Mestrado em Informática e Sistemas

Análise de dados e *Machine Learning* na Mobilidade Urbana

Trabalho de projeto apresentado para a obtenção do grau de Mestre em Informática e Sistemas
Especialização em Desenvolvimento de Software

Autor

João Pedro Fernandes Simões

Orientador

Doutora Ana Cristina da Costa Oliveira Alves
Professora Adjunta do Departamento de Informática e Sistemas
Instituto Superior de Engenharia de Coimbra

Doutor Rui Jorge Reis Gomes
Professor Auxiliar Convidado do
Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologias da Universidade de Coimbra

Coimbra, Abril, 2019

AGRADECIMENTOS

Em primeiro, gostaria de agradecer à professora Doutora Ana Alves e ao professor Doutor Rui Gomes, pelos conhecimentos, apoio e disponibilidade constantes ao longo de todo o projeto.

Aos meus pais e a todos os colegas e amigos que encontrei ao longo de todo o percurso.

Ao meu primo e amigo Nelson, pela ajuda que sempre me deu e pelo incentivo constante, desde sempre.

Aos meus colegas e amigos de trabalho, pela ajuda no trabalho diário, amizade e pela flexibilidade existente na conciliação do projeto de mestrado com o trabalho na empresa.

Dedico esta tese de mestrado aos meus avós, pelos valores que sempre me quiseram transmitir, pelo apoio, pela coragem e humildade que sempre tiveram ao longo das suas vidas.

RESUMO

A mobilidade tornou-se num dos desafios mais difíceis que as cidades têm de enfrentar. Mais de metade da população mundial reside em áreas urbanas e com o contínuo aumento da população é imperativo que as cidades usem os seus recursos de forma eficiente. Exige-se por isso, que cada vez mais, a gestão e o planeamento da oferta de transportes, tenha de ser realizada de uma forma racional e eficaz de modo a satisfazer as necessidades dos cidadãos. Obter e reunir dados a partir de diferentes fontes de dados pode ser extretamente importante para apoiar novas soluções que podem ajudar a construir uma melhor mobilidade. O crowdsensing tornou-se uma conhecida forma de partilhar dados extraídos por dispositivos, que capturam dados através dos seus sensores, como o *smartphone* com o objetivo de atingir um bem comum. Nesta tese de mestrado é proposta uma metodologia que analisa dos dados extraídos, identifica áreas de maior procura e as possíveis razões pra este fenómeno. Esta metodologia pretende auxiliar o melhoramento da gestão e oferta da rede de transportes de uma dada cidade em estudo, neste caso a área metropolitana do Porto, considerando dados recolhidos da utilização da técnica de crowdsensing.

Palavras-chave: mobilidade, sistemas ubíquos, *clustering*, *crowdsensing*, planeamento de transportes, *sensor systems*, sistemas de transportes públicos, computação urbana.

ABSTRACT

Mobility has become one of the most difficult challenges that cities must face. More than half of world's population resides in urban areas and with the continuously growing population it is imperative that cities use their resources more efficiently. This requires, that more and more, the management and the planning of the network offer, must be realized in a rational and effectively way, to meet citizens needs. Obtaining and gathering data from different sources can be extremely important to support new solutions that will help building a better mobility for the citizens. The crowdsensing has become a popular way to share data collected by devices that capture the data using sensors, like smartphone with a goal to achieve a common interest. In this master thesis is proposed a methodology which analyzes the extracted data, identifies areas of greater demand and possible reasons for this phenomenon. This methodology is intended to help the improvement of the management and supply of the transport network of the metropolitan area of Porto, considering data collected from the use of crowdsensing technique.

Keywords: mobility, ubiquitous systems, clustering, crowdsensing, transportation planning, sensor systems, public transport systems, urban computing.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

ACRÓNIMOS

API - Interface de Programação de Aplicações

BIRCH- *Balanced Iterative Reducing and Clustering using Hierarchies*

CLARA- *Clustering LARge Applications*

CLARANS- *Clustering Large Applications based on RANdomized Search*

CLIQUE- *Clustering in QUEst*

CSV - *Comma-Separated Values*

CURE- *Clustering Using REpresentatives*

DBSCAN - *Density-based spatial clustering of applications with noise*

DENCLUE - *DENsity-based CLUstEring*

FEUP – Faculdade de Engenharia da Universidade do Porto.

GIS – *Geographic Information System* é um sistema informação geográfica que engloba o hardware, software, informação espacial, procedimentos computacionais e recursos humanos que permite e facilita a análise, gestão ou representação do espaço e dos fenómenos que nele ocorrem.

GPS – *Global Positioning System* é um sistema de posicionamento por satélite. Este sistema permite que qualquer pessoa, na detenção de um dispositivo com esta capacidade consiga saber a sua localização em formato de coordenadas de latitude, longitude e altitude.

HDSBCAN *- *Density-Based Clustering Based on Hierarchical Density*

JSON - *JavaScript Object Notation*

OPTICS - *Ordering Points to Identify the Clustering Structure*

PAM – Particionamento entre *medoids*

POI – *Point Of Interest* é um ponto numa localização que alguém considera útil ou interessante. Termo usado em cartografia com aplicações em sistemas de navegação, GIS e GPS.

NSE- *National Science Experiment*

STING- *Statistical Information Grid-based*

STCP – Sociedade de Transportes Colectivos do Porto

SVM – *Support Vector Machine*

1 INTRODUÇÃO

A mobilidade tem-se tornado num dos maiores desafios que as cidades enfrentam nos dias de hoje, cada vez mais, influenciada por fatores que dificultam a circulação (congestionamento de tráfego, informação não atualizada, eventos, acidentes, supressão de vias, obras, entre outros). Mais de metade da população, cerca de 54% reside nas áreas urbanas e em 2050 é previsto que este número alcance os 67% [1]. Todo este crescimento cria um conjunto de vários desafios para os quais as cidades necessitam estar preparadas. A necessidade constante de melhoramento das infraestruturas e da adaptação da rede de transportes capazes de responder às necessidades dos cidadãos é hoje uma realidade. O aumento da poluição, que advém do excesso de veículos em circulação, através das emissões de dióxido de carbono para a atmosfera é por si só, um fator preocupante. A venda de automóveis tem continuado a crescer tendo-se como estimativa um aumento de 125 milhões de vendas em 2025 sendo que alguns analistas indicam que tal aumento poderá mesmo duplicar no ano de 2030 [2]. Segundo a Organização Mundial de Saúde, já mesmo no ano de 2014, sete milhões de mortes prematuras ocorreram devido à poluição do ar e uma parte significativa dessa causa foi diretamente relacionada com o trânsito urbano [3].

A falta de serviços integrados de mobilidade leva a que haja uma menor utilização dos transportes coletivos, a um aumento do trânsito nas áreas urbanas e, consequentemente, a congestionamentos, maior poluição e diminuição da eficiência energética. São então cada vez mais necessários sistemas de apoio à decisão relativamente a planeamento urbano para melhorar os serviços de mobilidade disponíveis, detetar padrões de utilização e procura de serviços a fim de sugerir futuras localizações de novos recursos e instalações face às necessidades. Pretende-se que o planeamento realizado pela rede de sistemas de transporte seja eficiente e consiga corresponder às necessidades dos cidadãos. Entender por isso, determinados fatores e padrões que motivam as pessoas a realizar determinadas viagens é fulcral para conseguir oferecer uma oferta inteligente aos cidadãos. Perceber por exemplo, em cenários mais complexos, como os horários fora da rotina (à noite e aos fins-de-semana), que regiões e que linhas devem ser reforçadas é essencial para se conseguir oferecer aos cidadãos uma alternativa útil.

Neste momento, a possibilidade de ter mais informação atualizada, melhor integração e uma maior coordenação dos transportes é fulcral para uma melhor mobilidade, inclusiva, inteligente e ajustada às necessidades dos utentes. Considera-se que, a mobilidade do futuro terá de ter em conta diferentes fontes de dados, com diversos níveis de granularidade e frequência de atualização adequados. Por esse motivo, consideramos que um cenário de mobilidade urbana sustentada requer a integração de várias fontes de dados heterogéneas para fornecer informação acerca do trânsito, condições ambientais, sentimentos das pessoas e suas opiniões.

Para analisar a mobilidade urbana é necessário recolher informação do maior número de indivíduos possível, e uma das formas mais fáceis de o fazer é recorrer à utilização de sensores dos *smartphones* capazes de recolher informação georreferenciada. Devido ao aumento exponencial dos *smartphones*, o estudo da mobilidade humana tem mudado significativamente. Estamos cada vez mais dependentes deles e hoje é impensável pensar como podemos viver sem eles.

As pessoas tratam os *smartphones* como uma "segunda pele" e estão com eles em contacto permanente [4]. Atualmente, cerca de 68% da população mundial tem um smartphone e no ano de 2019 é expectável que este número se situe na região dos 72% [5]. Acompanhando este crescimento, os próprios sensores do *smartphone* têm se tornado uma boa fonte de dados para analisar o comportamento humano. Com eles podemos perceber onde estamos, o que estamos a fazer e em alguns casos mesmo conseguirmos inferir o porquê de estarmos a fazer algo. Se se pretende estudar a mobilidade humana, uma opção válida para encontrar padrões e fornecer soluções capazes de melhorar a mobilidade, poderá ser a utilização dos *smartphones* para captação de dados relativos à mobilidade. Quem tem acompanhando este crescimento é também a técnica do *mobile crowdsensing*. Esta técnica baseia-se na existência de um grupo de indivíduos, que possuem um *smartphone* ou um dispositivo capaz de extrair dados, que através dos seus sensores partilham dados armazenados com o objetivo de atingir um interesse em comum [6]. Um exemplo prático desta técnica é a utilização do sistema de navegação *Waze*[1], em que as pessoas partilham dados sobre o trânsito com o objetivo de todos terem informação em tempo real sobre a melhor rota para o seu destino.

Através dos dados recolhidos pela aplicação *SenseMyFeup* pretende-se, com este trabalho, tentar perceber se existe correlação entre os destinos das viagens que as pessoas reportaram na aplicação, na zona metropolitana do Porto, com a existência de eventos e pontos de interesse (POIs) que ocorram fora da rotina diária. Esta correlação permitirá identificar quais as zonas que a Sociedade de Transportes Coletivos do Porto (STCP) terá de reforçar, quando ocorrem eventos em horário fora da rotina (à noite e aos fins-de-semana), numa determinada zona da cidade do Porto, ou determinadas áreas que possuem POIs relevantes, que são visitados por muitas pessoas e necessitam de ter melhor reforço de linhas. O *SenseMyFeup* é uma aplicação de *mobile crowdsensing* que se dedica à extração de dados relativos à mobilidade do utilizador através dos sensores do *smartphone*, tais como posição GPS (*Global Positioning System*), acelerómetro, giroscópio, entre outros.

[1] https://www.waze.com/pt-PT/

1.1 Motivação

O projeto URBY.SENSE (Referência P2020-PTDC/ECM-TRA/6803/2014[2]) teve início em junho de 2016 e tem como principal objetivo extrair padrões de mobilidade fora da rotina (de lazer, sociais, etc.) a partir de múltiplas fontes de dados através da recolha, fusão e análise destes dados. Os padrões que se pretende estabelecer são diversos, tais como locais de interesse, modos de transporte, rotas comuns e atividades baseadas em localização. No projeto URBY.SENSE foram elaboradas soluções de *crowdsensing*, que permitiram obter dados geográficos sobre viagens, que os cidadãos realizaram na zona metropolitana da cidade do Porto.

No contexto do projeto URBY.SENSE, insere-se este projeto de mestrado, que tem como objetivo a análise da informação georreferenciada extraída, através da aplicação *SenseMyFeup*. A análise a efetuar durante a elaboração deste trabalho permitirá identificar na zona metropolitana do Porto, as zonas que contém maior concentração de POIs, as áreas de maior procura na zona metropolitana do Porto monitorizadas pela aplicação *SenseMyFeup*, bem como a existência de zonas com um maior número de eventos. Com a identificação destas áreas, os operadores de transportes poderão planear melhor a oferta de modo a corresponder às necessidades e eventuais lacunas existentes.

Tipicamente, no planeamento de transportes os dados utilizados provêm, na sua maioria, de métodos tradicionais, como questionários e censos, que para além de dispendiosos, são morosos e necessitam da participação ativa, fornecendo às entidades responsáveis um mero retrato histórico da mobilidade. A utilização massificada de dispositivos interativos de computação (telemóveis, cartões inteligentes, dispositivos GPS, câmaras digitais, etc.) e os registos dos sistemas de transporte (por exemplo, a contagem de validação de bilhetes) fornecem 'pegadas digitais' sem precedentes, revelando onde estão os utilizadores e quando, permitindo traçar dinamicamente o perfil de mobilidade [7].

Entender a forma como o espaço nas cidades é usado é um dos pilares da gestão e planeamento urbano. Por isso, dados relativos a POIs são importantes para definir uma estratégia de planeamento e gestão de recursos da cidade. Dados semânticos sobre os POIs em áreas urbanas são difíceis de visualizar e nesse contexto foi implementada a plataforma *web*, *CityClusters* [8], que tem como objetivo facilitar a visualização deste tipo de dados. Com a realização deste projeto de mestrado, pretende-se também desenvolver uma ferramenta de *software* que permita a partir de dados georreferenciados sobre os POIs existentes de uma dada cidade estimar e criar diferentes camadas de *clusters* para visualização e alimentar a plataforma *CityClusters*. Pretende-se testar esta ferramenta com dados geográficos sobre os POIs existentes na área metropolitana da cidade do Porto.

[2] https://www.cisuc.uc.pt/projects/show/217 (acedido em 2018/08/10)

1.2 Objetivos

Com a realização deste trabalho pretende-se responder às seguintes questões, bem como atingir os seguintes objetivos:

- Verificar a existência de uma correlação entre os destinos das viagens reportadas na aplicação do *SenseMyFeup* e a ocorrência de eventos na zona metropolitana do Porto;
- Verificar a existência de uma correlação entre os destinos das viagens reportadas na aplicação do *SenseMyFeup* e a existência de POIs na zona metropolitana do Porto;
- Fornecer dados geográficos sobre os POIs existentes na cidade do Porto à plataforma do *CityClusters*;
- Identificar de forma automática:
    - As áreas do Porto que são mais procuradas durante o fim de semana e semana no horário fora da rotina;
    - As áreas do Porto que oferecem mais serviços e que podem servir como locais de atratividade elevada à semana em horas fora da rotina e ao fim de semana;
    - As áreas do Porto que oferecem eventos regularmente durante o fim de semana e à semana em horas fora da rotina;
    - Se as áreas com maior procura estão atualmente cobertas pela rede de transportes do STCP durante o fim de semana e à semana em horas fora da rotina.

A existência ou não da correlação será visível numa ferramenta de análise geográfica, o ArcMap[3], capaz de analisar diversos conjuntos geográficos. Pretende-se que as diversas questões sejam respondidas na etapa de visualização dos resultados na ferramenta referida.

1.3 Estrutura do relatório

Este documento está organizado da seguinte forma: no capítulo 2 é realizado um estudo do estado de arte, sobre as aplicações *crowdsensing* e técnicas de *clustering* usadas para captar e analisar, respetivamente, dados sobre mobilidade urbana. Ainda neste capítulo são apresentados também os diversos sensores usados para captar dados.

[3] http://desktop.arcgis.com/en/arcmap/ (Último acesso em 03/12/18)

No capítulo 3 é apresentado o projeto URBY.SENSE onde este trabalho se insere. Para além disto são também referidas as principais contribuições, que este trabalho pretende atingir, para com o projeto URBY.SENSE.

No capítulo 4 é realizado o estudo sobre reconhecimento de padrões na escolha de destinos dos utilizadores da aplicação *SenseMyFeup*. Neste estudo, pretender-se-á propor uma metodologia genérica para responder às questões e objetivos propostos deste trabalho de mestrado na presença de dados recolhidos de diferentes formas.

O capítulo 5 tem como finalidade a apresentação de resultados obtidos e sua respetiva validação, durante o estudo realizado no capítulo anterior, neste caso aplicado à zona metropolitana do Porto.

Por fim, no capítulo 6 são apresentadas as conclusões globais deste projeto de investigação e sugestões, que poderão ser realizadas como trabalho futuro.

2 ESTADO DA ARTE

Neste capítulo é realizado um estudo do estado de arte, apresentando diferentes aplicações *crowdsensing*, que foram estudadas e desenvolvidas pelos seus autores com o intuito de recolher dados úteis sobre a mobilidade e que possam auxiliar os sistemas de transportes na tomada de decisões. Nesta análise, para além das aplicações é também fornecida uma visão geral dos grupos de sensores usados atualmente para captar dados relativos à mobilidade. E uma vez perante tal quantidade de dados, pretende-se estudar como diferentes projetos de investigação abordam de forma exploratória a análise destes dados, onde é dada particular ênfase aos algoritmos de *clustering* aplicados à mobilidade, que permitem inferir e auxiliar na tomada de decisões.

2.1 Mobile Crowdsensing

O *mobile crowdsensing* tem se tornado numa das formas mais promissoras para captar dados válidos da mobilidade do utilizador, que possam melhorar o planeamento de sistemas de transportes e gestão nas diferentes áreas e escalas geográficas e temporais [9]. Além dos dados disponíveis sobre serviços na cidade, existe atualmente um grande foco no desenvolvimento de métodos de recolha colaborativa para determinar os padrões de mobilidade urbana no uso de transportes públicos. Entender estes padrões permite aos operadores de transportes planear a oferta voltada para suprir as necessidades e eventuais lacunas que não conseguem dar respostas aos utentes destes transportes. O *mobile crowdsensing* tem assim potencial necessário para se tornar numa importante fonte de dados para os sistemas inteligentes de transporte urbano [9].

Para além do uso de *smartphones* para captação de dados, a popularidade e o uso de plataformas sociais oferecem também uma promissora fonte de dados, que ajuda a compreender os locais mais visitados e quais as tendências de visita das pessoas. Este tipo de plataformas enquadra-se no *crowdsensing* implícito, onde milhares de utilizadores capturam e armazenam dados voluntariamente, sem qualquer requisito externo ou definitivo, ou seja, sem intenção ou objetivo específico de o fazer. Por outro lado, quando os utilizadores estão explicitamente instruídos a intencionalmente capturar dados, o *crowdsensing* é categorizado como explícito [10]. A consciência do utilizador na realização do *crowdsensing* apresenta assim uma nova dimensão [11].

No contexto do *crowdsensing* explícito, podemos ainda subdividir as aplicações *crowdsensing*, através da forma, com que o utilizador interage com elas: participatórias ou oportunísticas. O *crowdsensing* é participatório quando a aplicação necessita estritamente do *input* do utilizador para obter dados sobre ele. Geralmente o sistema requer que o utilizador reporte ou subscreva a tarefas, normalmente através do

preenchimento de questionários [4]. Por sua vez, quando o *crowdsensing* é oportunístico, a aplicação extrai diretamente os dados do utilizador através dos sensores do dispositivo que esta a usar (*smartphone, tablets, ipads*, entre outros), sem requerer (ou quase nenhuma) interação direta com o utilizador. Neste tipo de *crowdsensing*, o utilizador muitas vezes nem sabe que a aplicação está a recolher dados, dada a baixa e até em alguns casos, ausência de intrusividade para com o utilizador.

2.1.1 A utilização de sensores dos *smartphones* na mobilidade urbana

Se o crescimento dos *smartphones* mudou principalmente a forma como as pessoas interagem umas com as outras, a tecnologia dos sensores tem se tornado ubíqua e é considerada uma fonte de dados capaz de fornecer informação relativa à rotina diária e atividades do utilizador. Perceber e entender os movimentos e deslocações que os cidadãos realizam na cidade poderá ser uma mais valia para apoiar um sistema inteligente de transportes. Nas seguintes subsecções estão descritos os diversos grupos de sensores caracterizados por movimento, fisiológico, ambiente, rede e social.

2.1.1.1 Sensores de movimento

Estes sensores pertencem ao grupo mais comum de sensores usado para detetar movimento e para os estudos que no geral têm como objetivo detetar padrões sobre as viagens que os cidadãos realizam. Acelerómetro, giroscópio, magnómetro e o GPS são os sensores que se inserem nesta categoria. O acelerómetro mede aceleração, o giroscópio mede rotação e o magnómetro mede a força do campo magnético ao longo dos eixos *x*, *y*, e *z* [12]. O GPS é o sensor que fornece aos *smartphones* a localização geográfica do utilizador, mais concretamente dados sobre as coordenadas longitude, latitude e em alguns casos, altitude [13]. De seguida, seguem-se alguns exemplos de casos de uso usando estes sensores: Em [14], os autores investigaram como este tipo de sensores podem ser usados para detetar eventos de risco durante a prática da condução e desenvolveram uma plataforma que monitoriza os hábitos de condução. Este tipo de sensores permitem inferir qual a atividade que o utilizador está a realizar no momento. Em [4] um dos objetivos a que a aplicação do *SenseMyCity* se propôs foi a medição do consumo de combustível para um condutor em tempo real. Usaram o GPS para perceber a localização do utilizador e com base nisso determinaram o gradiente da estrada. Após determinarem esse gradiente usaram o sensor do acelerómetro do *smartphone* para saber qual a velocidade a que o condutor seguia e por fim calculavam o consumo de combustível com base nessas duas variáveis, o gradiente da estrada onde seguiam e a velocidade a que seguiam. O sistema de reconhecimento de atividades da *Google*, o

Google Activity Recognition[4], também usa este grupo de sensores para inferir em tempo real qual a atividade que o utilizador está a realizar.

2.1.1.2 Sensores Fisiológicos

Os sensores fisiológicos são geralmente usados para inferir os níveis de *stress* e emocionais do utilizador. Capturam sinais elétricos através do corpo, tal como um eletrocardiograma. Este tipo de dados pode servir como indicador para, por exemplo, capturar o nível de stress durante a condução [13]. O estado psicológico é um fator muito importante em qualquer condutor e com a captura destes dados é possível identificar zonas geograficamente críticas ou entendidas como perigosas para a maioria dos utilizadores. A identificação destas zonas pode então ser conseguida através do índice de *stress* dos condutores, o que pode indicar um elevado volume de trânsito ou uma dificuldade maior na prática da condução naquele local. A definição da melhor rota a seguir, por exemplo, para os sistemas de transporte de emergência pode apoiar-se sobre este tipo de dados e, desta forma, tentar que o percurso a efetuar seja realizado pela maneira mais segura e rápida.

2.1.1.3 Sensores de ambiente

Os sensores de ambiente integram sensores que medem e monitorizam as condições do ambiente, tais como, iluminação, temperatura ambiente, aceleração do vento e pressão atmosférica. Os sensores que se podem enquadrar nesta categoria são: sensores de luz (luminosidade), microfones e camaras. Os microfones e camaras podem ser úteis para reconhecimento de lugares particulares, ou simplesmente para identificar o contexto, no qual os utilizadores se encontram no momento [13].

2.1.1.4 Sensores de rede

Os sensores de rede são principalmente compostos pelas tecnologias *Wi-Fi* e *Bluetooth* e podem ser usados para estudos que identificam padrões de mobilidade. Em [13], os autores demonstram que é possível usar este grupo de sensores para inferir e entender os movimentos e interações que uma pessoa realiza por dia. Baseado no número de pontos de acesso à rede detetados por estas duas tecnologias podemos concluir alguns aspetos interessantes. Na figura 1 encontra-se representado o exemplo de duas rotinas diárias, $P_1(a)$ e $P_2(b)$. Em ambas, os utilizadores utilizam o carro como meio de transporte e as áreas cinzentas indicam quando os participantes estão a comunicar.

[4] https://developers.google.com/location-context/activity-recognition/ (Último acesso em 27/11/2018)

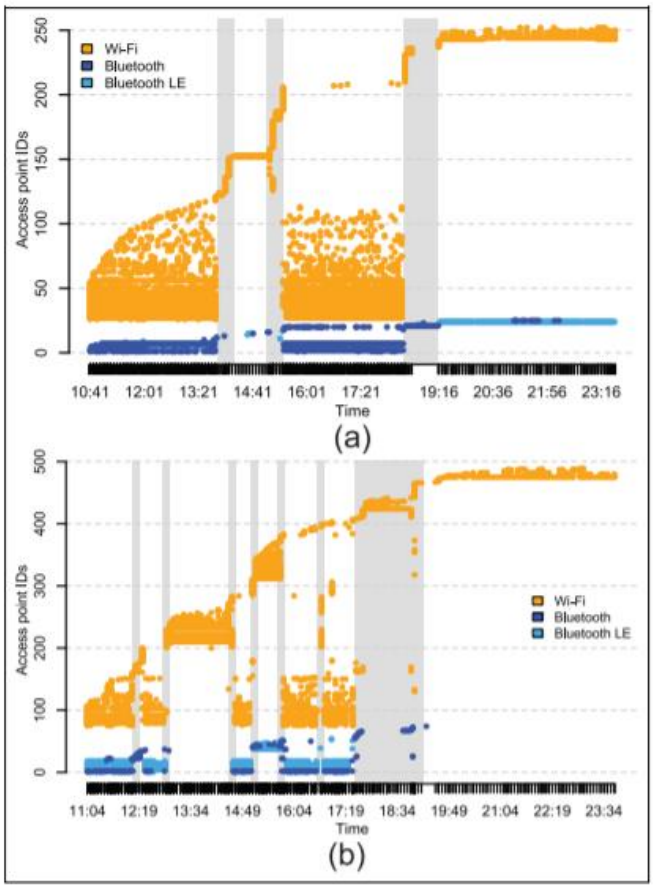


Figura 1 - Representação das rotinas $P_1(a)$ e $P_2(b)$, com base no número de pontos de acesso à rede identificados pelos sensores de rede: Wi-Fi, Bluetooth e Bluetooth LE (Bluetooth de baixo consumo energético) Fonte [13].

Podemos concluir que ambas as rotinas tendem a encontrar um grande número de pontos de acesso. A rotina $P_2(b)$ aparenta visitar mais lugares do que $P_1(a)$, porque o número de acessos ao longo do dia é menos constante do que a rotina $P_1(a)$. Logo, podemos afirmar que $P_2(b)$ move-se muito menos no seu lugar de trabalho do que a outra rotina. Observamos também uma mudança no número de pontos de acesso que ocorre durante as 13:30h e 14:30h que deverá ser devida à hora de almoço, uma vez que em ambas as rotinas se deslocam durante essas horas [13].

2.1.1.5 Sensores sociais

Os sensores sociais são aqueles que fornecem dados extraídos diretamente das redes sociais, como o *Facebook*[5], o *Instagram*[6] e o *Twitter*[7]. Hoje em dia é inquestionável a importância e influência da existência das redes sociais na vida quotidiana. A sua utilização na extração de dados sobre a mobilidade poderá ser um bom indicador para deteção de padrões de mobilidade interessantes. Por exemplo, a utilização de *geolocated tweets*, no caso particular da plataforma do *Twitter* poderá ser uma fonte de dados válida capaz de fornecer informação geográfica relevante do utilizador. No *Twitter* cada utilizador poderá escrever *posts* ou *tweets* com um limite de 140 carateres. Utilizadores que escrevem *posts* poderão indicar a sua localização GPS precisa no momento, estes são os chamados *geolocated tweets*. Jurdak et al. em [15] realizaram um estudo onde analisaram um *dataset* composto por mais de 6 milhões de *geotagged tweets* postados na Austrália. No estudo realizado foi demonstrado como os *geotagged tweets* podem capturar e extrair importantes factos sobre a mobilidade, tais como a diversidade de movimentos entre indivíduos e movimentos entre cidades [15].

2.1.2 *Crowdsensing* Oportunístico

Tipicamente, o *crowdsensing* oportunístico requer maior trabalho de processamento, uma vez que é necessário filtrar e inferir informação a partir da captação de dados. Por outro lado, este tipo de sistemas têm como objetivo a captação de dados em larga escala, o que comparativamente ao método participatório, essa obtenção é mais facilitada, uma vez que requer a mínima intervenção possível do utilizador [4]. Esta não intrusividade é conseguida através dos sensores do *smartphone* permitindo a extração de dados sem que o utilizador tenha de realizar uma tarefa específica.

2.1.2.1 Future Mobility Survey

O *Future Mobility Survey* é uma aplicação *crowdsensing* desenvolvida e testada em Singapura, que visa a recolha de informações durante viagens realizadas por utilizadores [16]. A aplicação realiza um sistema de inquéritos de viagem disponível em dispositivos móveis que recolhe dados de forma voluntária e ubíqua sobre a utilização de transportes públicos. Este sistema recolhe dados em 4 fases: no registo do utilizador onde o seu perfil é indicado, pré-questionário onde são introduzidos dados socioeconómicos do agregado familiar, diário de atividades preenchido durante o período em estudo e modos de

[5] https://www.facebook.com (Último acesso em 11/06/2018)

[6] https://www.instagram.com/ (Último acesso em 11/06/2018)

[7] https://twitter.com (Último acesso em 11/06/2018)

transporte detetados automaticamente pelo aplicativo, e finalmente um inquérito final onde o utilizador responde a questões sobre a utilização do sistema.

A principal funcionalidade nesta aplicação é permitir que os *smartphones* atuem como *data loggers*, ou seja pretende-se que o registo de informação a recolher seja sobretudo informação geográfica das posições visitadas e percorridas pelos utilizadores relativas às suas viagens. A implementação desta aplicação foi realizada com o objetivo de permitir também a integração de outro tipo de dispositivos que possam recolher informação geográfica. No entanto, dados extraídos do acelerómetro, Wi-Fi e GSM (*Global System for Mobile communication*) também são alvo de estudo nesta aplicação. O uso destes dados, em conjunto com a informação geográfica extraída pelo sensor de GPS permite que a aplicação consiga inferir qual o meio de transporte a que o utilizador está a utilizar. Na figura 2 podemos analisar a arquitetura desta aplicação:

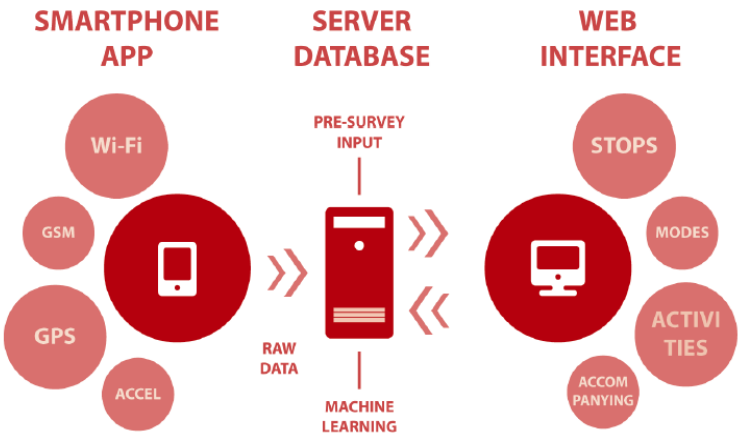


Figura 2 - Arquitetura da aplicação Future Mobility Survey. Fonte: [16]

Esta aplicação poderá ser então dividida em três diferentes módulos: do lado do cliente temos a aplicação *smartphone* responsável pela captura e envio de dados para o servidor. O servidor, por sua vez, recebe os dados da aplicação, mapeia-os, filtra-os e limpa-os. Depois dessa fase eles tornam-se prontos para ser analisados e acessíveis num *website*, onde ao utilizador é também sugerido que confirma os dados obtidos sobre as viagens realizadas. Dados sobre os locais onde parou, a duração da paragem, atividades que foram realizadas durante os locais por onde passou, o meio de transporte que usou durante as viagens e o custo associado ao modo de viagem escolhido são os dados que a aplicação *Web* pede ao utilizador que confirme. Para que a interação do utilizador para com o *website* seja mínima, foram desenvolvidas técnicas de análise de dados, que determinam se o utilizador está parado ou não e qual o meio de transporte que está a usar.

2.1.2.2 SWIPE

O *SWIPE* [17] é uma aplicação *open-source*, que tem como objetivo capturar e processar dinâmicas humanas que possam auxiliar os sistemas de mobilidade na tomada de decisões

usando *smartwatches* e *smartphones*. A arquitetura da aplicação presente na figura 3 assenta na existência de uma aplicação *android*, que regularmente extrai dados simultaneamente de um *smartphone* e de um *smartwatch*.

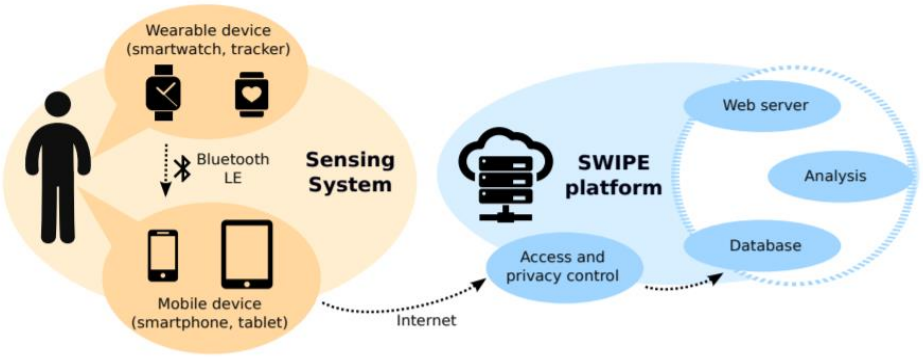


Figura 3 - Arquitetura da aplicação open-source SWIPE. Fonte: [17]

O *smartwach* regularmente envia dados para o *smartphone*, para que este último sirva de *gateway* de acesso para o servidor. O *smartwach* captura dados como o número de batimentos cardíacos, pulsação e número de pontos de acesso à rede que encontra no momento e envia-os para o *smartphone*. Para além da receção de dados do *smartwach*, o *smartphone*, extrai também dados do utilizador usando sensores de localização, GPS, acelerómetro e também do microfone, com o intuito de detetar locais com mais, ou menos barulho. Os dados capturados pelo *smartphone* pretendem também indicar o número de passos e percursos percorridos. Para além desta fase de recolha, os dados são posteriormente armazenados no servidor e analisados. O SWIPE disponibiliza também uma interface *web*, na qual os utilizadores poderão visualizar a análise e resultados obtidos.

