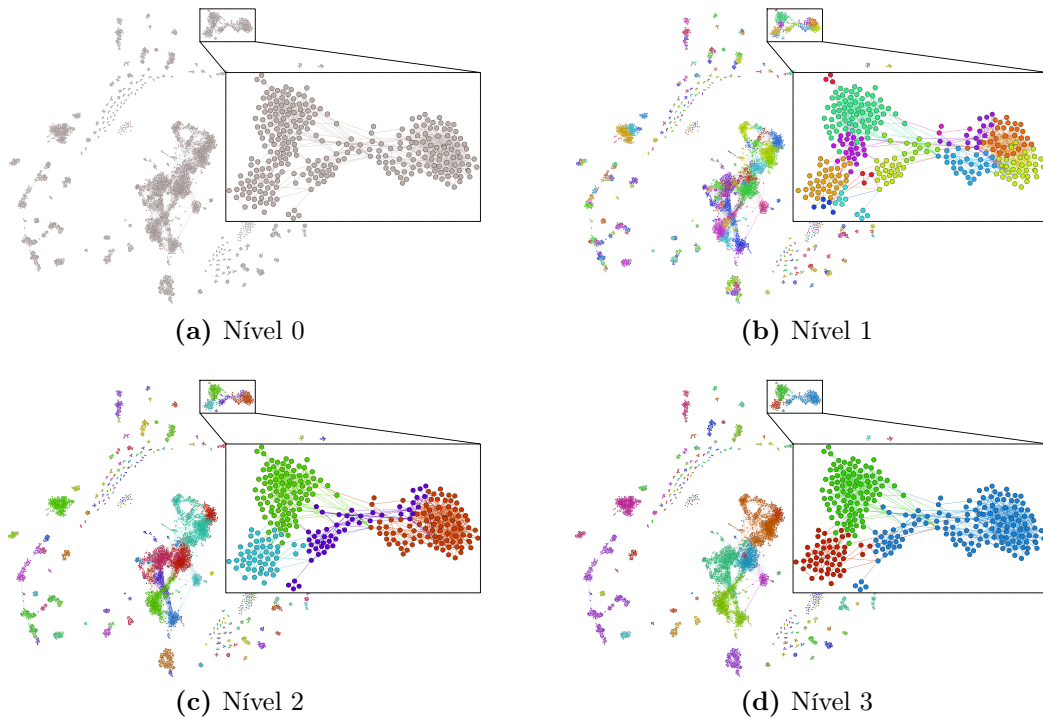


Figura 5.1: Representação das observações em forma de grafo. Cada cor representa uma comunidade (1^a abordagem). Assim que o nível da hierarquia vai aumentando, as comunidades tendem a ficar cada vez mais agregadas (Utilizador A - 91 dias).

entrada. Nesse nível inicial cada nodo é tratado como uma comunidade. À medida que se vai subindo na hierarquia, o número de comunidades vai diminuindo, devido à existência de uma maior agregação nos níveis superiores. A Figura 5.1 mostra a evolução do processo de agregação de comunidades ao longo da hierarquia. Quanto maior é o nível hierárquico, maior é a sua agregação em comunidades. Como não existe uma grande diferença no número de comunidades detetadas entre o nível 3 e o nível 4, este último acabou por não ser incluído na figura.

Para além da abordagem a escolher, interessa sobretudo determinar o nível que apresenta a melhor decomposição em comunidades. Como descrito na subsecção 3.2.2, uma maior modularidade não significa propriamente uma melhor decomposição. Portanto, para tomar a decisão da melhor abordagem

e nível hierárquico, é necessário verificar qual dos níveis apresenta uma melhor eficácia na deteção de locais.

5.3 Avaliação

De modo a avaliar quantitativamente a eficácia do algoritmo, para cada um dos níveis hierárquicos, foi adotada uma abordagem de validação utilizada por outros autores (D. Kim et al., 2010) (Figura 5.2). Os autores definiram um conjunto de locais significativos e errados, onde os locais registados no diário são denominados de locais recordados e os locais visitados detetados pelo algoritmo são chamados de locais descobertos.

Fazem parte dos locais recordados os locais esquecidos, que são registados no diário mas não foram detetados pelo algoritmo. Os locais falsos e interessantes fazem parte dos locais descobertos. Se um local não foi registado no diário mas é descoberto pelo algoritmo e o utilizador confirma que realmente visitou esse local mas esqueceu-se de o registar, então esse local é classificado de interessante. Caso contrário, é considerado como um falso local.

A interseção dos locais recordados com os descobertos contém os locais que foram anotados no diário e posteriormente detetados com sucesso pelo algoritmo e são categorizados como locais corretos, unidos e divididos. Um local unido refere-se a dois locais distintos, no diário, que foram descobertos como apenas um único local. Se um único local registado no diário é detetado como dois ou mais locais, então é classificado como dividido.

Utilizando esta classificação de locais, o algoritmo de deteção de locais desenvolvido é avaliado em termos de *precision* e *recall*. A *precision* é definida pela percentagem de locais descobertos que realmente foram visitados. A *recall* é a percentagem de locais recordados que realmente foram descobertos corretamente pelo algoritmo. Estas duas métricas são definidas como:

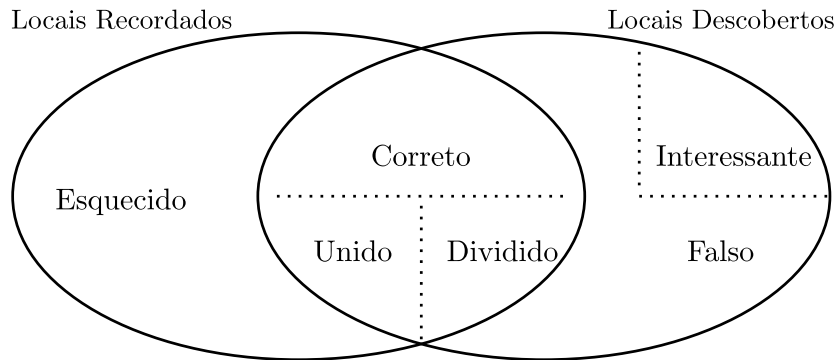


Figura 5.2: Locais recordados pelo utilizador no diário e locais visitados descobertos pelo algoritmo de deteção. Uma maior existência de locais corretos e interessantes indicam um melhor desempenho do algoritmo na deteção de locais (adaptado de D. Kim et al. (2010)).

$$Precision = \frac{\#Corretos + \#Interessantes}{\#Descobertos}, Recall = \frac{\#Corretos}{\#Recordados}$$

O desempenho do algoritmo de deteção de locais desenvolvido é avaliado pela utilização das métricas acima definidas. Realçando novamente, no contexto das métricas utilizadas, os locais presentes nos diários, definidos na secção 5.1, são os locais recordados, e os locais visitados detetados pelo algoritmo (classificados pela aplicação do primeiro filtro) são os locais descobertos.

A decisão de avaliar os locais que foram visitados e não os locais potenciais, deveu-se sobretudo ao diário realizado pelos utilizadores. Foi pedido que anotassem somente os locais que permanecessem pelo menos durante 5 minutos. Os locais potenciais incluem todos os locais por onde o utilizador passou e que pode não ter realmente permanecido. Como a solução final, os MMP, baseia-se somente nos locais visitados e relevantes, não interessa validar os locais potenciais.

As Tabelas 5.3 e 5.4 mostram os resultados obtidos para as métricas com base nos dados recolhidos pelo algoritmo e aos diários realizados por dois

Tabela 5.3: Avaliação quantitativa dos locais descobertos a partir de dados recolhidos pelo utilizador A.

	Utilizador A							
	1 ^a Abordagem				2 ^a Abordagem			
	1	2	3	4	1	2	3	4
Locais								
Correto	16	10	10	10	16	10	9	9
Interessante	1	1	1	1	1	1	1	1
Unido	1	8	8	8	0	8	9	9
Dividido	1	0	0	0	1	0	0	0
Esquecido	0	0	0	0	1	0	0	0
Falso	0	0	0	0	0	0	0	0
<i>Recall</i>	0.89	0.56	0.56	0.56	0.89	0.56	0.50	0.50
<i>Precision</i>	0.89	0.58	0.58	0.58	0.94	0.58	0.53	0.53

utilizadores (A e B) durante um período de 15 dias.

