FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
DA UNIVERSIDADE DE COIMBRA

Classificação de Perfis de Utilizador nos Transportes Públicos

Dissertação de Mestrado em Engenharia Informática

Luís da Silva Perdigão
perdigao@student.dei.uc.pt

Orientadores:
Ana Maria de Almeida
Fernando Amílcar Cardoso

Setembro de 2013

Resumo

Os centros urbanos tendem cada vez mais a ser locais de concentração de poluição no nosso planeta, sendo um dos principais fatores a poluição automóvel e o desperdício energético que estes veículos tendem a causar devido à diminuta ocupação dos mesmos. Uma das soluções para este problema passa por atrair mais utentes aos transportes públicos. O objetivo desta dissertação é o de conseguir um sistema que pode ser aplicado na melhoria dos sistemas de transporte público, mais especificamente os autocarros. A ideia concreta é a de, através da análise de dados de bilhética e fontes externas de contextualização de locais geográficos, conseguir desenhar um sistema que permita a identificação dos utilizadores e uma caracterização e encaixe num perfil de tipicidade (estereótipo).

O presente trabalho insere-se no âmbito do projeto para fornecimento de serviços de mobilidade centrados no utilizador: Tice.Mobilidade, em particular, o tema da dissertação insere-se no âmbito de um estágio de investigação no PPS 5 BUSCA – Bus Context Awareness, que pretende aumentar o nível de conforto e a perceção de ganho por parte do utilizador pelo facto de usar transportes públicos.

A informação que nos é apresentada no dia-a-dia, através de painéis informativos ou publicitários nem sempre é a mais apelativa e chama a nossa atenção. Para tornar esta informação um fator de redução do tempo psicológico de viagem nos transportes públicos é necessário, em primeiro lugar, perceber quem viaja no veículo. Esta dissertação irá desenhar e implementar um modelo de classificação de perfis típicos de utilizador de transportes públicos baseado em algoritmos de *Data Mining*, *clustering* de dados e classificadores baseados em regras e contextualização geográfica através de pontos de interesse. Este é um passo fundamental para permitir que outras aplicações dedicadas possuam a informação suficiente para gerar conteúdos informativos adaptados a quem os irá receber. A principal contribuição deste trabalho consiste na construção de um gerador de estereótipos de utentes de transportes públicos com sistemas de bilhética, a partir da recolha de dados e classificação de perfis.

Palavras-chave

Análise de Dados
Perfis de Utilizador
Classificação de Características
Pontos de Interesse
Transportes Públicos

Agradecimentos

Durante o último ano, várias pessoas me apoiaram na realização desta dissertação de mestrado, e às quais não posso deixar de agradecer:

Em primeiro lugar, agradeço aos meus orientadores, em especial à minha orientadora Ana Almeida, por todo o apoio que me proporcionou durante a dissertação. Agradeço pela sua disponibilidade e partilha de conhecimentos.

Ao Professor Carlos Bento e ao grupo de estagiários, por todos os conselhos e orientação que me facultaram durante as reuniões quinzenais que ocorreram ao longo da dissertação.

Aos meus colegas de curso, que estiveram sempre disponíveis para qualquer dúvida e debate de ideias como também toda a ajuda na preparação das defesas da dissertação.

A terminar, mas não menos importante agradeço aos meus pais, ao meu irmão e toda a minha família pelo apoio incondicional. O incentivo para melhorar a cada dia que passa, e a nunca desistir dos meus sonhos e objetivos, foram muito importantes ao longo do desenvolvimento deste trabalho.

Índice

Resumo i
Palavras-chave i
Agradecimentos ii
Índice iii
Lista de Figuras v
Lista de Tabelas vi
Lista de Abreviaturas vii
1. Introdução 1
1.1. Contexto e Motivação 2
1.2. Objetivos 2
1.3. Organização do Documento 3
2. Estado da Arte 5
2.1. Perfis de Utilizador 5
2.2. Abordagens ao Problema 9
2.3. Métodos de Classificação 9
3. Dados de Bilhética 13
3.1. Introdução 13
3.2. Base de Dados de Validações de Utentes 14
3.3. Exploração e Tratamento dos Dados 15
3.4. Escolha de Métodos de Classificação 16
4. Dados Geográficos 19
4.1. Sistemas de Informação Geográfica 19
4.2. Organização de Categorias de POIs 19
4.3. Método de Recolha de Informação 22
4.4. Método de Classificação 22
5. Solução Integrada 25
5.1. Perfis de Utilizador Integrados 25
5.2. Algoritmo Geral de Classificação 26
6. Resultados 29
6.1. Estatísticas Gerais 29
6.2. Estatísticas de Utilização 35
6.3. Resultados Gerais de Classificação de Perfis de Utilizador 40
6.4. Validação do Algoritmo 41
6.5. Análise aos resultados 42
7. Conclusões e Próximos Passos 47
8. Referências 49
9. Anexos 53
9.1. Outras Tabelas 53

Lista de Figuras

Figura 1 - Exemplo de Recomendações 5
Figura 2 - Sistema de informação *Smart Card* 6
Figura 3 - Distribuição de acordo com o tipo de cartão 6
Figura 4 - Geopulse-Context 8
Figura 5 - Esquema de recolha de dados de bilhética 13
Figura 6 - Comparação de Clusters 17
Figura 7 - Resultados Clusters 18
Figura 8 - Caixa de recolha OSM e círculo de interesse 22
Figura 9 - Arquitetura do Algoritmo de classificação 27
Figura 10 - Paragens com maior afluência 29
Figura 11 - O-D mais comuns 30
Figura 12 - Censos 2011 Movimentos Pendulares [53] 31
Figura 13 –Viagens em função da hora do dia (C1) ao domingo (esq.) e nos dias úteis (dir.). 31
Figura 14 - Número de Viagens em fução da hora (C2) domingo (esq.) e dias uteis (dir.) 32
Figura 15 - Número de Viagens em função da hora (C1) ao domingo (esq); e ao sábado (dir.)..... 32
Figura 17 - Número de Viagens aos Feriados em função da hora do dia 33
Figura 16 - Número de Viagens em função da hora (C2) ao domingo (esq) e ao sábado (dir.)..... 33
Figura 18 - Utilização de 4 carreiras no Conjunto C1: dias 1 a 8 de dezembro 34
Figura 19 - Utilização por dia de Semana (C2) 34
Figura 20 - Subperfil por Utilização (C1) 35
Figura 21 - Subperfil por Utilização (C2) 35
Figura 22 - Espaços Temporais 36
Figura 23 - Perfil Horário 36
Figura 25 - Redução de Bilhete 37
Figura 24 - Exemplo de Atribuição de Mascara 37
Figura 26 - Subperfil de Redução de Custo 37
Figura 27 - SubPerfil por tipo de utilização 38
Figura 28 - SubPerfil por viagens disponíveis 38
Figura 29 - Calculo de distância percorrida 39
Figura 30 – Subperfil por Distância Percorrida (C2) 39
Figura 31 - Subperfil por Distância Percorrida (C1) 39
Figura 32 - Perfis de Utilizador (POI) 40
Figura 33 - Perfis de Utilizador Finais 40
Figura 34- Distribuição dos perfis de utilizador pelas paragens de autocarro 42
Figura 35 - Perfis de utilizador ao longo das paragens de autocarro 43
Figura 36 - Zona característica da linha 50 44
Figura 37 - Distribuição dos perfis de utilizador na linha 50 44
Figura 38 - Distribuição dos perfis de utilizador na linha 53 44
Figura 39 - Zona característica da linha 53 44
Figura 40 - Distribuição dos perfis de utilizador na linha 51 45
Figura 41 - Distribuição dos perfis de utilizador por hora da linha 51 45