2.1.2.3 SenseMyCity

O *SenseMyCity* [4] é uma aplicação *mobile crowdsensing* [4], que tem como objetivo o estudo da mobilidade humana numa área urbana. A aplicação é responsável por extrair dados através dos sensores embebidos: GPS, *WiFi*, acelerómetro, giroscópio e conta ainda com o auxílio de sensores externos como o *bluetooth*. A aplicação possui uma interface minimalista, possuindo apenas 5 botões que permitem as seguintes funcionalidades: dar o início da recolha de dados, terminar a recolha de dados, sincronização dos dados que são obtidos para o servidor e um acesso às preferências e definições da aplicação. Na figura 4 encontra-se representada a arquitetura do *SenseMyCity* [4].

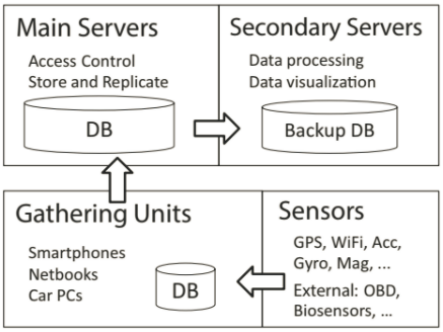


Figura 4 - Arquitetura da aplicação SenseMyCity. Fonte: [4].

Do lado do cliente, a aplicação é instalada e recolhe dados dos sensores do *smartphone*, podendo também ser integrada com sensores captados a partir de fontes externas. Todos os dados captados são armazenados localmente numa base de dados *SQLite*[8]. Posteriormente, quando o utilizador indica na aplicação que pretende sincronizar os dados com o servidor, a aplicação envia-os para o componente "*Main Servers*". Os dados recebidos são replicados em servidores secundários para permitir que a análise dos dados seja feita do lado dos "*Secondary Servers*" e deste modo aumentar a eficiência e aproveitamento dos recursos.

Recentemente em [9], a versão do *SenseMycity* foi melhorada na nova versão *SenseMyFeup* [9]. Esta versão do *SenseMyFeup* foi construída em cima da plataforma do *SenseMyCity*, com o objetivo de analisar a mobilidade de padrões das pessoas da FEUP (Faculdade de Engenharia da Universidade do Porto) e estudar a sustentabilidade ecológica da comunidade. Para a realização do estudo foi aos participantes que contribuíssem com dados anónimos, em contrapartida a aplicação mostrava as estatísticas de cada viagem realizada pelo utilizador: a distância percorrida, o tamanho da pegada ecológica, bem como a media da comunidade. Na figura 5 é apresentada a interface da aplicação.

[8] https://www.sqlite.org/index.html (Último acesso em 20/08/2018)

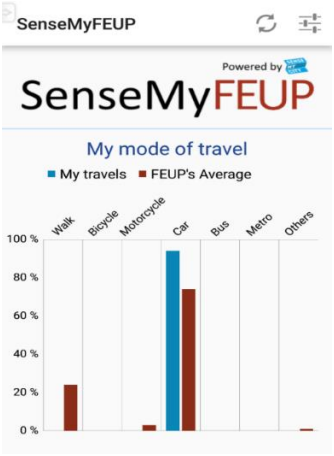


Figura 5 - Interface da aplicação SenseMyFeup

Nesta versão foram melhorados mecanismos, que permitem que a aplicação seja menos intrusiva para com o utilizador. Como melhoramentos, nesta nova versão, a aplicação consegue detetar movimento, sem que o utilizador tenha que explicitamente indicar à aplicação para iniciar a recolha de dados. O uso desta nova funcionalidade permitiu, também que o uso dos recursos do *smartphone* sejam feitos de forma mais eficiente, permitindo a poupança de mais bateria. A outra funcionalidade, que foi introduzida nesta nova versão foi a possibilidade oportunística de envio de dados para o servidor. Este envio, que anteriormente teria de ser realizado manualmente pelo utilizador, nesta nova versão o envio ocorre assim que uma ligação à rede esteja disponível, por isso a ação do utilizador deixa de ser desnecessária.

2.1.2.4 Travel Mode Detection

Shafique et al em [18] propuseram um modelo que usa dados extraídos dos sensores do *smartphone*, como o acelerómetro, sensor de orientação e GPS para inferir que atividades, o utilizador esta a realizar. Este modelo foi testado com 50 participantes em *Kobe*, Japão. Eles contribuíram para esta fase de extração de dados durante 1 mês. Quando iniciam uma viagem, os participantes selecionam a opção na aplicação para indicar o início da extração de dados. Quando alcançam o destino final da viagem, o utilizador seleciona na aplicação o fim da mesma, para parar a extração. No fim do dia, um pequeno questionário é ainda dirigido aos utilizadores com o objetivo de validar as análises obtidas pelo sistema. Através deste *feedback* vindo dos utilizadores, o sistema "afina" a inferência usada para determinar se uma determinada atividade que o utilizador realizou corresponde à real. Neste estudo seis atividades foram alvo de estudo: caminhar, andar de bicicleta, andar de comboio e autocarro.

2.1.2.5 NSE (National Science Experiment)

O projeto NSE [19] tem como objetivo analisar e entender a mobilidade de jovens estudantes, nomeadamente a determinação de padrões no meio de transporte usado, através da recolha exclusiva de pontos de acesso à rede extraídos pelo Wi-Fi. Estudos

existentes, com a mesma abordagem apresentam pontos fracos em comum. Em primeiro lugar, a quantidade de dados, que usaram não é suficiente para que se possa atingir determinadas conclusões e em segundo os resultados que são obtidos não correspondem inteiramente só de pontos de acesso, mas sim resultado de outros tipos de dados. Pelo contrário, o projeto NSE foi usado entre 2015 e 2016, em cerca de 90 000 participantes em 100 escolas espalhadas por toda a cidade de Singapura, que permitiram obter padrões precisos, exclusivamente apenas de dados sobre o número de pontos de acesso [19]. Na figura 6 é apresentada a arquitetura do projeto NSE.

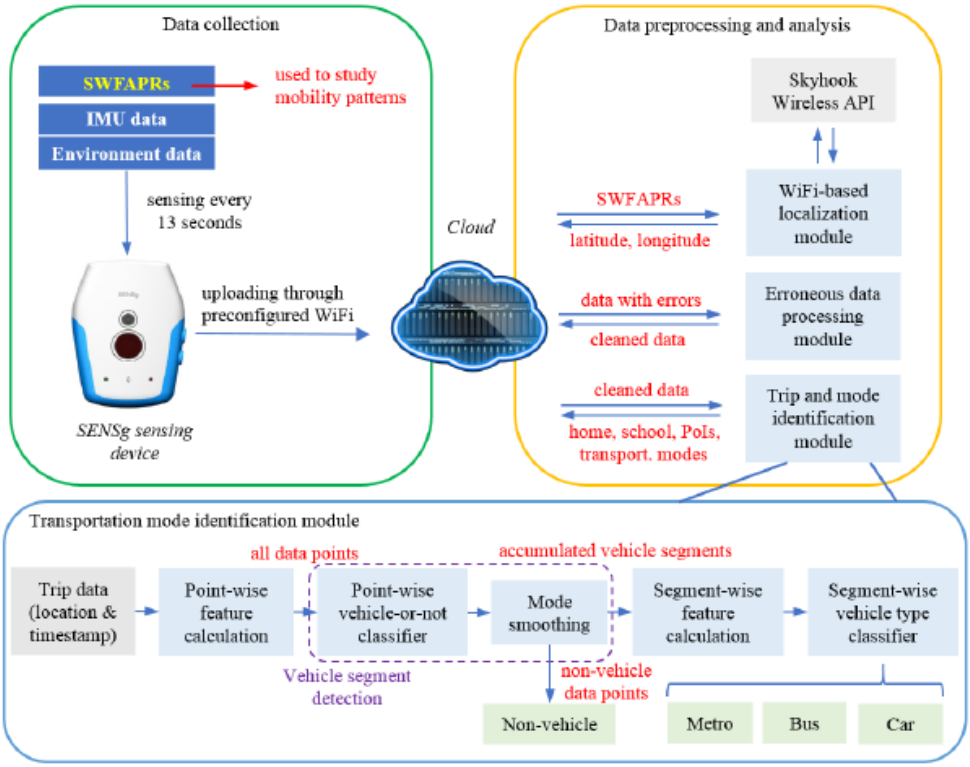


Figura 6 - Arquitetura do Projeto NSE. Fonte: [19]

Com o intuito de gerar menor intusividade possível para com o utilizador, os autores desenvolveram um dispositivo de recolha de número de pontos de acesso, chamado *SENSg*. Durante a fase de pré-processamento de dados os pontos de acesso que são detetados são alvo de um mapeamento de dados entre a API do *Skyhook Wireless*[9]. Este mapeamento permite saber a localização geográfica do utilizador (coordenadas latitude e longitude), com base nos pontos de acesso à rede, que naquela determinada posição o dispositivo deteta. Com o uso desta API conseguiu-se evitar o uso do sensor de GPS e desta forma saber na mesma a localização geográfica do utilizador. Depois de se

[9] https://www.skyhook.com/ (Último acesso em 22/08/2018)

conseguir obter os dados geográficos sobre os percursos que cada utilizador realizou procede-se à fase de análise onde os objetivo do NSE passam por: deteção do meio de transporte usado diariamente, a identificação do lugar onde mora e a escola, identificação de viagens e pontos de interesse (POIs) e a identificação de lugares onde o utilizador esteve parado. Todas estas métricas foram usadas para a análise final, onde se pretendeu saber, para cada aluno, desde casa até à escola qual o modo de transporte que usaram, consoante a distância que a sua casa dista até à escola. Com o estudo realizado concluiu-se que quando a distância da casa à escola é mínima, cerca de 1km e 600 metros a opção predominante dos estudantes é ir a pé. A partir dos 5km de distância, os estudantes optam por ir de autocarro. Verifica-se que, a partir dos 15km os estudantes optam pelo metro, como meio de transporte. Um dado também alvo de análise é a utilização baixa de carro, como meio de transporte, que se manteve sempre relativamente constante independentemente da distância.

2.1.2.6 City

A aplicação *smartphone City*[10] consiste na monitorização e recolha de atividades realizadas pelo utilizador, tais como: andar a pé, andar de carro, estar parado, entre outras. Para além disso, a aplicação identifica os locais mais visitados durante a rotina diária e fornece alguns dados estatísticos, como por exemplo, o tempo que o utilizador gasta em cada local. O estudo realizado em [20] utilizou a *City* como ferramenta de recolha de dados, no qual se pretendia compreender as atividades e dinâmicas que um idoso tem na sua rotina. Em qualquer sociedade, a população idosa é um elemento chave, que requer particular importância na prestação de serviços de mobilidade. A noção das rotinas diárias de cada idoso, através da identificação dos locais que mais visita e das regiões da cidade, que cada idoso interage mais é um importante passo, que possibilita a criação de soluções capazes de melhorar a mobilidade. O estudo realizado teve como principal objetivo a identificação das regiões e POIs da população idosa na cidade de Singapura. Envolvendo cerca de 100 participantes, com idade superior a 50 anos, a *City* recolheu os seguintes dados: coordenadas geográficas (latitude e longitude) captadas pelo sensor de GPS da aplicação, a atividade que está a ser realizada pelo utilizador inferida pelo uso da tecnologia da *Google Activity Recogntion*[11], nível de barulho sentido no ambiente à volta, nível de bateria do dispositivo e luminosidade detetada pelo sensor de luz. Na figura 7 encontra-se representado o procedimento seguido durante a captação de dados.

[10] https://apkpure.com/city/sutd.dev.testapp (Último acesso em 23/08/2018)

[11] https://developers.google.com/location-context/activity-recognition/ (Último acesso em 23/08/2018

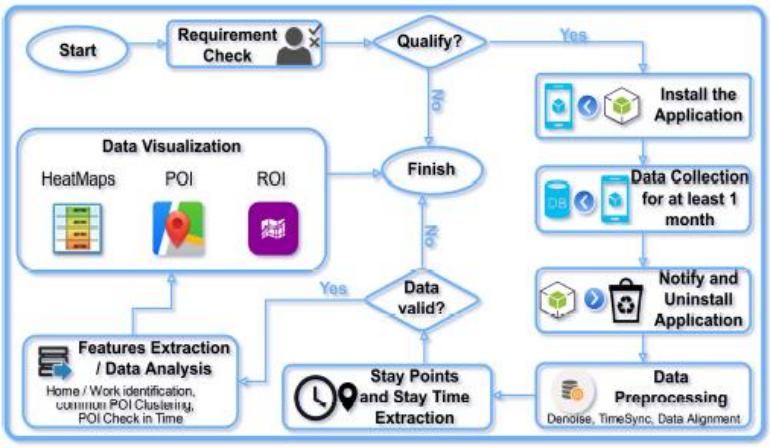


Figura 7 - Procedimento da extração de dados. Fonte: [20]

A fase de análise de dados realizou-se recorrendo a uma aplicação *back-end* em *Java* e a um *package* de *software* MATLAB. Depois dos dados serem alvo de análise, ferramentas como *Google Maps*, *Voronoi Diagrams* e *Heat Maps* foram usados para visualizar as conclusões obtidas, de modo a facilitar a identificação e compreensão das regiões e POIs.

2.1.3 *Crowdsensing* Participatório

O *crowdsensing* participatório não requer tanto processamento e análise de dados como o oportunístico requer. Este tipo de *crowdsensing* pode ser usado com o objetivo de detetar de uma forma rápida, os eventos de maior importância, no entanto falham a detetar eventos com menor importância ou padrões não conhecidos pelo ambiente que os utilizadores não considerem importantes no ato de reportar dados [4].

2.1.3.1 PublicSense

O *PublicSense* [21] é uma *mobile crowdsensing framework* que propõe a utilização de dados vindos de plataformas de reclamação, para compreender os problemas atuais da cidade (mobilidade, construções, entre outros) e dessa forma apoiar a tomada de decisões, no que diz respeito à gestão dos recursos da cidade para resolução de problemas. Poucos estudos prestam atenção aos dados recolhidos das plataformas de reclamação, mas a verdade é que estes dados, permitem-nos obter uma opinião geral, distinta e única, à acerca dos problemas e dinâmicas que acontecem na cidade e desta forma poder também melhorar o sistema de mobilidade [21]. As potenciais aplicações que poderiam usar esta *framework* poderiam ser sistemas e serviços de monitorização e transporte de emergências. Na figura 8 é possível visualizar a arquitetura proposta por esta *framework*:

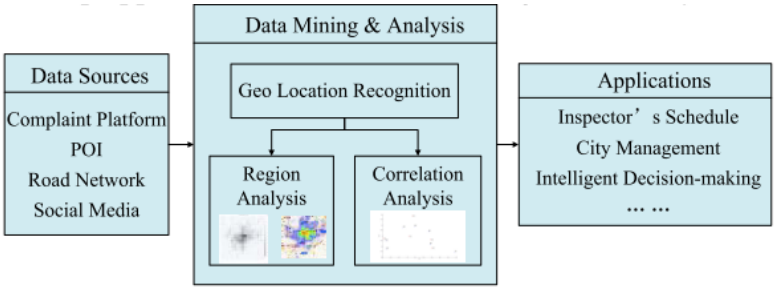


Figura 8 - Arquitetura da framework PublicSense. Fonte: [21]

Usando apenas dados vindos de uma plataforma de reclamações sem conectá-los com informação geográfica pode não ser suficiente para atingir conclusões. Então, esta *framework* propõe um modulo de reconhecimento de localização responsável por detetar informação geográfica em falta nas reclamações. No caso de estudo usado na cidade de *Xi'an* no Japão, foi criado um dicionário com nomes de todas as estradas existentes na cidade e com base nesses dados as reclamações foram mapeadas com coordenadas de acordo com o problema a retratar. Os dados das reclamações foram fornecidos pela câmara municipal da cidade referida, no entanto, como a *framework* propõe, estes poderiam ter sido obtidos através de comentários existentes nos POIs da cidade, ou através das redes sociais. Após este mapeamento foi criado um *heatmap* correspondente aos problemas retratados e dispostos graficamente. Este *heatmap* permitiu identificar visualmente os principais problemas registados e auxiliar na tomada de decisões.

2.1.3.2 GeoLife

O projeto *GeoLife* [22] é focado na visualização e análise de percursos captados pelo sensor de localização geográfica, o GPS. Neste projeto é fornecida uma plataforma, que permite aos seus utilizadores gerir os seus dados geográficos e também entender o seu histórico de dados. A visualização dos dados é realizada sobre mapas digitais, que permitem a procura de percursos geográficos realizados pelo utilizador num determinado período de tempo e/ou numa determinada distância ou área geográfica. A figura 9 representa a arquitetura do *GeoLife*.

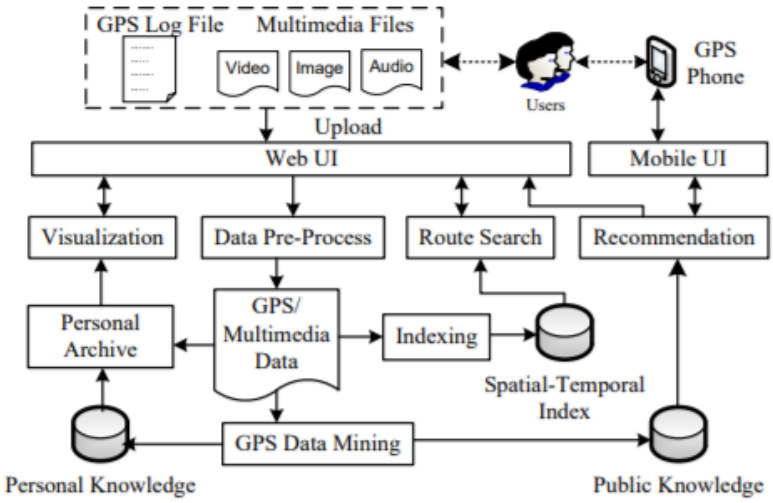


Figura 9 - Arquitetura do projeto GeoLife. Fonte: [22]

Os utilizadores podem fazer o *upload* dos seus dados geográficos, bem como associá-los a ficheiros multimédia, imagens ou vídeo, que serão categorizados com as coordenadas geográficas (latitude e longitude), correspondentes aos locais de onde foram recolhidos/capturados. A análise dos dados que é feita na fase seguinte depende da vontade explícita do utilizador. A análise poderá ser realizada apenas para uso pessoal ou também para uso público. No contexto do uso pessoal, apenas para o utilizador, o *GeoLife* arquiva todo o histórico geográfico do utilizador e analisa-o inferindo quais as rotinas de transporte mais usadas, os lugares mais visitados e de maior importância para o utilizador, bem como o padrão de vida associado. O objetivo é ajudar também os utilizadores a terem uma análise das suas viagens e a adquirirem hábitos saudáveis nas suas vidas. Do ponto de vista do uso público, o uso destes dados permitirá analisar os destinos de viagens mais populares e o volume de trânsito em diferentes locais e dias. O uso destas variáveis pode auxiliar decisores locais e sistemas inteligentes de transportes na tomada de decisões.

2.2 Técnicas de aprendizagem e análise de dados aplicadas à mobilidade

Em [19], os autores identificam 7 métodos de análise de dados usados para compreender a mobilidade urbana: visualização, estatística, heurística lógica, teoria de grafo, otimização, *clustering* e classificação.

A visualização é um método que permite compreender e apresentar padrões de mobilidade urbana. Diversos tipos de gráficos conseguem apresentar dados que podem permitir tirar conclusões importantes. Nesta categoria podemos incluir os gráficos de dispersão, gráficos de barras ou histogramas e os *heatmaps*. Os gráficos de dispersão são frequentemente usados para visualizar trajetórias, percursos e localizações de lugares

importantes. Os gráficos de barras ou histogramas são frequentemente usados para estudar valores de atributos de mobilidade, que são acumulados em intervalos de tempo consecutivos. Por sua vez, os *heatmaps* podem revelar conhecimento sobre distribuições temporais ou espaciais de um determinado tipo de atividades humanas. A distribuição espacial representada em [21] é um exemplo prático da utilização deste tipo de gráfico.

A análise estatística, que se efetua em estudos de mobilidade humana distingue-se em três etapas: em primeiro lugar, através da extração de estatísticas conclusivas de atributos de mobilidade complexa ou resultados de pesquisa, como as matrizes de avaliação de classificação apresentadas em [16]. Em segundo, a criação de modelos estatísticos para descrever os atributos de mobilidade, como em [23] . Em terceiro lugar, a criação de testes estatísticos para avaliar as conclusões de determinadas pesquisas, como por exemplo, o uso de simulações de *Monte Carlo* para testar a análise de *wavelet* realizadas em [24].

A heurística lógica significa neste contexto, a criação de regras de processamento de dados, com base na lógica ou conhecimento do domínio do problema. Um dos exemplos práticos da utilização desta análise é a deteção de POIs realizada em [25], em que os parâmetros indicados ao algoritmo de procura são determinados através do conhecimento e domínio do problema.

A teoria de grafos é uma técnica de análise de mobilidade, que consiste na construção de modelos em grafo, com o objetivo de determinar ligações entre pessoas e locais que tenham sido visitados por elas. Em [24] um modelo de grafo é construído para investigar as ligações que existem nos lugares visitados pelas mesmas pessoas.

As técnicas de otimização são geralmente usadas quando existe necessidade de realizar algum ajuste no modelo estatístico, de forma a maximizar ou a minimizar algum método ou variável. Um exemplo prático da aplicação desta técnica é a quantificação dos parâmetros de movimento humano. Em [24] foi usada a estimativa por máxima verossimilhança para estimar os parâmetros do modelo estatístico.

Por sua vez, o *clustering* é usado em projetos de mobilidade urbana para: encontrar habituais locais de visita de pessoas [25] e agrupar diferentes perfis de pessoas em grupos, de acordo com os seus padrões de mobilidade [26]. Os algoritmos de *clustering* mais populares são o *k-means* e o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN).

A classificação é usada maioritariamente em projetos de mobilidade que têm como objetivo identificar categorias de lugares onde as pessoas visitam [23], a classificação de pessoas em diferentes perfis e a identificação dos modos de transporte das pessoas. Nesta categoria inserem-se muitos classificadores, tais como: *tree-based models*, *support vector machines (*SVM*)* e *artificial neural networks*.

Na tabela 1 apresentam-se as técnicas de aprendizagem que foram utilizadas nas aplicações, projetos e *frameworks* anteriormente referidos na secção 2.1.

Tabela 1 - Técnicas de aprendizagem utilizadas nas aplicações, projetos e frameworks crowdsensing

Nome	**Tipo de *Crowdsensing***	**Técnicas aplicadas na análise de dados**
Future Mobility Survey [16]	Oportunístico	*Support Vector Machine* *Rule-based algorithm*
SWIPE [17]	Oportunístico	*Support Vector Machine*
PublicSense [21]	Participatório	*Clustering*
Travel Mode Detection [18]	Oportunístico	*Random Forest*
GeoLife [22]	Participatório	*Deep neural network*
SenseMyFeup [9]	Oportunístico	Heurística (método de deteção de movimento)
NSE [19]	Oportunístico	*Clustering* *Random Forest*
City[12]	Oportunístico	*Clustering* *Voroni Diagrams*

2.2.1 Support Vector Machine

Estudos desenvolvidos com a aplicação SWIPE [17], contando com a participação de 13 pessoas foram realizados, com diferentes géneros e idades, que trabalhavam no mesmo edifício da Universidade do Luxemburgo. O objetivo deste estudo foi perceber se a aplicação seria capaz de identificar as seguintes tarefas: em primeiro lugar, as atividades que um dado utilizador realiza ao longo do dia: se estaria sentado, levantado, a caminhar, a correr, ou a jogar ténis. Em segundo lugar, detetar qual o meio de transporte que usava em cada viagem que realizava: carro, autocarro, comboio, moto ou carro. Por fim, identificar o contexto, no qual os utilizadores se enquadravam ao longo do dia: se estavam

[12] https://apkpure.com/city/sutd.dev.testapp (Último acesso em 04/12/2018)

a trabalhar, em reunião, no *shopping*, no intervalo, ou se estariam em casa. Para isso, ambos os dados extraídos do *smartwach* e da aplicação móvel foram utilizados. No caso do *smartwach* foram utilizados a aceleração, o número de passos e o número de batimentos cardíacos. No caso da aplicação móvel foram usados a aceleração e o número de passos. Para além destes, foram também usados dados do microfone, número de pontos de acesso detetados pelas tecnologias Wi-Fi e *Bluetooth*, estado de rede do telefone e a velocidade obtida pelo sensor de GPS. Usando estes dados recorreu-se à utilização de um classificador, o algoritmo SVM. Com o uso deste classificador conseguiu-se atingir os objetivos pretendidos e a aplicação foi capaz de identificar com sucesso a maior parte das atividades que o utilizador realizou, bem como o contexto diário em que ele se inseriu. Foram realizados testes usando apenas o dispositivo móvel e a aplicação *smartwach*, cada uma de forma isolada. De acordo com os resultados obtidos concluiu-se que a opção que garante melhores resultados é a combinação do uso dos dois dispositivos ao mesmo tempo. Nesta configuração que apresentou melhores resultados, o algoritmo SVM foi capaz de identificar com sucesso em 100% as atividades de corrida, jogo de ténis, andar de comboio e moto. Relativamente ao contexto diário que o utilizador se encontra, o algoritmo foi capaz de obter uma *accuracy* de 100%, quando o utilizador está em casa. De seguida no âmbito das atividades seguiram-se o andar, o levantar, andar de comboio e andar de carro com 99.5%, 95.3%, 95.8%, 88.9%, 93.2%, respetivamente. No geral, o algoritmo obteve uma *accuracy* de 93.2% para as atividades. No âmbito do contexto, o algoritmo apresenta uma *accuracy* geral de 90.4%. Estar a trabalhar, em reuniões, no *shopping* ou em intervalo apresentaram *accuracies* de 95.3%, 94.7%, 86.6% e de 95.5%, respetivamente.

Para além da aplicação anterior referida, o *Future Mobility Survey* também aplica o algoritmo de *machine learning*, SVM. A aplicação usa os dados obtidos através do acelerómetro e do GPS para identificar o meio de transporte do utilizador no momento. Se está a andar de carro, autocarro, a pé, de bicicleta ou de moto são alguns dos exemplos, no qual a aplicação, através da aplicação do algoritmo SVM identifica.

2.2.2 Rule-based algorithm

A análise de dados que é realizada na aplicação *Future Mobility Survey* [16], identifica quais os momentos em que o utilizador esteve parado, através da implementação de um *rule-based algorithm*. Numa primeira fase, a aplicação verifica quais as janelas temporais que contêm potenciais posições geográficas, que correspondem a locais onde o utilizador esteve parado. Depois disso utiliza dados do Wi-Fi, GSM e acelerómetro para identificar os períodos em que o utilizador esteve efetivamente parado. Também usa validações que tenham sido realizadas pelo utilizador no passado, em que identificou como estando parado, como por exemplo, casa e trabalho.

2.2.3 Clustering

No caso de estudo apresentado na *framework* do *PublicSense* [21] realizado na cidade de *Xi'an* no Japão, recorreu-se ao uso de *clustering*, nomeadamente o *k-means*, com o objetivo de identificar as áreas que continham mais reclamações e desta forma analisar as áreas que requerem maior atenção. O uso do *clustering* permitiu que a análise efetuada permitisse retirar as seguintes conclusões: a maior parte das queixas acontecem em áreas com uma densidade de população elevada ou com elevado volume de trânsito. Problemas com tampas de esgotos levantadas acontecem com maior frequência em zonas, fora do centro da cidade. Para além destas conclusões, também foi possível concluir quais as zonas com maior probabilidade de acontecer determinado tipo de reclamações.

No projeto NSE, a identificação de POIs e de viagens é realizada, através da aplicação do algoritmo DBSCAN.

No estudo realizado com a aplicação *City*, depois da fase de pré-processamento é aplicado *clustering* sobre os dados de geolocalização (latitude e longitude), correspondentes aos locais visitados pelo idoso. O algoritmo *k-means* foi o escolhido pelos autores, com o intuito de obter um conjunto de dados, que resumisse a contribuição geral da participação de cada idoso. No total foram gerados 1781 *clusters* relativos a 50 utilizadores. Na tabela gerada, para além do identificador do utilizador, idade e género, registaram-se também a localização geográfica do *cluster* obtido, o propósito da visita do idoso na visita ao local, o tempo que esteve presente no local, o número de visitas que realizou ao local durante o tempo de recolha de dados e o tempo da primeira e última visita ao local.

Com base nos dados recolhidos relativos aos POIs visitados por cada utilizador em [20] efetuou-se *clustering* sobre os POIs, com o objetivo de agrupar por distância as posições registadas a cada POI. Utilizou-se o algoritmo de DBSCAN com um raio de 100 metros para que não existissem POIs que distanciassem menos do que esse valor. Depois do processo de *clustering* terminado, obtiveram-se 980 POIs a partir dos dados extraídos. A identificação de POIs com maior interesse foi de seguida determinada, através dos seguintes dados: número de entradas em cada POI (*check-ins*) e o tempo registado por cada participante. Com este estudo os autores conseguiram identificar o período do dia, que tipo de POIs são mais visitados, qual o maior número de afluência a determinadas horas do dia. Uma das conclusões obtidas foi de que, quando os participantes realizam viagens maiores do que 10km, tinham como destino propósitos religiosos (visitar igrejas, santuários, entre outros).

2.2.4 Random Forest

Na aplicação *Travel Mode Detection*, o classificador usado para inferir as atividades do utilizador foi o *Random Forest*. Os resultados demonstraram uma *accuracy* de 99.96%,

na identificação das atividades realizadas pelo utilizador, que englobava: andar a pé, de bicicleta, de carro, de autocarro, de comboio e de metro.

O modo de transporte usado pelos participantes no projeto NSE foi inferido, através da aplicação do algoritmo *Random Forest*. A *accuracy* obtida foi de 81%.

2.2.5 Deteção de Movimento

Uma das formas de análise e comuns nas aplicações descritas é a sua capacidade de deteção de movimento, sem que o utilizador tenha de indicar à aplicação para que por exemplo ela tenha de começar a recolher dados. Uma das aplicações em que esta funcionalidade foi mais visível ocorreu na nova versão do *SenseMyFeup* [9]. A aplicação consegue detetar movimento, sem que o utilizador tenha que explicitamente indicar à aplicação para iniciar a recolha de dados. Quando o utilizador começa a realizar uma atividade, os dados extraídos pelo *smartphone* são analisados e é verificado se a partir do conteúdo dos mesmos se é possível inferir que o utilizador está em movimento, parado ou se não foi possível inferir algum tipo de resultado. "*Stopped*", "*undefined*" e "*moving*" são os possíveis estados que o algoritmo de classificação retorna [9]. O uso desta nova funcionalidade permitiu, também que o uso dos recursos do *smartphone* sejam feitos de forma mais eficiente, permitindo a poupança de mais bateria.

2.2.6 Voronoi Diagrams

A identificação de regiões de interesse em [20] foi realizada em três etapas: a primeira, com a identificação do local onde cada participante vive, bem como o tempo em que cada utilizador passa em casa, em relação ao tempo total da recolha de dados. Em segundo, com base nos dados recolhidos na etapa anterior, segue-se a identificação de regiões de interesse na vizinhança relativamente ao local onde cada participante vive. Esta identificação utilizou duas variáveis: a percentagem de visitas a determinadas zonas e o tempo que permaneceram nas mesmas. A análise e determinação dos locais, com mais interesse na vizinhança foram conseguidos através da distribuição de zonas de *Voronoi*. O mesmo procedimento seguiu-se na terceira etapa, que teve como objetivo a identificação de regiões de interesse na região de Singapura.

2.2.7 Deep neural networks

O estudo realizado em [27] procurou identificar o meio de transporte usado nos percursos registados no *dataset* do projeto *GeoLife* [22]. Determinaram e usaram os seguintes dados: velocidade, velocidade média, distância, variância da velocidade, taxa de paragem (*stop rate*) e da cabeça (*head change rate*), e a taxa de variação de velocidade (*velocity change rate*). Com base nos dados obtidos aplicaram um modelo de *Deep neural network*. Os autores realizaram também uma comparação entre os resultados obtidos do modelo de *Deep neural network* e modelos de árvore de decisão, *Logistic Regression* e *Support*

vector machine. O melhor resultado obtido, com melhor *accuracy*, pertenceu ao modelo de *Deep neural network*, com cerca de 74%.

2.3 Algoritmos de *Clustering*

Neste capítulo é realizado um estudo do estado de arte relativamente aos algoritmos de *clustering*. Neste estudo procurou-se encontrar e averiguar principais vantagens e desvantagens nos algoritmos existentes, de modo a poder-se escolher aquele ou aqueles que melhor se adequam ao projeto a desenvolver.

O motivo pelo qual se realizou este estudo sobre *clustering* é porque este projeto irá analisar coordenadas geográficas obtidas pela aplicação descrita na secção anterior, o *SenseMyFeup*. Como não existem dados já previamente identificados como pertencentes a determinadas classes, o modo de aprendizagem é não supervisionado. Este modo de aprendizagem identifica os padrões de cada classe de forma heurística [28], sem precisar de exemplos que permitam a aprendizagem. Como já foi referido anteriormente, não temos dados de treino nem qualquer outro conhecimento prévio de dados que nos leve a concluir qualquer tipo de resultados, portanto, a análise a efetuar terá por base uma aprendizagem não supervisionada. Dentro das técnicas de aprendizagem não-supervisionada como as regras de associação, cadeias de *markov* e *clustering*, uma vez que entre os trabalhos estudados previamente verificou-se a utilização desta última decidimos aprofundar a utilização destas técnicas no contexto de encontrar grupos de localizações (recolhidos de forma voluntária), grupos de atividades (com base em informação recolhida em diretórios de empresas) e grupos de ocorrências de eventos sociais.

O *clustering* é o processo de organização ou particionamento de um conjunto de dados, usualmente representados como um vetor de medidas ou um ponto num espaço multidimensional em *subsets* (ou *clusters*) baseados na sua semelhança [29] [30]. Um *cluster* é um conjunto de objetos, que são semelhantes entre si, partilhando características em comum e diferentes ou não relacionados entre objetos pertencentes a outros clusters [30].