Pela observação da Tabela 5.3, os resultados do primeiro nível, das duas abordagens, são bastante superiores aos obtidos para os restantes níveis. Para a primeira abordagem o algoritmo recorda corretamente 89% dos locais visitados com igual valor de *precision*. Na segunda abordagem, verifica-se uma ligeira diferença entre os valores da *precision* para o nível 1 comparativamente ao mesmo nível da primeira abordagem. Esta diferença aconteceu porque na primeira abordagem um determinado local estava unido com outro, enquanto que na segunda esse local não foi detetado (i.e., esquecido pelo algoritmo). Assim, para a segunda abordagem, o algoritmo foi capaz de recordar 89% dos locais visitados com uma *precision* de 94%.

Com valores idênticos, a partir dos dados recolhidos pelo utilizador B, os

Tabela 5.4: Avaliação quantitativa dos locais descobertos a partir de dados recolhidos pelo utilizador B.

	Utilizador B							
	1 ^a Abordagem				2 ^a Abordagem			
	1	2	3	4	1	2	3	4
Locais								
Correto	12	10	10	10	12	9	9	–
Interessante	3	2	2	2	3	1	1	–
Unido	1	3	3	3	1	4	4	–
Dividido	0	0	0	0	0	0	0	–
Esquecido	0	0	0	0	0	0	0	–
Falso	0	0	0	0	0	0	0	–
<i>Recall</i>	0.92	0.77	0.77	0.77	0.92	0.69	0.69	–
<i>Precision</i>	0.94	0.80	0.80	0.80	0.94	0.71	0.71	–

resultados do primeiro nível das duas abordagens são bastante superiores dos obtidos para os restantes níveis. Segundo a Tabela 5.4, independentemente da abordagem utilizada, os resultados obtidos para a *recall* e *precision* com o primeiro nível hierárquicos são superiores aos restantes níveis. Para o nível 1, o algoritmo desenvolvido recordou 92% dos locais visitados com uma *precision* de 94%, utilizando qualquer uma das duas abordagens.

Para os dois utilizadores, independentemente da abordagem utilizada, de todos os níveis existentes, o nível 1 é o que apresenta melhores resultados. Contudo, apesar de já ser conhecido qual o nível que apresenta melhores resultados, ainda não é possível afirmar qual das abordagens é superior, visto que os valores obtidos são idênticos.

Como os resultados sobre os dados dos locais não foram suficientes, foram

Tabela 5.5: Resultados por estadias em locais.

	B		A	
	1 ^a Abord.	2 ^a Abord.	1 ^a Abord.	2 ^a Abord.
Estadias				
Diário		59	105	
Algoritmo	59	59	105	104
Falhas	0	0	0	1
Falsos Positivos	3	3	1	1

analisadas as estadias detetadas nesses locais em cada uma das abordagens para o nível 1 (Tabela 5.6). Para o utilizador A, com a primeira abordagem o algoritmo foi capaz de detetar todas as estadias. Na segunda abordagem, uma estadia presente no diário não foi corretamente detetada pelo algoritmo. Para ambas as abordagens foi detetado um falso positivo. Em relação ao utilizador B, das 59 estadias registadas no diário, todas elas foram detetadas pelo algoritmo. Foram detetados três falsos positivos, que dizem respeito às estadias nos três locais interessantes. Neste caso, os resultados são os mesmos para as duas abordagens.

Como é possível verificar, também nas estadias, não existe uma diferença relevante entre as duas abordagens. Por isso, não é ainda possível concluir qual das abordagens é melhor. A última análise passará pela observação da diferença dos tempos de entrada e saída em locais pela comparação das estadias detetadas pelo algoritmo com o *ground truth* recolhido.

Para os dois utilizadores foram então sobrepostas as estadias reais com as estadias estimadas para um período de 15 dias. A Figura 5.3 pretende mostrar essa sobreposição utilizando os dados recolhidos pelo utilizador B. No eixo dos Y estão todos os locais corretamente detetados pelo algoritmo e

Tabela 5.6: Resultados por estadias em locais.

	A		B	
	1ª Abord.	2ª Abord.	1ª Abord.	2ª Abord.
Estadias				
Diário	105		59	
Algoritmo	105	104	59	59
Falhas	0	1	0	0
Falsos Positivos	1	1	3	3

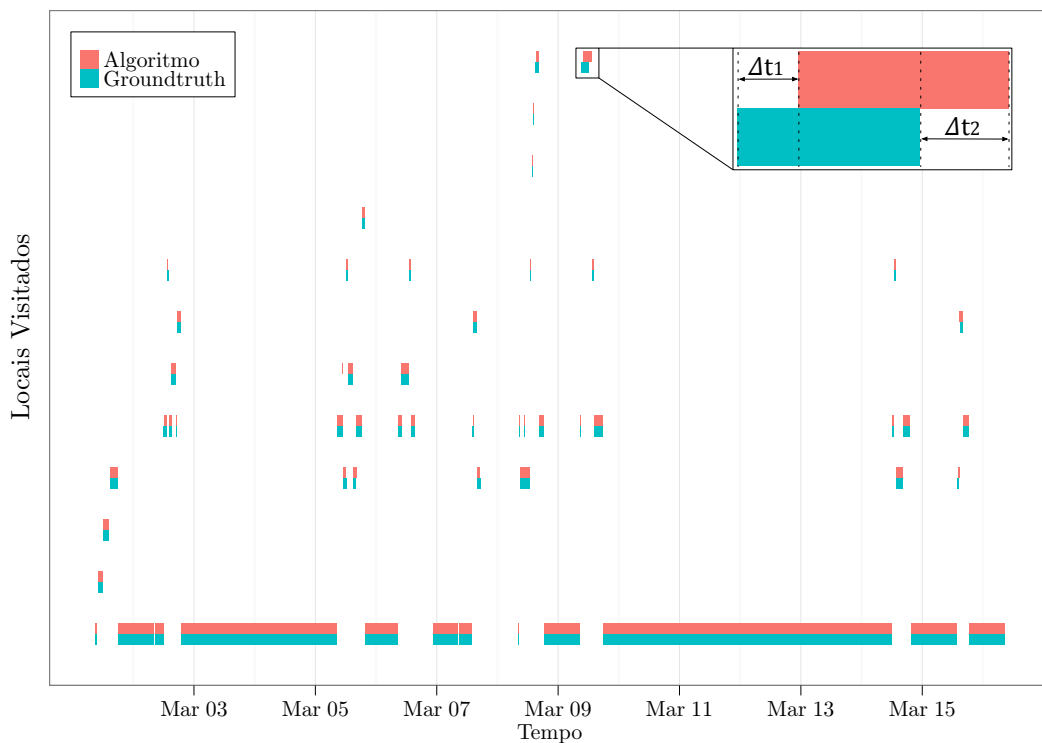


Figura 5.3: Estadias detetadas pelo algoritmo vs. Estadias reais presentes no diário (Utilizador B). A vermelho estão representadas as estadias provenientes do algoritmo e a azul as estadias anotadas pelo utilizador no diário. A diferença entre elas permite conhecer os atrasos ou avanços temporais na deteção de locais.

o eixo dos X corresponde ao tempo. O destaque presente na figura realça a existência de uma disparidade temporal no tempo de entrada (Δt_1) e saída (Δt_2) entre os dados do algoritmo e do *ground truth*. O Δt_1 é negativo se o algoritmo detetar a entrada num local antes do tempo anotado no diário. Se o algoritmo detetar com adiantamento relativamente ao diário, então o Δt_1 é positivo. No caso do Δt_2 , é negativo se o algoritmo detetar a saída de um local com um atraso temporal relativamente ao diário. Quando o algoritmo deteta a saída após o tempo anotado no diário, então o Δt_2 é positivo.

O conjunto dos valores de Δt extraídos é utilizado para estimar a raiz do erro quadrático médio (RMSE¹). A RMSE permite conhecer com que rigor, em média, é estimado o instante de entrada e saída de um local.

Para o utilizador A, para a primeira abordagem, foi estimado um RMSE de 3.24 minutos para a entrada e um RMSE de 2.76 minutos para a saída. Para a segunda abordagem, estimou-se um RMSE de 3.23 minutos para a entrada e um RMSE de 2.80 minutos para a saída. Em relação ao utilizador B, para a primeira abordagem, foi estimado um RMSE de 2.26 minutos para a entrada e um RMSE de 2.98 minutos para a saída. Utilizando a segunda abordagem, estimou-se um RMSE de 3.33 minutos para a entrada e um RMSE de 2.80 minutos para a saída.

Calculou-se ainda a probabilidade do algoritmo detetar uma entrada e/ou saída dentro de um intervalo de tempo definido. Como Δt_1 e Δt_2 são variáveis aleatórias contínuas, a sua estatística pode ser descrita pela função densidade de probabilidade (FDP).

Dada a FDP, é possível calcular a probabilidade de ser detetar a entrada e a saída de um local dentro de um dado intervalo de tempo. Inicialmente considerou-se o intervalo $[-3, 3]$, centrado no instante de tempo real.