Lista de Tabelas

Tabela 1 - Exemplo de inconsistências dos dados..... 16
Tabela 2 – Divisão de Categorias de Pontos de interesse por Perfil de Utilizador e seus POIs e utilizadores associados. 21
Tabela 3 - Características dos perfis de utilizador e classes de POI associadas..... 26

Tabela A 2 - Descrição dos Dados (continuação) 54
Tabela A 1 - Descrição dos Dados 53

Lista de Abreviaturas

API	Application Programming Interface
CARRIS	Companhia Carris de Ferro de Lisboa
CNPD	Comissão Nacional de Proteção de Dados
EM	Expectation Maximization
HCA	Hierarchical Clustering Algorithm
NUTS III	Unidades territoriais para fins estatísticos
O-D	Origem – Destino
POI	Ponto de Interesse / Point of Interest
STCP	Sociedade de Transportes Colectivos do Porto
SOM	Self Organization Map
TfL	Transport for London
TP	Transporte Público
GIS	Geographic Information System
O-D	Origem Destino
AAFID	Autonomous Agents For Intrusion Detection
EMERALD	Event monitoring enabling responses to anomalous live disturbances

1. Introdução

Na atualidade, um dos maiores problemas das (grandes) cidades é o excesso de tráfego nos centros metropolitanos, o que leva a congestionamentos, dificuldade de estacionamento e a um impacto ambiental que se manifesta através de um consumo energético elevado levando à poluição do centro urbano. [1]. Outra situação preocupante são os acidentes de viação que matam todos os anos cerca de 1.3 milhões de pessoas e causam cerca de 50 milhões de vítimas nas estradas [2] sendo estes acidentes proporcionais ao número de veículos que circulam nas estradas. Todos estes fatores contribuem para a diminuição da qualidade de vida das pessoas, principalmente nos tempos atuais em que "cerca de 60% da população vive em zonas urbanas"[3]. No caso da União Europeia, são perdidos "anualmente perto de 100 mil milhões de euros, ou seja, 1% do PIB da UE, devido aos problemas de mobilidade"[3]. Em Portugal, segundo o Inquérito à Mobilidade da População Residente (2000), cerca de 50% das deslocações são realizadas em automóveis privados, em que 70% dessas deslocações ocorrem apenas com um ocupante, fazendo com que a taxa de ocupação dos veículos seja fraca [4]. Uma das possíveis medidas para resolver estes problemas passa por cativar mais pessoas para os transportes públicos.

Esta dissertação insere-se no âmbito das áreas de investigação da Faculdade de Ciências e Tecnologia da Universidade de Coimbra, nomeadamente na construção da plataforma de mobilidade One.Stop.Transport, projeto QREN TICE-Mobilidade[1], que tem como objetivo a disponibilização de serviços de mobilidade centrados no utilizador. Este projeto agrega 10 diferentes subprojetos ou PPS (processos, produtos ou serviços) que, em geral, têm como função oferecer um conjunto de serviços direcionados ao utilizador final. Em particular este estudo foi efetuado no âmbito de estágio do PPS5 BUSCA - BUS Context-Awareness. O BUSCA tem como objetivo principal o desenvolvimento de um conjunto de soluções que facilitem a interação entre o utilizador de transportes públicos, o meio urbano e a dinâmica entre estes dois. Com o desenvolvimento de serviços no âmbito deste PPS espera-se que a experiência de viagem dos utilizadores melhore e que se torne mais apelativo o uso de transportes públicos, aumentando assim o número de utilizadores deste tipo de transportes [5] [6]. Assim, o trabalho desenvolvido tem como fim contribuir diretamente para os objetivos do PPS 5, através de um gerador de estereótipos de perfis de utilizador de transportes públicos. Mais especificamente este gerador de estereótipos é composto por: Análise de dados de validação e extração de métricas de utilização dos utentes deste serviço a partir de técnicas de *clustering*, classificador através de regras. Para além da base de dados de validações dos títulos dos autocarros, utilizaremos ainda um sistema GIS para ajudar a classificação. Com base na área geográfica das paragens de autocarro serão utilizados os de pontos de interesse nas proximidades para classificar essas áreas e consecutivamente os utilizadores.

Com esta abordagem através de junção de três técnicas de derivação de perfis de utilizador (*clustering*, classificação por regras e extração de perfil geográfico das paragens de autocarro) espera-se trazer inovação científica para a área em questão. A classificação automática de utilizadores através de dados de bilhética e dados de contexto está ainda pouco explorada ou é quase inexistente. Pensamos que a exploração destas técnicas possa trazer benefício às

[1] Projeto agregador Tice.Mobilidade – Sistema de serviços centrados no utilizador constituído por vários subprojetos que pretendem fornecer serviços de mobilidade particulares através de uma plataforma denominada One.Stop.Transport.

transportadoras uma vez que pode ajudar a melhorar as suas redes de transporte e aumentar o número de utilizadores.

1.1. Contexto e Motivação

Entre as razões principais para as pessoas não utilizarem os transportes públicos (TP) encontra-se o desconforto que estes podem proporcionar ao utilizador [7] [8]. Para diminuir este desconforto pode-se melhorar o serviço, em particular os próprios veículos e as paragens (locais de entrada /saída dos veículos). No entanto, esta solução pode tornar-se muito dispendiosa e obrigar a custos mais elevados na cobrança de deslocações aos utentes, tornando-se noutra razão para desmotivar o uso dos TP. Uma alternativa para aumentar o conforto do utilizador passa por diminuir o tempo psicológico da viagem, [5], ou como claramente exposto em [9] : "*The passengers are normally bored when they ride a public transport system*". Esta é uma área de estudo atual, onde, se preconiza, por exemplo, a utilização de painéis digitais, quer nos veículos, quer nas respetivas paragens, possibilitando a transmissão de informação útil aos utilizadores. O objetivo prático consiste em abstrair o utilizador da passagem do tempo, melhorando a sua experiência pessoal de utilização de transportes públicos ao eliminar a memória de tempo de espera. No entanto, o sucesso desta aproximação reside numa efetiva atracão do utente à informação disponibilizada e, de forma fulcral, no conteúdo exposto estar ou não adaptado aos utilizadores. Por exemplo, se um autocarro estiver a transportar maioritariamente crianças, não faz sentido, ou é mesmo inadequado, apresentar publicidade sobre bebidas alcoólicas. Para adaptar esta informação é necessário saber que utentes estão ou vão estar num dado local, veículo ou paragem, numa determinada janela temporal de forma a perceber que conteúdos devem ser apresentados de acordo com o estereótipo maioritário desses utentes.