Na figura 10 é possível observar as etapas que o *clustering* executa:

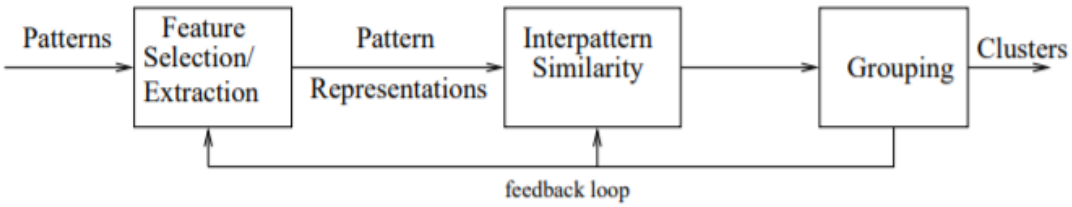


Figura 10 - Etapas de clustering. Fonte: [31]

- Representação de padrões e seleção de funcionalidades [31]: refere-se ao número de classes, ao número possível de padrões e ao número, tipo e escala das funcionalidades disponíveis no algoritmo de *clustering*. A extração de funcionalidades baseia-se no processo de identificar o subconjunto de dados mais eficaz a usar no *clustering*.
- Definição da medida de proximidade do padrão apropriado [31]: nesta etapa procede-se à identificação da medida de distância a usar. A distância euclidiana é a medida mais comum para que se consiga refletir a semelhança ou não semelhança entre dois padrões.
- *Clustering* ou agrupamento [31]: execução do algoritmo em si.
- Abstração de dados, se necessário [31]: é o processo de extrair um conjunto de dados representativo, ou seja, uma amostra.
- Avaliação do *output*, se necessário [31]: validação dos resultados obtidos.

O suporte de grandes volumes de dados, habilidade para lidar com diferentes tipos de atributos, descoberta de *clusters* com diferentes tipos de formas e capacidade de lidar com *outliers* são algumas das funcionalidades, que um método de *clustering* deve suportar [32]. A utilização de *clustering* para análise de dados tem vindo a ser empregada em diversos contextos, tais como:

- Biologia – Classificação de animais e plantas de acordo com as suas funcionalidades [29].
- *Business Intelligence* – Organização / agrupamento de clientes com comportamentos semelhantes para atividades de *marketing* [29] [33] [34].
- Pesquisas *Web* – *Clustering* pode ser usado para agrupar os resultados de pesquisas que se fazem na *Web*. Por exemplo, uma pesquisa que se realize com a palavra "filme" poderá retornar páginas relacionadas, como por exemplo, as *reviews*, *trailers*, atores entre outros. Cada categoria poderá ser dividida em subcategorias produzindo uma estrutura hierárquica que permite ao utilizador uma melhor exploração dos resultados [29].
- Reconhecimento de padrões em imagens – Descobrir clusters em sistemas de reconhecimento de dígitos escritos por humanos [29] [35].

- Clima: Para se entender o clima do planeta necessita-se que se encontre padrões na atmosfera e oceanos. Para esse fim, o *clustering* pode ser usado para encontrar padrões na pressão atmosférica de regiões polares e áreas de oceano que têm um impacto significativo no clima [29].
- Psicologia e Medicina: Uso na deteção de padrões numa distribuição espacial ou temporal de uma doença. Identificação de diferentes tipos de depressão [29].
- Planeamento de cidades – Identificar grupos de casas de acordo com o tipo de casa ou localização geográfica. A utilização de *clustering* neste âmbito tem um papel importante no entendimento de padrões de mobilidade humana [29] [36].

Em [37] [38] [39] [40], os autores sugerem que os métodos de *clustering* usados para análise de dados espaciais podem ser classificados dentro das seguintes categorias: particionados, hierárquicos, baseado em densidade e em grelha.

2.3.1 Métodos particionados

Provavelmente a classe de algoritmos de *clustering* mais conhecida serão os métodos particionados. Dado o número de partições a construir, o algoritmo cria uma estrutura de particionamento inicial. Depois usa uma realocação iterativa, que tenta melhorar o particionamento movendo objetos de um grupo para o outro [41]. O objetivo é encontrar um particionamento que, para um determinado número de clusters especificado como parâmetro de entrada, otimiza o critério de particionamento.

De uma forma mais formal, dado um conjunto de dados, *D*, contendo *n* objetos e *k*, como o número de clusters a formar, um algoritmo de particionamento organiza os objetos (*n*) em *k* partições, onde cada partição representa um cluster [42].

2.3.1.1 K-Means

O *K-Means* é um dos mais simples algoritmos não supervisionados para *clustering*. Para além de um conjunto de dados, o algoritmo também necessita de saber o número de *clusters* que terão de ser gerados. Este algoritmo particiona os dados em *k* clusters, em que cada cluster é representado pelo seu centroide ($C_1, \ldots, C_k$). O centroide pode ser definido como o ponto médio representativo de todas as instâncias pertencentes ao *cluster*. O algoritmo tem como objetivo minimizar a função objetiva: $J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^j - C_j||^2$, onde $||x_i^j - C_j||^2$, representa a medida de distância escolhida entre um ponto, x_i^j, e o centro do *cluster*, C_j [43]. Estes são os principais passos do algoritmo *K-Means*:

1. Atribuição do valor *k*, especificado pelo utilizador, que corresponde ao número de *clusters* que se pretendem encontrar
2. Criação de *centroides*. São criados *k* centroides em localizações aleatórias para serem os pontos representativos de cada *cluster*.

3. Cada ponto é atribuído ao *cluster*, que contém o centroide mais próximo dele, de acordo com a distância euclidiana.
4. Os centroides são de novo recalculados com base nos novos *clusters*.
5. Os passos 3 e 4 são repetidos iterativamente, até que um estado de convergência seja encontrado. Quando esse estado for atingido, por mais iterações que haja, os centroides permanecem na mesma posição.

Supondo um exemplo prático com $k = 2$, significa que se pretende gerar 2 *clusters*, logo o algoritmo cria aleatoriamente dois pontos, P_1 e P_2, denominados também como centroides. Na figura 11 podemos observar o próximo passo do algoritmo, a identificação dos pontos mais próximos em relação a P_1 e P_2 (assinalados como triângulos na figura 11, amarelo e vermelho, respetivamente). Com base na distância euclidiana, o algoritmo calcula as distâncias e determina os pontos mais próximos de P_1 e P_2.

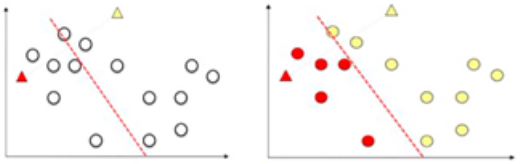


Figura 11 - Identificação dos pontos mais próximos de cada centroide. Fonte:[44]

Depois destes passos temos então no final desta iteração, os primeiros dois *clusters* formados. De seguida o algoritmo reformula a posição do centróide para ambos os *clusters*. A nova posição do centróide tem por base, a média as posições de cada ponto existente em cada cluster internamente. Para cada cluster o algoritmo calcula a posição central em relação aos pontos associados. De seguida, o algoritmo volta a medir as distâncias para cada centróide e a identificar quais os pontos mais próximos de cada um. Por fim, o algoritmo termina quando verifica que após reformular os centros do centróides e a calcular as distancias para determinar quais os pontos mais próximos de cada um, os pontos associados aos *clusters* anteriormente identificados no passo anterior são os mesmos. Após algumas iterações, o algoritmo fica num estado em que independentemente das iterações que possa fazer a seguir, os mesmos centroides permanecerão no mesmo sítio (estado de convergência), visível na figura 12.

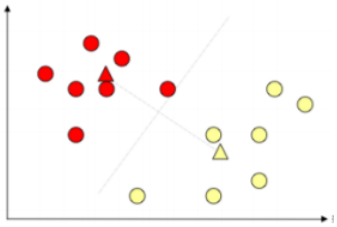


Figura 12 - K-Means: Estado de Convergência. Fonte:[44]

A utilização deste algoritmo apresenta algumas desvantagens como:

- ser sensível à configuração inicial, pois diferentes partições que sejam criadas inicialmente podem gerar resultados diferentes de *clustering*;
- a dificuldade de determinar o número de *clusters* que terão de ser gerados, nos casos de estudo em que não se sabe *à priori*, o número de *clusters* pretendidos;
- apresenta dificuldades para identificar *clusters*, que não tenham formas esféricas;
- não consegue lidar também, com a presença de *outliers*.

Como principais vantagens, o algoritmo é simples de implementar e permite uma fácil interpretação dos resultados [45].

2.3.1.2 *K-Medoids*

O *K-Medoids* ou PAM (Particionamento Através de *Medoids*) [46] utiliza *medoids* ao contrário do método apresentado anteriormente, que utiliza centroides. Os *medoids* são objetos representativos de um conjunto de dados, ou um cluster com um conjunto de dados cuja dissimilaridade média para todos os objetos no cluster é mínima. A principal diferença de um *medoid* para um centroide é que os *medoids* são sempre restritos a serem membros do conjunto de dados. Para além de representarem o cluster pertencem também ao conjunto de dados, ao contrário do centroide.

Como dados de entrada, o algoritmo necessita de saber o número de *clusters*, *k*, a serem gerados e um conjunto de dados numéricos. De seguida, dentro dos pontos existentes que existem no conjunto de dados, são escolhidos *k* pontos aleatoriamente como representativos do grupo de clusters (*medoids*). Para cada *medoid* são calculadas as distâncias euclidianas em relação ao resto dos pontos existentes com o intuito de verificar quais os pontos mais próximos de cada *medoid*. Depois de já se possuir os pontos para cada cluster e respetivo *medoid* é usada uma função de custo, que tem como objetivo medir a distância total entre os pontos, internamente nos *clusters*. Desta forma o algoritmo mede o *clustering* obtido, para conseguir quantificar o resultado obtido se utilizar um determinado ponto como *medoid*. O algoritmo prossegue usando outro ponto como *medoid*, até que todos tenham sido utilizados. No fim de cada iteração o algoritmo decide, através da função de custo, se o resultado obtido utilizando aquele determinado ponto como *medoid* apresenta melhor resultados que alguma das soluções já conseguidas. Se o resultado for melhor, o algoritmo atualiza as suas melhores soluções até ao momento.

O algoritmo *K–Medoids*, apresenta duas vantagens comparativamente ao *K-Means*, em primeiro lugar é mais robusto na presença de *noise* e *outliers*, porque o *medoid* é menos influenciado por *outliers* ou outros valores, do que no caso do centroide, em que é calculado baseado na média dos pontos pertencentes ao cluster. Em segundo, o *K-Medoids* não apresenta limitações no tipo de atributos. No entanto, o processamento requer mais custo em termos de performance e em *datasets*, com volume de dados elevado e, o algoritmo não é escalável [43].

2.3.1.3 CLARA

O algoritmo CLARA (*Clustering LARge Applications*) [47] é um algoritmo particionado, que em vez de usar todo o conjunto de dados, extrai subconjuntos aleatórios, aplicando o PAM a cada subconjunto e retorna o melhor resultado de *clustering* como *output*. As principais diferenças entre o PAM e o CLARA é que o PAM procura pelo melhor *K-Medoids* ao longo de todo o conjunto de dados, enquanto que neste algoritmo procura pelo melhor *K-Medoids* ao longo do subconjunto escolhido [42] [43].

2.3.1.4 CLARANS

O CLARANS (*Clustering Large Applications based on RANdomized Search*) [48] é uma versão melhorada do CLARA, que usa múltiplas amostras. Enquanto que o CLARA extrai um conjunto de amostras no início de cada pesquisa, o CLARANS extrai dinamicamente amostras de vizinhos e o processo de *clustering* pode ser apresentado com a procura de um grafo, onde cada nodo é uma potencial solução. Esta diferença tem como principal vantagem não limitar a pesquisa a uma determinada área. Requer mais dois parâmetros de entrada em relação ao PAM: *numlocal* e *maxneighbor*. O *numlocal* indica o número de amostras que devem ser extraídas. O *maxneighbor* representa o número de nodos vizinhos que cada nodo tem de ser comparado [43].

A principal desvantagem do CLARANS é assumir que todos os dados estão em memória. No entanto é mais eficiente e escalável que o PAM e o CLARA [43].

2.3.2 Métodos hierárquicos

A utilização de métodos hierárquicos tem como finalidade a decomposição hierárquica de um conjunto de dados em *n* objetos, como por exemplo a formação de uma árvore de *clusters* com diferentes níveis de particionamento a partir de um conjunto de dados base. Poderão ser classificados em duas categorias, que tem por base a forma de como a decomposição hierárquica é formada: aglomerativos (*bottom-up*) ou divisivos (*top-down*).

Quando a decomposição hierárquica é aglomerativa os objetos vão sendo atribuídos a grupos (*clusters*) e recursivamente esses mesmos grupos são combinados entre si com base na semelhança existente entre ambos. O processo continua iterativamente a formar e a juntar *clusters* até que uma determinada condição de paragem do algoritmo seja atingida. Por outro lado, quando a decomposição hierárquica é considerada divisiva, o algoritmo começa por atribuir todos os objetos a um único *cluster*. Posteriormente vai dividindo recursivamente o *cluster* em *clusters* mais pequenos, até que também uma determinada condição de paragem seja atingida, como por exemplo, se um determinado número de *clusters* for atingido, ou se o diâmetro de cada *cluster* atingir um determinado valor.

2.3.2.1 DIANA

O DIANA (*DIvisive ANAlysis Clustering*) [49] é um método hierárquico divisivo. Tal como indica um método hierárquico divisivo, inicialmente todos os objetos de dados usados são atribuídos a um único *cluster*. A divisão começa pelo dado que é menos semelhante em relação a todos os outros, sendo assim selecionado para pertencer a um novo *cluster*. Posteriormente, são de seguida selecionados os objetos que são mais semelhantes a este novo *cluster,* do que ao que foi inicialmente criado. Aqueles que forem mais semelhantes são transpostos para este novo *cluster*. Como principais vantagens este algoritmo, não necessita de previamente indicar o número de clusters a serem formados e é de fácil implementação. Como principais desvantagens, o algoritmo não consegue desfazer passos que tenham sido dados anteriormente e é sensível à presença de ruído.

2.3.2.2 CURE

O CURE (*Clustering Using REpresentatives*) [50] é um método hierárquico aglomerativo, que em vez de usar um centroide para representar um *cluster*, usa um determinado número fixo de pontos para representar esse *cluster*, os denominados pontos representativos. O uso destes pontos permite que o algoritmo consiga capturar *clusters* com formas aleatórias. As principais funcionalidades deste algoritmo são:

- Capacidade de reconhecer clusters com formas aleatórias.
- Robustez na presença de *outliers*
- Capacidade para lidar com um volume grande de dados.

O algoritmo começa por identificar um conjunto de pontos (mais afastados possível entre si dentro do *cluster*), que possam representar um *cluster*. De seguida, redefine a posição dos pontos representativos para uma posição mais central entre ambos, de acordo com o valor. No próximo passo, o algoritmo combina os dois *clusters* mais próximos num só, considerando que os *clusters* mais próximos são aqueles cujos pontos representativos se encontram mais próximos um do outro. Para cada cluster são novamente redefinidos os pontos representativos e o processo é repetido novamente até que o número de *clusters* a encontrar seja atingido. Na figura 13 é apresentado um exemplo da aplicação do algoritmo, acima descrito.

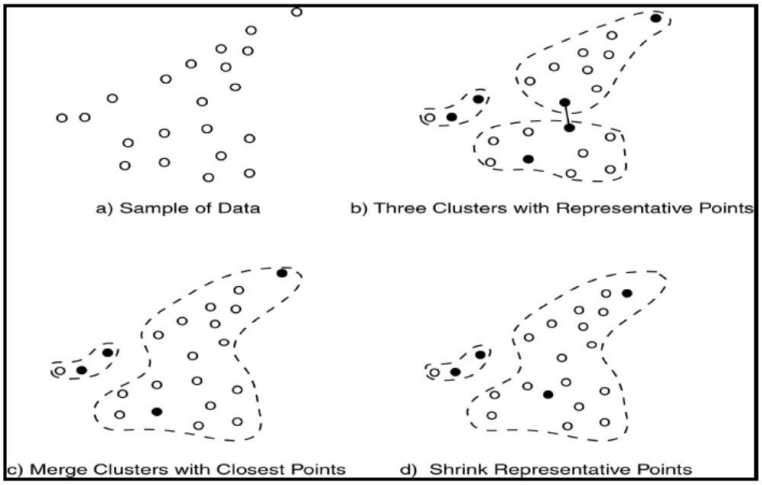


Figura 13 - Representação do algoritmo CURE. Fonte: [51]

2.3.2.3 BIRCH

O BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*) [52] é um algoritmo hierárquico aglomerativo que foi implementado com o principal objetivo de ser capaz de lidar com um grande volume de dados. Este algoritmo aplica um *clustering* dinâmico e incremental sobre objetos e apenas analisa os dados uma vez, sendo o suficiente para realizar *clustering*. Constrói de forma incremental, o CF (*Clustering Feature*), uma estrutura hierárquica, em forma de árvore. O algoritmo é composto por duas fases. A primeira é responsável por criar a *clustering feature* em que cada entrada corresponde a um *cluster* de objetos caracterizados por três variáveis: número de pontos do *cluster*, soma linear dos pontos e a soma quadrática dos pontos. Esta estrutura é então responsável por armazenar dados estatísticos sobre cada *cluster*. A figura 14 é representativa de uma *clustering feature.*

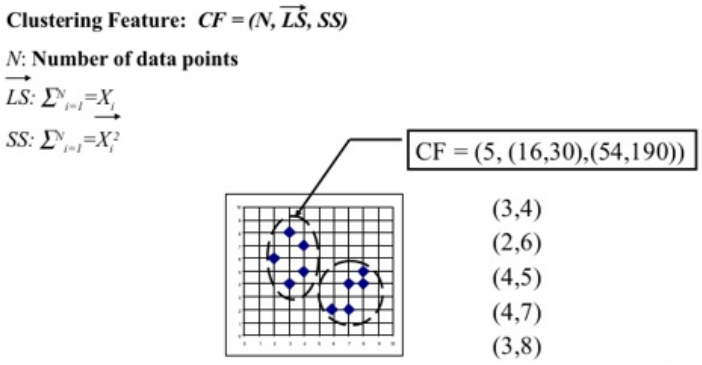


Figura 14 - Representação de uma clustering feature. Fonte: [32]

A segunda fase tem como objetivo a construção das folhas restantes da *Clustering Feature* usando um algoritmo de *clustering* arbitrário. Para cada ponto o algoritmo encontra a folha mais próxima e adiciona-o à folha, tendo que posteriormente reorganizar a *clustering feature*. A *clustering feature* detém dois parâmetros, o número máximo de

filhos e o diâmetro máximo de *sub-clusters* nas folhas. Se com a entrada atingir o diâmetro máximo da folha, esta tem de ser dividida entre os pais e assim sucessivamente. O BIRCH é sensível à ordem de inserção dos dados e os *clusters* que são formados tendem a ser esféricos.

2.3.2.4 ROCK

O algoritmo ROCK (RObust Clustering using linKs) [53] é um algoritmo aglomerativo que se baseia na noção de *links* (ligações), como forma de medir a semelhança entre um par de pontos com atributos nominais. Em [53], os autores demonstraram que, para lidar com um conjunto de dados que contêm atributos categóricos, a utilização dos tradicionais métodos de *clustering* que utilizam distâncias entre pontos não são apropriados para este tipo de dados. Por isso, propuseram o algoritmo ROCK, que aplica ligações (*links*) e não distâncias entre pontos para formar *clusters*. Neste algoritmo a semelhança entre clusters é baseada no número de pontos de diferentes *clusters*, que têm pontos vizinhos em comum [54]. Um determinado par de pontos é considerado vizinho se a semelhança entre eles exceder um determinado *threshold*, que é previamente definido. Pontos que pertençam ao mesmo cluster terão um grande número de vizinhos em comum. Este algoritmo pode ser dividido em 3 etapas, como se pode observar na figura 15.

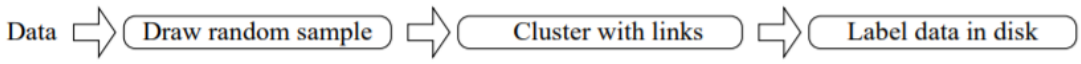


Figura 15 - Etapas do clustering ROCK. Fonte: [53]

Numa primeira fase, a denominada *random sampling*, uma amostra é retirada do *dataset*, ou seja, um determinado grupo de pontos é escolhido de forma aleatória. De seguida, o algoritmo ROCK é aplicado ao conjunto de dados extraído na fase anterior. Na última fase, os clusters que foram formados são usados para distribuir os pontos restantes que não foram selecionados anteriormente na amostra aleatória. Este algoritmo permite também uma boa escalabilidade em conjuntos de dados com bastante volume de dados. Necessita dos seguintes parâmetros de entrada: um conjunto de dados, o número de *clusters* que terão de ser encontrados e o *threshold* usado para considerar um determinado par de pontos como vizinhos.

2.3.3 Métodos baseados em densidade

Os métodos de *clustering* baseados em densidade permitem a identificação de grupos de objetos que formam regiões de densidade elevada. Uma das principais razões do seu desenvolvimento tem a ver com a capacidade de determinar clusters com formas aleatórias. Consideram a existência de *clusters* em regiões densas de objetos que são separados por outras de menor densidade, o denominado *noise* ou ruído [30]. Este método de *clustering* é bastante usado para projetos de *data mining* com o intuito de extrair e identificar padrões [55].

2.3.3.1 DBSCAN

O algoritmo DBSCAN (*Density Based Spatial Clustering of Applications with Noise*) [56] foi implementado com o objetivo de ser capaz de identificar *clusters* com diversas formas e tamanhos aleatórios, nomeadamente em conjuntos de dados espaciais. Regiões com elevada densidade de pontos indicam a presença de *clusters*, por sua vez, regiões com uma baixa densidade apontam para a existência de ruído. Este algoritmo necessita para além de um conjunto de pontos, duas variáveis, o épsilon e o número mínimo de pontos. O *épsilon* representa a distância máxima entre dois pontos para que sejam considerados vizinhos, enquanto que o número mínimo de pontos indica ao algoritmo, o número mínimo de pontos vizinhos que devem originar um *cluster*. A figura 16 fornece um exemplo de clusters que foram obtidos usando o algoritmo DBSCAN.

Figura 16 - Clusters obtidos usando o algoritmo DBSCAN. Fonte:[56]

Para qualquer ponto existente, a densidade pode ser medida através do número de pontos que existem dentro de uma determinada área que existe à volta do ponto de acordo com o raio (épsilon) especificado.

De acordo com o algoritmo, um ponto pode ser classificado, como sendo: *Core*, *border* ou *noise point.* Um ponto é definido como *core* se tem, dentro dos limites de distância máxima, de acordo com o raio definido, o número mínimo de pontos especificado. Um *border point* é aquele que, apesar de conter menos pontos do que aqueles que são estabelecidos como o mínimo para formar um cluster dentro do raio especificado, está contido na vizinhança de um *core point*. Por fim, um *noise point* é qualquer ponto que não esteja incluído nas duas classificações anteriores, todos os pontos que não sejam alcançáveis são considerados como *outliers* ou *noise*. A figura 17 representa as três classificações então usadas para classificar um ponto segundo o algoritmo DBSCAN.

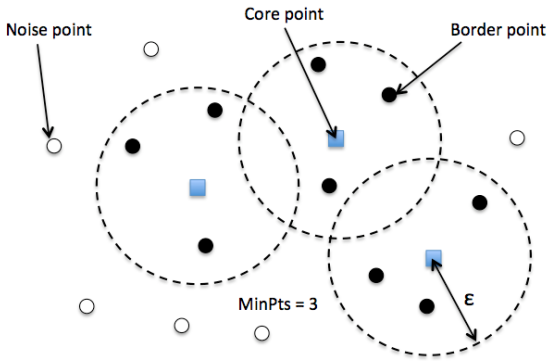


Figura 17 - Representação de um core, noise e border point [13]

Um *cluster* gerado pelo algoritmo satisfaz duas propriedades: todos os pontos de um cluster estão conectados entre si. Se um ponto é alcançável por um *core point*, então ele também pertence ao *cluster*.

Antes de apresentar uma breve descrição do algoritmo em si é importante primeiro definir dois conceitos: *directly density-reachable* e *density-reachable.* Dados dois pontos, *p* e *q*, em que, *q* é um *core point* é diretamente alcançável pela densidade ou *directly density-reachable* em relação a *p*, se *p* estiver presente na vizinhança de *q*. O conceito de *density-reachable*, dados os mesmos pontos, *p* e *q* do exemplo anterior, o ponto *p* é alcançável pela densidade em relação a *q*, se existir uma cadeia de pontos, de *p* a *q* diretamente alcançáveis pela densidade. A figura 18 representa os dois conceitos, à esquerda o conceito de *directly density-reachable* e à direita o *density-reachable*.

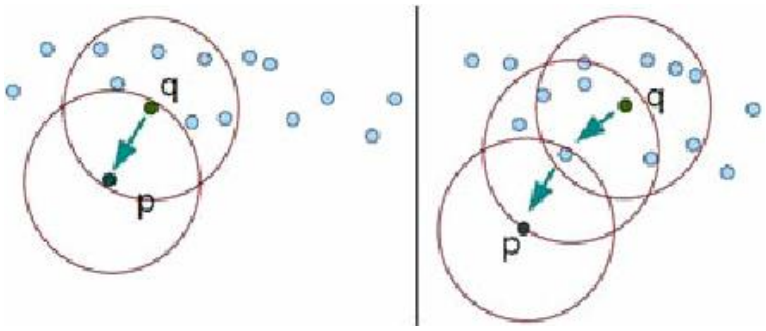


Figura 18 - Conceitos de directly density reachable e density-reachable. Fonte: [57]

O algoritmo baseia-se nos seguintes passos:

1. Aleatoriamente escolhe um ponto, para melhor entendimento definimos *p*, como o ponto escolhido.

[13] http://dev.wode.ai/repo/scipy-2016-sklearn/notebooks/21%20Unsupervised%20learning%20-%20Hierarchical%20and%20density-based%20clustering%20algorithms.ipynb (Último acesso em: 10/12/2018)

2. De seguida, o algoritmo extrai os pontos alcançáveis pela densidade (*density-reachable*).
3. Se *p* é um *core point*, um *cluster* é formado.
4. Se *p* é um *border point*, nenhum ponto é diretamente alcançável pela densidade e o algoritmo visita aleatoriamente outro ponto dentro do conjunto de dados existente.
5. O processo continua até que todos os pontos sejam processados.

A possibilidade de encontrar clusters com formas aleatórias, a noção e capacidade de lidar com *outliers* e *noise*, o facto de não se ter de especificar o número de *clusters* que terão de ser gerados são vantagens úteis para projetos em que não se tem informação relativa aos dados que se pretendem obter. No entanto, este algoritmo apresenta desvantagens tais como a dificuldade em lidar com *datasets*, que possuam diferentes níveis de densidade e em saber quais os parâmetros ideais a usar (e*psilon* e o número mínimo de pontos), de acordo com o contexto do problema, uma vez que a sua escolha tem um impacto determinante na obtenção de resultados.

2.3.3.2 OPTICS

O algoritmo OPTICS (*Ordering Points to Identify the Clustering Structure*) [58] é uma extensão do algoritmo DBSCAN, apresentado na secção anterior, que requer igualmente a existência das duas variáveis, o *épsilon* e o número mínimo de pontos, para além de um conjunto de pontos. Uma das principais diferenças do OPTICS, para com o DBSCAN, advém da menor sensibilidade à parametrização definida, ou seja, diferentes valores de épsilon e de número mínimo de pontos produzem menores diferenças entre os resultados de *clustering*. A ideia base do OPTICS começa por inicialmente identificar e processar os pontos com maior densidade. O algoritmo produz clusters com base em duas variáveis: a *core distance* e a *reachability distance*. Estas duas variáveis já anteriormente descritas na secção anterior, permitem a implementação e visualização do gráfico de *reachability distance*, que permite ao OPTICS a identificação de *clusters*.

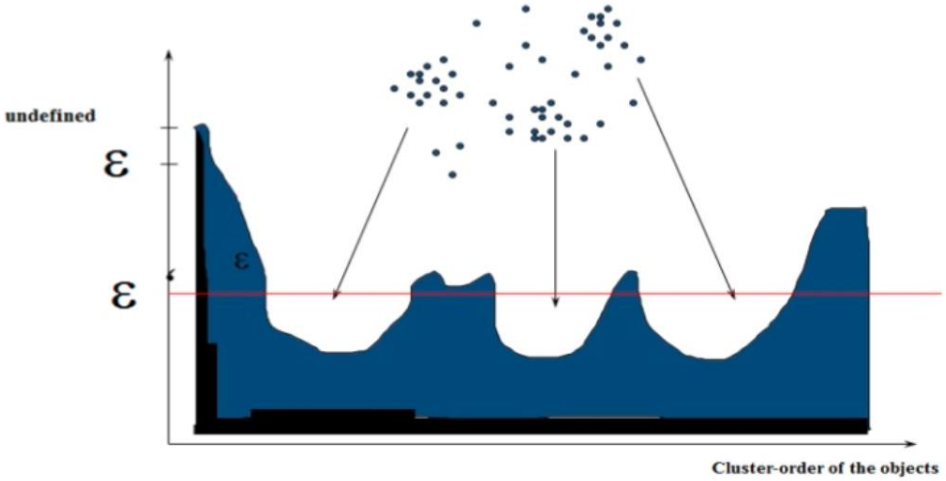


Figura 19 - Exemplo de um gráfico de reachability distance [59]

A existência de vales percetíveis na figura 19 acima representada, significa a presença de *clusters*. A *core distance* é a distância mínima do raio que um ponto necessita para encontrar à sua volta o número mínimo de pontos indicado ao algoritmo. Dado um ponto, *p* escolhido aleatoriamente do conjunto de pontos, o algoritmo verifica qual a distância mínima necessária para encontrar o número mínimo de pontos definido. A *core distance* pode ter dois valores distintos, *undefined*, se para a distância máxima definida não foi possível encontrar o número mínimo de pontos pretendido para formar um *cluster*, ou então o valor mínimo encontrado que tenha sido suficiente para encontrar à volta o número mínimo de pontos pretendido. Em relação à *reachability distance*, dados dois pontos, *p* e *q*, o algoritmo procura definir o valor mínimo de raio, que permita que *p* seja alcançável pela densidade até *q*. Poderá também assumir valor de *undefined*, se *q* não for um *core point.* Caso não o seja, o valor da variável de *reachability distance* será o valor máximo destas duas componentes: a *core distance* de *p*, entre a distância de *p* a *q*.

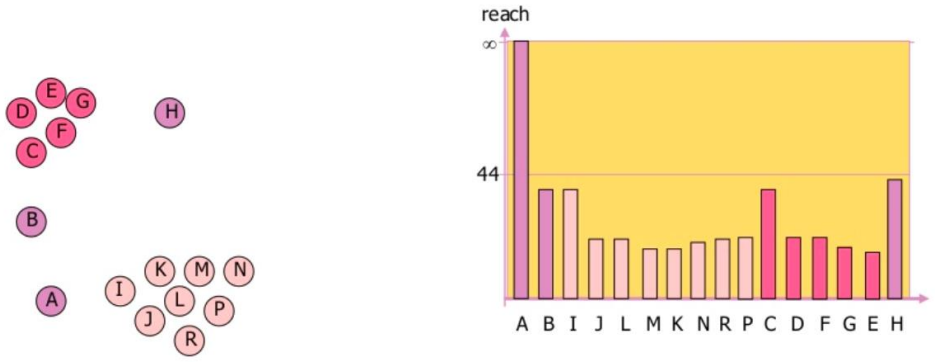


Figura 20 - Aplicação do algoritmo OPTICS [59]

Na figura 20 é representado um exemplo prático da aplicação do algoritmo OPTICS. Podemos observar que existe a presença de dois *clusters*: D,E,G,F e C correspondem a

um *cluster* e que K,L.M,N,P,R,J e I pertencem a outro. A, B e H são classificados como ruído.

Para além de também conseguir identificar clusters com diversas formas aleatórias, tal como o DBSCAN, o OPTICS consegue lidar com conjuntos de dados com diversos níveis de densidade, sendo desta forma igualmente capaz de identificar bons *clusters* [60].

2.3.3.3 HDBSCAN*

O HDBSCAN* (*Density-Based Clustering Based on Hierarchical Density*) [61] é um algoritmo que converte o DBSCAN, que é um algoritmo baseado em densidade, num algoritmo de *clustering* hierárquico, podendo ser visto também como um melhoramento no algoritmo OPTICS. Combina duas aproximações, densidade e hierarquia. Baseia-se no princípio de densidade, na identificação/separação de *clusters* e baseia-se na hierarquia ao formar uma hierarquia de *clustering*, em vez de partições, num nível de densidade simples. O algoritmo tem como parâmetro de entrada, apenas o número mínimo de pontos necessário para gerar um *cluster* gerando posteriormente uma hierarquia de *clustering* baseada em densidade, que contém todos os limites de densidade, sendo por isso capaz de identificar *clusters* com diferentes níveis de densidade (diferentes valores de épsilon). A hierarquia pode ser visualizada usando, por exemplo um dendrograma. Na figura 21 está presente o algoritmo do HDBSCAN*.

Algoritmo 1 – *HDBSCAN** Passo Principal

Entrada: Seja **X** um conjunto de dados multidimensional (ou **D** a matriz de proximidade entre os objetos de **X**) e m_{pts} o número mínimo de objetos em uma determinada vizinhança.