Para o utilizador A, para a primeira abordagem foi obtida uma probabili-

¹Do inglês, *root-mean-square error*.

dade de 67% na detecção de entradas e 78% na detecção de saídas. Na segunda abordagem, a probabilidade de detetar uma entrada foi de 66% e 75% na detecção de saídas. Em relação ao utilizador B, e utilizando os dados da primeira abordagem, a probabilidade de detetar uma entrada dentro do intervalo definido foi de 85%, e de 71% na detecção de saídas. Utilizando os dados da segunda abordagem, verificou-se que não houve alterações nas probabilidades de detecção.

Os resultados das probabilidades obtidas estão diretamente ligados com a qualidade do *ground truth* recolhido. A tarefa de realizar um diário de todos os locais visitados torna-se desconfortável e desagradável após alguns dias. Esta situação leva os utilizadores a esquecerem-se de anotar as visitas atempadamente. Um exemplo frequente ocorre quando um utilizador entra num local e esquece-se de registar no diário. Passado uns largos minutos, o utilizador lembra-se que tem de fazer o registo, e coloca uma hora de entrada mais ou menos aproximada. Ao longo do tempo, estes desvios vão sendo cada vez mais frequentes, originando um *ground truth* de baixa qualidade e por conseguinte conduz a baixas probabilidades na detecção de entradas e saídas.

De modo a atenuar o problema que é realizar um diário, foi aplicada uma tolerância de 2 minutos ao intervalo anterior. Sendo assim, foram calculados novos valores de probabilidade com um erro temporal de 5 minutos. Novamente para o utilizador A, utilizando a primeira abordagem, existe uma probabilidade de 86% de detetar uma entrada e 89% de detetar a saída de um local dentro do intervalo definido. Para a segunda abordagem, a probabilidade de entrada e saída é de 87% para ambas. No que diz respeito ao utilizador B, independentemente da abordagem utilizada, a probabilidade de detetar uma entrada é de 95%, e de 86% para as saídas.

Mais uma vez, a diferença entre os resultados obtidos para cada uma das abordagens é mínima. Deste modo, a Figura 5.4 inclui apenas os gráficos

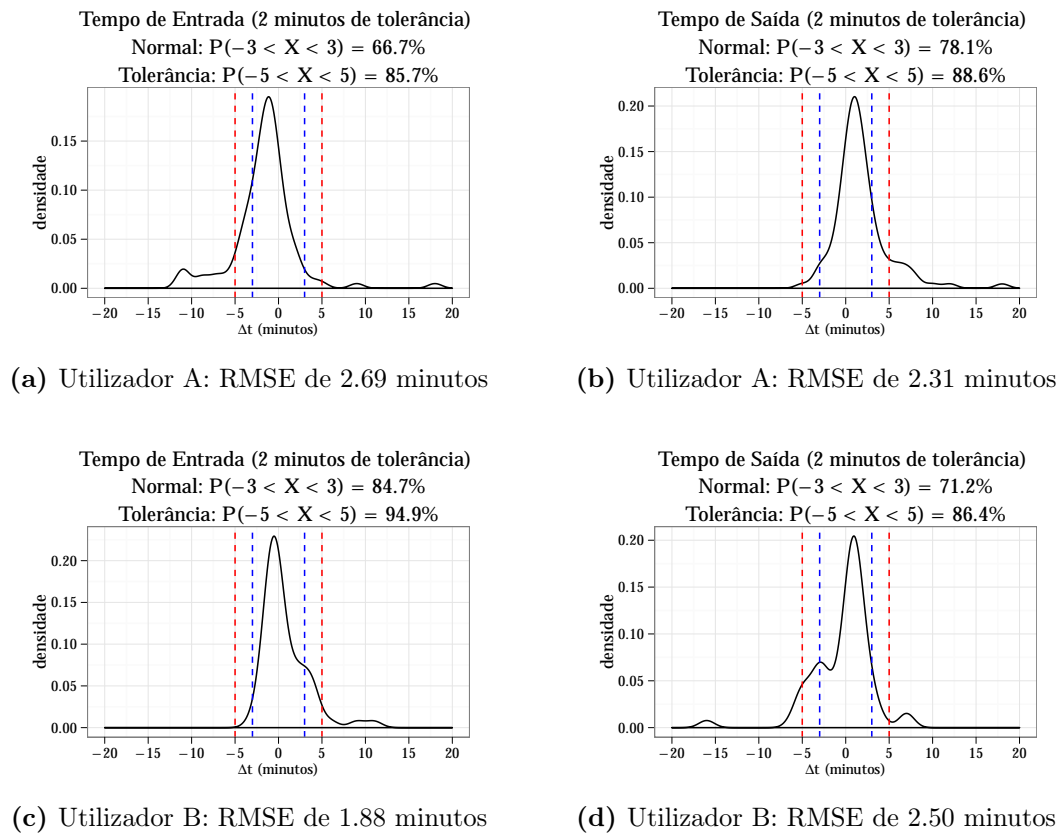


Figura 5.4: Função densidade de probabilidade da diferença de tempos de entrada e saída resultante da comparação dos dados produzidos pelo algoritmo com os registados no diário pelo utilizador (1^a abordagem).

da FDP, ou da curva da densidade, para a primeira abordagem. As linhas verticais representam os intervalos definidos: a azul o erro temporal de 3 minutos e a vermelho o de 5 minutos.

A avaliação do algoritmo através da utilização das métricas de *precision* e *recall*, das estadias pela comparação das anotadas no diário com as descobertas pelo algoritmo, bem como a probabilidade de detetar as entradas e saídas em locais, não foram suficientes para mostrar que a utilização de uma abordagem em detrimento da outra originaria um melhor desempenho do algoritmo de deteção de locais. Isto é, para o nível 1 e utilizando qualquer uma das duas abordagens, os resultados finais são praticamente os mesmos. Deste modo, como a diferença entre as duas abordagens está no número de arestas do grafo ponderado, a utilização da abordagem 1, como possui um menor número de arestas, requer um menor poder computacional, um menor espaço de armazenamento, bem como um menor tempo de computação do processo de deteção de locais.

5.3.1 Confronto com outros Algoritmos

De modo a avaliar a eficácia do algoritmo na deteção de locais, foram comparados os resultados obtidos, pela utilização das métricas de *precision* e *recall*, com outros algoritmos do mesmo género. Os resultados dos outros algoritmos, SensLoc (D. Kim et al., 2010), PlaceSense (D. Kim et al., 2009) e Kang et al. (2005), são provenientes de uma avaliação realizada por D. Kim et al. (2010). O SensLoc, em média, obteve uma *precision* de 94% com igual *recall*. O PlaceSense obteve 89% de *precision* e 84% de *recall*. Por fim, Kang et al. (2005) teve uma *precision* de 38% com uma *recall* de 42%.

Comparando com os resultados obtidos pelo algoritmo desenvolvido, em média uma *precision* de 92% e uma *recall* de 91% para a primeira abordagem e 94% de *precision* e 91% de *recall* com a segunda abordagem, apenas o

SensLoc foi superior. Contudo, as condições com que os resultados foram obtidos são um pouco diferentes das do algoritmo desenvolvido no âmbito desta dissertação: (i) o SensLoc foi avaliado com apenas uma semana de dados, (ii) a frequência de recolha foi definida e era sempre constante ao longo do período de recolha e (iii) a utilização de funções de similaridade que têm como base as variações da força do sinal Wi-Fi ao longo do tempo. No caso do algoritmo aqui desenvolvido, (i) foram avaliados os dados relativos a 2 semanas, (ii) a aplicação MOVE possuía uma frequência de recolha aleatória dentro do intervalo de 5 a 50 segundos e (iii) foi utilizada uma abordagem mais simples, recorrendo à força do sinal Wi-Fi apenas na construção do grafo ponderado.

Pela análise das diferenças entre os dois algoritmos, o SensLoc apresenta condições mais favoráveis para a descoberta de locais com base em dados Wi-Fi, mas exige um maior esforço e tempo de processamento. Avaliar uma semana é diferente de avaliar duas semanas. Uma frequência de recolha regular garante por si só um conjunto de dados de boa qualidade com observações constantes ao longo do tempo. O cálculo de uma função de similaridade para comparar cada observação com outra exige muito mais em termos computacionais e temporais do que a identificação do AP com maior sinal de uma observação em conjunto com a deteção de comunidades.