Como referimos anteriormente achamos que os perfis de utilizador podem ainda ajudar a melhorar a rede de transportes em si. Sabendo os utilizadores que utilizam o serviço e como o fazem torna-se muito importante para as transportadoras tomarem decisões tais como saberem onde centrar o investimento, onde criar novos circuitos, onde aumentar o número viagens ou que/quantos veículos colocar numa determinada zona.

Desta forma, nesta dissertação pretende-se conseguir especificar um gerador de perfis que possa servir para contextualizar informação a disponibilizar aos utentes em função da caracterização maioritária destes em cada momento com base no histórico dos sistemas de bilhética de carreiras e fusão de informação contextualizada.

1.2. Objetivos

Como objetivo principal desta dissertação propõem-se desenhar um sistema capaz de identificar os perfis de utilizador que estarão maioritariamente dentro de um autocarro (ou transporte público) num dado instante.

Para o desenvolvimento deste sistema é necessário concretizar os seguintes objetivos particulares:

1. **Analisar dados de bilhética e extrair padrões de utilização**
   Consiste na análise de uma base de dados de validações de utentes em autocarros e extração de padrões de utilização associados.
2. **Extração de perfis de utilizador usando algoritmos de *Data Mining***
   Aplicação de algoritmo de análise de dados para extração de padrões dos dados, mais propriamente algoritmos de classificação/agrupamento.
3. **Extração de informação Geoespacial (POIs)**
   Pesquisa e extração de informação geográfica com objetivo de ajudar a classificar os utilizadores de acordo com os espaços geográficos através dos pontos de interesse[2] (POIs) que caracterizam esses espaços.
4. **Pesquisa de Eventos por Área Geográfica**
   Pesquisa de Eventos que ocorrem perto das paragens de autocarro de modo a prever-se perfis de utilizador esporádicos numa dada altura.
5. **Identificar os perfis de utilizador a identificar e identificar a sua caracterização**
   Através da aplicação dos diferentes algoritmos aos dados de bilhética/ geográficos identificar quais os perfis de utilizador possíveis de caracterizar.

Para além dos objetivos acima, este trabalho tem ainda como objetivo estudar a viabilidade de um sistema destes e a sua aplicação num sistema real de transportes. No caso de sucesso será implementado para a empresa AMI (empresa de tecnologias para transportes), um software de *backoffice* com base no estudo desta dissertação.

1.3. Organização do Documento

O Capítulo 2 analisa o estado da arte, resultante da pesquisa efetuada, e é apresentada uma revisão sumariada da literatura descrevendo os principais trabalhos realizados anteriormente na área de perfis de utilizador. Apresenta-se ainda, a nossa análise crítica da importância relativa destes trabalhos para o nosso desenvolvimento.

No Capítulo 3 é inicialmente apresentado o estudo e tratamento pormenorizado feito aos dados disponibilizados e a escolha do *dataset*. Uma vez que os principais resultados estão diretamente relacionados com a qualidade dos dados este capítulo torna-se um dos mais importantes desta dissertação. É então apresentada a fonte de dados utilizada, são descritos, discutidos e comparados os métodos de classificação escolhidos. É também iniciada a pormenorização do algoritmo de classificação de perfis de utilizador relativa à análise e classificação a partir de dados de bilhética.

No Capítulo 4 é discutido o sistema de informação geográfica a utilizar, como devem ser organizadas as categorias de pontos de interesse de acordo com os perfis de utilizador alvo. De seguida é detalhado o método de recolha de pontos de interesse de acordo com as paragens de autocarro a classificar e por fim é descrito o método de classificação dos utentes de acordo com esta metodologia.

[2] Ponto de interesse (PDI, POI em inglês) é uma localização específica, normalmente georreferenciada, que de forma geral é útil para a sociedade. Podem ser considerados POIs por exemplo hotéis, restaurantes, bombas de gasolina.

O Capítulo 5 diz respeito à integração entre as duas abordagens discutidas no Capítulo 3 e Capítulo 4, detalhando quais foram os perfis de utilizador escolhidos para a classificação, assim como foram integrados os dois algoritmos de classificação num só consistindo então no classificador final de perfis de utilizador de transportes públicos.

No Capítulo 6, estão documentados todos os resultados obtidos ao longo do trabalho desenvolvido que resultou nesta dissertação desde as estatísticas gerais sobre os dados de bilhética e os seus utilizadores. Por fim são apresentados quais os perfis de utilizador identificados e a sua distribuição assim como os testes e validação desses resultados.

Por fim, no Capítulo 7, apresentam-se as Conclusões do trabalho realizado e quais se esperam ser os próximos passos deste trabalho.

2. Estado da Arte

Neste capítulo é apresentada uma análise breve relativamente a trabalho relacionado que tem vindo a ser realizado na área de criação/identificação de perfis de utilizador. É inicialmente descrito em que áreas, projetos e estudos os perfis de utilizador estão a ser utilizados, seguindo-se de uma breve descrição da nossa abordagem e como esta difere das restantes.

2.1. Perfis de Utilizador

Um *perfil de utilizador* consiste na estereotipagem que se faz a uma pessoa, atribuindo-lhe um conjunto de elementos distintivos, que permitem identificá-la segundo uma determinada característica ou comportamento [10] [11]. Segundo o trabalho [1] os perfis de utilizador são então "uma representação digital explícita da entidade de uma pessoa", ou de um grupo de pessoas.

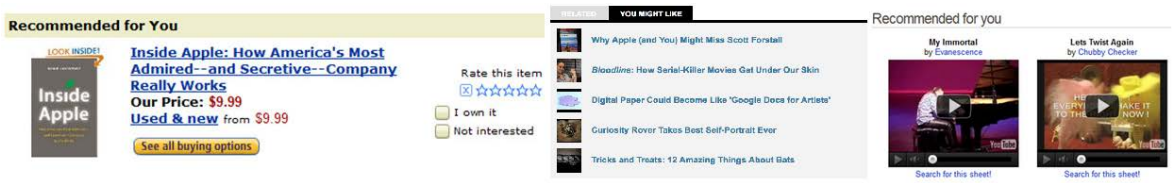

Figura 1 - Exemplo de Recomendações

A área de personalização de utilizadores tem vindo a ser cada vez mais explorada por sistemas de recomendação, [12][13], por exemplo em sites *Web* ou aplicações que pretendem ajustar o conteúdo apresentado ao utilizador com base em perfis de utilização (apresentamos alguns exemplos na Figura 1). Ao fazer este tipo de personalização Haider Gerald et al, [12], afirmam que esta técnica transmite ao utilizador uma sensação familiar enquanto este utiliza os serviços personalizados, podendo esta personalização ser direta perguntando ao utilizador as suas preferências ou subtil, por exemplo analisando o comportamento de pesquisa do utilizador. Alguns problemas em relação à personalização são levantados ainda por estes autores, defendendo que esta deve ser feita de modo a que o utilizador não sinta que a sua privacidade está a ser violada, uma vez que más experiências podem fazer os utilizadores deixem de utilizar estes serviços. Afirmam ainda que os perfis de utilizador são informação pessoal que facilmente pode ser usada para fins que não foram consentidos pelo utilizador tornando-se um entrave a este tipo de tecnologia As técnicas de personalização utilizadas nesta área passam por filtros colaborativos que podem ser interativos (o utilizador diz o que gosta ou partilha um interesse com outro utilizador) ou automáticos. Quando automáticos, estes filtros podem ainda ser baseados num sistema de memória orientado ao utilizador, guardando os interesses / padrões de utilização ou orientados aos produtos guardando as interações dos utilizadores com determinado produto. Ao contrário dos autores anteriores que assumem uma personalização mista (interativa e automática), Soltysiak e Crabtree [13] defendem que o *feedback* pedido ao utilizador deve ser mínimo para que o esforço seja feito maioritariamente pelo agente de personalização. Mais especificamente estes autores focam-se na construção de um agente de personalização baseado na análise de E-mails e páginas Web visitadas, construindo vetores que os descrevem. Estes são depois agrupados em *clusters* defendendo que um grupo com um tamanho (dito) significativo represente o interesse do utilizador.