Saída: Hierarquia de *HDBSCAN**

1: Calcula a distância *core* (utilizando m_{pts}) para todos objetos do conjunto de dados **X**.
2: Calcula a *MST* de um grafo $G_{m_{pts}}$ (baseado na distância de accessibilidade mútua).
3: Estende a *MST* para obter MST_{est}, adicionando para cada vértice (objeto) um "*self-loop*"com a distância *core* do objeto correspondente.
4: Ordena MST_{est} em ordem decrescente de peso.
5: Para a raiz da árvore, atribua a todos os objetos o mesmo rótulo.
6: **enquanto** Existem arestas em MST_{est} **faça**
7: Remove-se arestas de MST_{est} de maior pesos.
8: Antes da remoção, define-se a escala do dendrograma com o valor corrente do nível hierárquico com o peso da aresta a ser removida.
9: Após a remoção, atribui-se rótulos para os componentes conectados (*CC*) que contêm os vértices finais da aresta removida para obter o próximo nível hierárquico:
10: **se** componente conectado em *CC* contém pelo menos uma aresta **então**
11: Atribui-se um rótulo de novo grupo ao componente.
12: **senão**
13: Atribui-se ao componente um rótulo nulo (*ruído*).
14: **fim se**
15: **fim enquanto**

*Figura 21 - Algoritmo HDBSCAN**

Uma das características mais vantajosas do HDSBCAN* é a sua capacidade de identificar *clusters* com diferentes densidades, o que não acontece com o DBSCAN, O facto de para além de um conjunto de dados, só necessitar do número mínimo de pontos para formar

um *cluster* é também uma característica interessante, que melhora a seleção de parâmetros tornando-a mais intuitiva para o utilizador.

2.3.3.4 DENCLUE

O DENCLUE (*DENsity-based CLUstEring*) [62] é um algoritmo de densidade que estima a densidade local de um conjunto de dados usando uma função matemática muito semelhante aos estimadores da função de densidade de probabilidade do *kernel*. No DENCLUE, os autores introduziram uma nova definição: a *influence function*, que tem como finalidade descrever o impacto de um ponto na sua vizinhança. A *influence function* é copiada para cada ponto produzindo assim a função de densidade. Os clusters podem ser determinados matematicamente identificando os denominados "atractores de densidade" que representam o máximo local da densidade total da função (a soma da *influence function* em todos os pontos de dados).

O algoritmo tem por base duas etapas fundamentais. A primeira etapa corresponde à fase de pré-processamento, no qual é contruído um mapa de dados usando hipercubos. Cada hipercubo contém: número de pontos, ponteiros para pontos e a soma dos valores de dados. Por fim, apenas os cubos que contêm esta informação são armazenados numa *B+ tree*. Na segunda etapa, o algoritmo foca-se apenas nos cubos da etapa anterior, que contém bastante informação, bem como aqueles que estão ligados entre si. De seguida determina os "atractores de densidade" para todos os pontos usando *hill climbing* [32].

Como principais vantagens este algoritmo apresenta a capacidade de identificar *clusters* com diversas formas aleatórias e uma sólida base matemática. Também é considerado um bom algoritmo para lidar com conjuntos de dados que tenham um elevado número de pontos não pertencentes a nenhum *cluster* (*noise*) [63].

O algoritmo apresenta como desvantagens a difícil seleção de valores para as variáveis de densidade e *threshold* de *noise*. Ambos necessitam de ser cuidadosamente calculados, uma vez que as escolhas dos seus valores afetam significativamente a qualidade dos resultados obtidos [63].

2.3.4 Métodos baseados em grelha

Os métodos baseados em grelha exploram a utilização de uma estrutura de dados em grelha. Têm como objetivo, o particionamento do espaço de dados, num determinado número de colunas, formando a denominada estrutura em grelha. Após a criação desta estrutura todas as operações de *clustering* serão realizadas para cada coluna de forma isolada. A principal vantagem da utilização desta aproximação, baseia-se na sua alta eficiência e escabilidade, uma vez que o algoritmo de procura de *clusters* se concentra apenas no grupo de objetos identificado em cada coluna. Depois da criação desta estrutura, os algoritmos baseados em grelha, calculam a densidade de cada coluna, de

seguida, ordenam-nas de acordo coma sua densidade, identificam os centros de cada cluster e por fim, efetuam cruzamento com colunas vizinhas [38].

2.3.4.1 STING

O STING (*Statistical Information Grid-based*) [64] é um método baseado em grelha usado para efetuar *clustering* em bases de dados espaciais, nomeadamente para facilitar diversos tipos de pesquisas espaciais. O espaço de dados é dividido em colunas retangulares. Existem diferentes níveis de colunas retangulares correspondentes a diferentes resoluções formando uma estrutura hierárquica. A figura 22 representa um exemplo de uma estrutura hierárquica formada por este algoritmo.

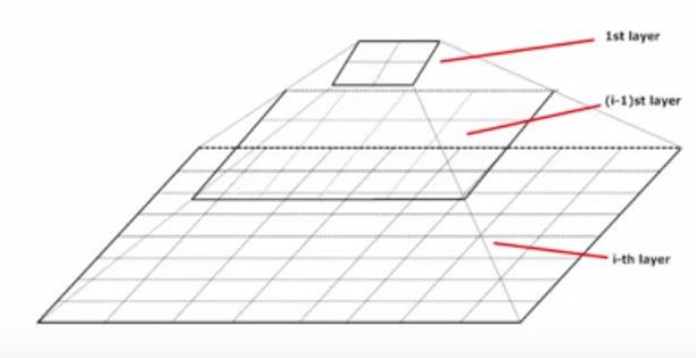


Figura 22 - Estrutura hierárquica formada pelo algoritmo STING [14]

Cada coluna pertencente a um nível mais alto é particionada de forma a agregar um determinado número de colunas num nível mais inferior. A informação estatística de cada coluna é calculada (como por exemplo, a média e o número de pontos existentes na coluna), armazenada e usada para responder a determinadas pesquisas. O algoritmo é executado segundo uma abordagem *top-down* (cima para baixo), em que para cada coluna é calculado um intervalo de confiança, com base na informação estatística da coluna. O valor calculado é usado para saber se a coluna é relevante para a pesquisa ou não. Para as colunas que forem relevantes o algoritmo continua a sua execução utilizando as células-filhas até chegar ao nível mais baixo da hierarquia [65].

2.3.4.2 CLIQUE

O CLIQUE (*Clustering in QUEst*) [41] é um método baseado em grelha, mas também se pode considerar num mesmo nível, um método baseado em densidade. Ele é baseado em grelha, porque divide o espaço de dados numa estrutura em grelha e estima a sua densidade calculando o número de pontos em cada coluna existente na grelha.

---

14 https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/9781783982103/6/ch06lvl1sec62/recommendation-system-and-sting (Último acesso em: 12/12/2018)

O algoritmo tem como parâmetros de entrada 2 valores, um número máximo de intervalos, *xi*, e um *threshold* de densidade, *tau*. O CLIQUE identifica *clusters* começando primeiro por dividir cada dimensão em *xi* intervalos de largura igual. Começando por usar apenas uma dimensão, são guardados os intervalos, onde a densidade é maior do que o *threshold* previamente definido. De seguida, passando a utilizar duas dimensões, cada conjunto é analisado: se houver dois intervalos de interseção nessas duas dimensões e a densidade na interseção desses intervalos for maior que *tau*, a interseção será novamente identificada como um *cluster*. Este processo é repetido para todos os conjuntos de três, quatro, cinco...dimensões. Após cada etapa, os clusters adjacentes são substituídos por um *cluster* comum e, no final, todos os clusters são gerados.

As principais características deste algoritmo estão relacionadas com a capacidade de encontrar *clusters* com formas aleatórias, capacidade de encontrar qualquer número de *clusters* usando qualquer tipo de dimensão e o número de *clusters* a determinar não é pré-determinado. Como principais desvantagens, bem como em todas as aproximações em grelha, a qualidade dos resultados, depende muito da escolha apropriada do número das partições e tamanho da coluna de grelha [65].

3 URBY.SENSE

Neste capítulo é apresentado o projeto de investigação URBY.SENSE, que tem como objetivo estudar a mobilidade em situações fora da rotina (lazer, social, entre outros), extraindo padrões a partir de múltiplas fontes de dados. Os seguintes padrões de mobilidade são de grande interesse: locais com significância importante, modos de transporte, padrões de trajetórias e escolha do meio de transporte para chegar ao destino. Dados obtidos através de dispositivos ubíquos, combinados com dados extraídos das plataformas de redes socias fornecem uma aproximação mais próxima da realidade, que pode ser combinada com dados a partir das fontes tradicionais (*surveys*, registos de sistemas de transporte e dados estáticos), para o planeamento e gestão eficiente da mobilidade. Quando consideramos estes dados de forma isolada, cada uma destas fontes de dados tem algo em falta, então a combinação e a integração de diversas fontes de dados pode permitir uma melhor análise aos sistemas de transporte, que por sua vez, permite melhorar a oferta dada aos cidadãos.

De seguida são apresentadas diversas tarefas que foram realizadas pela equipa de investigação do projeto de investigação URBY.SENSE, nomeadamente como foi realizada a recolha de dados, fusão, análise e fase de modelação de dados. O trabalho desenvolvido neste projeto de mestrado insere-se na componente de análise de dados do projeto URBY.SENSE e por fim são apresentadas as principais contribuições, que este projeto de mestrado tem para com o projeto de investigação URBY.SENSE.

3.1 Recolha dos dados

A campanha de recolha de dados, denominada *SenseMyFeup*, que utilizou a aplicação *SenseMyFeup*, ocorreu durante o mês de Abril de 2016 tendo participado nela estudantes, professores, investigadores e empregados da Faculdade de Engenharia da Universidade do Porto (FEUP). Aos participantes foi pedido que instalassem a aplicação *SenseMyFeup* nos seus *smartphones*. A aplicação automaticamente recolhe dados da mobilidade do utilizador durante as suas viagens e lança questionários opcionais inquirindo os participantes sobre o meio de transporte usado. Uma vez que a aplicação lida com informação sensível, procedimentos de anonimização de dados foram desenvolvidos com o intuito de proteger a privacidade de cada utilizador. Para incentivar a recolha, aumentando o número de participantes, a aplicação forneceu aos seus utilizadores estatísticas relacionadas com as suas viagens: as distâncias percorridas, duração e pegada digital *vs*. a média da comunidade. Para além destas funcionalidades, como forma de atrair mais participantes foram também atribuídos prémios baseados nos dias que cada utilizador contribuiu para o estudo.

A campanha conseguiu recolher dados sobre 301 participantes da FEUP. Na tabela 2 encontram-se representados os tipos de utilizadores que participaram na recolha e respetiva contagem.

Tabela 2 - Tipo de utilizadores na campanha SenseMyFeup

Tipo de Utilizador	Número	Percentagem (%)
Estudante de Licenciatura	90	30
Estudante de Mestrado	159	53
Estudante de Doutoramento	12	4
Professor	5	2
Pessoal não técnico	16	5
Investigador	18	6
Total	**301**	**100**

No que diz respeito às viagens realizadas por cada utilizador, os dados foram extraídos e armazenados em duas entidades: viagens e segmentos. Uma viagem é constituída por um ou mais segmentos e um novo segmento começa quando algum utilizador muda o seu meio de transporte. Cada viagem é constituída pela sua posição inicial geográfica (latitude e longitude), ou seja, onde começou, e por fim a posição geográfica final onde terminou (latitude e longitude). Cada segmento é igualmente constituído por uma posição geográfica inicial e outra final. As estimativas calculadas pela aplicação relativas ao modo de transporte foram automaticamente fornecidas com base na aceleração e velocidade de acordo com o perfil de cada utilizador. Para cada viagem realizada pelo utilizador é calculada a probabilidade da mesma ter sido realizada de carro, de bicicleta, de metro, autocarro ou a pé.

Depois da recolha de dados a fase que se seguiu foi a realização do pré-processamento de dados. Nesta fase foram removidos numa primeira fase, dados relativos a viagens, que tivessem sido realizadas fora da cidade metropolitana do Porto. De seguida foram removidos dados considerados como ruído. Na figura 23 podemos ver como a fase de pré-processamento se realizou. De seguida, com base nos dados já filtrados, a aplicação analisa cada segmento e deteta o meio de transporte usado pelo participante.

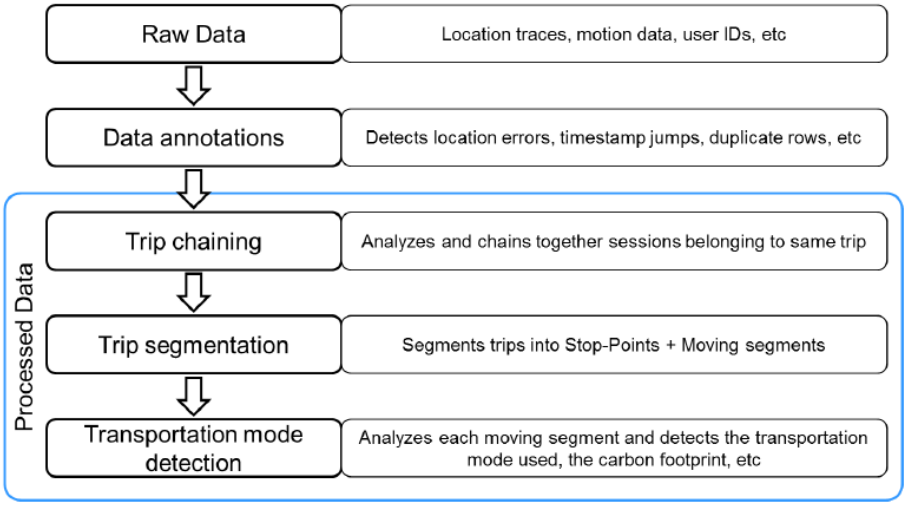


Figura 23 - Fase de pré-processsamento

O *dataset* final contém cerca de 9700 viagens, correspondendo no total a mais de 400.000 quilómetros recolhidos durante 3800 horas de processamento. Relativamente ao número de segmentos, foram extraídos cerca de 193696, em que cada um contém uma probabilidade sobre o meio de transporte que o utilizador usou durante o percurso. A tabela 3 representa o número de viagens, o meio de transporte utilizado e o tempo despendido em cada viagem por cada tipo de utilizador.

Tabela 3 - o número de viagens, o meio de transporte utilizado e o tempo despendido em cada viagem por cada tipo de utilizador

User	*Trips*		*Distance Travelled (average-km)*					*Time Spent (average-min)*				
	No.	%	Car	Bus	Metro	Bicycle	Foot	Car	Bus	Metro	Bicycle	Foot
Student 1st cycle	727	28.21	5.690	3.196	3.741	0.000	0.960	21.82	34.30	16.68	0.00	18.57
Student 2nd cycle	1234	47.89	5.665	3.853	7.420	4.586	0.925	25.60	17.55	33.06	44.18	16.76
Student 3rd cycle	244	9.47	5.752	5.483	1.628	0.000	1.456	21.12	22.19	9.57	0.00	23.05
Researcher	89	3.45	3.393	3.318	4.978	0.000	1.327	19.94	24.30	26.22	0.00	18.08
Teacher	35	1.36	15.227	14.835	0.000	0.000	0.668	49.89	84.50	0.00	0.00	11.63
Non-teaching personnel	102	3.96	8.421	4.069	0.000	0.000	0.952	31.00	50.52	0.00	0.00	18.92
Unknown	146	5.67	-	-	-	-	-	-	-	-	-	-
Total	2577	100										

Os participantes responderam ainda a 5300 questionários sobre o modo de transporte usado nas suas viagens. Para além da recolha de dados anteriormente referida foram extraídos dados das redes sociais, de forma a melhorar o estudo desenvolvido, usando para isso as *API's* associadas a cada plataforma, *Facebook*, *Instagram* e *Foursquare*. Nesta extração foram também incluídos dados do *Factual.com* do *infoPorto.pt*, que contêm dados sobre pontos de interesse (POIs) e de eventos da cidade metropolitana do Porto, respetivamente. O Factual é uma plataforma de dados aberta que tem foco em dados de localização geográfica sobre POIs. O Factual encoraja o *crowdsensing* e outras

aproximações colaborativas que permitam a recolha e partilha de dados. O *infoPorto* é um *website* que publica dados sobre eventos na cidade do Porto.

Depois de algum pré-processamento também realizado na extração destes dados conseguiu-se obter cerca de 224 eventos que se realizaram desde abril de 2016. As categorias correspondentes a cada evento, bem como a sua localização mais precisa pode ser visualizada na tabela 4.

Tabela 4 - Número de eventos e sua localização no Porto

	Total	*Weekdays*	*Weekends*	*Arts and Entertainment*	*Dance and Night Club*	*Market*	*Music Concert*	*Sports*
Espinho	0	0	0	0	0	0	0	0
Gondomar	4	0	4	3	0	0	1	0
Maia	0	0	0	0	0	0	0	0
Matosinhos	5	4	1	2	0	0	3	0
Porto	206	130	76	89	46	9	60	2
Póvoa de Varzim	1	1	0	1	0	0	0	0
Santo Tirso	2	0	2	2	0	0	0	0
Trofa	0	0	0	0	0	0	0	0
Valongo	0	0	0	0	0	0	0	0
Vila do Conde	1	0	1	0	0	0	1	0
Vila Nova de Gaia	5	3	2	3	0	2	0	0
Greater Porto	224	138	86	100	46	11	65	2

Dados obtidos sobre o ambiente foram também alvo de recolha. O projeto do *UrbanSense* [66] armazena dados sobre o clima e qualidade do ar. Obtidos através de 19 estações localizadas em locais estratégicos na cidade do Porto, para monitorizar o clima e extrair características do ambiente.

Todos os dados descritos nesta secção foram armazenados num servidor único e base de dados gerida pelo projeto URBY.SENSE, sem acesso público. O sistema de base de dados é o *PostegreSQL*, versão 9.6 com a extensão geográfica de análise (*postgis*).

O modelo de dados a analisar é composto pela existência de quatro esquemas: *Sensemycity, social_network, Public* e *envirionment*.

3.1.1 SenseMyCity

O esquema *SenseMyCity* armazena os dados filtrados e processados da aplicação *SenseMyFeup* usada na fase de recolha das viagens dos utilizadores na cidade do Porto. Na tabela 5 é possível perceber quais as tabelas presentes no esquema e respetiva função associada.

Tabela 5- Esquema SenseMyCity

Esquema	Tabelas
SenseMyCity	*Sessions*: informação anónima da recolha de dados por sessão, incluindo o *user daily ID*, primeira e última localizações, *timestamps*, locais, *wifi-beacons* e informação sobre o reconhecimento de atividades.
	Trips: criado pelo algoritmo do *trip_chainning*. Esta tabela agrega os percursos dos participantes por viagem, incluindo a primeira e última localização, *timestamps* e o total da distância percorrida. Mais informação pode ser obtida a partir da lista de sessões de recolha que foram identificadas como sendo parte da mesma viagem.
	Segments: contém informação dos segmentos percorridos na viagem, onde cada segmento corresponde a um único modo de transporte. Contém a primeira e última posição e *timestamp* associado, assim como a distância percorrida, métricas de velocidade e aceleração e o modo de viagem detetado com um nível de confiança associado a cada um.
	Travelmode_survey: lista dos modos de viagem usados em cada sessão de recolha, reportado pelo participante.
	Activity: contém a atividade realizada pelo utilizador detetada pelo serviço do reconhecimento de atividades da *Google*, contendo um intervalo de confiança para as atividades detetadas que poderão ser: Condução num veículo, andar de bicicleta, correr, andar e sentar.

As tabelas associadas ao esquema *SenseMyCity* podem ser consultadas na secção ANEXOS-B.

3.1.2 Social_network

O esquema *Social_network* é responsável pelo armazenamento de dados sobre POIs e eventos extraídos através das redes sociais. Na tabela 6 é possível perceber quais as tabelas e respetiva função associada.

Tabela 6 – Esquema Social_network

Esquema	Tabelas

Social_network	*Facebook_places*: armazena dados sobre locais no Porto obtidos a partir da API do *Facebook Places*[15].
	Factual_pois: contém dados sobre locais na cidade do Porto, obtidos diretamente a partir do factual.com
	Foursquare_venues: esta tabela contém dados sobre locais de interesse ou *venues* obtidos a partir do *foursquare*.
	Infoporto_events: contém informação sobre eventos na cidade do Porto a partir de abril de 2016, obtidos diretamente a partir do *infoporto*.
	Facebook_events: contém os dados associados aos eventos criados na rede social *Facebook*.

As tabelas associadas ao esquema *Social_network* podem ser consultadas na secção ANEXOS-C.

3.1.3 Public

O esquema *Public* é responsável pelo armazenamento de dados sobre localizações das paragens de metro e autocarro, bem como dados sobre a geometria da região. Na tabela 7 é possível perceber, quais as tabelas existentes no esquema e respetiva função associada.

Tabela 7 – Esquema Public

Esquema	Tabelas
Public	*Stctstops*: localizações das paragens de autocarro e correspondentes linhas de serviço.
	Metrostops: localizações de paragens de metro.
	Subregions: geometria da região usada para análise estática.

[15] https://developers.facebook.com/docs/places/web (Último acesso 08/12/2018)

Os atributos das tabelas associadas ao esquema *Public* podem ser consultadas na secção ANEXOS-D.

3.1.4 Environment

O esquema *Environment* armazena dados relativos ao clima e poluição da cidade do Porto obtidos pelo projeto *UrbanSense* [66]. Na tabela 8 é possível observar as principais tabelas e funções das mesmas.

Tabela 8 – Esquema Environment

Esquema	Tabelas
Environment	*Node:* lista dos identificadores físicos possíveis das unitades de dados de recolha *(DCUs)*
	Deployment: contém todos os *deployments* realizados na cidade bem como todas as estações de alta qualidade disponíveis.
	Location: todas as localizações disponíveis com medidas da plataforma do *UrbanSense* e estações.
	Noise: esta tabela incorpora dados a partir dos sensores de ruído (deteção de níveis de poluição sonora).
	Weather: esta tabela incorpora dados dos sensores de ambiente, como velocidade e direção do vento, níveis de precipitação e radiação solar.
	Airquality: esta tabela contém dados sobre a qualidade do ar.
	Basic_environment: esta tabela incorpora dados a partir dos sensores de Temperatura, Humidade e Luminosidade.
	External_gt_sensors: esta tabela contém dados dos sensores das estações *Ground truth* que não pertencem à plataforma do *UrbanSense.*

Os atributos das tabelas associadas ao esquema *Environment* podem ser consultadas na secção ANEXOS-E.

3.2 Fusão

Dada a diversidade de diversas fontes de dados existem vários problemas que podem tornar a fusão de dados, uma tarefa difícil, como podemos observar na figura 24.

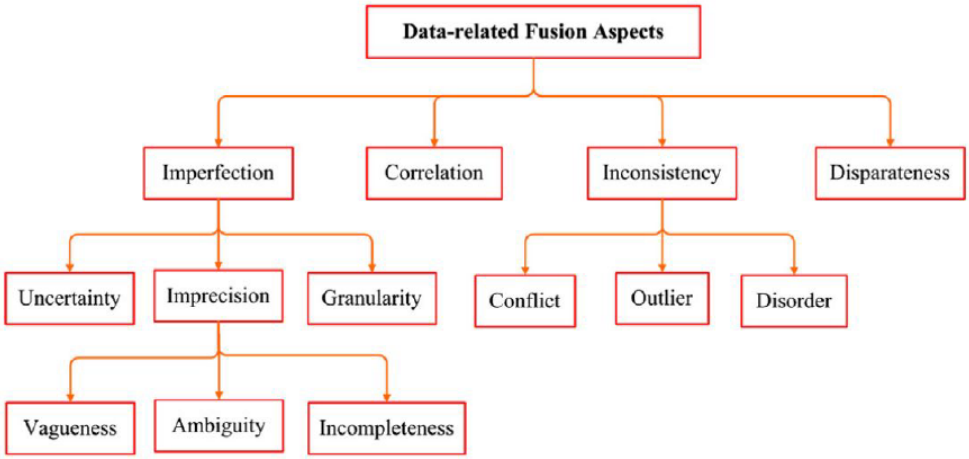


Figura 24 - Problemas na fusão de dados. Fonte:[67]

Um dos principais problemas é a imperfeição dos dados. Alguma parte dos dados que foram recolhidos, nomeadamente aqueles que resultaram da extração de fontes como as redes sociais é muito incompleto e vago. Alguns dos problemas foram encontrados e investigados, no entanto não existe nenhum algoritmo de fusão de dados capaz de agregar todos os problemas. Diferentes métodos descritos na literatura focam-se na existência de um *subset* e são determinados com base na aplicação em mãos.

3.3 Análise e Contextualização da dissertação

Após a extração e recolha dos dados, o projeto URBY.SENSE, entrou na fase de análise dos mesmos. Tendo em conta os dados que foram conseguidos poderá ser possível efetuar diversas análises. A figura 25 representa as diferentes análises, que o projeto considerou tendo em conta a diversidade de dados que conseguiu obter.

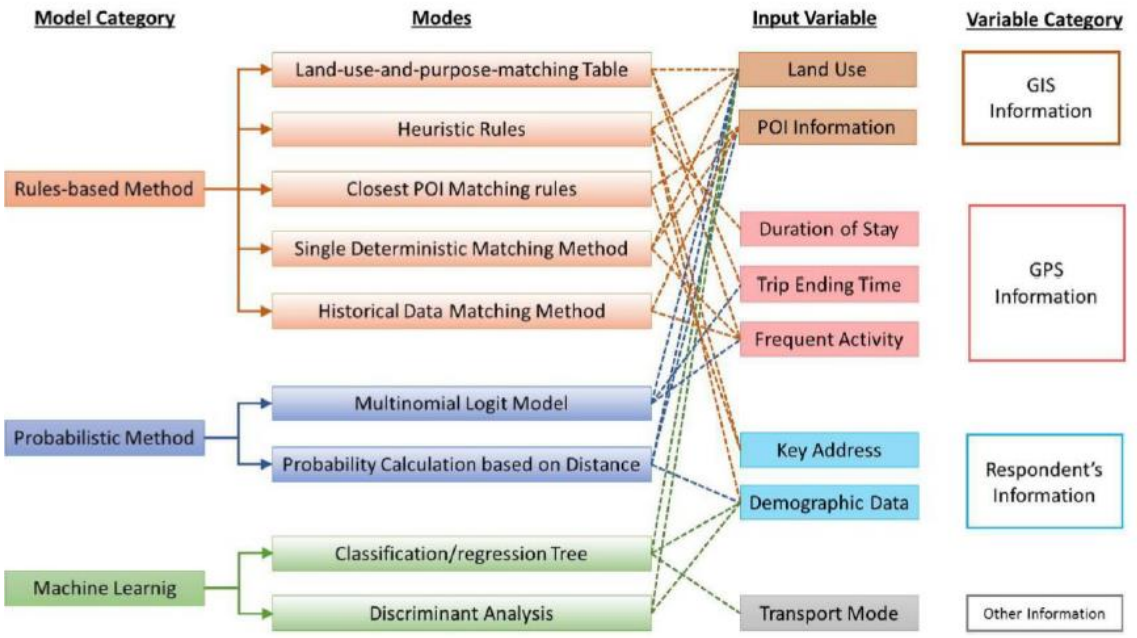


Figura 25 – Possíveis análises mediante dos dados recolhidos. Fonte: [68]

Nesta secção insere-se este projeto de mestrado. Com a elaboração deste projeto pretende-se apoiar o operador público de transportes da zona metropolitana do Porto, STCP, na definição de rotas. Com base nas viagens recolhidas pela aplicação do *SenseMyFeup* pretende-se analisar a possível existência de correlação entre as viagens reportadas e a existência de eventos e, POIs, nessas determinadas zonas em horas fora da rotina diária das pessoas. Desta forma, pretende-se indicar onde é possível melhorar a cobertura em zonas de maior atenção por parte dos decisores urbanos.

3.4 Fase de Modelação

O objetivo desta fase no projeto URBY.SENSE passou por gerar duas matrizes origem-destino, uma para um dia da semana e outra para um fim de semana, onde se criaram dois modelos. O primeiro, o modelo de regressão logística binomial, com base na escolha do meio de transporte. Em segundo, o modelo de regressão logística multinomial, com base na escolha de destinos. As matrizes origem-destino são usadas para indicar a probabilidade de uma escolha de um determinado destino, dada a origem da viagem. Nessas matrizes, as origens correspondem então, ao começo da viagem. Os destinos são as freguesias ou concelhos, onde estão localizados os eventos ou POIs. Os dias 21 e 23 de abril, respetivamente quinta e sábado, foram escolhidos para a geração das matrizes origem-destino. As viagens consideradas correspondem ao período entre as 19h00 e 7 da manhã do dia seguinte.

O modelo de regressão logística binomial foi implementado com o intuito de obter resposta à pergunta: O tipo de utilizador, a localização do evento, o POI, a categoria de evento ou POI, o trabalho em relação ao dia do fim de semana, o tempo de viagem e a distância percorrida influenciam a escolha do modo de transporte? O modo de transporte foi divido em dois grupos: modos de atividade (autocarro, metro, de bicicleta ou a pé) e carro.

Foram considerados vários tipos de variáveis, características sociodemográficas relacionadas à viagem e atributos do evento ou ponto de interesse. Com relação aos recursos relacionados à viagem, foram incluídas variáveis nominais, como o modo utilizado para a viagem e se a viagem foi em um dia de trabalho ou de fim de semana. O tempo de viagem e distância percorrida foram considerados como variáveis. Na tabela 9 são apresentados os resultados obtidos da utilização deste modelo.

Tabela 9 - Resultados obtidos do modelo de regressão logística binomial

Observed			*Mode*		*Percentage Correct*
			Active Modes	Car	
Step 1	Mode	Active Modes	37	8	82.2
		Car	12	40	76.9
		Overall Percentage			79.4

Podemos observar que, dentro dos modos ativos, onde foram observados 45 casos, 37 foram corretamente classificados, enquanto apenas 8 não foram, levando a uma percentagem correta (*accuracy*) de 82,2%. Por outro lado, para a categoria de automóveis, dos 52 valores observados, 40 estavam corretos e 12 foram classificados incorretamente, levando a uma *accuracy* de 76,9%. A distância parece ser a variável mais significativa, que determina o uso do modo carro, ou seja, quanto maior a distância percorrida, maior a probabilidade de o carro ser escolhido para essa viagem.

O segundo modelo anteriormente referido teve como finalidade obter respostas sobre a seguinte pergunta: O papel do utilizador, o modo de transporte, o trabalho *versus* o dia do fim de semana, tempo de viagem, a distância percorrida e o número de *check-ins* num evento ou ponto de interesse influenciam a escolha do tipo de destino?

O modelo considerou diversas categorias importantes de eventos ou POIs, nomeadamente: arte, exposição, mercado, cafés, bares, restaurantes, clubes de dança e noturnos, concertos e teatros. O modelo usou as mesmas variáveis que o anterior, no entanto, em relação aos eventos ou POIs, neste segundo modelo apenas foi considerado o número de *check-ins*, em vez da sua localização geográfica. O estudo realizado com este modelo revelou que quatro das seis variáveis totais são particularmente significativas para o modelo, ou seja, a distância, o check-in, o modo e o trabalho versus o dia do fim

de semana. Por outro lado, variáveis como tempo e o tipo de utilizador não apresentaram resultados significativos para a escolha do destino.

4 RECONHECIMENTO DE PADRÕES NA ESCOLHA DE DESTINOS

Neste capítulo é apresentada a metodologia aplicada durante a elaboração deste trabalho, explicando de forma detalhada os passos seguidos e a justificação das opções tomadas. Este capítulo está organizado da seguinte forma: numa primeira fase é apresentado o modelo de dados utilizado, evidenciando quais os dados que foram alvo de análise, de seguida, na análise exploratória de dados são apresentados os algoritmos utilizados no projeto e quais os motivos que levaram à sua utilização. Posteriormente é descrita a forma como foi realizada a fase de visualização de resultados e a análise de correlações realizada.

A zona metropolitana do Porto é constituída por 17 municípios: Arouca, Espinho, Gondomar, Maia, Matosinhos, Oliveira de Azeméis, Paredes, Porto, Póvoa de Varzim, Santa Maria da Feira, Santo Tirso, São João da Madeira, Trofa, Vale de Cambra, Valongo, Vila do Conde e Vila Nova de Gaia, com uma área aproximada de $2\,041 km^2$, apresenta características únicas que conferem ao território metropolitano a sua diversidade cultural. A oferta existente de programas gastronómicos, desportivos, de natureza e culturais oferece aos visitantes e aos conterrâneos vivências e experiências únicas. O caso de estudo aplicado neste projeto foca-se na área metropolitana do Porto. A cidade do Porto é a segunda maior do país, tendo sido designada em 2001, como a capital europeia da cultura. Esta cidade apresenta uma grande diversidade de museus, igrejas e espaços de elevado interesse turístico [69].

O objetivo deste capítulo baseou-se na criação de uma metodologia capaz de identificar em primeiro lugar, as principais zonas que são mais procuradas em termos de destinos fora da rotina, e uma tentativa de inferir as possíveis finalidades destas viagens, que levaram os utilizadores da aplicação *SenseMyFeup* a viajarem para um determinado destino. Pretende-se que a metodologia implementada consiga dar resposta a perguntas como, quais as áreas do Porto mais procuradas e em quais horas, que correspondam ao período fora da rotina diária das pessoas, durante a semana depois das 18:30h até às 07:00h do dia seguinte e fim de semana. Perceber também quais as áreas da zona metropolitana que mais oferecem serviços e que por esse modo possam ser consideradas como regiões de elevada atratividade. Por fim, com os dados das rotas das linhas dos STCP, a análise permitirá identificar se as áreas com maior procura estão atualmente cobertas pela rede de transportes no horário de fora de rotina pretendido.

4.1 Modelo de Dados

O estudo a desenvolver baseou-se na existência de três entidades: os eventos, os destinos das viagens reportadas na aplicação do *SenseMyFeup* e os pontos de interesse (POIs). Estas três entidades foram recolhidas em diferentes fontes e estão representadas pelos esquemas de base de dados: *public* e *SenseMyCity,* apresentados nas secções: 3.1.1 e 3.1.2, respetivamente.

Em relação ao esquema *SenseMyCity* extraiu-se da tabela *trips*, as coordenadas geográficas (latitude e longitude), respeitantes aos destinos das viagens reportados pela aplicação. Do esquema *Social_Network*, foram extraídos dados de todas as suas tabelas, com informação relativa a eventos e POIs. A tabela 10 representa os dados usados para cada entidade.