É certo que o SensLoc avaliou um conjunto de cinco diários. Mas pelas diferenças observadas, e caso as condições fossem as mesmas para os dois algoritmos, provavelmente os resultados obtidos, tanto para o SensLoc como para o algoritmo desenvolvido, poderiam ser bem diferentes e o julgamento dos resultados da avaliação pela utilização das métricas seria mais justo.

5.4 Classificação de Locais

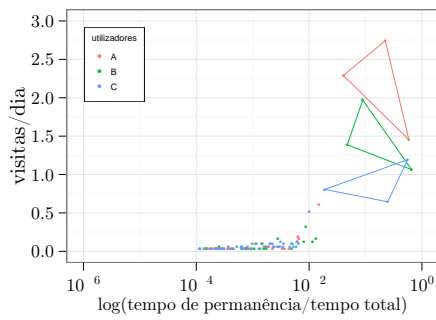
A classificação dos locais potenciais em visitados, é realizada apenas para obter apenas os locais realmente visitados pelo utilizador, e remover os locais onde existiu apenas uma passagem. Esta classificação é somente um passo intermédio para a classificação dos locais relevantes, e não apresenta resultados importantes de serem apresentados.

A classificação dos locais visitados em relevantes envolve a análise das estadias do utilizador em cada um dos locais visitados. Utilizando os dados recolhidos pelos voluntários, durante um período de 3 meses, foram definidos os dois critérios que alimentam o segundo filtro de classificação dos locais: número de visitas por dia maior que um determinado *threshold* e percentagem de permanência maior que um determinado *threshold*.

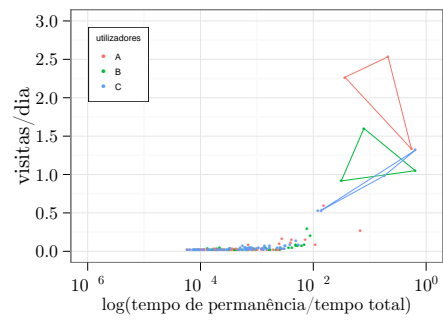
Contudo, antes de realizar a classificação foi necessário descobrir um *threshold* mínimo para cada um dos critérios. Esse *threshold* é a fronteira que permite distinguir os locais relevantes dos locais visitados.

De modo a descobrir a referida fronteira, observou-se a distribuição dos locais visitados segundo os dois critérios estabelecidos. A Figura 5.5 mostra essa distribuição. Cada utilizador tem uma cor associada e cada ponto do gráfico representa um local visitado. No eixo dos X está representada a percentagem de tempo total de estadia num local, e no eixo dos Y a frequência de visitas a um local (usando um dia como unidade de tempo).

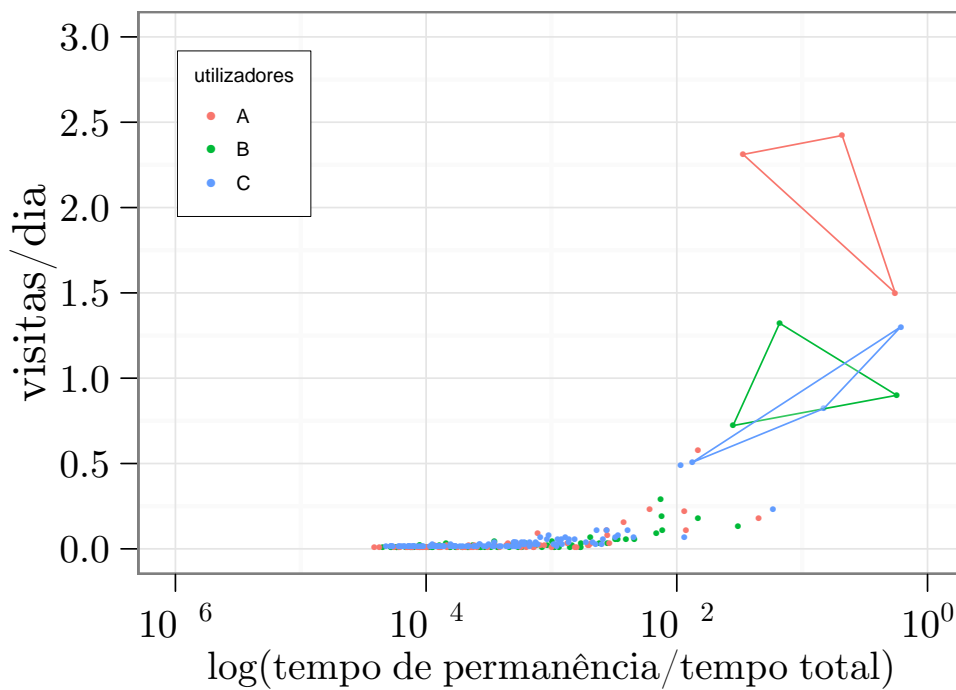
Analisando a distribuição dos locais de cada utilizador ao longo dos três meses, é observável a existência de três locais que se destacam dos restantes. Isto significa que, embora os utilizadores visitem um grande número de locais no seu dia-a-dia, na verdade eles tendem a retornar frequentemente aos mesmos três locais. Todos eles situam-se acima das 0.5 visitas por dia. Isto é, os utilizadores visitam estes locais pelo menos uma vez em cada dois dias. No caso do utilizador C, apenas no primeiro mês de dados é que é possível



(a) 1 mês de dados



(b) 2 meses de dados



(c) 3 meses de dados

Figura 5.5: A forma dos locais mais relevantes. Os resultados dos gráficos das três figuras foram obtidos a partir da acumulação dos dados de 1 mês (a), de 2 meses (b) e por fim de 3 meses (c).

observável a existência de três locais que se destacam. Ao fim de três meses, o utilizador C apresenta apenas dois locais que se destacam dos restantes.

Outro fato sobre as 0.5 visitas por dia é que os locais acima das 0.5 visitadas por dia são os mesmos ao longo do tempo. No final dos três meses (Figura 5.5c) é observada uma clara diferença, isto é um vazio, entre os locais acima e abaixo das 0.5 visitas por dia.

Este vazio já era esperado, visto que, ao longo do tempo, os locais que não fazem parte da rotina diária tendem a mover-se para longe do vazio que separa os dois tipos de locais. A questão fundamental estava em descobrir essa fronteira, que pela análise dos dados de três utilizadores, situa-se por volta de uma visita a cada dois dias (0.5).

De acordo com a Figura 5.5, a união dos três locais que mais se destacam forma um triângulo. Esta forma geométrica permite ter uma visão geral acerca dos hábitos de mobilidade do utilizador. Se a forma não variar muito, significa que o utilizador continua a frequentar regularmente esses locais com uma permanência constante ao longo dos meses. Se a forma mudar de aspeto, significa que o utilizador deixou de frequentar com a mesma regularidade os mesmos três locais, i.e., o utilizador começou a passar mais tempo num dos locais e menos noutro, ou alterou a frequência com que os visita.

Pela análise da Figura 5.5, ao longo dos três meses, a forma do utilizador A varia muito pouco. É o típico indivíduo que apresenta uma alta regularidade na sua mobilidade diária (i.e., visita e permanece com a mesma frequência e regularidade os mesmos locais). O utilizador B apresenta uma forma bastante regular, excepto entre o primeiro e o segundo mês. Contudo, entre o segundo e o terceiro mês, o seu padrão de mobilidade estabilizou pois a sua forma não sofreu grandes alterações. Em relação ao utilizador C é muito idêntico ao utilizador B. A sua forma varia entre o primeiro e o segundo mês, estabilizando a partir desse momento. A frequência de visitas a um local nem sempre apre-

senta uma certa regularidade ao longo do tempo. Contudo, segundo o eixo dos X, para todos os utilizadores, a percentagem de permanência é praticamente a mesma ao longo dos três meses. Deste modo, apesar dos utilizadores não apresentarem uma regularidade no número de visitas realizadas, estes apresentam uma regularidade no tempo em que permanecem em cada um dos locais acima das 0.5 visitas por dia.

Até ao momento, ainda não foi possível descobrir a fronteira relativamente à percentagem de tempo de permanência que distingue locais relevantes dos locais visitados. Os resultados atuais sugerem um valor, contudo são necessários mais dados e mais utilizadores de modo a realizar uma validação correta. No caso do eixo dos Y, apesar do valor 0.5 visitas por dia ser também provisório, nota-se claramente um vazio a partir desse *threshold*. De modo a fortalecer a escolha deste valor, foi analisada mais pormenorizadamente a distribuição dos locais para cada utilizador. A Figura 5.6 mostra a variação dos locais visitados ao longo dos três meses para o utilizador A, o mais regular dos três existentes. Pela observação da figura, nota-se claramente que, desde o terceiro dia que o vazio entre os locais já é identificável. Em relação à distribuição dos locais, também a partir do terceiro dia de recolha (Figura 5.6a) até ao último mês (Figura 5.6f), o número de locais acima do *threshold* é constante ao longo do tempo, sejam eles os mesmos locais ou não.