É ainda comum utilizarem-se perfis de utilizador em sistemas de deteção de intrusão [14], em que, utilizando os padrões de utilização dos utentes, é possível identificar situações fraudulentas. No trabalho de Hall et al. [14], o objetivo constituiu em criar perfis de utilizador baseado na utilização de telemóveis (padrões de telefonemas, utilização do serviço, padrões de mobilidade) de forma a identificar anonimamente intrusões, como por exemplo um telemóvel roubado, uma vez que o ladrão provavelmente teria um padrão de utilização diferente. São exemplos de sistemas de deteção de intrusão o AAFID de Balasubramaniyan [15] e EMERALD de Porras e Neuman [16].
Mais especificamente na área dos transportes públicos foram explorados alguns trabalhos na área de perfis de utilizador:

1. No trabalho de Agard et al ,[17], foram conduzidas experiências onde foi implementado um sistema baseado em perfis de utilizador de transportes públicos com o objetivo de melhorar o próprio sistema de transportes. As experiências ocorreram em Gatineau, Quebeque e contaram com um sistema de transportes públicos de média dimensão operando cerca de 200 autocarros e servindo cerca de 240.000 habitantes. Desde de 2001 que os chamados *Smart Cards* (cartões pré-comprados de controlo de acessos) fazem parte desta companhia de transportes sendo que 80% de todos os passageiros já os possuem. Para além disso os veículos estão equipados com sistemas de GPS que em cada paragem registam a sua localização e o percurso feito até aí. Toda a privacidade necessária relativamente aos dados é assegurada uma vez que os dados dos são anónimos (Podemos ver na Figura 2 o modulo que garante essa privacidade assim como todo o sistema de informação inerente).

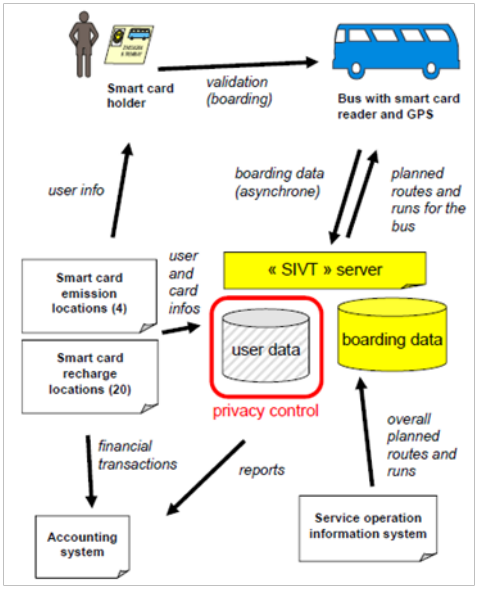

Figura 2 - Sistema de informação *Smart Card*

Card type	Gr1	Gr2	Gr3	Gr4	TOT
Adult	58,8%	13,9%	9,2%	18,1%	100%
Student	21,0%	17,7%	26,4%	34,8%	100%
Elderly	6,2%	6,4%	7,9%	79,5%	100%

Figura 3 - Distribuição de acordo com o tipo de cartão

Utilizando este tipo de títulos, os autores propuseram-se a verificar se a utilização de técnicas de *Data Mining* podiam ser utilizadas para estudar o comportamento dos utilizadores. Utilizando diversas técnicas de *Data Mining*, através de técnicas de *clustering* (*K-Means* e *HCA*) e bem como de caracterização de grupos, os autores chegaram a conclusão que os utilizadores podiam ser divididos em 4 grupos principais, independentemente do tipo de bilhete: No primeiro grupo existe uma relação dos utilizadores com viagens regulares e atividades constantes (tais como ir/vir do trabalho) com base no facto que estes utentes apenas se deslocam durante as horas de pico. O

segundo grupo refere-se a pessoas com atividades regulares especialmente durante o período da manhã. Os restantes dois grupos apresentam características de utilizadores com poucos hábitos de utilização destes transportes (os autores chegam mais tarde à conclusão de que se trata de utilizadores tipicamente idosos). Na Figura 3 apresenta-se a comparação entre o tipo de cartão e os *clusters* obtidos. É de salientar que esta caracterização apenas teve em conta os dias úteis da semana. Notando que se trata de uma primeira impressão sobre o tema, alegam que as técnicas de *Data Mining* ajudam na identificação e caracterização dos segmentos de mercado dos utilizadores de transporte público e que muitas estatísticas podem ser extraídas através deste método.

2. Também na área dos transportes públicos foi implementado um sistema com o objetivo de informar os utilizadores qual o tipo de títulos de transportes mais indicado para o seu padrão de utilização, [18]. Este trabalho faz sentido no caso em questão, uma vez que o sistema de bilhetes de Londres (mais especificamente da operadora *Transport for London* ou *TfL*) inclui 7 temporadas diferentes, 9 zonas geográficas e 12 tipos de utilizador diferentes. Os dados usados neste trabalho dizem respeito ao uso dos transportes públicos de Londres (TfL) durante 83 dias divididos em 2 períodos (Maio a Julho e Outubro a Janeiro). Os dados dizem respeito a 5% de todos os utilizadores registados no sistema nos dois períodos, sendo estes anónimos para prevalecer a privacidade necessária. Também neste trabalho aos dados de bilhética (*Smart Cards*) são aplicadas técnicas de *Data Mining*. Usando os dados das validações, é feita uma análise aos padrões de utilização de cada utente permitindo assim prever os futuros hábitos de consumo de cada utilizador. Medidas de poupança para cada utilizador são ponderadas, procurando pelo título de transporte ótimo para as necessidades previstas de cada utilizador.