Tabela 10 - Dados usados das entidades eventos, destinos e pontos de interesse

<table>
<tr><th colspan="2">Eventos</th><th colspan="2">Destinos</th><th colspan="2">Pontos de Interesse</th></tr>
<tr><td>Latitude</td><td>Longitude</td><td>Latitude</td><td>Longitude</td><td>Latitude</td><td>Longitude</td></tr>
<tr><td colspan="2">Data e hora de início do evento</td><td colspan="2" rowspan="5">Data e hora da chegada ao destino</td><td colspan="2" rowspan="2">Categoria</td></tr>
<tr><td colspan="2">Número de pessoas que indicaram na rede social Facebook que irão ao evento</td></tr>
<tr><td colspan="2">Número de pessoas que indicaram na rede social Facebook que talvez irão ao evento</td><td colspan="2" rowspan="3">Horário de funcionamento</td></tr>
<tr><td colspan="2">Número de pessoas que indicaram na rede social Facebook que estão interessadas em ir ao evento</td></tr>
<tr><td colspan="2">Número de pessoas que indicaram na rede social Facebook que não responderam se iam ao evento</td></tr>
</table>

De modo a poder-se identificar quais as localizações que contêm uma maior probabilidade de atrair mais pessoas para eventos, utilizaram-se os dados da fonte de dados *Facebook*, presente na tabela *facebook_events*. As variáveis extraídas foram as seguintes: o número de pessoas que indicaram na rede social *Facebook*, que iriam ou que estavam interessadas em ir ao evento, ou que até ao momento não tinham respondido se iriam ou não, por fim o número de pessoas que foram convidadas para o evento, mas que até ao momento não tinham dado uma resposta. A variável popularidade de cada posição georreferenciada foi calculada da seguinte forma:

$$\sum_{i=0}^{n} X_i = A_i * 0.45 + I_i * 0.3 + M_i * 0.15 + N_i * 0.10$$

Para cada localização i, composta pelas cordenadas geográficas, latitude e longitude é calculada a sua popularidade, X_i, sendo A_i, o número total de pessoas que indicaram que iriam para eventos daquela localização, I_i, o número total de pessoas que indicaram que estariam interessadas em ir a eventos naquela localização, M_i, o número total de pessoas que indicaram que talvez iriam a eventos situados naquela localização e por fim, N_i, o número total de pessoas que não responderam ao convite de adesão a eventos daquela localização geográfica. Atribuiu-se um peso a cada dado extraído, sendo que o número de pessoas que indicam que vão ao evento tem maior relevância em relação aos restantes. Desta forma foi possível ter-se uma noção de quais as localizações de eventos que poderão ter maior probabilidade de atrair mais pessoas.

Neste estudo, a popularidade dos POIs não foi usada porque esta informação não está disponível em muitos conjuntos de POIs (como é o caso da fonte *Factual*) e gostaríamos que o nosso modelo fosse mais generalizável e não dependesse de dados que apenas está presente numa pequena parte dos dados recolhidos para uma dada entidade. No entanto, concordamos que a popularidade dos lugares é uma propriedade importante que pode beneficiar estudos relacionados desde que se consiga obter esta variável de forma mais consistente das fontes externas.

Para além dos dados já referidos, também foi utilizada informação relativa às linhas STCP presentes na área metropolitana do Porto. O esquema *stcpstops* serviu essencialmente para representar as paragens STCP presentes na área metropolitana do Porto e auxiliar na fase de visualização dos resultados obtidos.

Como um dos objetivos desta metodologia foi perceber quais as zonas com maior procura em distintas fases temporais, criaram-se três filtros: o filtro de fora da rotina geral, o filtro de fora da rotina – fim de semana e o filtro de fora da rotina – semana. O primeiro seleciona dados sobre destinos de viagens que tenham sido realizadas durante a semana e ao fim de semana. Durante a semana, a partir das 18:30h até às 07:00h e ao fim de semana, eventos que tenham começado durante a semana a partir das 19:00h e terminado até às 07:00h e

POIs que estejam abertos durante a semana entre as 19:00h até às 07:00h do dia seguinte e que tenham estado pelo menos em funcionamento cerca de 1h. Todos os dias da semana que tenham sido feriados são englobados pelo dia todo, sem restrição de horário, ou seja a existência de feriado é considerado como um dia fora da rotina. Todos os dados respeitantes ao fim de semana das três entidades são englobados no filtro de rotina geral, ou seja, eventos que tenham sido realizados no fim de semana, POIs que estejam a funcionar ao fim de semana e destinos de viagens que tenham sido realizados também no fim de semana são considerados para este filtro. O filtro de rotina - fim de semana, contém os dados das três entidades respeitantes apenas ao fim de semana. Por fim, o filtro de rotina – semana, contém os dados que dizem respeito apenas à semana e incluindo os feriados.

De forma a identificar com maior precisão, o tipo/categoria de POIs que têm maior atratividade em horário fora de rotina, os POIs foram divididos nos seguintes tipos mais genéricos de acordo com a taxonomia definida pela fonte de dados *Factual* (6.2Anexo F: : social, saúde, *automotive*, *landmarks*, negócios e serviços, comunidade e governo, desporto, transportes, viagens, retalho e transportes. De seguida são aqui apresentados exemplos de POIs de cada categoria:

- Negócios e Serviços: fornecem serviços em geral, relacionados com a banca, seguros, entre outros. Exemplo: Sirusa: Sociedade de Investimentos Rurais e Urbanos.
- Saúde: fornecem serviços relacionados com a saúde. Exemplo: *Lugus* – Prótese Dentária.
- *Automotive*: fornecem serviços relacionados com a indústria automóvel. Exemplo: Ecc - *Exclusive Classic Cars*.
- Social: fornecem serviços relacionados com o entretenimento, restauração, bares, clubes noturnos, entre outros. Exemplo: Foz - Café
- Retalho: fornecem serviços relacionados com o mercado de retalho. Exemplos: *Lidl* e Continente.
- Comunidade e Governo: fornecem serviços relacionados com a comunidade em geral e governo. Exemplo: junta de freguesia de São Pedro.
- Desporto: fornecem serviços relacionados com desporto. Exemplo: *One Soul CrossFit*
- Transportes: fornecem serviços relacionados com transportes. Exemplo: DHL – Transportes Rápidos Internacionais.
- Viagens: fornecem serviços relacionados com viagens, como hotéis. Exemplo: *Park* Hotel Porto Gaia.
- *Landmarks*: conhecidos como pontos de referência. São recursos naturais ou artificiais reconhecidos, usados para navegação, um recurso que se destaca do seu

ambiente próximo e geralmente é visível de longas distâncias. Exemplo: Praia do Marreco.

Dentro de cada categoria existem também outras sub-categorias. Na secção ANEXOS – F, encontram-se representadas todas as categorias e sub-categorias presentes na base de dados do Factual. Durante o trabalho todas as sub-categorias foram englobadas nas dez categorias principais.

Na tabela 11 é apresentado o número de registos por tipo de filtro de rotina e camada.

Tabela 11 - Número de registos por camada e por filtro de rotina

Camada		**Filtro de Rotina – Geral**	**Filtro de Rotina – Fim de Semana**	**Filtro de Rotina - Semana**
Destinos		4711	2158	2576
Eventos		390	272	149
Pontos de Interesse	*Automotive*	5572	5571	1355
	Negócios e Serviços	47303	46237	4061
	Landmarks	694	694	0
	Social	11786	11778	9229
	Transportes	725	643	233
	Viagens	3172	3176	625
	Saúde	11786	11778	9229
	Desporto	1159	1158	40
	Comunidade e Governo	5795	5784	2091
	Retalho	30718	30189	4941
	Total	118710	117008	26863

4.2 Análise Exploratória dos Dados

De acordo com a literatura existente apresentada na secção 2.2, diferentes técnicas podem ser utilizadas para a análise de dados na mobilidade urbana, podendo ser classificadas como não supervisionadas, por exemplo no caso do *clustering*, ou supervisionadas como as redes neuronais, SVMs e classificadores *Naive Bayes*, entre outros. Neste projeto optou-se por seguir uma abordagem não supervisionada, uma vez que não se conhece nada à partida sobre os dados, nem qual o *output* que terá de ser gerado. Não se sabe também, se os destinos de viagens que foram reportados dizem respeito a visitas a determinados pontos de interesse ou presença em eventos, pelo que o agrupamento deste tipo de dados em regiões poderá identificar a importância de algumas zonas e possivelmente os motivos das viagens. Além disso, não se pretende estudar viagens individuais de utilizadores, mas sim as zonas mais procuradas e a concentração de serviços ou eventos que possam justificar esta procura. Desta forma, utilizou-se a técnica de *clustering*, como o método a usar para a análise de dados.

4.2.1 Técnicas de *Clustering*

De forma a poder-se optar pelas técnicas que poderão ter melhor desempenho, realizou-se uma análise aos algoritmos já existentes, apresentados na secção 2.3. Na análise realizada, as seguintes métricas foram alvo de estudo durante a comparação das técnicas:

1. Característica A: capacidade de identificar *clusters* com formas aleatórias.
2. Característica B: capacidade para identificar *clusters* em *datasets* com elevado volume de dados.
3. Característica C: boa *performance* na obtenção de resultados, principalmente também em *datasets* com considerável volume de dados.
4. Característica D: capacidade de lidar com ruído e conseguir que a sua presença não tenha impacto nos resultados obtidos.
5. Característica E: parametrização/configuração inicial do algoritmo (a não obrigatoriedade de indicar o número de *clusters* finais que terão de ser gerados e estimação dos parâmetros iniciais).
6. Característica F: lidar com valores numéricos.
7. Característica G: sensibilidade ao modo de ordenação dos dados.

Estas métricas foram escolhidas, porque têm em conta, as necessidades e objetivos que pretendemos atingir, mediante o contexto deste projeto. Como não sabemos à partida, nem temos dados sobre o número de *clusters*, que devem ser gerados, a possibilidade de não indicar esse valor é uma métrica que é interessante para o projeto a desenvolver. Existem diversos métodos que poderíamos utilizar para determinar previamente esse valor e dessa forma contornava-se essa dificuldade, no entanto isso requeria mais processamento de dados para efetuar essa análise. A capacidade de o algoritmo ser capaz de identificar *clusters* com diversas formas aleatórias e não ser capaz só de identificar *clusters* esféricos, por exemplo

é também uma característica que pretendemos, uma vez que não sabemos à partida que tipo de formas os *clusters* poderão ou deverão ter. A característica do algoritmo conseguir lidar com ruído é importante neste projeto, uma vez que a probabilidade de existir devido aos dados serem recolhidos por uma plataforma crowdsourcing poderá ser alta. A capacidade de lidar com valores numéricos é imprescindível, sendo que os dados a analisar são essencialmente numéricos (e.g. contagens, popularidade). A sensibilidade ao modo de ordenação inicial dos dados é também uma métrica a ter em conta, uma vez que não existe qualquer tipo de ordem. Apesar de no estudo em concreto, não termos um elevado volume de dados, o bom desempenho na obtenção de resultados e na capacidade de identificação de *clusters* são duas métricas importantes, para que no futuro, se possa usar a mesma metodologia em *datasets* com um volume maior de dados.

A tabela 12 representa a análise comparativa realizada, sendo de realçar que o HDSBCAN* poderia também estar apenas entre os métodos hierárquicos.

Tabela 12 - Análise comparativa das técnicas de clustering

Algoritmo		**Característica**						
		A	**B**	**C**	**D**	**E**	**F**	**G**
Métodos particionados	*K-Means*	✗	✗	✗	✗	✗	✓	✗
	PAM	✓	✗	✗	✗	✗	✓	✗
	CLARA	✓	✗	✗	✗	✗	✓	✗
	CLARANS	✓	✓	✓	✗	✗	✓	✗
Métodos baseados em densidade	DBSCAN	✓	✓	✓	✓	✓	✓	✗
	OPTICS	✓	✓	✓	✓	✓	✓	✗
	DENCLUE	✓	✓	✓	✓	✓	✓	✗
	HDBSCAN*	✓	✓	✓	✓	✓	✓	✗
Métodos baseados em grelha	STING	✓	✓	✓	✓	✓	✓	✗
	CLIQUE	✓	✓	✓	✓	✓	✓	✗
	BIRCH	✗	✓	✓	✓	✗	✓	✓
	ROCK	✓	✓	✓	✗	✗	✗	✗

Métodos hierárquicos	CURE	✓	✓	✓	✓	✓	✓	✗
	DIANA	✓	✓	✗	✗	✓	✓	✗

Os algoritmos pertencentes aos métodos particionados são conhecidos por não serem capazes de identificar *clusters* com formas aleatórias, sendo apenas capazes de determinar *clusters* esféricos, como no caso do *k-means*. Além disso são bastante sensíveis ao ruído e a sua presença influencia os resultados finais. Devido a estes motivos e ao facto de principalmente se ter de indicar o número prévio de *clusters* que terão de ser gerados, optou-se por não se utilizar no projeto nenhum algoritmo deste tipo.

Tendo em conta a análise comparativa apresentada na tabela 12 os algoritmos de *clustering* que melhor se adaptam para este projeto pertencem aos métodos baseados em grelha e em densidade. À excepção do CURE, os restantes algoritmos pertencentes aos métodos hierárquicos apresentam características não tão vantajosas para o projeto. Em resumo, os algoritmos que apresentam as características pretendidas para analisar informação georreferenciada são: DSBCAN, OPTICS, DENCLUE, HDBSCAN*, CLIQUE, STING e o CURE.

Dentro dos algoritmos considerados candidatos, ou seja, que reúnem todas as características necessárias, analisamos algumas vantagens e desvantagens específicas de cada algoritmo, de modo a poder-se encontrar alguns pormenores que possam nos levar a optar por uns em detrimento de outros.

O STING poderia ser uma boa escolha, no entanto devido à sua natureza probabilística pode implicar uma *accuracy* menor. No caso do CLIQUE, tal como em todas as aproximações de *clustering* baseadas em grelha, a qualidade dos resultados depende muito da escolha apropriada do número e tamanho de células da grelha. O mesmo acontece com os algoritmos HDSBCAN, OPTICS, DBSCAN e DENCLUE, em que a parametrização escolhida tem bastante impacto nos resultados obtidos. Como o HDSBCAN*, OPTICS e DBSCAN partilham as mesmas variáveis de entrada (à excepçao da variável do épsilon que não é necessário indicar ao HDSBCAN*), a estimativa dos melhores parâmetros pode ser realizada de forma paralela e a aplicação dos algoritmos também. Por esta razão optou-se por seguir com o estudo aplicando os algoritmos HDBSCAN*, OPTICS e DBSCAN.

Para usar os três algoritmos referidos, utilizou-se a biblioteca *scikit-learn*[16], *software* de código aberto escrito em *python*, para projetos de *machine learning*. Escolhidos os

[16] https://scikit-learn.org/ (Último acesso em 12/12/2018)

algoritmos e a biblioteca a usar, seguiu-se a implementação de uma aplicação em *python*, que extrai da base de dados, referida em 3.1, o modelo de dados apresentado na secção 4.1. Para interagir com o motor de base de dados *PostegreSQL* usou-se a biblioteca *psycopg2*[17]. Na aplicação foram criadas as diversas pesquisas de modo a extrair os dados pretendidos de acordo com o filtro de rotina a aplicar. De seguida após se ter os dados pretendidos para cada filtro procedeu-se ao uso dos algoritmos.

À excepção do HDBSCAN*, o OPTICS e o DBSCAN necessitam do valor do *épsilon*, que é um valor em ângulo. Para melhor compreensão dos dados, efetuou-se uma conversão dos graus decimais para metros, tendo em conta que, em diferentes latitudes pode haver diferentes distâncias entre dois pontos. Esta diferença depende da curvatura da Terra, apresentada na figura 26, por exemplo, na zona da linha do equador, dois pontos com a mesma diferença em graus estarão mais distantes entre si, pois trata-se de uma zona mais plana, do que nos trópicos, com uma curvatura mais acentuada. Por exemplo, 1 grau decimal de distância na cidade do Porto, sabendo que possui as coordenadas: latitude 41.15, longitude: -8.16, corresponde aos seguintes metros:

- Tendo em conta que o perímetro da Terra corresponde a 40075 km, o tamanho em metros de um grau de longitude corresponde a 40075 km * *cos*(41.15)/360, que corresponde a aproximadamente 84 metros.

[17] https://pypi.org/project/psycopg2/ (Último acesso em 12/12/2018)

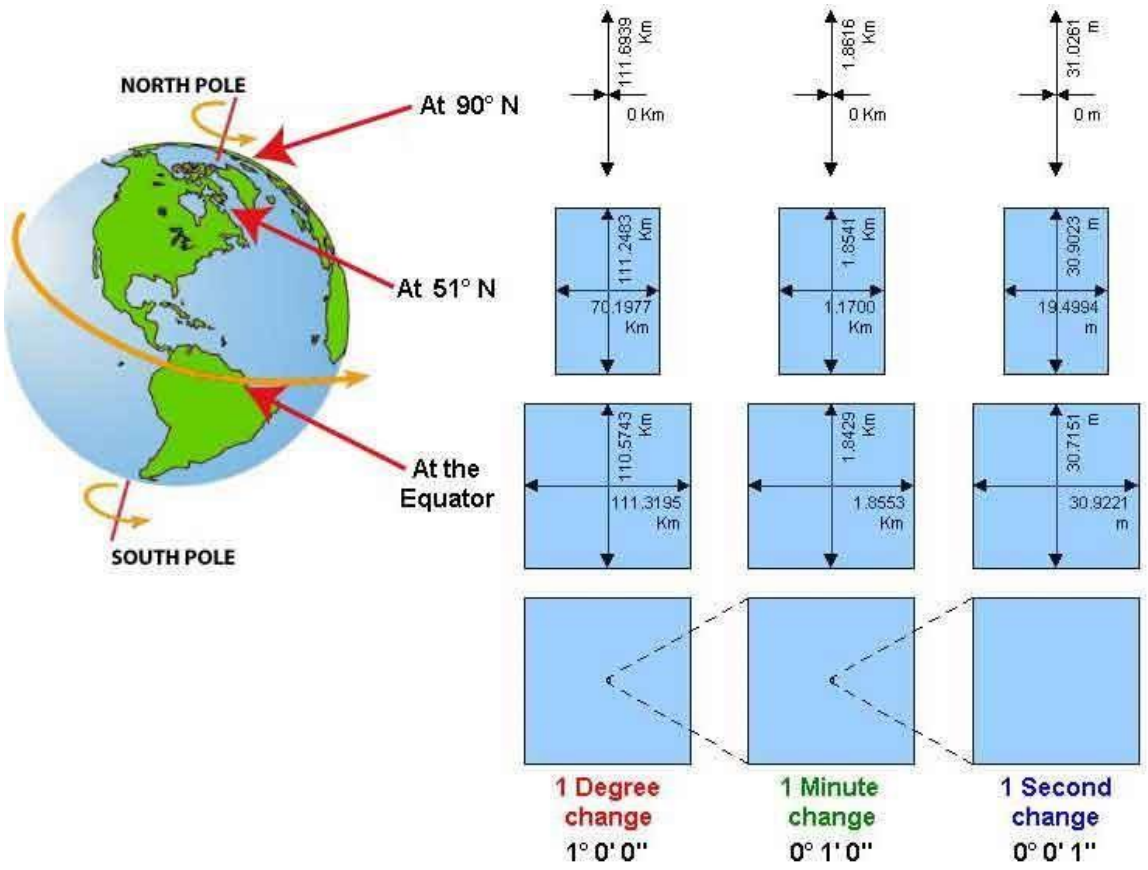


Figura 26 – Diferências de distâncias de acordo com a posição geográfica onde nos situamos [18]

De modo a ser uma distância passível de percorrer a pé, definiu-se que o valor do *épsilon* teria de corresponder de 40 a 200 metros, o que corresponde a um valor de aproximadamente 0.0005 e 0.0024, respetivamente.

A forma como foi realizada a escolha do melhor valor do *épsilon*, bem como a escolha do valor do número mínimo de pontos econtra-se descrita na secção 5.1.

4.2.2 Correlação

Depois de se terem originado os diversos *clusters* foi necessário verificar algum tipo de correlação, que nos possa permitir tirar conclusões. O coeficiente de *Person* [70], r (ou r^2), mede a força da relação linear entre duas variáveis quantitativas. A fórmula do coeficiente pode ser entendida da seguinte forma:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{\bar{x}}}{S_x}\right)\left(\frac{y_i - \bar{\bar{y}}}{S_y}\right)$$

[18] http://www.longitudestore.com/how-big-is-one-gps-degree.html (Último acesso em: 12/12/2018)

O valor do coeficiente varia de -1 a 1, onde -1 significa uma perfeita relação negativa e 1, uma perfeita relação positiva. Valores próximos de 0, significam que não foi possível encontrar qualquer tipo de correlação. Na tabela 13 podemos verificar mais detalhamente o significado de cada intervalo de valores:

Tabela 13 - Singificado para os diferentes valores de r

Valores	**Significado**
r > 0.7	Relação linear forte (positiva)
0.5 < r < 0.7	Relação linear moderada (positiva)
0.3 < r < 0.5	Relação linear fraca (positiva)
0 < r < 0.3	Ausência de qualquer relação
-0.3 < r < 0	Ausência de qualquer relação
-0.5 < r < -0.3	Relação linear fraca (negativa)
-0.7 < r < -0.5	Relação linear moderada (negativa)
r < -0.7	Relação linear forte (negativa)

Associando uma anologia, para um melhor entendimento, na agricultura agrícola, à medida que a vegetação aumenta, a quantidade de refletância no infravermelho próximo aumenta. Um exemplo de uma relação negativa na agricultura agrícola é que, à medida que a vegetação aumenta, a quantidade de refletância do comprimento de onda visível diminui (até a saturação)[19]. Uma característica importante do coeficiente r de *Pearson* é que transmite se uma relação é positiva, negativa ou se não encontrou qualquer tipo de correlação.

Para além de r, a correlação de *Pearson* fornece uma probabilidade de confiança, p, sobre os resultados obtidos de r. Indica a probabilidade da relação encontrada ser zero, ou seja, a probabilidade de não haver correlação, mesmo que o valor de r seja alto. Fortes correlações têm baixos valores de p. As correlações podem ser consideradas significantes se contiverem uma probabilidade de confiança, p, abaixo de 0.05.

Neste estudo, implementaram-se três correlações, que tiveram como objetivo identificar quais as camadas que tivessem maior relação entre destinos e pontos de interesse ou eventos. Essencialmente, estas correlações têm como intuito verificar quais os possíveis motivos, que

[19] http://www.gisagmaps.com/about-r-and-r-squared/ (Último acesso em: 12/12/2018)

levaram os utilizadores da aplicação *SenseMyFeup* a viajarem para um determinado lugar na área metropolitana do Porto. De seguida são apresentadas as três correlações alvo de estudo neste projeto:

- Correlação A: entre destinos e eventos ou pontos de interesse. Baseia-se na soma do número de pontos da camada destinos (para cada *cluster*) à volta da camada pontos de interesse ou eventos, num raio de 500 metros entre a soma do número de pontos (para cada *cluster*) da camada de pontos de interesse ou eventos um raio de 500 metros.

Esta correlação teve como objetivo permitir identificar, quais os possíveis motivos pelos quais determinados destinos de viagens foram realizados para um determinado lugar. Pretendeu-se verificar, se à medida que a soma do número de pontos da camada destinos à volta das camadas pontos de interesse ou eventos aumenta, a soma do número de pontos da camada de pontos de interesse ou eventos também aumentam num raio de 500 metros. Essencialmente, o que a correlação poderá transmitir é que, quanto maior é a oferta (através do maior número de eventos ou POIs), maior é a procura (através do maior número de destinos próximos a eventos/POIs).

- Correlação B: entre eventos e destinos. Popularidade dos eventos entre o número de destinos à distância de 500 metros.

A correlação B teve como finalidade a verificação da existência ou não de uma relação que possa existir entre a popularidade dos eventos existentes, com o número de destinos perto desses lugares. Pretendeu-se verificar, se à medida que a popularidade das regiões com eventos aumenta se o número de destinos perto dessas regiões também aumenta. A existência desta correlação permite identificar as regiões que de facto, contém maior atratividade.

- Correlação C: entre eventos e destinos. Popularidade de eventos *outliers* entre o número de destinos a 500 metros de distância.

A implementação da correlação C teve como objetivo verificar, se os pontos que apesar de não terem sido selecionados pelos métodos de *clustering*, como pertencentes a um determinado *cluster*, devido à sua não proximidade com outros eventos, apresentam, no entanto, um valor de popularidade acima da média. Apesar de não estarem próximos de outras regiões onde ocorram muitos eventos, poderão só por si, atrair pessoas com elevada atratividade.

4.3 Relação entre procura e oferta (STCP)

Nesta fase, o que se pretendeu foi analisar visualmente as relações existentes entre a procura e oferta que existe atualmente nas linhas STCP. Após se ter anteriormente identificado quais

as correlações que teriam melhores resultados, foi necessário verificar quais as paragens por onde passam linhas dos STCP que estão em serviço em horário fora da rotina de acordo com os filtros definidos. De seguida, visualizaram-se na ferramenta do ArcMap, quais as regiões com maior interesse.

4.3.1 Dados a analisar

Depois de se terem identificados os principais focos de procura, por parte dos utilizadores da aplicação *SenseMyFeup*, procurou-se perceber se os locais com maior procura estão cobertos atualmente pela oferta que os STCP oferecem. Para isso, para cada filtro fora da rotina (geral, apenas fim de semana e apenas semana) foram extraídas somente as paragens que estão abrangidas por cada filtro, ou seja, paragens que sejam abrangidas por linhas que fazem o seu percurso em horário fora da rotina apenas aos fim de semana, apenas durante a semana depois das 18:30h e até às 7:00h e durante a semana e fim de semana pelo horário referido anteriormente. Como atualmente na base de dados usada, não havia dados sobre os horários das linhas e paragens que estas linhas percorriam, cada horário de cada linha e respectivas paragens foi extraído através do próprio site *web* dos STCP respeitante ao horário das linhas[20]. As paragens existentes na base de dados foram exportadas para um ficheiro CSV (*Comma-separated values*), para que posteriormente podessem ser visualizadas na ferramenta de visualização.

4.3.2 Visualização

De seguida, após se determinar quais os melhores parâmetros a usar e aplicar as diferentes técnicas de *clustering* seguiu-se a fase de criação dos *shapefiles* associados a cada camada por filtro de rotina. De modo a facilitar a posterior visualização de dados, foram formados polígonos, correspondendo cada *cluster* a um polígono. Para a criação do *shapefile*, bem como para a criação dos polígonos, usou-se a biblioteca *Shapely*[21]. Nesse ficheiro *shapefile*, para além dos vérticies do polígno também é passada a informação de quantos pontos contém aquele polígono, ou seja, quantos pontos pertencem ao *cluster*. Na fase de visualização dos dados obtidos, utilizou-se a ferramenta *ArcMap*. Para os três tipos de horário fora da rotina, importou-se para a ferramenta, as combinações de *clusters* (*shapefiles*) que apresentaram as melhores correlações. Para além da importação de *clusters*, também se importaram as paragens STCP existentes na área metropolitana do Porto. A visualização dos clusters procedeu-se através da importação dos *shapefiles* criados. A cada polígono visualizado no mapa correspondeu a um *cluster*. As coordenadas geográficas

[20] https://www.stcp.pt/pt/viajar/linhas (Último acesso em: 12/12/2018)

[21] https://pypi.org/project/Shapely/ (Último acesso em: 12/12/2018)

respeitantes a eventos, destinos, pontos de interesse e paragens STCP foram importados consoante a norma EPSG:3857 - WGS84 *Web Mercator*[22].

4.4 Importação de *clusters* para a plataforma do *City Clusters*

A plataforma *Web*, *CityClusters* [8] fornece um mapa interativo onde se pode visualizar a informação de cada ponto de interesse, contido num determinado *cluster*, incluindo dados sobre a sua localização geográfica e a sua categoria associada. A plataforma recebe dois ficheiros com dados relativos aos *clusters* em formato JSON (*JavaScript Object Notation*) e cria um ficheiro com o *convex hull* de cada *cluster* (polígono que contém todos os pontos de um *cluster* para o representar).

Com estes dois ficheiros (o *input* e *output* da ferramenta) a plataforma consegue criar um mapa interativo onde são apresentados os vários *clusters* gerados agrupados pelo tipo de *cluster*, tal como apresentado na figura 27, também com a capacidade de incluir vários níveis de zoom de acordo com a aproximação mostrar dados mais específicos.

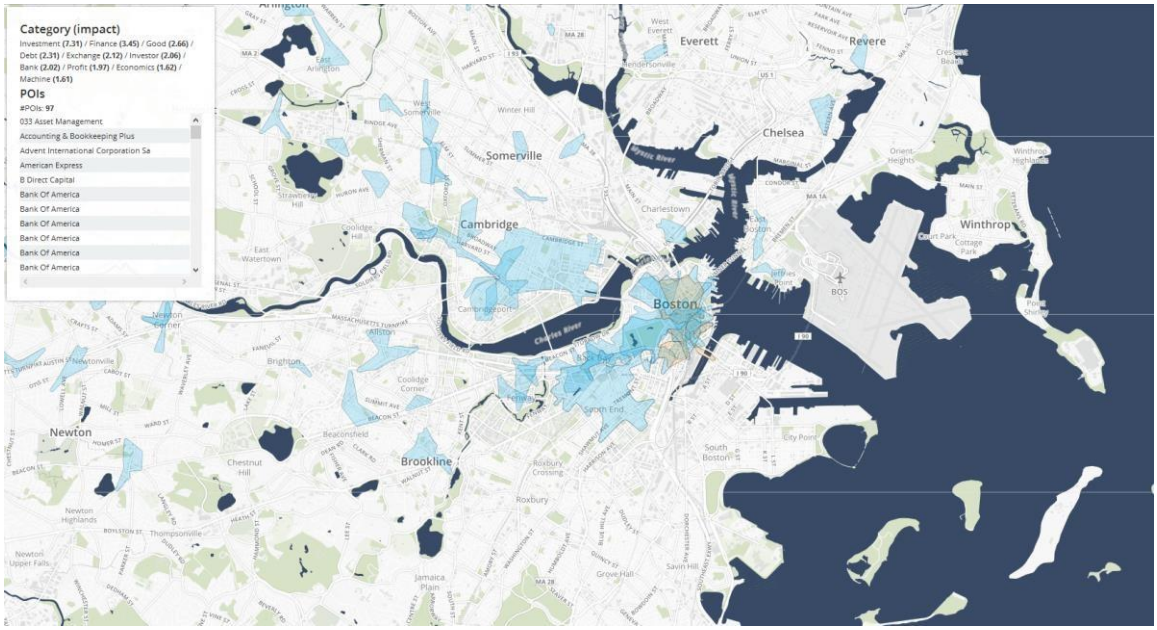


Figura 27- Exemplo da visualização de clusters no CityClusters

Na primeira fase procedeu-se então à criação dos *clusters* relativos aos pontos de interesse de cada categoria generica identificada: comunidade e governo, desporto, social, retalho, negócios e serviços, viagens, transportes, *landmarks*, saúde e *automotive*. Os *clusters* foram gerados de acordo com os três filtros aplicados neste projeto. Uma vez que a plataforma

[22] http://spatialreference.org/ref/sr-org/epsg3857-wgs84-web-mercator-auxiliary-sphere/ (Último acesso em: 12/12/2018)

CityClusters permite a introdução de 3 níveis de *zoom*, cada nível teve associado diferentes valores do *épsilon* e do número mínimo de pontos. Para o nível de *zoom* mais próximo usaram-se as mesmas configurações obtidas na fase de estimação dos parâmetros de clustering que serão apresentados na seção 5.1. Nos restantes níveis, 2 e 3, definiu-se previamente um valor máximo e mínimo do *épsilon*, 250 a 350 metros (aproximadamente, de 0.0029 a 0.0042) e 350 a 500 metros (aproximadamente, de 0.0042 a 0.006), respetivamente. De acordo com estes limites, seguiu-se a mesma metodologia abordada em 5.1, mantendo o mesmo raciocínio para a variável do número mínimo de pontos. Depois de definidos os parâmetros a aplicar para cada algoritmo, procedeu-se à criação de clusters para cada filtro de rotina. Após a criação de clusters transformou-se os dados obtidos em objetos com formato JSON de modo a permitir a importação para a plataforma do *CityClusters*.

5 RESULTADOS E VALIDAÇÃO

Neste capítulo são apresentados os resultados obtidos, bem como os métodos que foram aplicados durante a validação dos mesmos.

5.1 Parametrização e Validação

Uma das principais etapas do *clustering*, passa como avaliar os resultados. Uma aproximação comum baseia-se na utilização de índices de validação [71]. A avaliação de c*lustering*, pode ser dividida em três categorias principais: internas, externas e relativas. A aproximação externa, tem por base haver um conhecimento prévio dos dados. As utilizações de índices externos baseiam-se na existência de uma medida de concordância entre duas partições, em que a primeira partição é a estrutura de um *cluster* conhecido *à priori* e a segunda resulta do procedimento de agrupamento [72]. Por sua vez, os índices internos são usados para medir a qualidade de *cluster* sem informações externas [73]. As aproximações relativas são usadas para comparar diferentes *clusterings* ou *clusters*. Tendo em conta estas aproximações, optou-se por seguir a utilização de um índice interno, uma vez que o objetivo passa por determinar quais as melhores configurações para cada algoritmo: HDSBCAN*, DBSCAN e OPTICS. Uma vez, que a visualização manual de cada configuração possível, seria de todo impossível utilizou-se um índice interno para validar quais as melhores configurações de cada algoritmo. De forma a atingir esse objetivo usou-se o índice interno, *silhouette score* [74]. Este índice baseia-se no príncipio da máxima coesão interna e máxima separação entre *clusters*, ou seja, mede o quanto os objetos estão próximos entre si, pertencentes ao mesmo *cluster* e o quanto estão afastados relativamente a objetos de outros *clusters*. A fórmula do *silhouette score* para um elemento, x_i, pode ser definida da seguinte forma:

$$s(\bar{\bar{x}}_i) = \frac{b(\bar{\bar{x}}_i) - a(\bar{x}_i)}{max\{a(\bar{x}_i),\ b(\bar{x}_i)\}}$$

Os valores do *silhouette score* variam de -1 a 1. Valores que sejam menores do que 0.25, significa que não foi encontrada nenhuma estrutura substancial, valores que variem de 0.26 até 0.50, significa que a estrutura encontrada é fraca e pode ser artificial, valores que variem de 0.51 a 0.70, significam que uma estrutura razoável foi encontrada. Por fim, valores acima de 0.7, significa que foi encontrada uma estrutura forte. De modo a atingir um maior volume de amostras para a análise de correlação, optou-se por selecionar configurações que obtivessem bons resultados de acordo com o índice, mas que também conseguissem originar mais do que 30 *clusters* para uma maior significância estatística. Para cada camada, usando o OPTICS e o DBSCAN executou-se cada algoritmo com valores de *épsilon* que variaram de 0.0005 e 0.0024. Para cada valor de *épsilon*, usou-se um valor mínimo de pontos para

formar um *cluster*, desde 4 até 170. O valor mínimo estabelecido foi 4, porque é o número mínimo de pontos mais usado na literatura para formar um polígono. O valor máximo de 170 pontos foi usado como teste, se se verificasse que não se obteria resultados, este valor seria aumentado. Para o HDSBCAN*, como não utiliza o *épsilon* executou-se o algoritmo para cada camada com o valor do número mínimo de pontos a variar de 4 a 170.