Como a criação dos MMP está diretamente ligado com a classificação dos locais, será utilizado o *threshold* de 0.5 visitas por dia como o único critério que permite diferenciar os locais relevantes dos restantes locais visitados. Assim, todos os locais que atinjam esse valor serão classificados como locais relevantes.

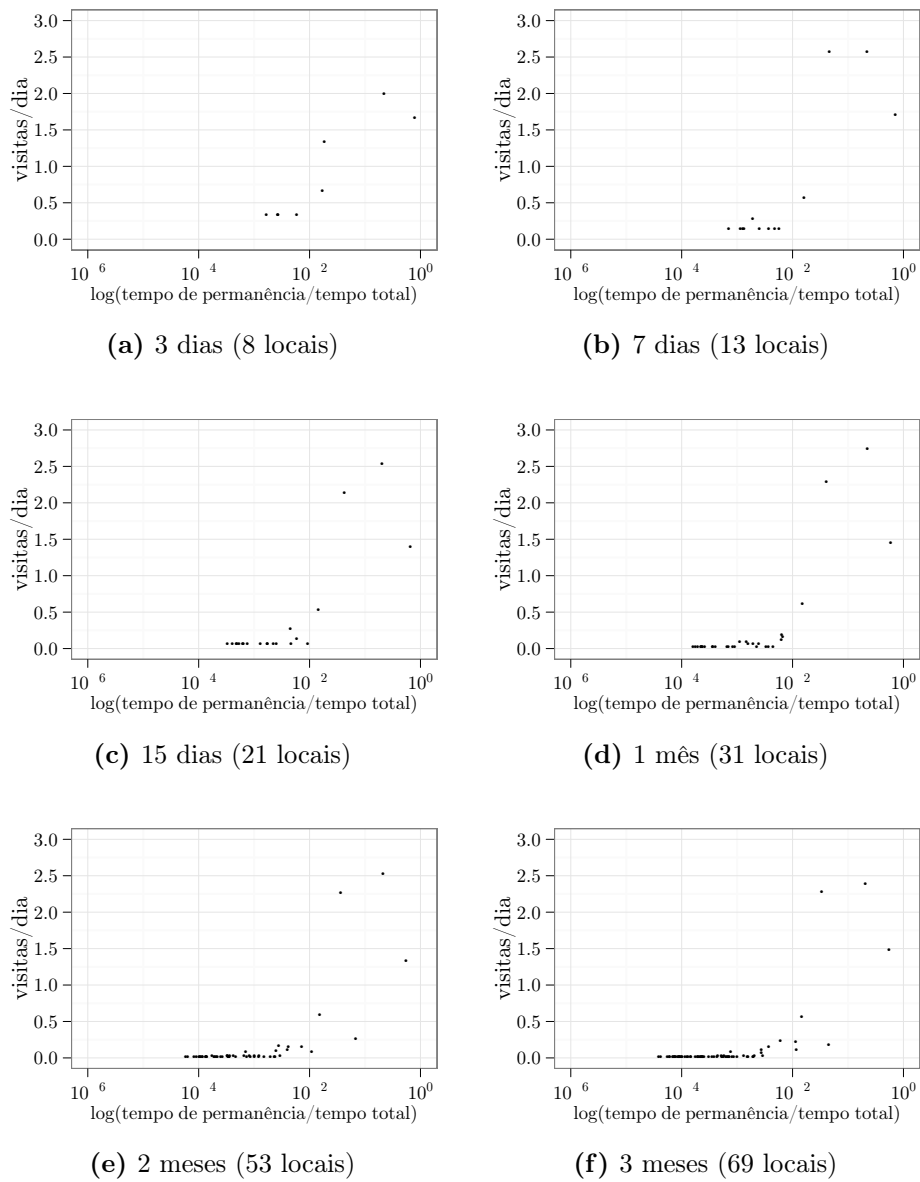


Figura 5.6: Distribuição dos locais visitados ao longo de 3 meses (Utilizador A).

Tabela 5.7: Caracterização temporal dos locais relevantes (Utilizador A).

Local ID	Visitas / Dia	Tempo Total de Permanência (%)	Tempo de Permanência Médio / Visita (min)
6	2.62	21	113.64
7	3.14	3	15.43
10	0.64	1	32.92
16	1.55	54	507.94

5.5 Mapas de Mobilidade Pessoal

Após ser atingida a última camada do processo de descoberta de locais, é possível finalmente reproduzir a mobilidade dos utilizadores através da construção dos MMP. Seguindo a mesma linha de orientação da secção anterior, utilizaram-se os dados recolhidos por três utilizadores ao longo de três meses para desenvolver e enriquecer os MMP.

A Figura 5.7 apresenta os MMP obtidos para os três utilizadores. Do lado esquerdo estão os mapas ainda sem a aplicação do segundo filtro de classificação dos locais (locais visitados) e do lado direito os mapas com a classificação já realizada (locais relevantes).

Cada nodo e aresta do grafo é caracterizado temporalmente. Para cada um deles foi obtido um conjunto de informação caracterizadora descrita na secção 3.2.5. As Tabelas 5.7 e 5.8 apresentam a informação caracterizadora do utilizador A obtida para o MMP dos locais relevantes e respetivas trajetórias. Utilizando cada um dos atributos obtidos, os MMP deixam de ser somente um conjunto de nodos e arestas, e passam a ser caracterizados de acordo com o atributo desejado.

No caso dos MMP presentes na Figura 5.7, os nodos de maior diâmetro são aqueles com maior percentagem de tempo de permanência. Os que apresentam

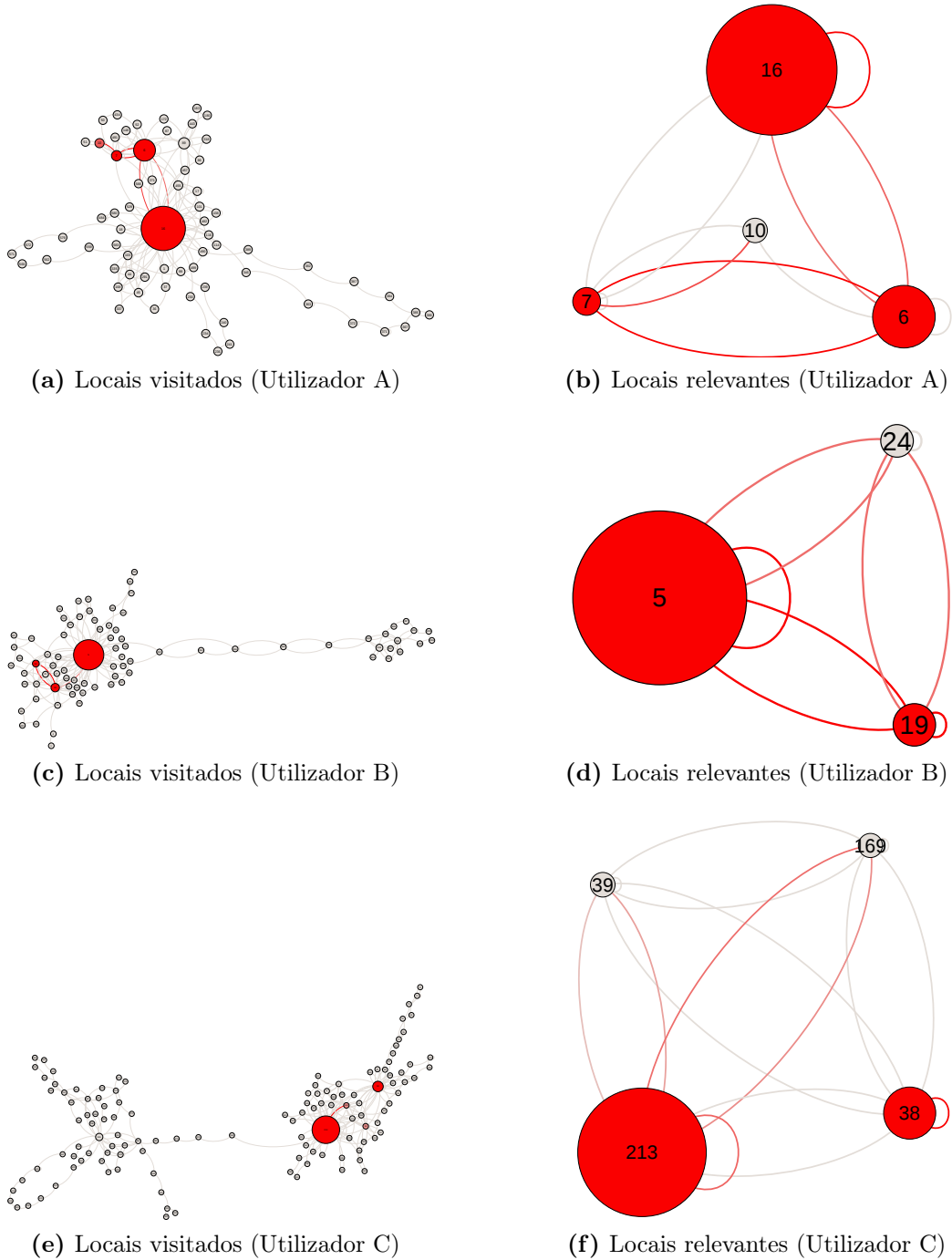


Figura 5.7: Mapas de Mobilidade Pessoal dos locais visitados (esquerda) e dos locais relevantes (direita) dos três utilizadores.