3. Por fim, ainda na mesma área, no artigo [19], também os padrões de utilização são utilizados mas neste caso no sentido de identificar potenciais clientes para entidades externas. É referido [19] que estes padrões de utilização nos transportes públicos podem dizer-nos muito sobre aspetos como a rotina de trabalho, hábitos de consumo ou mesmo como as pessoas passam os seus tempos livres. O principal foco deste trabalho passa pelo desenvolvimento de um método que facilite a recolha de comportamento e dinâmica do utente. São utilizados métodos de medida de similaridade entre diferentes utilizadores de modo a poder comparar os estes com modelos preestabelecidos. Foi utilizada uma base de dados de validações com uma janela temporal de quatro semanas consecutivas com cerca de 74.000 validações por dia. Cada utilizador é analisado e elementos como dia da semana, rota, direção, local, tipo de cartão e coordenadas são organizados num vetor que mais tarde ajudará a organizar os utilizadores com os mesmos padrões. Os resultados da identificação correta dos utilizadores neste artigo [19] apresentam valores de 56.9%, 40.3%, 68.8% e 79.4% respetivamente para os 4 tipos de bilhetes que foram analisados, dando uma média de acerto de 61%. Os autores defendem que os resultados ajudam a identificar as diferenças dos padrões de utilização dos utilizadores principais e noutro ponto de vista o facto de identificar zonas de transferência de linhas pelos utilizadores pode ajudar a melhorar o serviço de transportes por exemplo criando novas rotas onde mais transferências são identificadas.

Outros trabalhos exploraram ainda a personalização de utilizadores através de classificação de áreas ou pontos geográficos. Uma vez que os pontos de interesse que pessoas as pessoas frequentam estão sempre a mudar, especialmente nos centros urbanos, também diferentes tipos de utilizador frequentam diferentes locais em alturas diferentes, desta forma Josh Jia-Ching Ying et all [20], decidiram explorar uma análise destes pontos de interesse com base nas aplicações de *check-in*, onde os utilizadores dizem espontaneamente onde estão. Desta forma uma mudança de utilizadores de um local para outro é facilmente identificada. Foi também encontrado um serviço

bastante interessante para a caracterização demográfica chamado Geopulse Context [21] que fornece dados relativos á distribuição da população dependente do ponto de análise escolhida pelo utilizador (como se pode observar na Figura 4).

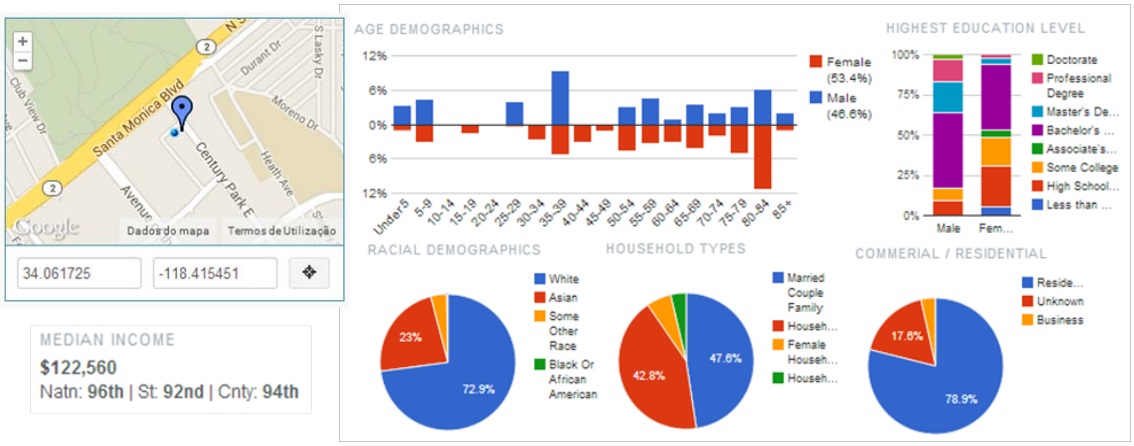

Figura 4 - Geopulse-Context

São disponibilizados a partir deste serviço dados como:

- Idade
- Nível de Educação da População
- Distribuição de Etnias
- Estado Social da População
- Distribuição de edifícios (Residencial, Comercial…)
- Rendimento Médio da população

Embora seja um serviço muito interessante apenas está disponível para certas zonas geográficas e, mais especificamente no continente Americano.

Noutro trabalho [22], com base nos pontos de interesse visitados por um turista, os utilizadores foram caracterizados de forma a poderem ser sugeridos novos locais de interesse para esse utilizador. Embora a personalização associada a áreas ou pontos geográficos esteja pouco explorada parece muito interessante para a caracterização dos utilizadores mais prováveis perto de um ponto de interesse, como um centro comercial, uma escola, um edifício público, um museu ou até mesmo uma discoteca.

Apesar dos trabalhamos descritos, a personalização associada aos transportes públicos é uma área ainda pouco explorada. Os trabalhos publicados sobre este assunto manifestam-se principalmente sobre a forma de trabalhos estatísticos (como em [23][24][25][26][27]), o que faz com que os sistemas de transportes sejam ainda pouco dinâmicos e pouco apelativos para o utilizador. As necessidades dos utilizadores estão em constante mudança obrigando a que estes sistemas de personalização necessitem de mecanismos de resposta mais rápidos e dinâmicos.

2.2. Abordagens ao Problema

Uma das formas de caracterização de utilizador que parece mais fiável consiste na consulta do registo de validação de transporte usando os cartões de identificação ou *Smart Card*. Este método é muito interessante uma vez que, apesar de não ter informação pessoal do utilizador, tem todo o comportamento do respetivo utente registado [28]. Outro método que parece promissor foi a exploração de plataformas GIS para classificar zonas de acordo com os perfis de utilizador dos utentes. A partir dos pontos de interesse que abrangem a área das paragens de transportes públicos associados aos horários desses pontos de interesse pretendemos induzir ou enriquecer os perfis de utilizador mais prováveis numa dada zona num dado momento. Para além disso a ocorrência de eventos públicos georreferenciados permite perceber qual será a classe de utilizadores mais esperada para essa(s) zona(s) nesses períodos ou por exemplo prever quando um desvio dos utilizadores típicos numa carreira pode acontecer. Utilizadores pouco regulares possibilitam também aqui uma forma alternativa de serem identificados.

Uma vez que os *Smart Cards* são essenciais para uma abordagem eficaz ao problema, a eles é dedicado o próximo capítulo da dissertação sobre dados de bilhética.

2.3. Métodos de Classificação

O problema da aprendizagem e classificação dos dados de forma não supervisionada, isto é, sem uma atribuição à partida dos dados a classes, insere-se numa área da inteligência artificial chamada *Machine Learning*, na subcategoria de algoritmos não supervisionados. Tem como principal objetivo obter conhecimento através de uma fonte de dados [29][30].

Em geral, a identificação de padrões pode ser feita de várias formas [31] [32], recorrendo aos seguintes métodos:

Clustering (Agrupamento) – Utiliza-se quando não existe informação associada a cada objeto que o classifique de alguma forma. As técnicas de *clustering* agrupam objetos de acordo com a similaridade entre objetos sendo esta similaridade medida através de uma função apropriada de distância entre os objetos.

Recolha de Informação - Nesta técnica são recolhida objetos que têm informação semelhante à de um objeto de pesquisa. Isto pode ser feito identificando objetos que estão mais perto da característica de pesquisa. Objetos que estejam mais perto da característica de pesquisa são mais prováveis de ser similares.

Deteção de anomalias - Ao contrário das restantes técnicas o objetivo na deteção de anomalias reside em encontrar objetos que não pertencem a um determinado grupo. Este é normalmente um grupo pequeno comparado com o restante universo.