De seguida são apresentadas as tabelas de 14 a 25, as melhores configurações obtidas para o filtro de rotina geral, de 26 a 37, para o filtro de rotina ao fim de semana e de 38 a 47 para o filtro de rotina durante a semana.

Tabela 14 – Configuração obtida da camada destinos, no filtro de rotina geral

Camada	**Algoritmo**		
Destinos	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	9	6	6
Épsilon	0.0007	0.0005	N/A
Número de *Clusters*	120	140	181
Silhouette Score	0.79	0.82	0.774

Tabela 15 – Configuração obtida da camada eventos, no filtro de rotina geral

Camada	**Algoritmo**		
Eventos	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	4	4
Épsilon	0.001	0.0011	N/A
Número de *Clusters*	23	18	23
Silhouette Score	0.75	0.75	0.7

Tabela 16 – Configuração obtida da camada Automotive, no filtro de rotina geral

Camada	**Algoritmo**		
Automotive	**DBSCAN**	**OPTICS**	**HDBSCAN**

Número mínimo de pontos	8	6	8
Épsilon	0.001	0.001	N/A
Número de *Clusters*	38	43	34
Silhouette Score	0.709	0.807	0.70

Tabela 17 – Configuração obtida da camada Negócios e Serviços, no filtro de rotina geral

Camada	**Algoritmo**		
Negócios e Serviços	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	73	66	42
Épsilon	0.001	0.001	N/A
Número de *Clusters*	26	25	35
Silhouette Score	0.61	0.6	0.714

Tabela 18 - Configuração obtida da camada LandMarks, no filtro de rotina geral

Camada	**Algoritmo**		
LandMarks	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	4	5
Épsilon	0.0023	0.0024	N/A
Número de *Clusters*	18	9	3
Silhouette Score	0.812	0.92	0.707

Tabela 19 – Configuração obtida da camada Social, no filtro de rotina geral

Camada	**Algoritmo**

Social	DBSCAN	OPTICS	HDBSCAN
Número mínimo de pontos	14	11	18
Épsilon	0.0005	0.0005	N/A
Número de *Clusters*	35	37	40
Silhouette Score	0.6	0.6	0.75

Tabela 20 - Configuração obtida na camada Transportes, no filtro de rotina geral

Camada	Algoritmo		
Transportes	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	4	4
Épsilon	0.0025	0.0024	N/A
Número de *Clusters*	19	7	12
Silhouette Score	0.648	0.688	0.67

Tabela 21 - Configuração obtida na camada Viagens, no filtro de rotina geral

Camada	Algoritmo		
Viagens	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	8	6	6
Épsilon	0.0005	0.0005	N/A
Número de *Clusters*	31	37	80
Silhouette Score	0.6	0.6	0.62

Tabela 22 - Configuração obtida na camada Saúde, no filtro de rotina geral

Camada	Algoritmo

Saúde	DBSCAN	OPTICS	HDBSCAN
Número mínimo de pontos	8	7	9
Épsilon	0.0005	0.0005	N/A
Número de *Clusters*	50	44	74
Silhouette Score	0.7	0.7	0.681

Tabela 23 - Configuração obtida na camada Desporto, no filtro de rotina geral

Camada	Algoritmo		
Desporto	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	5	5	4
Épsilon	0.002	0.0023	N/A
Número de *Clusters*	31	16	37
Silhouette Score	0.711	0.736	0.67

Tabela 24 - Configuração obtida na camada Comunidade e Governo, no filtro de rotina geral

Camada	Algoritmo		
Comunidade e Governo	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	6	5	8
Épsilon	0.0005	0.0005	N/A
Número de *Clusters*	49	33	49
Silhouette Score	0.82	0.86	0.65

Tabela 25 - Configuração obtida na camada Retalho, no filtro de rotina geral

Camada	**Algoritmo**		
Retalho	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	29	20	32
Épsilon	0.0008	0.0007	N/A
Número de *Clusters*	43	56	40
Silhouette Score	0.60	0.60	0.72

Tabela 26 - Configuração obtida na camada Destinos, no filtro de rotina - fim de semana

Camada	**Algoritmo**		
Destinos	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	10	8	11
Épsilon	0.001	0.001	N/A
Número de *Clusters*	37	41	37
Silhouette Score	0.87	0.89	0.79

Tabela 27 - Configuração obtida na camada Eventos, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
Eventos	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	4	4
Épsilon	0.001	0.0012	N/A
Número de *Clusters*	20	16	16
Silhouette Score	0.85	0.83	0.74

Tabela 28 - Configuração obtida na camada Automotive, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
Automotive	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	8	6	8
Épsilon	0.001	0.001	N/A
Número de *Clusters*	38	43	34
Silhouette Score	0.71	0.70	0.70

Tabela 29 - Configuração obtida na camada Negócios e Serviços, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
Negócios e Serviços	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	34	29	31
Épsilon	0.0005	0.0005	N/A
Número de *Clusters*	53	61	94
Silhouette Score	0.74	0.71	0.72

Tabela 30 - Configuração obtida na camada landMarks, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
LandMarks	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	4	5
Épsilon	0.0022	0.0024	N/A
Número de *Clusters*	18	9	3
Silhouette Score	0.82	0.92	0.71

Tabela 31 - Configuração obtida na camada Social, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
Social	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	14	12	19
Épsilon	0.0005	0.0005	N/A
Número de *Clusters*	35	29	30
Silhouette Score	0.6	0.64	0.78

Tabela 32 - Configuração obtida na camada Transportes, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
Transportes	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	4	4
Épsilon	0.0011	0.0021	N/A
Número de *Clusters*	5	5	11
Silhouette Score	0.88	0.72	0.67

Tabela 33 - Configuração obtida na camada Viagens, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
Viagens	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	6	5	7
Épsilon	0.0005	0.0005	N/A
Número de *Clusters*	60	55	60
Silhouette Score	0.57	0.58	0.601

Tabela 34 - Configuração obtida na camada Saúde, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
Saúde	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	12	11	11
Épsilon	0.0011	0.0011	N/A
Número de *Clusters*	37	33	45
Silhouette Score	0.51	0.52	0.7

Tabela 35 - Configuração obtida na camada Desporto, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
Desporto	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	4	4
Épsilon	0.0013	0.0023	N/A
Número de *Clusters*	33	28	42
Silhouette Score	0.86	0.69	0.66

Tabela 36 - Configuração obtida na camada Comunidade e Governo, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
Comunidade e Governo	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	10	6	8
Épsilon	0.001	0.001	N/A
Número de *Clusters*	28	47	53
Silhouette Score	0.64	0.52	0.65

Tabela 37 - Configuração obtida na camada Retalho, no filtro de rotina– fim de semana

Camada	**Algoritmo**		
Retalho	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	26	24	26
Épsilon	0.0007	0.0005	N/A
Número de *Clusters*	43	33	53
Silhouette Score	0.57	0.59	0.72

Tabela 38 - Configuração obtida na camada Destinos, no filtro de rotina– semana

Camada	**Algoritmo**		
Destinos	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	4	5
Épsilon	0.0006	0.0006	N/A
Número de *Clusters*	146	122	127
Silhouette Score	0.75	0.8	0.77

Tabela 39 - Configuração obtida na camada Eventos, no filtro de rotina– semana

Camada	**Algoritmo**		
Eventos	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	4	4
Épsilon	0.001	0.0011	N/A
Número de *Clusters*	15	12	11
Silhouette Score	0.87	0.86	0.78

Tabela 40 - Configuração obtida na camada Automotive, no filtro de rotina– semana

Camada	**Algoritmo**		
Automotive	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	5	5	5
Épsilon	0.002	0.002	N/A
Número de *Clusters*	27	12	11
Silhouette Score	0.7	0.759	0.723

Tabela 41 - Configuração obtida na camada Negócios e Serviços, no filtro de rotina– semana

Camada	**Algoritmo**		
Negócios e Serviços	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	8	5	7
Épsilon	0.0005	0.0005	N/A
Número de *Clusters*	47	71	84
Silhouette Score	0.71	0.71	0.67

Tabela 42 - Configuração obtida na camada Social, no filtro de rotina– semana

Camada	**Algoritmo**		
Social	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	11	9	14
Épsilon	0.0005	0.0005	N/A
Número de *Clusters*	33	38	42
Silhouette Score	0.67	0.7	0.744

Tabela 43 - Configuração obtida na camada Transportes, no filtro de rotina– semana

Camada	**Algoritmo**		
Transportes	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	-	4
Épsilon	0.0025	-	N/A
Número de *Clusters*	5	-	3
Silhouette Score	0.819	-	0.79

Tabela 44 - Configuração obtida na camada Viagens, no filtro de rotina– semana

Camada	**Algoritmo**		
Viagens	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	5	5	4
Épsilon	0.0018	0.0012	N/A
Número de *Clusters*	22	10	27
Silhouette Score	0.68	0.76	0.657

Tabela 45 - Configuração obtida na camada Saúde, no filtro de rotina– semana

Camada	**Algoritmo**		
Saúde	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	5	4	6
Épsilon	0.0005	0.0005	N/A
Número de *Clusters*	85	77	122
Silhouette Score	0.71	0.72	0.651

Tabela 46 - Configuração obtida na camada Comunidade e Governo, no filtro de rotina– semana

Camada	**Algoritmo**		
Comunidade e Governo	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	4	4	5
Épsilon	0.001	0.0011	N/A
Número de *Clusters*	58	30	42
Silhouette Score	0.68	0.71	0.66

Tabela 47 - Configuração obtida na camada Retalho, no filtro de rotina – semana

Camada	**Algoritmo**		
Retalho	**DBSCAN**	**OPTICS**	**HDBSCAN**
Número mínimo de pontos	9	6	9
Épsilon	0.0011	0.001	N/A
Número de *Clusters*	31	39	35
Silhouette Score	0.54	0.54	0.762

A camada POI, desporto não obteve qualquer configuração válida em nenhum dos algoritmos selecionados, no filtro de rotina - semana. A mesma camada transportes também não obteve resultados no algoritmo OPTICS. Pelos resultados obtidos, o índice detetou uma estrutura forte na camada de Destinos, Eventos, *Automotive* e *Landmarks* independentemente do algoritmo e do filtro utilizado. Quanto à categoria de Negócios e Serviços, assim como a de Transportes beneficiam se forem analisadas de forma separada entre a semana e ao fim-de-semana. Na categoria Comunidade e Governo, assim como na de Viagens, a estrutura apresenta-se forte apenas durante a semana.

5.2 Resultados

Depois de definidos os parâmetros a aplicar para cada camada no seu respetivo filtro, procedeu-se à execução das correlações já referidas na secção 4.2.2. Para cada camada a

analisar nas diferentes correlações, utilizaram-se as melhores configurações dos três algoritmos, com o objetivo de verificar qual a combinação, que evidencia uma correlação maior. As correlações que revelaram um valor elevado, acima ou igual a 0.65, foram alvo de análise de visualização no ArcMap. Foram obtidos os seguintes resultados, representados nas tabelas 48, 49 e 50:

- Correlação A: entre destinos e eventos ou pontos de interesse. Baseia-se na soma do número de pontos da camada destinos (para cada *cluster*) à volta da camada pontos de interesse ou eventos, num raio de 500 metros entre a soma do número de pontos (para cada *cluster*) da camada de pontos de interesse ou eventos um raio de 500 metros.

Tabela 48 - Resultados da correlação A

Filtro de Rotina						
Camada	**Geral**		**Fim de Semana**		**Semana**	
	r	p	r	p	r	p
Eventos	**0.65**	**0.003**	**0.96**	**0.00000018**	**0.91**	**0.004**
Automotive	0.46	0.001	0.18	0.22	**0.66**	**0.019**
Negócios e Serviços	**0.7**	**0.0001**	0.32	0.009	**0.83**	**0.0000927**
Landmarks	0.16	0.52	-0.22	0.39	-	-
Retalho	**0.72**	**0.00002**	**0.65**	**0.013**	0.55	0.000059
Comunidade e Governo	0.45	0.001	0.4	0.000017	0.37	0.04
Saúde	0.53	0.00005	0.63	0.000023	0.33	0.001
Viagens	**0.7**	**0.000016**	0.611	0.0000002	**0.80**	**0.000012**
Transportes	0.55	0.19	**0.68**	**0.04**	**0.67**	**0.20**
Social	**0.86**	**0.0000008**	**0.83**	**0.000000029**	**0.7**	**0.0000786**

De acordo com estes resultados obtidos, visualizou-se na ferramenta do ArcMap as melhores correlações (acima ou igual a 0.7), com o objetivo de identificar se os locais estão a ser cobertos atualmente pelas paragens STCP. As imagens obtidas podem ser consultadas nos ANEXOS – Anexo G.

- Correlação B: entre eventos e destinos. Popularidade dos eventos entre o número de destinos à distância de 500 metros.

Tabela 49 - Resultados da correlação B

Filtro de Rotina		
Geral	**Fim de Semana**	**Semana**
Algoritmo da camada destinos: DBSCAN	Algoritmo da camada destinos: DBSCAN	Algoritmo da camada destinos: DBSCAN
Algoritmo da camada eventos: HDBSCAN	Algoritmo da camada eventos: HDBSCAN	Algoritmo da camada eventos: DBSCAN
r = 0.17	r = - 0.3	r = -0.08
p = 0.43	p = 0.19	p = 0.78

Os valores obtidos na tabela 48 foram os mais significantes de acordo com o valor de p e r.

- Correlação C: entre eventos e destinos. Popularidade de eventos *outliers* entre o número de destinos a 500 metros de distância.

Tabela 50 - Resultados da correlação C

Filtro de Rotina		
Geral	**Fim de Semana**	**Semana**
Algoritmo da camada destinos: OPTICS	Algoritmo da camada destinos: OPTICS	Algoritmo da camada destinos: DBSCAN
Algoritmo da camada eventos: HDBSCAN (*Outliers* resultantes da aplicação do algoritmo):	Algoritmo da camada eventos: DBSCAN(*Outliers* resultantes da aplicação do algoritmo):	Algoritmo da camada eventos: OPTICS(*Outliers* resultantes da aplicação do algoritmo):
p=-0.09	p=0.42	p=0.38
r=0.6	r=0.01	r=0.03
Número de *Outliers*: 32	Número de *Outliers*: 35	Número de *Outliers*: 30

Os valores obtidos na tabela 48 foram os mais significantes de acordo com o valor de p e r.

5.3 Discussão

A correlação A, apresentou bons resultados e permitiu identificar diversos tipos de POIs e eventos que contém muitos destinos à sua volta. A correlação positiva observou-se em algumas categorias de POIs, sendo de realçar, que as camadas que obtiveram maior correlação foram: o social, em primeiro lugar, e os eventos em todos os filtros aplicados. Conclui-se, portanto, que os locais com maior procura por parte dos utilizadores da aplicação do *SenseMyFeup* foram as regiões que possuem pontos de interesse relacionados com a categoria social. A ida a eventos registou igualmente um valor de correlação bastante alto, conclui-se que as zonas de maior procura serão também as zonas onde existem eventos. POIs com categorias de negócios e serviços, retalho e viagens registaram igualmente bons valores, o que leva igualmente a concluir que são zonas com procura elevada.

A correlação B, não apresentou bons resultados e não foi possível encontrar qualquer tipo de relação com o nível de popularidade das localizações de eventos com o número de destinos à volta delas. O valor da variável p, na correlação foi bastante alto, o que indica que a probabilidade de a correlação não existir é alta. Para além disto, os valores de r, também foram relativamente baixos, próximos de zero. Esta correlação significa que a grande maioria dos utilizadores da aplicação *SenseMyFeup*, não visitaram eventos com elevada popularidade.

O facto da correlação C, ter resultados superiores à correlação B, indica que quando se pretende analisar este tipo de problema usando a técnica de *clustering*, neste caso concreto em relação a eventos, tem de se ter em conta também aqueles casos que embora não estejam a ser englobados pelo algoritmo, também podem ser representativos de lugares com atratividade elevada.

6 CONCLUSÕES E TRABALHO FUTURO

Neste capítulo são descritas as conclusões acerca do trabalho desenvolvido e dadas sugestões futuras para melhoria do trabalho.

6.1 Conclusões

A mobilidade é sem dúvida um dos principais desafios de hoje e com o contínuo crescimento populacional torna-se inevitável a criação de soluções capazes de satisfazer as necessidades dos cidadãos. Com a elaboração deste trabalho propôs-se uma metodologia, que pode ser usada para análise de oferta de sistemas de transportes, de forma a identificar quais os locais mais procurados atualmente na cidade e as possíveis razões para esta procura.

Apesar de atualmente não ser possível conseguir um acesso completo aos dados recolhidos por voluntários em plataformas deste tipo devido ao Regulamento Geral de Proteção de Dados. No contexto deste trabalho de projeto, nem mesmo a equipa que desenvolve a aplicação *SenseMyFEUP* consegue identificar um mesmo utilizador em dias diferentes, devido à política de proteção de dados adotada. Neste cenário, foi proposta uma metodologia de análise de *clustering* dos dados que representam a oferta (eventos e diferentes categorias de pontos de interesse) e procura (destinos dos utilizadores). Foi realizada uma análise de correlação entre a oferta e a procura de forma a identificar as possíveis razões para o motivo destas viagens.

Pelos resultados obtidos é possível concluir que uma utilização alargada de plataformas de *crowdsensing* por parte dos utilizadores é uma mais valia na análise e planeamento da oferta de transportes públicos. No problema em concreto em foco deste projeto, o horário fora da rotina, consegui-se encontrar padrões em dados que à partida não se se poderia prever se analisados de forma individual.

Ainda no âmbito deste projeto e como contribuição do estudo realizado, o resumo comparativo dos trabalhos relacionados que são apresentados no capítulo 2 deram origem a uma publicação e apresentação do artigo "*Urban Mobility: Mobile Crowdsensing Applications"* no *9th International Symposium on Ambient Intelligence (ISAmI 2018),* em Junho de 2018.

6.2 Trabalho Futuro

Como trabalho futuro, são apresentadas aqui, as seguintes sugestões para melhoria do trabalho:

- Obtenção de mais dados sobre a popularidade dos pontos de interesse e usar esse fator para realizar mais correlações, que possam auxiliar na identificação de zonas de maior volume atratividade de pessoas.
- Obtenção de mais dados sobre destinos e eventos na cidade metropolitana do Porto, através da realização de mais recolhas de dados.
- Adicionar as duas novas rotas STCP, que de momento, não presentes na base de dados do *SenseMyCity*: a linha 209 de Pasteleira até Prelada e a linha ZC, de Corujeira até Areias.
- Com base nas diversas variáveis apresentadas nas correlações usadas, construir uma árvore de decisão, que preveja o número de destinos que poderão ser realizados.
- Armazenar na base de dados os horários atuais das paragens STCP.
- Para além das linhas de transporte de autocarros, expandir a metodologia do trabalho apresentado também às linhas do metro.
- Utilização de modelos de regressão linear de modo a melhorar/adequar o peso de cada variável que constitui a popularidade dos eventos.

REFERÊNCIAS

[1] UN-Habitat, *Urbanization and Development: Emerging Futures*. 2016.

[2] J. Dargay, D. Gately, and M. Sommer, "Vehicle ownership and income growth, worldwide: 1960-2030," *Energy J.*, vol. 28, no. 4, pp. 143–170, 2007.

[3] World Health Organization, *World Health statistics 2014*. 2014.

[4] J. G. P. Rodrigues, A. Aguiar, and J. Barros, "SenseMyCity: Crowdsourcing an Urban Sensor," no. December, 2014.

[5] "The Statistics Portal." [Online]. Available: http://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide. [Accessed: 11-Dec-2017].

[6] T. Systems, "Transportation Systems," pp. 1–24, 2018.

[7] M. Freitas, "URBY.SENSE - Análise e previsão de mobilidade urbana fora da rotina com base em pegadas digitais," 2018. [Online]. Available: http://www.poci-compete2020.pt/noticias/detalhe/Proj16848_UrbySense. [Accessed: 22-Sep-2018].

[8] E. Polisciuc and A. Alves, "Understanding Urban Land Use through the Visualization of Points of Interest," no. September, pp. 51–59, 2015.

[9] J. G. P. Rodrigues, A. Aguiar, and C. Queiŕos, "Opportunistic mobile crowdsensing for gathering mobility information: Lessons learned," *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, no. 978, pp. 1654–1660, 2016.

[10] L. D. E. L. Idiap, "Computational Analysis of Urban Places Using Mobile Crowdsensing," vol. 7243, 2016.

[11] X. Guo, B., Zhang, D., Wang, Z., Yu, Z., and Zhou, "Mobile crowd sensing and computing: The review of an emerging humanpowered sensing paradigm," in *ACM Comput. Surv*, 2013.

[12] M. Ehatisham-ul-Haq *et al.*, "Authentication of Smartphone Users Based on

Activity Recognition and Mobile Sensing," *Sensors*, vol. 17, no. 9, p. 2043, 2017.

[13] S. Faye, W. Bronzi, I. Tahirou, and T. Engel, "Characterizing user mobility using mobile sensing systems," *Int. J. Distrib. Sens. Networks*, vol. 13, no. 8, 2017.

[14] C. Saiprasert, S. Thajchayapong, T. Pholprasit, and C. Tanprasert, "Driver Behaviour Profiling using Smartphone Sensory Data in a V2I Environment," no. June, 2016.

[15] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth, "Understanding human mobility from Twitter," *PLoS One*, vol. 10, no. 7, pp. 1–16, 2015.

[16] F. Pereira, C. Carrion, F. Zhao, C. D. Cottrill, C. Zegras, and M. E. Ben-Akiva, "The Future Mobility Survey: overview and preliminary evaluation," *Proc. East. Asia Soc. Transporation Stud.*, vol. 9, 2013.

[17] R. Frank and T. Engel, "Adaptive Activity and Context Recognition Using Multimodal Sensors in Smart Devices Adaptive Activity and Context Recognition using Multimodal Sensors in Smart Devices," no. November 2016, 2015.

[18] M. A. Shafique and E. Hato, "Travel mode detection with varying smartphone data collection frequencies," *Sensors (Switzerland)*, vol. 16, no. 5, 2016.

[19] Y. Zhou *et al.*, "Understand Urban Human Mobility through Crowdsensed Data," pp. 1–7, 2018.

[20] M. S. Hasala, B. Lau, P. Lik, and V. S. Kadaba, "Identifying Points of Interest for Elderly in Singapore through Mobile Identifying Points of Interest for Elderly in Singapore through Mobile Crowdsensing," no. February, 2017.

[21] J. Zhang, B. Guo, H. Chen, Z. Yu, J. Tian, and A. Chin, "Public sense: Refined urban sensing and public facility management with crowdsourced data," *Proc. - 2015 IEEE 12th Int. Conf. Ubiquitous Intell. Comput. 2015 IEEE 12th Int. Conf. Adv. Trust. Comput. 2015 IEEE 15th Int. Conf. Scalable Comput. Commun. 20*, no. August, pp. 1407–1412, 2016.

[22] Y. Zheng *et al.*, "GeoLife : Managing and Understanding Your Past Life over Maps," no. 49, pp. 4–5.

[23] S. Hasan, W. Lafayette, W. Lafayette, and S. V Ukkusuri, “Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media,” 2013.

[24] G. M. Vazquez-prokopec *et al.*, “Using GPS Technology to Quantify Human Mobility , Dynamic Contacts and Infectious Disease Dynamics in a Resource-Poor Urban Environment,” vol. 8, no. 4, pp. 1–10, 2013.

[25] S. Jiang, J. Ferreira, and M. C. Gonzalez, “Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data : A Case Study of Singapore,” pp. 1–11, 2016.

[26] M. K. El Mahrsi, E. Côme, L. Oukhellou, and M. Verleysen, “Clustering Smart Card Data for Urban Mobility Analysis,” vol. 18, no. 3, pp. 712–728, 2017.

[27] H. Wang, G. Liu, J. Duan, and L. Zhang, “Detecting Transportation Modes Using Deep Neural Network,” no. 5, pp. 1132–1135, 2017.

[28] R. Sathya and A. Abraham, “Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification,” *Int. J. Adv. Res. Artif. Intell.*, vol. 2, no. 2, pp. 34–38, 2013.

[29] C. Analysis and B. Concepts, “Cluster Analysis :”

[30] B. P., “Survey Of Clustering Data Mining Techniques. In Grouping Multidimensional Data.,” in *Springer Berlin Heidelberg*, 2006.

[31] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[32] F. José and N. Marques, “A Constraint-Based Clustering Algorithm for Detection of Meaningful Places,” 2014.

[33] S. J., *Business Intelligence: Making Decisions through Data Analytics*. 2011.

[34] and B. P. C. Shmueli G., Patel N. R., *Data Mining for Business Intelligence*. 2010.

[35] and P. P. Singh S., Singh M., Apte C., “Pattern Recognition and Image Analysis,” 2005.

[36] W. W.A.M., “Analysis of urban traffic patterns using clustering,” 2007.

[37] A. Mathematics, “ijpam.eu,” vol. 115, no. 8, pp. 425–430, 2017.

[38] S. Kaur, “SURVEY OF DIFFERENT DATA,” vol. 5, no. 5, pp. 584–588, 2016.

[39] E. Kolatch, “Clustering Algorithms for Spatial Databases : A Survey,” pp. 1–22, 2001.

[40] E. Chandra, “A Survey on Clustering Algorithms for Data in Spatial Database Management Systems,” vol. 24, no. 9, pp. 19–26, 2011.

[41] P. Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, *Automatic subspace clustering of high dimensional data for data mining applications*. 1998.

[42] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[43] K. P. Jacob, “Spatial Clustering Algorithms- An Overview,” no. August 2015.

[44] “IAML: K-means Clustering.”

[45] M. Kaushik and B. Mathur, “Comparative Study of K-Means and Hierarchical Clustering Techniques,” no. June 2014, 2016.

[46] and R. P. Kaufman L., “PAM.pdf,” in *Clustering by Means of Medoids*, 1987.

[47] P. . KAUFMAN, L and ROUSSEEUW, “Finding Groups in Data,” 1990.

[48] Ê. Ý. Ì. Æ. Ý, “CLARANS: A method for clustering objects for spatial data mining,” no. October, 2002.

[49] P. J. Kaufman, L., & Roussew, “Finding Groups in Data - An Introduction to Cluster Analysis,” 1990.

[50] M. Hill and M. Hill, “CURE : An Efficient Clustering Algorithm for Large E = c c IIP,” 1998.

[51] “CURE: An Efficient Clustering Algorithm for Large Databases.”

[52] S. Birch, “An Efficient Data Clustering Databases Method for Very Large,” vol. 1, pp. 103–114.

[53] M. Hill and M. Hill, “ROCK : A Robust Clustering Algorithm for Categorical Attributes.”

[54] J. A. S. Almeida, L. M. S. Barbosa, A. A. C. C. Pais, and S. J. Formosinho, “Improving hierarchical cluster analysis : A new method with outlier detection and automatic clustering,” vol. 87, pp. 208–217, 2007.

[55] G. H. Shah, C. K. Bhensdadia, and A. P. Ganatra, “An Empirical Evaluation of Density-Based Clustering Techniques,” no. 1, pp. 216–223, 2012.

[56] M. Ester, H. Kriegel, X. Xu, and D.- Miinchen, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” 1996.

[57] S. Kockara, M. Mete, B. Chen, and K. Aydin, “Analysis of density based and fuzzy c-means clustering methods on lesion border extraction in dermoscopy images,” no. October, 2010.

[58] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander, “OPTICS : Ordering Points To Identify the Clustering Structure,” 1999.

[59] G. Redinger and M. Hunner, “Visualization of the Optics Algorithm,” 2017.

[60] R. Chauhan, P. Batra, and S. Chaudhary, “A Survey of Density Based Clustering Algorithms 1,” vol. 8491, pp. 5–7, 2014.

[61] Ricardo J. G. B. CampelloDavoud MoulaviJoerg Sander, “Density-Based Clustering Based on Hierarchical Density Estimates,” 2013.

[62] A. H.-H. Gabriel, “DENCLUE 2.0: Fast Clustering Based on Kernel Density Estimation,” 2007.

[63] H. Shah, K. Napanda, and D. Lynette, “Density Based Clustering Algorithms,” no. 11, 2015.

[64] W. Wang and R. Muntz, “STING : A Statistical Information Grid Approach to Spatial Data Mining,” 1997.

[65] P. Rani, “A Survey on STING and CLIQUE Grid Based Clustering Methods,” vol. 8, no. 5, pp. 2015–2017, 2017.

[66] et al Luis, Yunior, “UrbanSense: An urban-scale sensing platform for the Internet of Things,” in *Smart Cities Conference (ISC2), 2016 IEEE International*, 2016.

[67] S. N. R. Bahador Khaleghi, Alaa Khamisa, Fakhreddine O.Karray, “Multisensor data fusion: A review of the state-of-the-art,” 2013.

[68] S. H. Gong L, Morikawa T, Yamamoto T, “Deriving personal trip data from GPS data: a literature review on the existing methodologies,” in *Procedia-Social and Behavioral Sciences*, 2014.

[69] M. Sousa, “Porto – cidade de cultura e lazer,” 2012. [Online]. Available: https://guia-viagens.aeiou.pt/porto-capital-europeia-da-cultura-em-2001-2394/. [Accessed: 11-Dec-2018].

[70] K. Pearson, “Mathematical contributions to the theory of evolution.On a form of spurious correlation which may arise when indices are used in the measurement of organs,” in *60 Proceedings of the Royal Society of London*.

[71] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, “Internal versus External cluster validation indexes,” vol. 5, no. 1, 2011.

[72] J. Dudoit, S., & Fridlyand, “A prediction-based resampling method for estimating the number of clusters in a dataset,” in *Genome biology*, 2002.

[73] W. H. Tseng, G. C., & Wong, “Tight clustering: a resampling-based approach for identifying stable and tight patterns in data,” in *Biometrics*, 2005.

[74] P. J. Rousseeuw, “Silhouettes : a graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math. 20*, vol. 20, pp. 53–65, 1987.

ANEXOS

Nesta secção apresenta-se a proposta que deu origem a este projeto. Desenvolvida pela Doutora Ana Cristina Oliveira Alves em Setembro de 2017 e pelo Doutor Rui Gomes.

Anexo A: PROPOSTA DE ESTÁGIO

PROPOSTA DE PROJETO

MESTRADO em INFORMÁTICA E SISTEMAS

Especialização em Desenvolvimento de Software

Ano Lectivo de 2017/2018

TEMA

Análise de Dados e *Machine Learning* aplicados à Mobilidade Urbana

Palavras-chave: Computação Ubíqua, Análise de Dados, Aprendizagem Automática, Mobilidade

1. ÂMBITO

A mobilidade é, cada vez mais, influenciada por fatores que dificultam a circulação (congestionamento de tráfego, informação não atualizada, eventos, acidentes, supressão de vias, obras, entre outros). Os constrangimentos do dia-a-dia dos cidadãos (por exemplo, morar a vários quilómetros do local de trabalho, o deixar os filhos na escola, entre outros) e a falta de serviços integrados de mobilidade levam a que haja uma menor utilização dos transportes coletivos, a um aumento do trânsito nas áreas urbanas e, consequentemente, a um aumento do tráfego e congestionamentos, maior poluição e diminuição da eficiência energética.

É claro, neste momento, que ter mais informação atualizada, mais integração e uma melhor coordenação dos transportes é fulcral para uma melhor mobilidade, inclusiva, inteligente e ajustada às necessidades dos utentes. Considera-se que, a mobilidade do futuro terá que ter em conta diferentes fontes de dados, com diversos níveis de granularidade e frequência de atualização adequados. Por esse motivo, consideramos que um cenário de mobilidade urbana requer a integração de várias

fontes de dados heterogéneas para fornecer informação acerca do trânsito, condições ambientais, sentimentos das pessoas e suas opiniões.

As fontes de dados externas podem ser classificadas em quatro tipos principais: a) dados estacionários - descrevem os elementos estáticos que compõem o ambiente alvo. Inclui informação geográfica, infraestruturas de transportes, pontos de interesse e outra informação que possa ser relevante para perceber a localização dos componentes dentro do cenário. b) dados espaciais dinâmicos - contém informação acerca de um específico ponto no tempo. Inclui informação sobre o trânsito georreferenciado, estado do trânsito e notícias, existência de eventos, e todos os tipos de informação temporal útil para medir as condições de mobilidade. c) dados ambientais - apresenta informação acerca das condições do ambiente assim como previsões do estado do tempo que possam ser importantes para perceber a mobilidade dos agentes; d) dados de redes sociais - fornecem fluxos de dados úteis para extrair informação acerca de sentimentos e eventos imprevisíveis.