Tabela 5.8: Caracterização temporal das transições mais frequentes (Utilizador A).

Origem	Destino	Frequência (dias)	Tempo de Viagem Médio (min)
6	7	1.48	2.05
6	16	0.51	64.49
6	6	0.26	89.31
6	10	0.22	13.38
7	6	1.74	2.93
7	10	0.36	1.16
7	7	0.26	9.14
7	16	0.21	58.29
16	16	0.76	148.69
16	6	0.46	135.84
16	7	0.25	83.53
10	7	0.58	7.08

uma cor avermelhada representam os locais com maior visitas por dia. Por outro lado, os nodos cinza são os locais com menos visitas. De notar, que os nodos cinza dos MMP dos locais relevantes têm pelo menos 0.5 visitas por dia e a cor cinza representa um número de visitas por dia baixo em comparação com os outros locais relevantes. Quanto às arestas, são representadas por uma curvatura que segue a orientação dos ponteiros do relógio para mostrar a direção, evitando-se assim o uso de setas nas pontas das arestas (Fekete, Wang, Dang, Aris, & Plaisant, 2003). As mais avermelhadas são aquelas com uma frequência diária aceitável (por exemplo, uma viagem entre os dois locais uma vez em cada dois dias). As arestas cinza representam viagens que acontecem esporadicamente, assumindo assim pouca relevância na mobilidade do utilizador.

Estas diferenças entre os MMP dos locais visitados e dos locais relevantes permitem perceber o que foi dito na secção da classificação dos locais. Pela observação dos mapas dos locais visitados, ao longo dos três meses, cada utilizador visitou diversos locais. Contudo, só alguns deles é que o utilizador volta a visitar com frequência e/ou permanece por grandes períodos de tempo. Esses locais, denominados de locais relevantes, são os que realmente interessa analisar em termos de mobilidade.

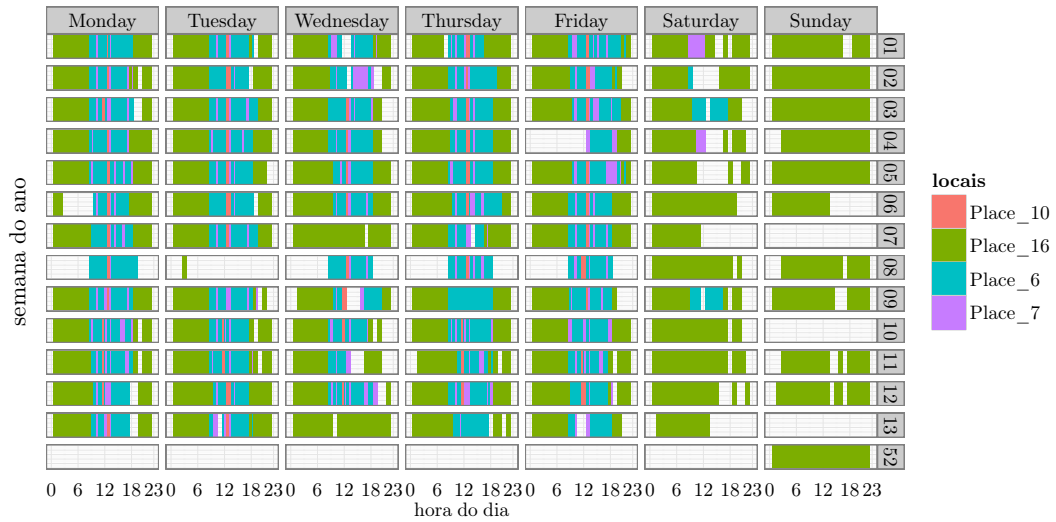
Para além do conjunto de nodos e arestas, e da informação que os caracteriza, o MMP tem também dois tipos de perfis temporais associados. Um deles retrata a ordem e regularidade nas visitas aos locais. Este perfil espelha a rotina do utilizador ao longo da sua vida diária. O segundo perfil temporal, mostra de forma agregada, qual a hora do dia em que o utilizador é visto com mais frequência num determinado local.

A Figura 5.8 mostra os dois tipos de perfis temporais do utilizador A. Através do primeiro perfil (Figura 5.8a) é possível confirmar a regularidade da forma obtida na classificação dos locais (Figura 5.6), e visualizar que o

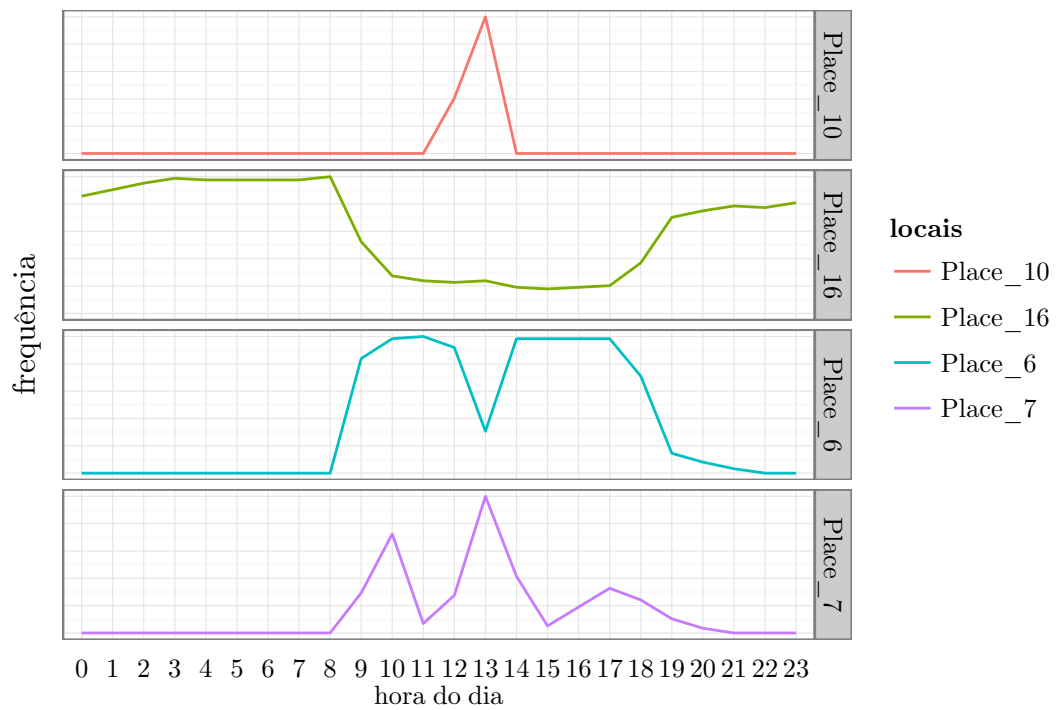
utilizador A é bastante regular nas visitas e no tempo que dispende nos locais relevantes.

Quanto ao segundo perfil temporal (Figura 5.8b), é observável a existência de locais que são visitados em diferentes períodos de tempo. No local a vermelho, *Place_10*, o utilizador é visto com maior frequência por volta das 13 horas. A verde, no *Place_16*, é perceptível que o utilizador passa a maior parte do seu tempo apenas durante a noite: sai por volta das 8 horas e regressa por volta das 18/19 horas. No local azul, *Place_6*, é visto com regularidade ao longo de todo o dia, exceto do período das 12 às 14 horas. Por fim, no local a roxo, *Place_7*, o utilizador visita-o em diferentes períodos do dia, com maior incidência por volta das 13 horas.

Pode ainda ser realizada uma análise semelhante no que respeita às transições entre locais. Como por exemplo, conhecer a ordem e regularidade das viagens realizadas, e a hora do dia em que essas viagens são mais frequentes.