Associação de Regras – Como o nome indica, um objeto para pertencer a um determinado grupo tem que corresponder positivamente a uma ou mais regras. Por exemplo todos os frutos que são redondos e amarelos são maçãs. Este tipo de classificação são normalmente utilizadas para num

contexto comercial, por exemplo para associar que itens são comprados ao mesmo tempo numa loja.

Na maior parte dos estudos relacionados encontrados na literatura relacionada foram utilizados algoritmos de *clustering*, uma vez que o seu objetivo principal é o da divisão de diferentes utilizadores em grupos distintos [13][17][33]. Nomeadamente encontrámos os seguintes algoritmos:

K-Means – Neste algoritmo é usada uma abordagem onde se especifica o número *K* de *clusters* ou agrupamentos desejados, que são representados pelo seu centroide[3] (*mean*). Iterativamente estes centros vão sendo ajustados de acordo com as distâncias aos pontos aos centroides [34]. É necessário definir o K previamente, este pode ser um problema, tendo em vista que normalmente não se sabe quantos *clusters* existem *a priori*. Quando bem afinado este algoritmo é de execução rápida, geralmente convergindo em poucas iterações.

HCA (*Hierarchical Cluster Algorithm*)– Neste algoritmo, os *clusters* são construídos a partir de una inicialização com agrupamentos singulares escolhidos aleatoriamente e,usando um método iterativo, os *clusters* vão sendo unidos por ordem de menor distância entre eles. A cada união, o centro do cluster é reajustado e repete-se o processo. Assim o número de *clusters* vai diminuindo a cada iteração, ficando os agrupamentos cada vez mais definidos em termos de características comuns [34]. Esta técnica é extremamente custosa em termos computacionais, pois exige que se faça um grande número de partições em cada iteração, o que pode levar um longo tempo de desempenho se o número de elementos for demasiado grande.

EM (*Expectation–maximization*)– Este algoritmo ao contrário dos anteriores não é baseado em distâncias. Assume que os dados podem ser modelados como uma combinação linear de distribuições normais, descobrindo assim os parâmetros da distribuição que maximizam um parâmetro de qualidade chamado '*likelihood*' (ou similaridade). Cada elemento pertence aos clusters de acordo com uma probabilidade que vai sendo alterada à medida que os parâmetros são estimados para cada objeto. Este algoritmo é muitas vezes utilizado nas áreas de visão por computador, processamento de fala e reconhecimento de padrões. As suas vantagens passam por ser robusto para dados com muito ruído e não ser necessário definir, à partida, o número de *clusters*. Uma vez que neste algoritmo têm que ser estimados mais parâmetros (centros, covariância e várias probabilidades) pode tornar-se computacionalmente pesado, principalmente quando falamos de base de dados de grande dimensão [35][36].

SOM (*Self-Organizing Map*) – O objetivo deste algoritmo é mapear os padrões dos dados numa tabela de n dimensões. A tabela gera o que se pode chamar de '*outputspace*'. Para preencher esta tabela é necessário encontrar o ponto com o vetor mais perto (menor distância) do vetor espacial dos dados. As vantagens deste algoritmo passam pelo facto de o mapeamento dos dados facilmente ser interpretado e ser capaz de executar grandes e complexas base de dados. Algumas dificuldades passam por determinar que pesos iniciais definir, pode resultar em clusters divididos e requer que pontos próximos tenham padrões similares [37] [38].

Contextualizando, no artigo [17], o objetivo das técnicas de *clustering* utilizadas foi caracterizar os utilizadores de transporte públicos no sentido de identificar parâmetros de mercado associados a estes utilizadores. O artigo [33] é também inserido no contexto de transportes públicos. O objetivo final deste trabalho era separar os utilizadores em 8 grupos de acordo com a sua atividade diária. O

[3] Centroide - Em geometria, um centroide é o ponto no interior de uma forma geométrica que define o seu centro geométrico e, portanto, implicitamente uma medida de distância.

artigo [39] está inserido num contexto mais técnico na qual o objetivo é apenas de comparação de algoritmos de *clustering* de dados.

Em trabalhos idênticos de análise a dados de *Smart Cards* [17], foram utilizados os algoritmos *K-Means* e *HCA* obtendo bons resultados quando comparados com os grupos de utilizadores conhecidos à partida. No artigo [33] afirma-se que o algoritmo *K-Means* é um algoritmo apropriado para analisar este tipo de dados. Dizem ainda que esta técnica é aplicada com sucesso na obtenção de clusters dos indivíduos, podendo este ser adaptado para os atuais simuladores urbanos.

No artigo [39] foi comparado o desempenho e a qualidade dos algoritmos SOM, *K-Means*, *EM*, *HCA*, segundo o tamanho dos dados e o número de clusters. As conclusões deste estudo mostram que os algoritmos que os algoritmos *K-Means* e *EM* apresentam um desempenho computacional melhor que o *HCA*, mas por sua vez apresentam também pior qualidade nos clusters finais. O algoritmo *SOM* apresenta uma qualidade superior nos *clusters* finais mas por sua vez o seu desempenho diminui com o número de clusters existentes. Para concluir os algoritmos *K-Means* e *EM* têm uma qualidade superior quando o tamanho da base de dados é maior. Estas conclusões estão consistentes com outros documentos consultados como por exemplo [47][35].

3. Dados de Bilhética

3.1. Introdução

Os *Smart Cards* têm vindo a ganhar popularidade nos transportes públicos pelo conforto que dão aos passageiros, uma vez que são facilmente recarregáveis, alguns são já personalizáveis, e, em caso de roubo, rapidamente canceláveis. Além disso, o facto de não existir interação com outro elemento humano, ou mesmo contacto direto prolongado, acelera o processo de validação, logo a entrada e acomodação do utente no veículo [40] [28] [41].

Para além deste tipo de conforto para o utilizador, estes títulos permitem otimizar a gestão de frota e serviço de transporte destas empresas, uma vez que estas podem ter acesso aos hábitos dos seus utilizadores e mais facilmente adequar a rede de transportes e respetiva frota ao uso real dos seus utentes.

Com a ajuda dos *Smart Cards* é possível saber "quem" vai dentro do veículo e perceber rotinas do utilização particulares (horas a que costuma entrar para o autocarro, em que zona origem-destino se desloca e a regularidade com que se desloca) possibilitando assim criar perfis de utilizador. Estes perfis de utilizador são essenciais entre outras aplicações, para alimentar sistemas de personalização [17].

Antes de existir a tecnologia dos *Smart Cards*, não era possível obter toda esta informação relativa aos utilizadores de uma forma rápida e dinâmica. Deduzir estatísticas de utilização destes serviços era um processo lento e pouco dinâmico. Para recolher esta informação, era necessário fazer questionários aos utentes, ou utilizar métodos de recolha de dados dispendiosos e rudimentares [42].

Os dados de bilhética usando *Smart Cards*, especificamente em autocarros, têm um processo de recolha de acordo com esquema apresentado na Figura 5.