Estes dados, principalmente os provenientes de fontes externas precisam também de ser tratados antes de serem armazenados. De forma resumida, as principais tarefas de tratamento de dados são: 1) limpeza de dados – preencher valores em falta, reduzir o ruído dos dados, identificar ou remover *outliers* e inconsistências; 2) integração de dados – estabelecimento de relação entre os dados das várias fontes; 3) transformação de dados – normalização e agregação; 4) redução de dados – obter representações menores em volume mas com o mesmo poder analítico; 5) Discretização de dados – técnicas de redução de dados com especial importância para dados numéricos. Após o tratamento, os dados podem ser usados para acesso ou para processamento. Dados derivados podem ser gerados através destes algoritmos ou mesmo de técnicas de *machine learning*. Estes dados podem ser armazenados no sistema para posterior consulta, análise e também para melhoria da performance de outros algoritmos de processamento.

2. OBJECTIVOS

O objetivo deste projeto industrial é estudar e aplicar diferentes técnicas de Análise de Dados e *Machine Learning* a partir de diferentes fontes de dados de mobilidade a fim de auxiliar: a determinação de métricas de utilização de transportes coletivos, compreender as necessidades dos utentes e otimizar os serviços dos operadores e entidades reguladoras/coordenadoras da mobilidade.

Este trabalho terá uma forte componente experimental, que deverá abranger: (i) estudo dos diferentes tipos de fontes de dados de Mobilidade disponíveis; (ii) estudo

e exploração de técnicas de aprendizagem aplicadas à Mobilidade (e.g. análise de dados para extração de características, seleção de algoritmos de *machine learning*); (iii) seleção das técnicas com melhor performance (iv) desenvolvimento de um serviço que integre as técnicas anteriores e permita a resposta conjuntos de dados de Mobilidade.

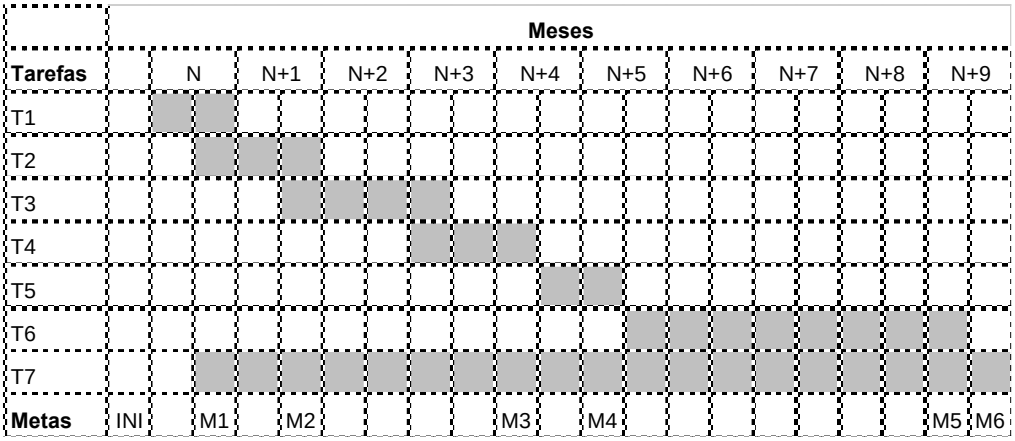
3. PROGRAMA DE TRABALHOS

O Projeto consistirá nas seguintes atividades e respetivas tarefas:

- T1 – Levantamento do estado da arte de fontes de dados de mobilidade disponíveis;
- T2 – Levantamento do estado da arte de técnicas de aprendizagem aplicada à mobilidade;
- T3 – Experimentação das técnicas de aprendizagem sobre os dados de mobilidade disponíveis;
- T4 - Seleção das técnicas de aprendizagem com melhor performance;
- T5 – Análise de requisitos de um serviço para aplicação das técnicas de aprendizagem;
- T6 – Desenvolvimento do serviço e testes;
- T7 - Elaboração do relatório final e escrita de artigo científico

4. CALENDARIZAÇÃO DAS TAREFAS

As Tarefas acima descritas, incluindo os testes de validação de cada módulo, serão executadas de acordo com a seguinte calendarização:



O plano de escalonamento dos trabalhos é apresentado em seguida:

INI		Início dos trabalhos
M1	(INI + 1 mês)	Tarefa T1 terminada

M2	(INI + 2 meses)	Tarefa T2 terminada
M3	(INI + 4,5 meses)	Tarefa T3 e T4 terminada
M4	(INI + 5,5 meses)	Tarefa T5 terminada
M5	(INI + 9,5 meses)	Tarefa T6 terminada
M6	(INI + 10 meses)	Tarefa T7 terminada

5. RESULTADOS

Os resultados do estágio serão consubstanciados num conjunto de documentos a elaborar pelo aluno de acordo com o seguinte plano:

M1: Relatório com o estado da arte de fontes de dados de mobilidade disponíveis

M2: Relatório com o estado da arte de técnicas de aprendizagem aplicada à mobilidade

M3: Relatório com os resultados das técnicas de aprendizagem e seleção das melhores sobre os dados de mobilidade disponíveis;

M4: Relatório técnico com o levantamento de requisitos do serviço;

M5: Relatório técnico com a descrição do serviço e plano de testes;

M6: Relatório final de projeto e artigo

6. LOCAL DE TRABALHO

DEIS

7. METODOLOGIA

Organização de um Dossier de Projeto na *cloud* e reuniões quinzenais.

8. ORIENTAÇÃO

Ana Cristina Oliveira Alves (aalves@isec.pt)

Professora Adjunta

Rui Gomes (ruig@dei.uc.pt)

Investigador (membro-pleno) do CISUC

Orientando: João Pedro Fernandes Simões (a21220126@alunos.isec.pt)

Aluno do Mestrado de Informática e de Sistemas

9. CARACTERIZAÇÃO

- Data de início: Outubro de 2017
- Data de fim: Setembro de 2018

Anexo B: ESQUEMA *SENSEMYCITY*

Nesta secção encontra-se descrito o esquema *SenseMyCity*.

Tabela: *Trips*	
Atributo	**Descrição**
ID	Identificador da viagem
Daily_user_id	Identificador diário do utilizador
Sessions_id	Lista de *id's* correspondente às sessões
Seconds_start	Data de início da viagem
Lat_start	Coordenada geográfica relativa ao início da viagem: Latitude (formato: *WGS84*)
lat_end	Coordenada geográfica relativa ao fim da viagem: Latitude (formato: *WGS84*)
Lon_start	Coordenada geográfica relativa ao início da viagem: Longitude (formato: *WGS84*)
Lon_end	Coordenada geográfica relativa ao fim da viagem: Longitude (formato: *WGS84*)
Distance	Distância percorrida
Foot	Booleano indicador do modo de viagem: 0 se não foi a pé, 1 se foi.
Bike	Booleano indicador do modo de viagem: 0 se não foi de bicicleta, 1 se foi.
Car	Booleano indicador do modo de viagem: 0 se não foi de carro, 1 se foi.
Bus	Booleano indicador do modo de viagem: 0 se não foi de autocarro, 1 se foi.
Metro	Booleano indicador do modo de viagem: 0 se não foi de metro, 1 se foi.
Travel_mode	Modo de viagem utilizado

Role	Representa o tipo de utilizador que registou a atividade. Exemplo: professor, estudante ou funcionário.

Tabela: *Sessions*	
Atributo	**Descrição**
Session_id	Identificador único da sessão
Daily_user_id	Identificação da sessão do utilizador do dia
Seconds_start	Data de início da recolha
Lat_start	Coordenada latitude correspondente à posição do começo da recolha.
Lon_start	Coordenada longitude correspondente à posição do começo da recolha.
Seconds_end	Fim da recolha
Lat_end	Coordenada latitude correspondente à posição do fim da recolha.
Lon_end	Coordenada longitude correspondente à posição do fim da recolha.
Distance	Distância percorrida durante a sessão

Tabela: *Segments*	
Atributo	**Descrição**
Session_id	Identificador da sessão
Segment_id	Identificador do segmento
Seconds_start	Início do segmento

Lat_start	Coordenada latitude correspondente à posição do início do segmento.
Lon_start	Coordenada longitude correspondente à posição do início do segmento.
Lat_end	Coordenada latitude correspondente à posição do fim do segmento.
Lon_end	Coordenada longitude correspondente à posição do fim do segmento.
Length	Comprimento do segmento
Speed_profile	Perfil de velocidade
Acceleration_profile	Perfil de aceleração
Walking	Se realizou o percurso de segmento a pé
Travelmodes_score	Valor obtido pela inferência de atividade

Tabela: *TravelModeSurvey*	
Atributo	**Descrição**
Session_id	Identificador da sessão
Nottraveled	Resposta do questionário: indicador que permite saber que na realidade a sessão reportada não correspondeu a uma viagem
Foot	Resposta do questionário: indica se o utilizador realizou o seu percurso a pé
Bicycle	Resposta do questionário: indica se o utilizador realizou o seu percurso de bicicleta
Motorcycle	Resposta do questionário: indica se o utilizador realizou o seu percurso de moto.
Car	Resposta do questionário: indica se o utilizador realizou o seu percurso de carro

Bus	Resposta do questionário: indica se o utilizador realizou o seu percurso de autocarro
Metro	Resposta do questionário: indica se o utilizador realizou o seu percurso de metro
Train	Resposta do questionário: indica se o utilizador realizou o seu percurso de comboio
other	Resposta do questionário: indica se o utilizador realizou o seu percurso por outros meios

Tabela: ***Activity***	
Atributo	**Descrição**
Session_id	Identificador da sessão
Seconds	Data de início da sessão
In_vehicle	Indica se o utilizador andou de carro
On_bicycle	Indica se o utilizador esteve a andar de bicicleta
On_foot	Indica se o utilizador esteve a andar a pé
Running	Indica se o utilizador esteve a correr
Walking	Indica se o utilizador esteve a andar
Still	Indica se o utilizador esteve parado
Unknow	Indica se o utilizador realizou outra atividade não reconhecida entre o sistema.

Anexo C: ESQUEMA *SOCIAL_NETWORK*

Nesta secção encontra-se descrito o esquema *Social_Network.*

Tabela: *Forsquare_venues*	
Atributo	**Descrição**
Id	Identificador do POI
Name	Nome do POI
Rating	*Rating* do POI (pontuação)
Stats_checkincount	Número de *check-ins* do POI
Stats_visitscount	Número de visitantes do POI
Likes_count	Número de gostos do POI
Description	Descrição do POI
Contact_instagram	Contacto do POI no *Instagram*
Contact_facebookname	Nome do POI no *Facebook*
Contact_phone	Contacto de telemóvel do POI
Location_city	Cidade do POI
Location_country	País do POI
Location_postalcode	Código postal do POI
Location_lat	Latitude do POI
Location_lng	Longitude do POI
Hours_status	Horário de funcionamento do POI
Hours_isopen	Horário de abertura do POI
Categories	Categorias do POI
URL	URL do POI

Tabela: *infoPorto_events*

Atributo	Descrição
ID	Identificador único do evento
Name	Nome do evento
Date_start	*Timestamp* com a data de início do evento
Date_end	*Timestamp* com a data de fim do evento
Location	Localização do evento
Location_longitude	Localização geográfica do evento: Longitude (formato: *WGS84*)
Location_latitude	Localização geográfica do evento: Latitude (formato: *WGS84*)
Website	*Website* do evento
Fb_event_id	Identificador do evento no *Facebook*
Category	Categoria do evento
Description	Descrição do evento

Tabela: *Facebook_events*	
Atributo	**Descrição**
ID	Identificador único do evento
Name	Nome do evento
Attending_count	Número de pessoas que indicaram que vão ao evento
Is_cancelled	Indicação do cancelamento do evento
Maybe_count	Número de pessoas que indicaram que talvez vão ao evento

Noreply_count	Número de pessoas que estão convidadas, mas que não responderam à ida ou não ao evento
Description	Descrição do evento
Date_start	Data de início do evento
Date_end	Data de fim do evento
Fb_place_id	Identificador do POI no *Facebook*

Tabela: *Factual_pois*	
Atributo	**Descrição**
Factual_id	*GUID* (*globally unique identifier*) – identificar único de cada POI
Name	Nome do POI
Tel	Número de telefone do POI
Locality	Localidade do POI
Region	Região do POI
Latitude	Coordenada geográfica: Latitude (formato: *WGS84*)
Longitude	Coordenada geográfica: Longitude (formato: *WGS84*)
Category_ids	Lista de *ID's* das categorias aos quais o POI pertence
Category_labels	Lista de nomes das categorias aos quais o POI pertence
Postcode	Código postal do POI
Address	Morada do POI
Website	*URL* do POI

Hours	Estrutura *JSON* com o horário de funcionamento
Neighborhood	Existência de POIS na proximidade (caso exista)
Email	Endereço de e-mail
Name_soundex	*ID* baseado no som do POI

Tabela: *Facebook_places*	
Atributo	**Descrição**
Id	Identificador único do POI
Name	Nome do POI
Fan_count	Número de pessoas que indicaram que gostam do POI
Hours	Horário de funcionamento do POI
Category_list	Lista de *ids* a que o POI pertence
Category	Lista de categorias a que o POI pertence
Checkins	Número de *checkins* registado no POI
Cidade	Cidade do POI
País	País do POI
Latitude	Localização geográfica do POI: Latitude
Longitude	Localização geográfica do POI: Longitude
Rua	Rua do POI
Código Postal	Código postal do POI
WebSite	*Website* do POI

Anexo D: ESQUEMA *PUBLIC*

Nesta secção encontra-se descrito o esquema *Public*.

Tabela: *Stcstop*	
Atributo	**Descrição**
Name	Nome da paragem
Stop	Descrição da paragem
Stop_lat	Coordenada geográfica da posição da paragem. Latitude (formato: *WGS84*)
Stop_lon	Coordenada geográfica da posição da paragem. Longitude (formato: *WGS84*)
pos	Identificador da ordem da paragem na linha.

Tabela: *Metrostops*	
Atributo	**Descrição**
Name	Nome da paragem
Lat	Descrição da paragem
lon	Coordenada geográfica da posição da paragem. Latitude (formato: *WGS84*)
Geo	Coordenada geográfica da posição da paragem. Longitude (formato: *WGS84*)
id	Identificador único da paragem

Anexo E: ESQUEMA *ENVIRONMENT*

Nesta secção encontra-se descrito o esquema *Environment*.

Tabela: *Airquality*	
Atributo	**Descrição**
Deployment_id	Identificador do *deployment*
Seconds	Data do *deployment*
O3_data	Quantidade de presença de oxigénio
Particles_p1	Quantidade de presença por partículas

Tabela: *Deployment*	
Atributo	**Descrição**
Deployment_id	Identificador do *deployment*
Description	Descrição do *deployment*
Node_id	Identificador do nodo do *deployment*
Location_id	Identificador da localização do *deployment*
Dep_start	Data de início da recolha dados
Dep_end	Data de fim da recolha de dados

Tabela: *BasicEnvironment*	
Atributo	**Descrição**
Deployment_id	Identificador do *deployment*
Seconds	Data do *deployment*

Temperature_data	Valores da temperatura
Humidity_data	Valores da humidade
Luminosity_lux	Valores de luminosidade

Tabela: ***Noise***	
Atributo	**Descrição**
Deployment_id	Identificador do *deployment*
Seconds	Data do deployment
Noise_data	Valor do *noise*

Tabela: ***Weather***	
Atributo	**Descrição**
Deployment_id	Identificador único do *deployment*
Seconds	Data do *deployment*
Windspeed_data	Velocidade do vento
Winddirection_data	Direção do vento
Precipitation_data	Nível de precipitação
Solarradiation_data	Radiação solar

Tabela: ***Location***	
Atributo	**Descrição**
Location_id	Identificador único da localização
Node_lat	Coordenada geográfica: Latitude (formato: *WGS84*)
Node_lon	Coordenada geográfica: Longitude (formato: *WGS84*)

Description	Descrição da morada

Anexo F: CATEGORIAS DO FACTUAL

Nesta secção apresenta-se a taxonomia do Factual pela qual os POIs são categorizados quanto ao seu tipo principal de serviço. As categorias estão dispostas segundo uma hierarquia. As categorias mais genéricas são as seguintes:

- *Automotive* (Automóvel)
- *Community and Government* (Comunidade e Governo)
- *Healthcare* (Saúde)
- *Landmarks (*Monumentos*)*
- *Retail* (Retalho)
- *Businesses and Services* (Negócios e Serviços)
- *Social* (Social)
- *Sports and Recreation* (Desportos)
- *Transportation* (Transportes)
- *Travel* (Viagens)

Foi descarregada utilizando a *API* do Factual pelo *endpoint Categories*, em Agosto de 2018.

ID	*Parent*	*Descrição EN*	*Descrição PT*	*Base*
1	0	Factual Places	Lugares Factual	1
2	1	Automotive	Automóvel	2
3	2	Car Appraisers	Peritos de Automóveis	2
4	2	Car Dealers and Leasing	Aluguer e Leasing de Automóveis	2
5	4	Used Cars	Carros Usados	2
6	2	Car Parts and Accessories	Partes de Automóveis e Acessórios	2
7	2	Car Wash and Detail	Lavagem e Limpeza Automóvel	2
8	2	Classic and Antique Car	Carros Clássicos e Antigos	2
9	2	Maintenance and Repair	Manutenção e Reparação	2
10	9	Oil and Lube	Óleo e Lubrificante	2
11	9	Smog Check	Controlo de Poluição	2
12	9	Tires	Pneus	2
13	9	Transmissions	Transmissões	2
14				
15				

	2	Motorcycles, Mopeds and Scooters	Motociclos, Ciclomotores e Velocípedes	2
	14	Repair	Reparação	2
16	14	Sales	Vendas	2
17	2	RVs and Motor Homes	Roulottes e Autocaravanas	2
18				
19	2	Salvage Yards	Ferro-velho	2
20	2	Towing	Reboque	2
21				
	1	Community and Government	Comunidade e Governo	20
22	20	Animal Shelters and Humane Societies	Abrigos de Animais e Associações Humanitárias	20
23				
24	20	Cemeteries	Cemitérios	20
	20	Day Care and Preschools	Creches e Jardins de Infância	20
25	20	Disabled Persons Services	Serviços de Apoio a Pessoas com Necessidades Especiais	20
26	20	Drug and Alcohol Services	Serviços de Tratamento de Alcoolismo e Toxicodependência	20
27				
	20	Education	Ensino	20
	26	Adult Education	Ensino para Adultos	20
28	26	Art Lessons and Schools	Escolas de Arte	20
29	26	Colleges and Universities	Faculdades, Universidades e Institutos Superiores	20
30	26	Computer Training	Formação em Informática	20
31	26	Culinary Lessons and Schools	Escolas de Culinária	20
32	26	Driving Schools	Escolas de Condução	20
33	26	Fraternities and Sororities	Repúblicas	20
34	26	Primary and Secondary Schools	Escolas Primárias e Secundárias	20
35	26	Tutoring and Educational Services	Cursos Particulares e Serviços Educativos	20
36	26	Vocational Schools	Escolas Profissionais	20
37	20	Government Departments and Agencies	Ministérios e Organismos Governamentais	20
38	20	Government Lobbyists	Lobbyistas Governamentais	20
39	20	Housing Assistance and Shelters	Assistência Habitacional e Abrigos	20
40	20	Law Enforcement and Public Safety	Forças de Segurança e Ordem Pública	20
41	40	Rescue Services	Serviços de Socorro	20
42	40	Fire Stations	Quarteis de Bombeiros	20
	40	Police Stations	Esquadras de Polícia	20

43	20	Libraries	Bibliotecas	20
44	20	Military	Forças Armadas	20
45	45	Bases	Bases Militares	20
46	20	Organizations and Associations	Organizações e Associações	20
47	47	Charities and Non-Profits	Organizações de Beneficiência e sem Fim Lucrativo	20
48	47	Environmental	Ambiente	20
49	47	Youth Organizations	Organizações Juvenis	20
50	20	Post Offices	Postos de Correio	20
51	20	Public and Social Services	Serviços Públicos e Sociais	20
52	20	Religious	Religião	20
53	53	Buddhist Temples	Templos Budistas	20
54	53	Churches	Igrejas	20
55	53	Hindu Temples	Templos Hindus	20
56	53	Mosques	Mesquitas	20
57	53	Synagogues	Sinagogas	20
58				
59	20	Senior Citizen Services	Serviços de Apoio a Pessoas da Terceira Idade	20
60	59	Retirement	Reforma	20
61	20	Utility Companies	Empresas de Serviços Públicos	20
62	1	Healthcare	Assistência Médica e Sanitária	62
63	62	AIDS Resources	Recursos relacionados com SIDA	62
64	62	Assisted Living Services	Serviços de Assistência Pessoal	62
65	62	Home Health Care Services	Início Serviços de Saúde	62
66	64	Facilities and Nursing Homes	Lares e Casas de Repouso	62
67	62	Blood Banks and Centers	Centros de Recolha e Bancos de Sangue	62
68	62	Chiropractors	Quiropráticos	62
69	62	Dentists	Dentistas	62
70	62	Emergency Services	Serviços de Emergência	62
71	70	Ambulance	Ambulância	62
72	62	Holistic, Alternative and Naturopathic Medicine	Medicinas Alternativas e Naturais	62
73	72	Acupuncture	Acupunctura	62
74	62	Hospitals, Clinics and Medical Centers	Hospitais, Clínicas e Centros Médicos	62
75	62	Medical Supplies and Labs	Material Médico e Laboratórios	62
76	62	Mental Health	Saúde Mental	62
77	76	Counseling and Therapy	Aconselhamento e Terapia	62
78	76	Psychologists	Psicólogos	62

79	62	Nurses	Enfermeiros	62
80	62	Pharmacies	Farmácias	62
81	62	Physical Therapy and Rehabilitation	Terapia Física e Reabilitação	62
82	81	Sports Medicine	Medicina Desportiva	62
83	62	Physicians	Médicos	62
84	83	Anesthesiologists	Anestesistas	62
85	83	Cardiologists	Cardiologistas	62
86	83	Dermatologists	Dermatologistas	62
87	83	Ear, Nose and Throat	Otorrinolaringologistas	62
88	83	Family Medicine	Médicos de Família	62
89	83	Gastroenterologists	Gastrenterologistas	62
90	83	General Surgery	Cirurgia Geral	62
91	83	Internal Medicine	Medicina Interna	62
92	83	Neurologists	Neurologistas	62
93	83	Obstetricians and Gynecologists	Obstetras e Ginecologistas	62
94	83	Oncologists	Oncologistas	62
95	83	Ophthalmologists	Oftalmologistas	62
96	83	Orthopedic Surgeons	Cirurgiões Ortopédicos	62
97	83	Pathologists	Patologistas	62
98	83	Pediatricians	Pediatras	62
99	83	Plastic Surgeons	Cirurgiões Plásticos	62
100	83	Psychiatrists	Psiquiatras	62
101	83	Radiologists	Radiologistas	62
102	83	Respiratory	Pneumologia	62
103	83	Urologists	Urologistas	62
104	62	Podiatrists	Podólogos	62
105	62	Pregnancy and Sexual Health	Gravidez e Saúde Sexual	62
106	62	Weight Loss and Nutritionists	Perda de Peso e Nutricionistas	62
107	1	Landmarks	Marcos	107
108	107	Buildings and Structures	Edifícios e Estruturas	107
109	107	Gardens	Jardins	107
110	107	Historic and Protected Sites	Lugares Históricos e Protegidos	107
	107	Monuments and Memorials	Monumentos Históricos e Comemorativos	107

111	107	Natural	Natureza	107
112	112	Beaches	Praias	107
113				
114	112	Mountains	Montanhas	107
115	112	Forests	Florestas	107
116	112	Lakes	Lagos	107
117	112	Rivers	Rios	107
118	107	Parks	Parques	107
119	118	Natural Parks	Parques Naturais	107
120	118	Picnic Areas	Parques de Merendas	107
121	118	Playgrounds	Parques Infantis	107
122	118	Urban Parks	Parques Urbanos	107
123	1	Retail	Venda a Retalho	123
124	123	Adult	Adulto	123
125	123	Antiques	Antiguidades	123
126	123	Arts and Crafts	Artesanato	123
127	123	Auctions	Leilões	123
128	123	Beauty Products	Produtos de Beleza	123
129	123	Bicycles	Bicicletas	123
130	123	Bookstores	Livrarias	123
131	123	Cards and Stationery	Artigos de Papelaria	123
132	123	Children	Crianças	123
133	123	Computers and Electronics	Informática e Electrónica	123
134	133	Cameras	Câmeras	123
135	133	Mobile Phones	Telemóveis	123
136	133	Video Games	Jogos de Vídeo	123
137	123	Construction Supplies	Material de Construção	123
138	123	Convenience Stores	Drogarias	123
139	123	Costumes	Máscaras e Disfarces	123
140	123	Dance and Music	Música e Dança	123
141	123	Department Stores	Lojas de Departamento	123
142	123	Fashion	Moda	123
143	142	Clothing and Accessories	Roupa e Acessórios	123
144	142	Jewelry and Watches	Jóias e Relógios	123
145	142	Shoes	Sapatos	123
146	142	Swimwear	Fatos de Banho	123
147	123	Flea Markets	Mercados	123
148	123	Florists	Floristas	123
149	123	Food and Beverage	Comida e Bebida	123
150	149	Beer, Wine and Spirits	Cerveja, Vinho e Licores	123
151	149	Candy Stores	Confeitarias	123
	149	Cheese	Queijo	123

152	149	Chocolate	Chocolate	123
153	149	Farmers' Markets	Mercados ao Ar Livre	123
154	149	Health and Diet Food	Comida Saudável e Dietética	123
155	149	Kosher	Kosher	123
156	123	Furniture and Decor	Mobiliário e Decoração	123
157	123	Gift and Novelty	Presentes e Lembranças	123
158	123	Glasses	Óptica	123
159	123	Hobby and Collectibles	Coleccionismo e Passatempos	123
160	123	Luggage	Bagagem e Malas	123
161	123	Music, Video and DVD	Música, Vídeo e DVD	123
162				
163	123	Newsstands	Quiosques	123
164	123	Nurseries and Garden Centers	Viveiros e Centros de Jardinagem	123
165	123	Outlet	Outlet	123
166	123	Pawn Shops	Lojas de Penhores	123
167	123	Pets	Animais Domésticos	123
168	123	Photos and Frames	Fotografias e Molduras	123
169	123	Shopping Centers and Malls	Centros Comerciais	123
170	123	Sporting Goods	Artigos Desportivos	123
171	123	Supermarkets and Groceries	Supermercados e Mercearias	123
172	123	Tobacco	Tabaco	123
173	123	Toys	Brinquedos	123
174	123	Vintage and Thrift	Clássicos e em Segunda Mão	123
175	123	Warehouses and Wholesale Stores	Armazéns e Feiras Grossistas	123
176	123	Wedding and Bridal	Casamentos e Acessórios para a Noiva	123
177	1	Businesses and Services	Empresas e Serviços	177
178	177	Business and Strategy Consulting	Consultoria em Negócios e Estratégia	177
179	177	Industrial Machinery and Vehicles	Máquinas e Veículos Industriais	177
180	177	Logging and Sawmills	Madeireiros e Serrações	177
181	177	Metals	Metalurgia	177
182	177	Packaging	Embalamento e Empacotamento	177
183	177	Petroleum	Petróleo	177
184	177	Plastics	Plásticos	177
185	177	Refrigeration and Ice	Gelo e Refrigeração	177
186	177	Rubber	Borracha	177
187	177	Scientific	Científico	177
188	177	Security and Safety	Segurança	177
189	177	Telecommunication Services	Serviços de Telecomunicações	177
	177	Textiles	Têxteis	177

190	177	Water and Waste Management	Gestão de Água e de Resíduos	177
191				
192	177	Welding	Soldagem	177
193	177	Advertising and Marketing	Publicidade e Marketing	177
194	193	Advertising Agencies and Media Buyers	Agências de Publicidade e Compradores de Meios de Comunicação	177
195	193	Creative Services	Serviços Criativos	177
196	193	Direct Mail and Email Marketing Services	Correio e Serviços de Marketing Postal	177
197	193	Market Research and Consulting	Estudo e Consultoria de Mercado	177
198	193	Online Advertising	Publicidade Online	177
199	193	Print, TV, Radio and Outdoor Advertising	Publicidade Impressa, para Televisão, Rádio e Outdoors	177
200	193	Promotional Items	Artigos Promocionais	177
201	193	Public Relations	Relações Públicas	177
202	193	Search Engine Marketing and Optimization	Optimização e Marketing para Motores de Pesquisa	177
203	193	Writing, Copywriting and Technical Writing	Escrita, Redacção e Redacção Técnica	177
204	177	Agriculture and Forestry	Agricultura e Silvicultura	177
205	177	Art Restoration	Restauro de Arte	177
206	177	Audiovisual	Audiovisual	177
207	177	Automation and Control Systems	Automação e Controlo de Sistemas	177
208	177	Chemicals and Gasses	Químicos e Gases	177
209	177	Computers	Informática	177
210	177	Corporate HQ	Sede de Empresa	177
211	177	Electrical Equipment	Equipamento Eléctrico	177
212	177	Employment Agencies	Centros de Emprego	177
213	177	Engineering	Engenharia	177
214	177	Entertainment	Entretenimento	177
215	214	Media	Meios de Comunicação	177
216	177	Equipment Rental	Aluguer de Equipamento	177
217	177	Events and Event Planning	Eventos e Planeamento de Eventos	177
218	221	ATMs	Caixas Automáticas Multibanco	177
219	177	Financial	Finanças	177
220	219	Accounting and Bookkeeping	Contabilidade	177
221	219	Banking and Finance	Bancos e Finanças	177
222	219	Business Brokers and Franchises	Corretores de Negócios e Franquia	177
223	219	Check Cashing	Depósito de Cheques	177
224	219	Collections	Colecções	177
225	219	Financial Planning and Investments	Planeamento Financeiro e Investimentos	177

219	Fund Raising	Angariação de Fundos	177
219	Loans and Mortgages	Empréstimos e Hipotecas	177
219	Stock Brokers	Corretores de Bolsa	177
219	Student Aid and Grants	Auxílio e Bolsas para Estudantes	177
177	Food and Beverage	Comida e Bebida	177
230	Catering	Catering	177
230	Distribution	Distribuição	177
177	Funeral Services	Agências Funerárias	177
177	Geological	Geologia	177
177	Home Improvement	Renovação Habitacional	177
235	Architects	Arquitectos	177
235	Carpenters	Carpinteiros	177
235	Carpet and Flooring	Tapetes e Pavimentos	177
235	Contractors	Empreiteiros	177
239	Bathrooms	Quartos de Banho	177
239	Deck and Patio	Deck e Pátio	177
239	Sewer	Esgoto	177
235	Doors and Windows	Portas e Janelas	177
235	Electricians	Electricistas	177
235	Fences, Fireplaces and Garage Doors	Vedações, Lareiras e Portões de Garagem	177
235	Hardware and Services	Ferramentas e Serviços	177
235	Heating, Ventilating and Air Conditioning	Aquecimento, Ventilação e Ar Condicionado	177
123	Home Appliances	Electrodomésticos	123
235	Home Inspection Services	Serviços de Inspecção de Edifícios	177
123	Housewares	Equipamento para Cozinha	123
235	Interior Design	Design de Interiores	177

	235	Kitchens	Cozinhas	177
	235	Landscaping and Gardeners	Jardinagem	177
254	235	Lighting Fixtures	Iluminação	177
255	291	Mobile Homes	Casas Móveis	177
256				
257	235	Movers	Mudanças	177
258	235	Painting	Pintura	177
259				
260	235	Pest Control	Controlo de Pragas	177
261	235	Plumbing	Canalização	177
262				
263	235	Pools and Spas	Piscinas e Spas	177
	235	Roofers	Conserto de Telhados	177
264				
265	177	Storage	Armazenamento	177
266				
267	235	Swimming Pool Maintenance and Services	Serviços e Manutenção de Piscinas	177
268	235	Tree Service	Serviço de Arborização	177
269				
270	235	Upholstery	Estofamento	177
	177	Human Resources	Recursos Humanos	177
271				
272	177	Import and Export	Importação e Exportação	177
273				
274	177	Leather	Couro	177
	177	Legal	Jurídico	177
	269	Credit Counseling and Bankruptcy Services	Aconselhamento de Crédito e Serviços de Falência	177
	269	Immigration	Imigração	177
	177	Insurance	Seguros	177
	177	Machine Shops	Oficinas Mecânicas	177
	177	Management	Administração	177
275	177	Manufacturing	Manufactura	177
276	177	Paper	Papel	177
277				
278	177	Personal Care	Cuidados Pessoais	177
	277	Dry Cleaning, and Ironing Laundry	Limpeza a seco, Engomadoria e Lavandaria	177
279				

280	277	Hair Removal	Depilação	177
281	277	Beauty Salons and Barbers	Salões de Beleza e Barbeiros	177
282	277	Manicures and Pedicures	Manicure e Pedicure	177
283	277	Massage Clinics and Therapists	Clínicas de Massagem e Terapeutas	177
284	277	Piercing	Piercing	177
285	277	Skin Care	Cuidados com a Pele	177
286	277	Spas	Spas e Termas	177
287	277	Tanning Salons	Solários	177
	277	Tattooing	Tatuagens	177
288	177	Printing, Copying and Signage	Impressão, Cópia e Sinalização	177
289	177	Professional Cleaning	Limpeza e Higiene Profissional	177
290	177	Publishing	Editoras	177
291	177	Real Estate	Ramo Imobiliário	177
292	291	Property Management	Gestão de Propriedades	177
293	291	Real Estate Agents	Agentes Imobiliários	177
294	291	Real Estate Appraiser	Avaliador Imobiliário	177
295	291	Real Estate Development and Title Companies	Empresas Imobiliárias e de Títulos	177
296	291	Apartments, Condos, and Houses	Apartamentos, Condomínios e Casas	177
297	291	Boarding Houses	Pensões	177
298	291	Building and Land Surveyors	Construção e Topógrafos	177
299	291	Commercial Real Estate	Comércio Imobiliário	177
300	291	Corporate Housing	Condomínio Empresarial	177
301	177	Renewable Energy	Energia Renovável	177
302	177	Repair Services	Serviços de Reparação	177
303	177	Shipping, Freight, and Material Transportation	Expedição e Transporte de Materiais	177
304	177	Tailors	Alfaiates	177
305	177	Veterinarians	Veterinários	177
306	460	Web Design and Development	Design e Desenvolvimento para a Web	177
307	177	Wholesale	Comércio por Grosso	177
308	1	Social	Social	308
309	308	Arts	Artes	308
310	309	Art Dealers and Galleries	Negociantes e Galerias de Arte	308
311	309	Museums	Museus	308
312	308	Bars	Bares	308
313	312	Hotel Lounges	Salões de Hotel	308
314	312	Jazz and Blues Cafes	Cafés de Jazz e Blues	308
	312	Sports Bars	Bares Desportivos	308