(a) Perfil temporal retratando a ordem e regularidade nas visitas aos locais



(b) Perfil temporal, retratando de forma agregada, qual a hora do dia em que o utilizador é visto com mais frequência num determinado local

Figura 5.8: Perfis temporais do utilizador A nos locais relevantes ao longo de três meses.

Capítulo 6

Conclusões e Trabalho Futuro

Ao longo deste último capítulo são apresentadas as principais conclusões do trabalho realizado. É realizada uma reflexão sobre os objetivos e implicações dos resultados obtidos. Por fim são apresentadas algumas considerações relativamente ao trabalho futuro.

6.1 Conclusões do Projeto

O trabalho realizado resultou no desenvolvimento de um algoritmo de descoberta de locais com base em dados sobre o ambiente Wi-Fi que rodeia o utilizador, recolhidos pelos sensores embutidos nos *smartphones*. O resultado do algoritmo permitiu a criação de modelos espaciais simbólicos, associados a cada utilizador, denominados de Mapas de Mobilidade Pessoal.

O contexto é mais do que a simples informação recolhida pelos sensores que permite caraterizar a situação de um indivíduo. Novas dimensões do contexto podem ser obtidas, seja pelo processo dos dados provenientes dos sensores ou pelo processamento de outras dimensões já inferidas. O processamento dos sinais emitidos pelos APs Wi-Fi permite a descoberta dos locais relevantes para o utilizador que caraterizam a dimensão do contexto sobre a localização

(neste caso simbólica, visto não existir qualquer referência a coordenadas geográficas). A utilização da informação temporal, por exemplo o conjunto dos tempo de entrada e saída nos diversos locais visitados, permite a criação de um histórico da informação de contexto obtida. Ordenada temporalmente, esta informação permite inferir os hábitos de mobilidade do indivíduo (padrões de mobilidade e perfis temporais) e prever contextos futuros (por exemplo, a probabilidade de um utilizador voltar a um local a uma determinada hora do dia).

Em termos da avaliação pelas métricas utilizadas, pode afirmar-se que o algoritmo desenvolvido está ao nível de outros algoritmos de descoberta de locais reconhecidos pela comunidade científica. Em média, o algoritmo permite a deteção de 92% dos locais que foram realmente visitados pelos utilizadores (*precision*). O algoritmo apresenta ainda, em média, uma probabilidade de deteção de 91% dos locais anotados no diário (*recall*). Apesar de só terem sido avaliados dois utilizadores, estes resultados podem ser considerados excelentes dadas as condições dos dados recolhidos e pela não utilização de qualquer conhecimento prévio da topologia da rede.

O processo de validação demonstrou a eficácia do algoritmo desenvolvido. Para além da excelente capacidade de detetar locais, também a maior parte das estadias em locais foi detetada com sucesso, sendo o algoritmo capaz de detetar atempadamente (intervalo de 5 minutos) a chegada a um local com uma probabilidade média de 90%, e 88% na deteção atempada das saídas de locais dentro do mesmo intervalo de tempo.

Apesar do algoritmo desenvolvido apenas detetar locais onde exista um ambiente Wi-Fi que o permita caracterizar contextualmente, o processo de validação mostrou que todos os locais anotados no diário foram detetados com êxito, sejam eles corretos, unidos ou divididos. Sendo assim, o que parecia ser ao início uma limitação, não deve ser considerada como tal. Este fato vem

reforçar a ubiquidade das redes Wi-Fi.

Os Mapas de Mobilidade Pessoal criados permitem a representação e caracterização dos locais relevantes no cotidiano do utilizador. No âmbito do PPS6, a construção destes mapas foram o início de todo um esforço que culminará nos relatórios de eficiência.

Ao longo da investigação realizada nesta dissertação, foi redigido, submetido, aceite e apresentado um artigo científico numa conferência internacional (Pérez-Penichet, Conde, & Moreira, 2012). O trabalho presente no artigo está relacionado com o descrito nesta dissertação, e integrado no mesmo projeto (PPS6).

6.2 Trabalho Futuro

A realização deste trabalho de investigação deixa muitas portas em aberto. A partir do estado atual do trabalho desenvolvido, para além da otimização dos resultados alcançados, existe um sem fim de investigações futuras, nomeadamente na área dos estudos da mobilidade humana.

O estado atual do trabalho desenvolvido pode ainda beneficiar de muitas melhorias. Um dos avanços mais imediatos que poderiam ser introduzidos diz respeito à descoberta do *threshold* que permite a distinção entre os locais visitados e os relevantes. Apesar do *threshold* de classificação definido ser apenas uma hipótese, a própria significância deste valor permite considerá-lo como aceitável (uma visita a cada dois dias). Para este propósito, a análise de dados provenientes de um largo conjunto de utilizadores poderá ser útil para confirmar, ou descobrir, o *threshold* ideal, bem como a introdução de outras variáveis para além das duas já em uso.

Uma das variáveis em uso, designada por percentagem do tempo de permanência num local, não foi devidamente explorada. Apesar de existirem

hipóteses, são necessários mais estudos, com mais utilizadores, de modo a descobrir a fronteira ideal para a classificação dos locais.

A totalidade da análise da classificação dos locais, dos padrões de mobilidade e perfis temporais, constitui a base para validar modelos de mobilidade existentes, nomeadamente aqueles que determinem a probabilidade do utilizador voltar a um local.

Novas dimensões do contexto podem ser utilizadas para enriquecer os atuais MMP, nomeadamente pela adição de contexto espacial utilizando os Mapas de Proximidade (Pérez-Penichet et al., 2012).

Devido ao tempo reduzido para a realização desta dissertação, ficou por realizar o teste de avaliação de outros algoritmos de deteção de comunidades. Apesar dos autores do método de Louvain afirmarem que o seu algoritmo é superior aos outros, o objetivo do teste a realizar é visto como uma tentativa de vir a melhorar ainda mais as partições das comunidades obtidas.

No domínio no qual esta dissertação se encontra inserida, a privacidade é uma questão delicada que levanta sempre grandes problemas. Contudo, no trabalho futuro que tirará partido dos MMP, todo o processamento de informação sensível será realizado nos *smartphones*, o que reduz significativamente quaisquer problemas relacionados com a privacidade.

Referências

- Abrahamse, W., Steg, L., Vlek, C., & Rothengatter, T. (2005). A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, *25*(3), 273–291.
- Ahmad, U., d’Auriol, B., Lee, Y., & Lee, S. (2007). The election algorithm for semantically meaningful location-awareness. In *Proceedings of the 6th international conference on mobile and ubiquitous multimedia* (p. 55–63). New York, NY, USA: ACM.
- Albert, R., & Barabási, A.-L. (2002, Jan). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, *74*, 47–97.
- Ampt, E., & Gleave, S. (2004). Understanding voluntary travel behaviour change. *Transport Engineering in Australia*, *9*, 53–66.
- Ashbrook, D., & Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, *7*(5), 275–286.
- Azizyan, M., Constandache, I., & Roy Choudhury, R. (2009). SurroundSense: mobile phone localization via ambience fingerprinting. In *Proceedings of the 15th annual international conference on mobile computing and networking* (p. 261–272). New York, NY, USA: ACM.
- Bahl, P., & Padmanabhan, V. (2000). RADAR: an in-building RF-based user location and tracking system. In *IEEE INFOCOM 2000. nineteenth annual joint conference of the IEEE computer and communications*

- societies. proceedings* (Vol. 2, p. 775–784). IEEE.
- Bahl, P., Padmanabhan, V., & Balachandran, A. (2000). *Enhancements to the RADAR user location and tracking system* (Technical Report MSR-TR-2000-12). Redmond, WA, USA: Microsoft Research.
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008, October). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
- Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075), 462–465.
- Brown, P., Bovey, J., & Chen, X. (1997). Context-aware applications: from the laboratory to the marketplace. *Personal Communications, IEEE*, 4(5), 58–64.
- Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., & Barabási, A. (2008, June). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22), 1–11.
- Chon, Y., & Cha, H. (2011). LifeMap: a Smartphone-Based context provider for Location-Based services. *Pervasive Computing, IEEE*, 10(2), 58–67.
- Chon, Y., Shin, H., Talipov, E., & Cha, H. (2012). Evaluating mobility models for temporal prediction with high-granularity mobility data. In *Pervasive computing and communications (percom), 2012 IEEE international conference on* (pp. 206–212).
- Chon, Y., Talipov, E., & Cha, H. (2012). Autonomous management of everyday places for a personalized location provider. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(4), 518–531.