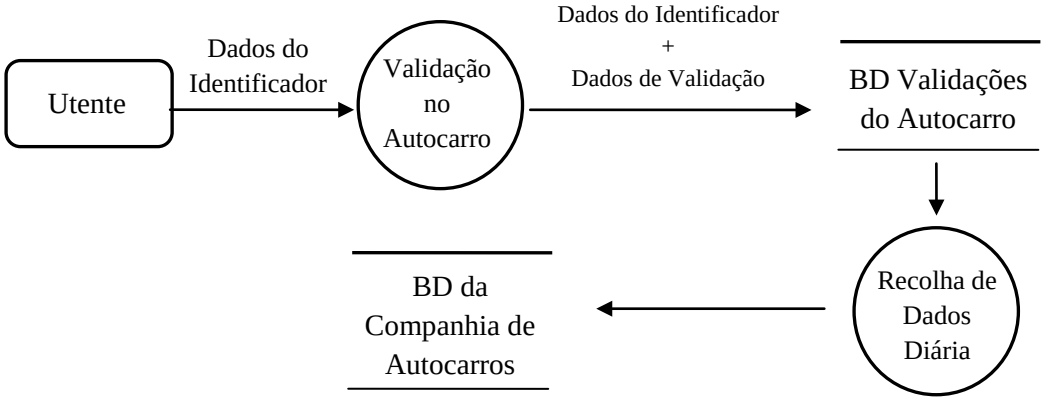

Figura 5 - Esquema de recolha de dados de bilhética

As validações acontecem dentro dos autocarros, quando os utilizadores entram no veículo. No momento da validação, é armazenada informação sobre o utente (cartão) que fez a validação,

nomeadamente: o número do cartão (que nos permite identificar um utente); o tipo de cartão; o número de viagens; descontos, assim como os dados de envolvência desta validação, como a paragem de validação, a linha de autocarro em causa, a zona, a data e hora da validação entre outros detalhes. Ao fim do dia toda esta informação é descarregada para a base de dados central do serviço de transportes.

Um grande problema associado aos dados de bilhética é tratarem de elementos muito sensíveis uma vez que podem pôr em causa a privacidade dos utilizadores [43]. Todas as empresas que trabalham sobre dados pessoais estão sujeitas à fiscalização da Comissão Nacional de Proteção de Dados - CNPD[4] e quando pretendem utilizar esses dados para fins externos devem ser pedidas autorizações para esses fins. A privacidade é garantida no presente caso uma vez que não temos acesso a qualquer informação pessoal associada aos dados disponibilizados, existindo apenas um identificador único para cada título que permite distinguir os cartões que foram utilizados ao longo de todas as viagens e os dados não pessoais (hora, local, etc.).

Para dar suporte à análise dos dados efetuada foi utilizada a ferramenta *Matlab*, nomeadamente para a visualização, manipulação e construção dos modelos de caracterização de perfis. Esta ferramenta é muito utilizada neste tipo de trabalhos por ter uma vertente de manipulação matricial e ser uma ferramenta muito madura. Foi inicialmente experimentada a ferramenta *R*, que é também muito conceituada, tendo sido mais tarde abandonada devido à maior experiência em ambiente *Matlab*.

3.2. Base de Dados de Validações de Utentes

No âmbito desta dissertação, foram disponibilizados para estudos dois conjuntos diversos de dados: dados propriedade da empresa AMI - Tecnologias para Transportes, referentes a informação de uma sua empresa cliente, ARRIVA/TUG (Guimarães), e dados da STCP - Sociedade de Transportes Colectivos do Porto. Depois de analisados os diferentes conjuntos de dados, foi excluído o conjunto da AMI, uma vez que este não disponibiliza qualquer informação sobre os utentes e suas viagens, mas apenas sobre o trajeto (rotas) dos autocarros.

O conjunto de dados escolhido pertence à STCP, uma empresa de transportes públicos com uma frota de 468 autocarros, servindo cerca de 900.000 pessoas em 51 freguesias e seis concelhos do Grande Porto [44]. Os dados relatam um período de 5 meses de validações de títulos dos utilizadores dos autocarros da STCP. Para facilitar a análise inicial, uma vez que se trata de um volume de dados muito grande, foi necessário escolher um subconjunto de dados para que fosse mais rápido e eficaz computar os resultados. Para isso foi escolhida a metodologia encontrada no trabalho de Agardet al. [17].

Surgindo esta necessidade a análise de dados foi dividida em dois conjuntos:

[4] CNPD – A comissão nacional de protecção de dados é uma entidade administrativa independente com poderes de autoridade, que funciona junto da Assembleia da República. Tem como atribuição genérica controlar e fiscalizar o processamento de dados pessoais, em rigoroso respeito pelos direitos do homem e pelas liberdades e garantias consagradas na Constituição e na lei. http://www.cnpd.pt/bin/cnpd/acnpd.htm

- Para o Conjunto 1 (C1) foram escolhidos os primeiros 8 dias de dados (desde 1 de dezembro de 2009 a dia 8 de dezembro de 2009). Assim temos 250.000 utilizadores ao longo de 8 dias com um total de 2 Milhões de validações efetuadas. Neste caso de estudo foi identificada uma frota de 458 veículos que efetuou um serviço de transporte ao longo de 2087 paragens de autocarro pertencentes a 82 carreiras diferentes.
- O conjunto 2 (C2), já com a janela temporal de 5 meses (desde dia 1 de dezembro de 2009 a dia 31 de maio de 2010) é caracterizado por 445.000 utilizadores efetuando um total de 36 Milhões de validações efetuadas por 463 veículos ao longo de 2094 paragens de autocarro pertencentes a 82 carreiras diferentes.

3.3. Exploração e Tratamento dos Dados

No sentido de tentar perceber como melhor tirar partido dos dados para o fim em vista foi efetuada uma análise exploratória preliminar utilizando o conjunto C1. Começou-se por fazer observações de pequenas porções de dados e à medida que se percebeu que informação relevante começou a surgir, aumentou-se progressivamente o tamanho do conjunto chegando numa fase final à totalidade dos dados (C2).

Numa primeira fase e de acordo com a metodologia em [45], algumas questões foram colocadas de modo a tentar caracterizar melhor o problema de como se poderia gerar estereótipos a partir desta informação.

- “Do que estamos à procura?”
- “Que relações se pretendem encontrar?”
- “Que tipos de dados estão disponíveis e que informação é que os atributos transmitem?”
- “A informação está em várias tabelas? Como se relacionam estas tabelas?”
- “Os dados são sazonais?”

Um dos aspetos mais importantes levantados nesta fase foi a sazonalidade dos dados. Uma vez que o período dos dados disponibilizados (C1) é muito específico (dezembro de 2009), e uma vez que o dia 1 e 8 de dezembro são feriados, estamos perante uma semana não padrão. Perante este aspeto, gradualmente fomos aumentando a janela temporal dos dados (até ao conjunto C2).

No sentido de preparar uma base de dados útil e fácil de trabalhar, várias operações foram efetuadas:

1. Limpeza e consolidação dos dados

Inconsistências nos dados foram tratadas, diferentes formatos dos dados foram uniformizados e manipulados de modo a não perder informação. Valores em falta ou incorretos foram corrigidos ou eliminados.