315	312	Wine Bars	Bares de Vinho	308
316	308	Entertainment	Entretenimento	308
317	317	Adult Entertainment	Entretenimento para Adultos	308
318	317	Amusement Parks	Parques de Diversões	308
319	317	Billiard and Pool	Salas de Bilhar	308
320	317	Bingo	Bingo	308
321				
322	317	Bowling	Bowling	308
323	317	Carnivals	Carnavais	308
324	317	Casinos and Gaming	Casinos e Jogo	308
325	317	Circuses	Circos	308
326	317	Dance Halls and Saloons	Salões de Dança	308
327	317	Fairgrounds and Rodeos	Feiras e Rodeios	308
328	317	Go Carts	Karting	308
329	317	Hookah Lounges	Salões de Narguilé	308
330	317	Karaoke	Karaoke	308
331	317	Miniature Golf	Mini-golfe	308
332	317	Movie Theatres	Cinemas	308
333	317	Music and Show Venues	Salas de Música e Espectáculos	308
334	317	Night Clubs	Clubes Nocturnos e Discotecas	308
335	317	Party Centers	Centros de Festa	308
336	317	Psychics and Astrologers	Videntes e Astrólogos	308
337	317	Ticket Sales	Bilheteiras	308
338	308	Food and Dining	Restauração	308
339	338	Bagels and Donuts	Bagels e Donuts	308
340	338	Bakeries	Padarias e Pastelarias	308
341	338	Breweries	Cervejarias	308
342	338	Cafes, Coffee and Tea Houses	Cafés e Salas de Chá	308
343	338	Dessert	Sobremesas	308
344	338	Ice Cream Parlors	Gelatarias	308
345	338	Internet Cafes	Cibercafés	308
346	338	Juice Bars and Smoothies	Bares de Sumos e Batidos	308
347	338	Restaurants	Restaurantes	308
348	347	American	Americano	308
349	347	Barbecue	Barbecue	308
350	347	Buffets	Buffet	308
351	347	Burgers	Hambúrgueres	308
352	347	Chinese	Chinês	308
353	347	Delis	Casas de Produtos Gourmet	308
354	347	Diners	Cantinas	308
355	347	Fast Food	Fast Food	308
	347	French	Francês	308

356	347	Indian	Indiano	308
357	347	Italian	Italiano	308
358	347	Japanese	Japonês	308
359	347	Korean	Coreano	308
360	347	Mexican	Mexicano	308
361	347	Middle Eastern	Do Médio Oriente	308
362	347	Pizza	Pizzarias	308
363	347	Seafood	Marisqueiras	308
364				
365	347	Steakhouses	Churrascarias	308
366	347	Sushi	De Sushi	308
367	347	Thai	Tailandês	308
368	347	Vegan and Vegetarian	Vegan e Vegetarianos	308
369	308	Country Clubs	Clube Privado	308
370	308	Wineries and Vineyards	Adegas e Vinhas	308
371	308	Zoos, Aquariums and Wildlife Sanctuaries	Jardins Zoológicos, Aquários e Parques Biológicos	308
372	1	Sports and Recreation	Desportos e Lazer	372
373	372	Athletic Fields	Campos de Atletismo	372
374	372	Baseball	Basebol	372
375	374	Batting Ranges	Centros de Treino de Basebol	372
376	372	Basketball	Basquetebol	372
377	372	Combat Sports	Desportos de Combate	372
378	372	Cycling	Ciclismo	372
379	372	Dance	Dança	372
380	372	Equestrian	Equitação	372
381	372	Football	Futebol Americano	372
382	372	Golf	Golfe	372
383	372	Gun Ranges	Campos de Tiro	372
384	372	Gymnastics	Ginástica	372
385	372	Gyms and Fitness Centers	Ginásios e Centros de Fitness	372
386	372	Hockey	Hóquei	372
387	372	Outdoors	Ar Livre	372
388	387	Campgrounds and RV Parks	Parques de Campismo e de Caravanas	372
389	387	Hiking	Caminhadas	372
390	387	Hot Air Balloons	Balões de Ar Quente	372
391	387	Hunting and Fishing	Caça e Pesca	372
392	387	Rock Climbing	Escalada	372
393	387	Skydiving	Pára-quedismo	372
	372	Paintball	Paintball	372

394	372	Personal Trainers	Treinadores Pessoais	372
395	372	Race Tracks	Pistas de Corrida	372
396	372	Racquet Sports	Desportos de Raquete	372
397	397	Racquetball	Raquetebol	372
398	397	Tennis	Ténis	372
399	372	Recreation Centers	Centros Recreativos	372
400	372	Running	Corrida	372
401	372	Skating	Patinagem	372
402	372	Snow Sports	Desportos de Inverno	372
403	372	Soccer	Futebol	372
404	372	Sports Clubs	Clubes Desportivos	372
405	372	Stadiums and Arenas	Estádios e Recintos Desportivos	372
406	372	Swimming Pools	Piscinas	372
407	372	Water Sports	Desportos Aquáticos	372
408	408	Boating	Passeios de Barco	372
409	408	Canoes and Kayaks	Canoas e Kayaks	372
410	408	Rafting	Rafting	372
411	408	Scuba Diving	Mergulho	372
412				
413	408	Swimming	Natação	372
414	372	Yoga and Pilates	Yoga e Pilates	372
415	1	Transportation	Transporte	415
416	415	Airlines and Aviation Services	Companhias Aéreas e Serviços de Aviação	415
417	415	Gas Stations	Postos de Combustível	415
418	415	Parking	Estacionamento	415
419	415	Public Transportation Services	Serviços de Transporte Público	415
420	415	Taxi and Car Services	Serviços de Automóveis e Táxi	415
421	420	Car and Truck Rentals	Aluguer de Carros e Camiões	415
422	420	Charter Buses	Autocarros de Serviço Ocasional	415
423	420	Limos and Chauffeurs	Limusinas e Motoristas	415
424	415	Transport Hubs	Estações de Transporte	415
425	424	Airports	Aeroportos	415
426	424	Bus Stations	Paragens de Autocarro	415
427	424	Heliports	Heliportos	415
428	424	Ports	Portos	415
429	424	Rail Stations	Estações Ferroviárias	415
430	1	Travel	Viagem	430
431	430	Cruises	Cruzeiros	430
432	430	Lodging	Alojamento	430
433	432	Bed and Breakfasts	Bed and Breakfast	430
434	432	Cottages and Cabins	Chalés e Cabanas	430
435	432	Hostels	Albergues	430
	432	Hotels and Motels	Hotéis e Motéis	430

436	432	Lodges and Vacation Rentals	Albergues e Alugueres para Férias	430
437	432	Resorts	Resorts	430
438	430	Tourist Information and Services	Serviços de Informação Turística	430
439	430	Travel Agents and Tour Operators	Agências de Viagens e Guias Turísticos	430
440	83	Geriatrics	Geriatria	62
441	123	Discount Stores	Lojas de Desconto	123
442				
443	149	Meat and Seafood	Carnes e Frutos do Mar	123
444	123	Office Supplies	Materiais de Escritório	123
445				
446	123	Party Supplies	Fontes do Partido	123
447	177	Career Counseling	Aconselhamento de Carreira	177
448				
449	177	Construction	Construção	177
450	269	Notary	Notário	177
451				
452	177	Photography	Fotografia	177
453	177	Translation Services	Serviços de Tradução	177
454				
455	382	Golf Courses	Campos de Golfe	372
	408	Surfing	Surfe	372
	37	Embassies	Embaixadas	20
	460	Infrastructure	Infra-estrutura	177
	460	Mobile	Móvel	177
456	460	Advertising	Publicidade	177
457	347	Asian	Asiático	308
458	347	Food Trucks	Food Trucks	308
459				
460	415	Rest Areas	Áreas de Descanso	415
461	177	Technology	Tecnologia	177
462				
463	118	Dog Parks	Parques Cão	107
464	425	International Airports	Aeroporto Internacional	415
465				
466	317	Arcades	Vídeojogos	308

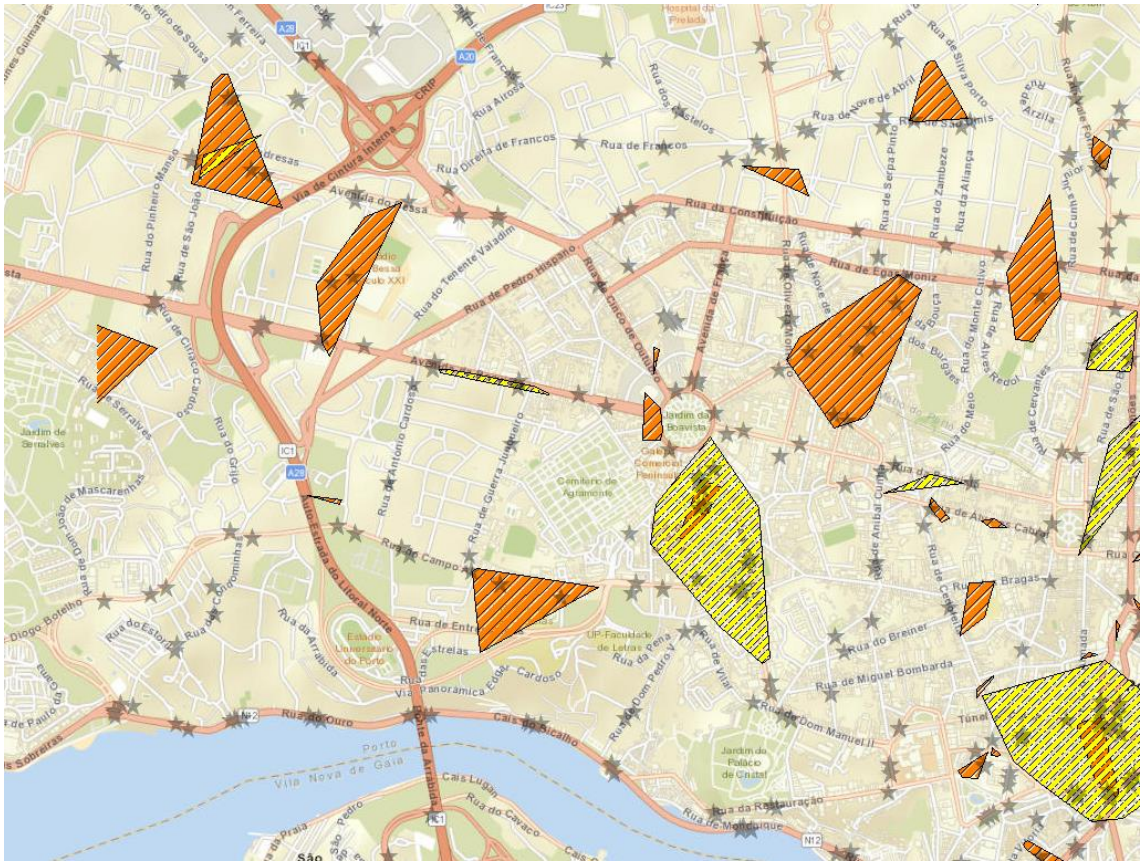
467	347	International	Internacional	308
	217	Convention Centers	Centros de Convenções	177
	62	Optometrist	Optometria	62
	1	NoExport	NoExport	467

Anexo G: ZONAS NÃO COBERTAS

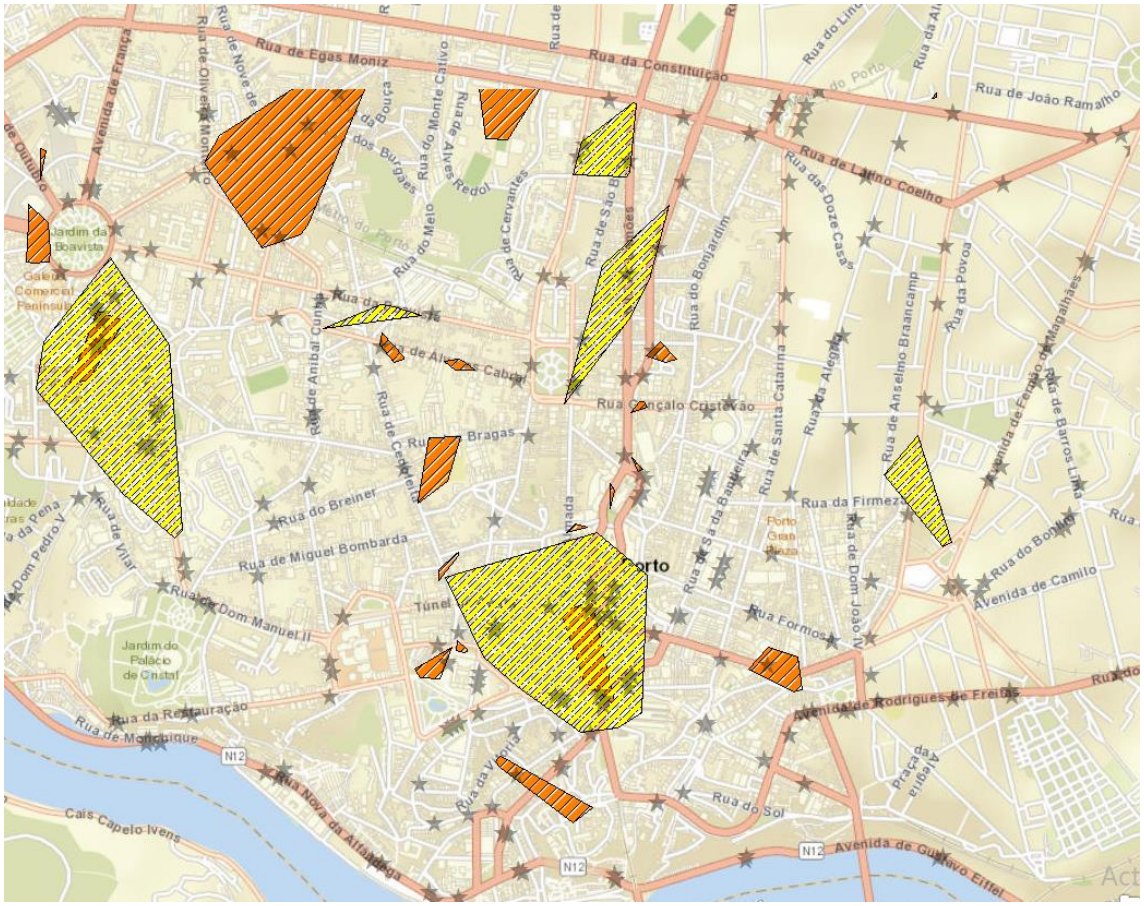
Nesta secção apresenta-se as várias imagens obtidas, sobre os locais com pouca oferta de transportes.

Todas as imagens utilizam a seguinte legenda: as paragens STCP estão assinaladas com uma estrela, a camada dos destinos esta assinalada como polígono a vermelho e as restantes camadas em comparação estão assinaladas como polígonos amarelos.

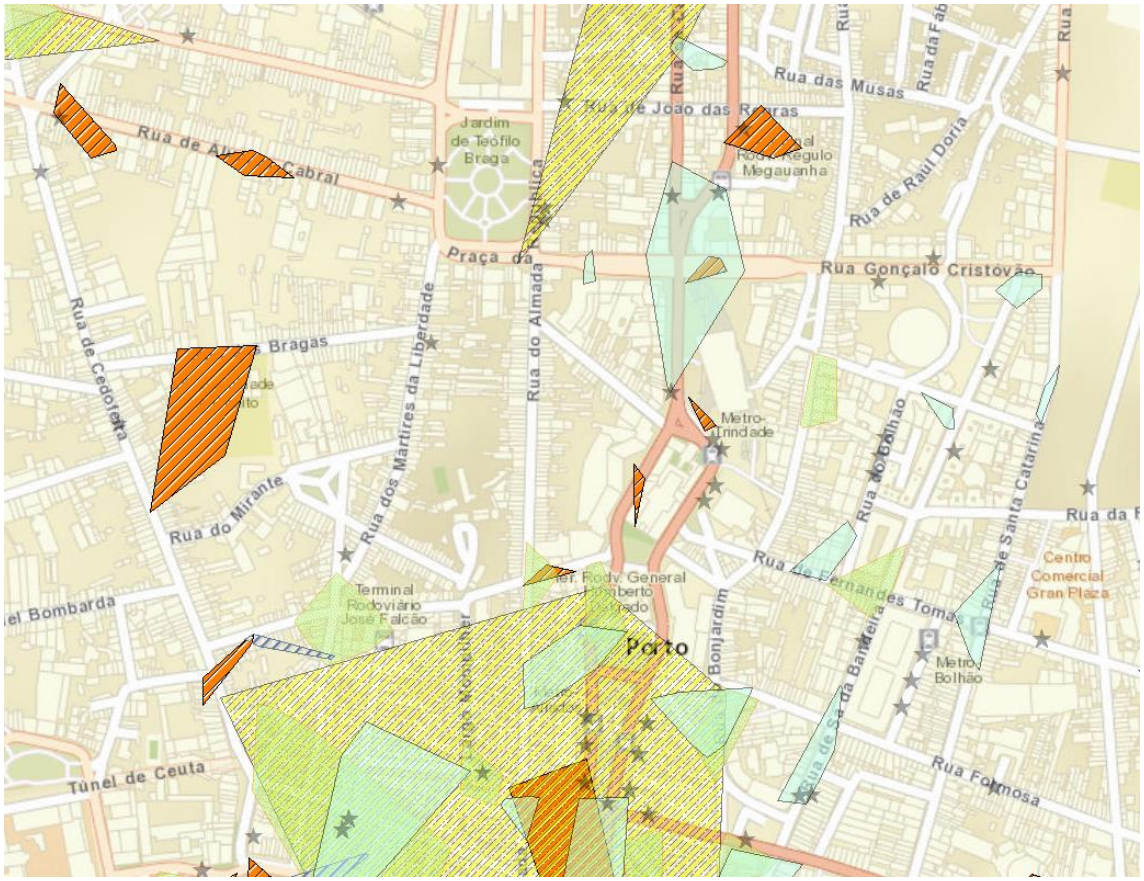
A seguinte imagem corresponde ao filtro de rotina-semana: entre destinos e viagens:



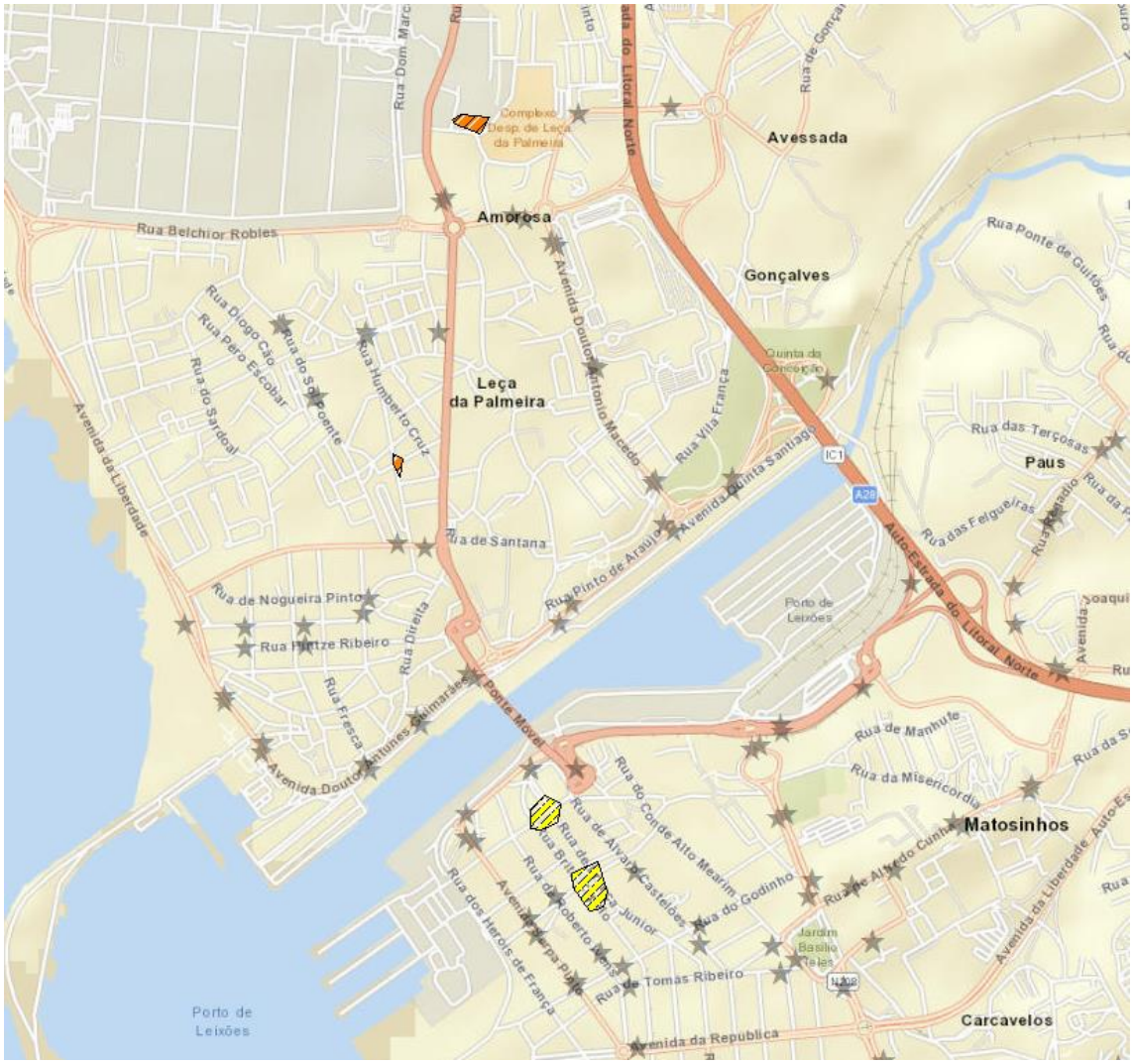
A seguinte imagem corresponde ao filtro de rotina-semana: entre destinos e viagens:

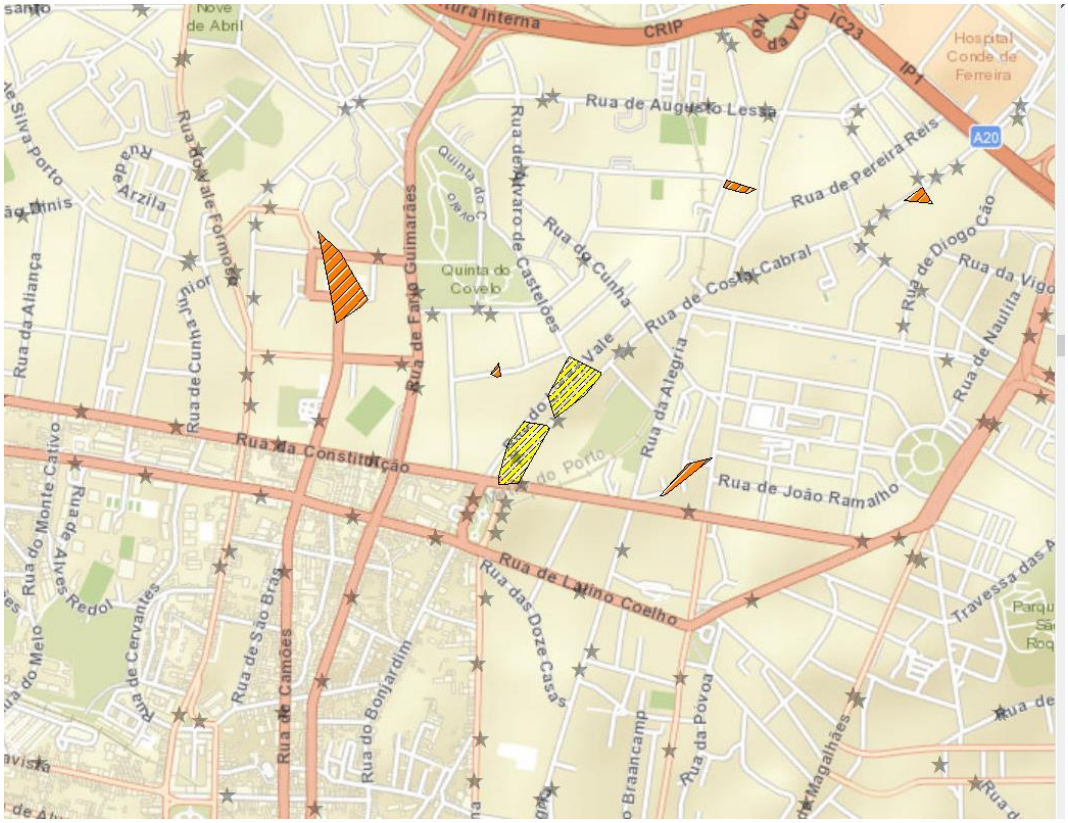


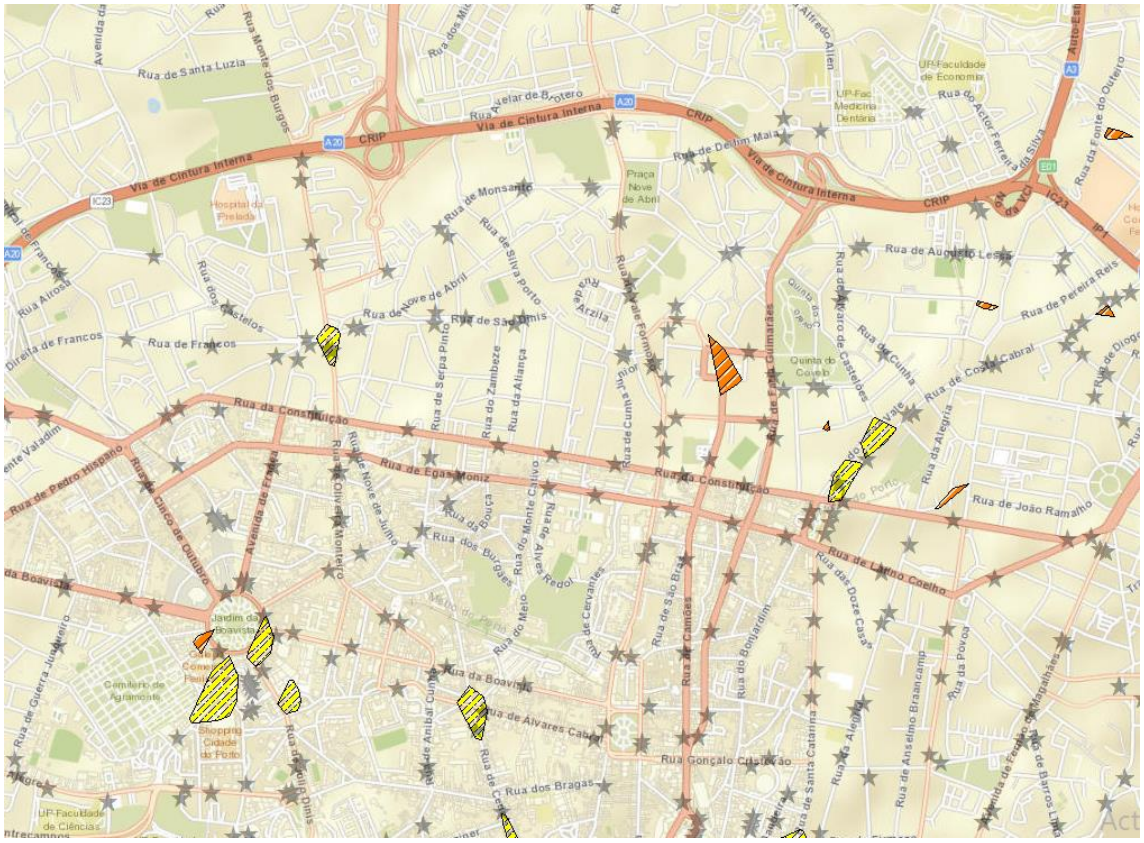
A seguinte imagem corresponde ao filtro de rotina-semana: entre destinos e eventos, em que a vermelho são os polígonos de destinos, as estrelas são as paragens STCP, os polígonos azuis são POIs social, a verde são POIs de negócios e serviços e a amarelo são POIs de categoria viagens:



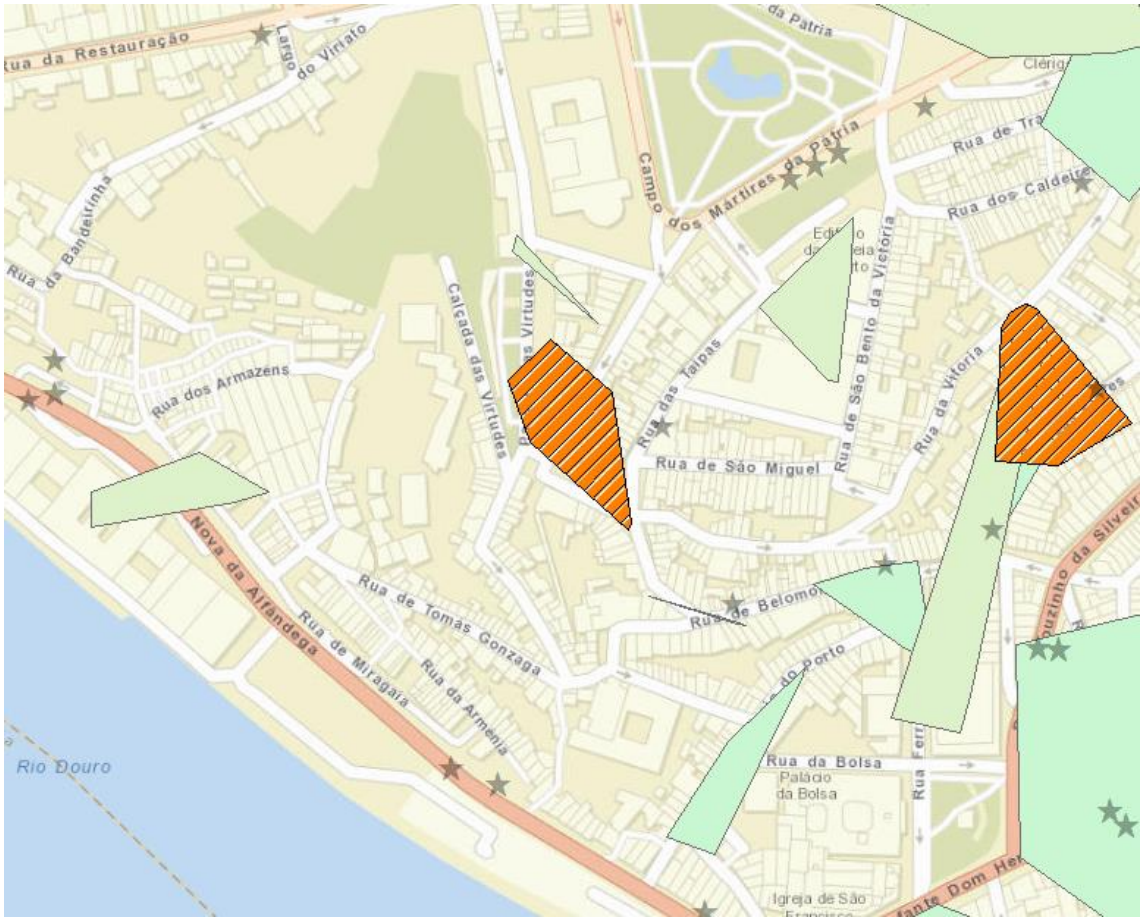
Durante o filtro de rotina fim de semana estas são as regiões menos cobertas pelos STCP, para as camadas destinos e retalho. Mais uma vez em ambos os três casos, os destinos são os polígonos a vermelho e os POIs de retalho são os amarelos.


Avessada
Amorosa
Gonçalves
Leça
da Palmeira
Paus
Matosinhos
Carcavelos
Porto de
Leixões
Rua Belchior Robles
Rua de Santana
Rua Pinto de Araújo
Rua de Nogueira Pinto
Rua Heitze Ribeiro
Rua Direita
Rua Fresca
Avenida Quinta Santiago
Rua Vila França
Rua das Terçosas
Rua das Felgueiras
Rua de Manhute
Rua da Misericordia
Rua de Tomas Ribeiro
Avenida da Republica
Avenida da Liberdade
Auto-Estrada do Litoral Norte
Ponte Movel
Rua do Conde Alto Mearim
Rua do Godinho
Rua dos Heróis de França
Jardim
Basilio
Teles


Rua de Augusto Lessa
Rua de Pereira Reis
Rua de Costa Cabral
Rua da Constituição
Rua de João Ramalho
Rua de Latino Coelho
Rua de Diogo Cão
Rua da Vigo
Rua de Naulila
Rua da Alegria
Rua da Cunha
Rua de Cunha Júnior
Rua de Faria Guimarães
Rua do Bonjardim
Rua de São Brás
Rua de Camões
Rua das Doze Casas
Rua da Póvoa
Rua da Aliança
Rua do Monte Cativo
Rua de Alves Redol
Rua de Cervantes
Rua de Arzila
Rua Nova de Formoselha
Rua de Álvaro de Castelões
Quinta do Covelo
Travessa das A
Hospital Conde de Ferreira
Nove de Abril
CRIP
IC23
IP1
A20



Durante o filtro de rotina fim de semana estas são as regiões menos cobertas pelos STCP, para as camadas destinos e social. Os destinos são os polígonos a vermelho e os POIs de social são os azuis.


Rua da Restauração
Largo do Viriato
Rua dos Bandeirinha
Rua dos Armazéns
Calçada das Virtudes
Campo dos Mártires da Pátria
Rua das Taipas
Rua de São Bento da Vitória
Rua da Vitória
Rua de São Miguel
Nova da Alfândega
Rua de Tomás Gonzaga
Rua de Miragaia
Rua da Arménia
Rio Douro
Rua da Bolsa
Palácio da Bolsa
Igreja de São