- Chon, Y., Talipov, E., Shin, H., & Cha, H. (2010). *Mobility prediction-based smartphone energy optimization for everyday location monitoring* (Tech. Rep. MOBED-TR-2010-3). Seoul, South Korea: Yonsei University.
- Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111.
- Constandache, I., Choudhury, R., & Rhee, I. (2010). Towards mobile phone localization without war-driving. In *Infocom, 2010 proceedings ieee* (pp. 1–9).
- Crockford, D. (2006). *The application/json media type for javascript object notation (json)*. Internet informational RFC 4627. Disponível em <https://tools.ietf.org/html/rfc4627>
- Dey, A. (2001). Understanding and using context. *Personal and ubiquitous computing*, *5*(1), 4–7.
- Dey, A., Abowd, G., & Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction*, *16*(2-4), 97–166.
- Eagle, N., & Pentland, A. (2009). Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, *63*(7), 1057–1066.
- Farrahi, K., & Gatica-Perez, D. (2008). What did you do today?: discovering daily routines from large-scale mobile data. In *Proceeding of the 16th ACM international conference on multimedia* (p. 849–852). New York, NY, USA: ACM.
- Farrahi, K., & Gatica-Perez, D. (2009). Learning and predicting multimodal daily life patterns from cell phones. In *Proceedings of the 2009 international conference on multimodal interfaces* (p. 277–280). New York, NY, USA: ACM.
- Fekete, J., Wang, D., Dang, N., Aris, A., & Plaisant, C. (2003). Overlaying graph links on treemaps. In *Ieee symposium on information visualization*

- conference compendium (demonstration)* (Vol. 5).
- Fischer, C. (2008). Feedback on household electricity consumption: a tool for saving energy? *Energy Efficiency*, 1(1), 79–104.
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3–5), 75–174.
- Fortunato, S., & Castellano, C. (2007). Community structure in graphs. *Arxiv preprint arxiv:0712.2716*.
- Gil, G., Jesus, A. de, & Lopez, J. (2011). inContexto: a fusion architecture to obtain mobile context. In *Proceedings of the 14th international conference on information fusion (FUSION)* (p. 1–8).
- Gonzalez, M., Hidalgo, C., & Barabasi, A. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196), 779–782.
- Hariharan, R., & Toyama, K. (2004). Project lachesis: parsing and modeling location histories. *Geographic Information Science*, 3234, 106–124.
- Hightower, J., Consolvo, S., LaMarca, A., Smith, I., & Hughes, J. (2005). Learning and recognizing the places we go. *UbiComp 2005: Ubiquitous Computing*, 3660, 159–176.
- Hightower, J., LaMarca, A., & Smith, I. (2006). Practical lessons from place lab. *Pervasive Computing, IEEE*, 5(3), 32–39.
- IEEE. (2002). IEEE standard for local and metropolitan area networks: Overview and architecture (IEEE std 802-2001). *The Institute of Electrical and Electronics Engineers, Inc.*
- Jiang, Y., Li, D., Yang, G., Lv, Q., & Liu, Z. (2011). Deliberation for intuition: a framework for energy-efficient trip detection on cellular phones. In *Proceedings of the 13th international conference on ubiquitous computing* (p. 315–324). New York, NY, USA: ACM.
- Kang, J., Welbourne, W., Stewart, B., & Borriello, G. (2005). Extracting places from traces of locations. *ACM SIGMOBILE Mobile Computing*

and Communications Review, 9(3), 58–68.

- Kim, D., Hightower, J., Govindan, R., & Estrin, D. (2009). Discovering semantically meaningful places from pervasive RF-beacons. In *Proceedings of the 11th international conference on ubiquitous computing* (p. 21–30). New York, NY, USA: ACM.
- Kim, D., Kim, Y., Estrin, D., & Srivastava, M. (2010). Sensloc: sensing everyday places and paths using less energy. In *Proceedings of the 8th ACM conference on embedded networked sensor systems* (p. 43–56). New York, NY, USA: ACM.
- Kim, M., Kotz, D., & Kim, S. (2006). Extracting a mobility model from real user traces. In *Proceedings of the 25th IEEE international conference on computer communications* (pp. 1–13).
- Klepeis, N., Nelson, W., Ott, W., Robinson, J., Tsang, A., Switzer, P., et al. (2001). The national human activity pattern survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Analysis and Environmental Epidemiology*, 11(3), 231–252.
- Krumm, J., & Horvitz, E. (2004). Locadio: Inferring motion and location from wi-fi signal strengths. In *First annual international conference on mobile and ubiquitous systems: Networking and services (MobiQuitous 2004)* (p. 4–13).
- Laasonen, K., Raento, M., & Toivonen, H. (2004). Adaptive on-device location recognition. *Pervasive Computing*, 3001, 287–304.
- LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., et al. (2005). Place lab: Device positioning using radio beacons in the wild. *Pervasive Computing*, 3468, 301–306.
- Lu, H., Pan, W., Lane, N., Choudhury, T., & Campbell, A. (2009). SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on mobile*

- systems, applications, and services* (p. 165–178). New York, NY, USA: ACM.
- Maibach, E., et al. (2003). Recreating communities to support active living: a new role for social marketing. *American Journal of Health Promotion*, 18(1), 114–119.
- Manson, N. (2006). Is operations research really research. *Orion*, 22(2), 155–180.
- March, S., & Smith, G. (1995). Design and natural science research on information technology. *Decision support systems*, 15(4), 251–266.
- Marmasse, N., & Schmandt, C. (2000). Location-aware information delivery with commotion. *Handheld and Ubiquitous Computing, 1927*, 361–370.
- Moreira, A., & Santos, M. Y. (2005). From GPS tracks to context - inference of high-level context information through spatial clustering. Estoril, Portugal: GIS Planet.
- Peixoto, J., & Moreira, A. (2012). Dealing with multiple source spatio-temporal data in urban dynamics analysis. In B. Murgante et al. (Eds.), *Computational science and its applications - iccsa 2012* (Vol. 7334, p. 450-465). Springer Berlin / Heidelberg.
- Pérez-Penichet, C., Conde, Â., & Moreira, A. (2012). Human mobility analysis by collaborative radio landscape observation. In *International conference on geographic information science* (pp. 153–158). Digital Editions.
- PNT, N. C. O., & Committee, C. G. S. I. (2012). *Global positioning system*. Disponível em <http://www.gps.gov/>
- Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In p. Yolum, T. Güngör, F. Gürgen, & C. Özturan (Eds.), *Computer and information sciences - ISCIS 2005* (Vol. 3733, pp. 284–293). Springer Berlin / Heidelberg.
- Raghunath, S., Malleswari, B., Sridhar, K., et al. (2011). Analysis of GPS

- errors during different times in a day. *International Journal Of Research in Computer Science*, 2(1), 45–48.
- Recordon, D., & Reed, D. (2006). OpenID 2.0: a platform for user-centric identity management. In *Proceedings of the second ACM workshop on digital identity management* (p. 11–16). New York, NY, USA: ACM.
- Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S., & Chong, S. (2011). On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)*, 19(3), 630–643.
- Ryan, N., Pascoe, J., & Morse, D. (1998). Enhanced reality fieldwork: the context-aware archaeological assistant. In *Computer applications in archaeology*. Oxford, UK: British Archaeological Reports.
- Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. In *First workshop on mobile computing systems and applications (WMCSA '94)* (p. 85–90).
- Schilit, B., & Theimer, M. (1994). Disseminating active map information to mobile hosts. *Network, IEEE*, 8(5), 22–32.
- Song, C., Qu, Z., Blumm, N., & Barabasi, A. (2010). Limits of predictability in human mobility. *Science*, 327(5968), 1018–1021.
- Song, L., Kotz, D., Jain, R., & He, X. (2006). Evaluating Next-Cell predictors with extensive Wi-Fi mobility data. *IEEE Transactions on Mobile Computing*, 5(12), 1633–1649.
- Takeda, H., Veerkamp, P., & Yoshikawa, H. (1990). Modeling design process. *AI magazine*, 11(4), 37.
- Vaishnavi, V., & Kuechler, W. (2007). *Design research in information systems*. Disponível em <http://desrist.org/design-research-in-information-systems/>
- Wakita, K., & Tsurumi, T. (2007). Finding community structure in mega-scale social networks. In *IADIS international conference on WWW/Internet*

2007 (pp. 153–162).

Weiser, M. (1991). The computer for the 21st century. *Scientific American*, 265(3), 94–104.

Zimmermann, A., Lorenz, A., & Oppermann, R. (2007). An operational definition of context. In *Proceedings of the 6th international and interdisciplinary conference on modeling and using context* (Vol. 4635, p. 558–571). Berlin, Heidelberg: Springer-Verlag.