2. Adição de dados externos

No sentido de resolver incoerências, alguns dados retirados do site da STCP [46] foram adicionados à fonte de dados. Para a recolha destes dados foi utilizada a tecnologia JSON e AJAX

para extrair informação adicional do site da STCP utilizando a linguagem Javascript. Utilizando a consola Javascript do browser *Chrome* foi possível extrair informação camuflada nos mapas que a STCP utiliza. Os dados mais importantes incluídos nesta fase foram as coordenadas das paragens de autocarro, a correção de parâmetros que estavam ambíguos, como nomes de paragens e identificador de paragens e, por fim, foram adicionados ainda identificadores que permitiram a ligação entre categorias nos dados de validação.

3. Escolha de informação relevante

Foi efetuada uma análise minuciosa de modo a identificar características menos precisas, relações entre características de difícil perceção e quais as características mais importantes para utilizar no estudo. Por exemplo, algumas categorias tinham dados que não continham nenhuma informação útil para a geração de perfis (informação da máquina de validação, ou vários formatos para as datas, ou até dados mais técnicos como o formato de armazenamento dos cartões).

4. Representação e cálculo de medidas padrão

Para ajudar à análise exploratória, mínimos, máximos e medianas foram calculados assim como foi definido o que cada categoria pretende representar, de modo a garantir que os dados fazem sentido e são úteis para o objetivo final.

Detalhes dos aspetos de preparação dos dados podem ser observados na Tabela A 1 e Tabela A 2, em anexo. Na Tabela 1, a seguir apresentada, descrevem-se algumas das incoerências com que se teve que lidar.

Paragens de validação a NULL	Derivação da paragem pela ordem da paragem
Demasiados elementos a NULL	Eliminação do registo
Número de viagens inconsistente	Derivação a partir do histórico do utilizador
Paragem de validação em falta	Derivação da paragem a partir do site da STCP pela ordem da paragem
Carateres de separação errados	Substituição dos mesmos
Categorias sem informação	Eliminação das Categorias
Datas inconsistentes	Uniformização das datas
Valores inteiros a NULL	Nalguns casos foram convertidos em NaN uma vez que são mais fáceis de analisar pelo Matlab

Tabela 1 - Exemplo de inconsistências dos dados

Depois desta análise dos 54 campos analisados e estudados, 24 foram excluídos (a cinzento na Tabela A 1 e na Tabela A 2, em anexo) uma vez que não têm informação relevante para definir um perfil de utilizador. Os restantes dados ficaram então prontos para utilizar num modelo de geração de perfis que especificaremos nas próximas secções.

3.4. Escolha de Métodos de Classificação

É necessário relembrar que cada caso tem as suas exigências o que faz com que estes algoritmos devam ser testados para o conjunto específico de dados. Para isso decidimos testar alguns algoritmos de clustering: o K-means, o HCA (ou HC), EM e SOM uma vez que estes apresentaram

bons resultados na literatura estudada. Para o caso do K-means fizemos também experiências com as inicializações dos centros (*uniform*, *cluster*, *modes*) e medidas de distância (*euclidian*, *cosine*, *correlation*), métodos disponíveis na ferramenta *Matlab*. Para comparar a qualidade dos clusters usámos o algoritmo NMI (Normalized Mutual Information) [48], cuja expressão é dada por (1), onde Ω e C são as classes esperadas e os resultados do agrupamento em clusters. $I(\Omega; C)$ é a informação mutua entre os dois conjuntos e $H(\Omega)$ e $\mathrm{H}(C)$ são as entropias respetivas.

$$NMI(\Omega, \mathbb{C}) = \frac{I(\Omega; C)}{\sqrt{H(\Omega) + H(C)}} \quad \textbf{(1)}$$

Este algoritmo é normalmente usado para determinar a similaridade dos resultados das técnicas de *clustering*. Os resultados esperados correspondem, neste caso, ao campo *layout do cartão*.

Ao executar os algoritmos EM e SOM sobre os nossos dados deparamo-nos com problemas de memória associados ao tamanho da nossa amostra o que nos fez exclui-los. Em relação ao algoritmo EM, já a partida sabíamos que seria uma das suas desvantagens, tendo esta sido identificada no estudo prévio. Em relação ao algoritmo SOM não foram obtidos resultados satisfatórios provavelmente devido a uma má escolha dos pesos iniciais.

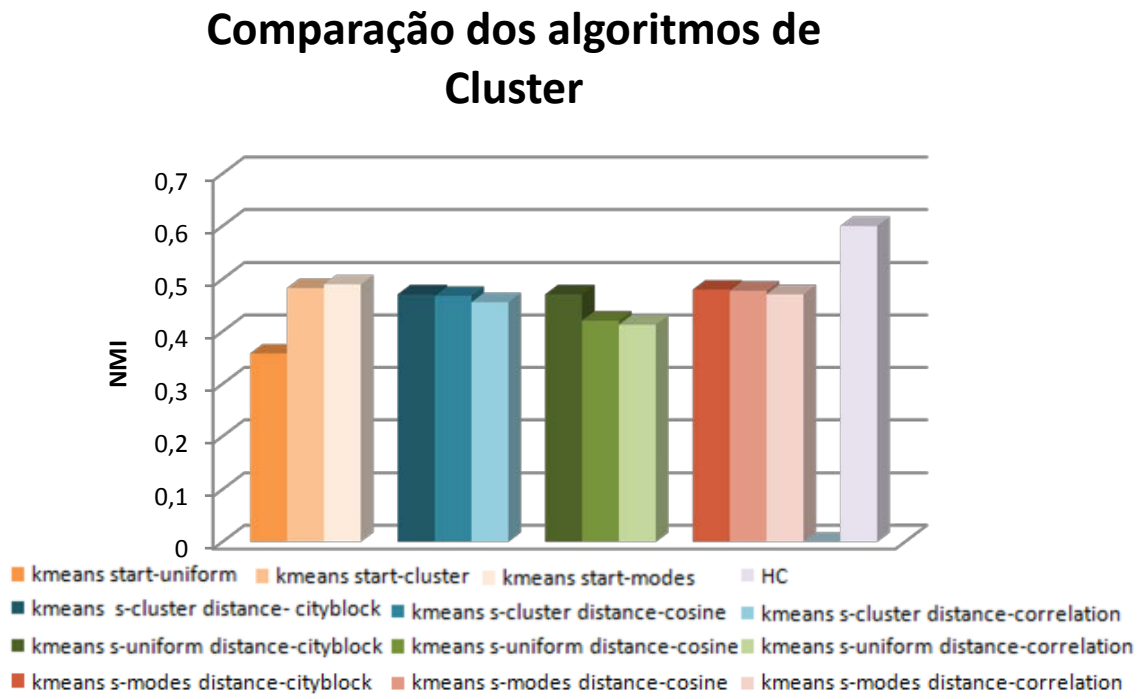

Figura 6 - Comparação de Clusters

Ao observar o gráfico de barras (Figura 6) podemos observar que os resultados mais favoráveis foram obtidos com o algoritmo *Hierachical Clustering* (última barra da figura) seguido do algoritmo *K-Means* (restantes barras). O algoritmo *K-Means* mostra-se bastante regular ao longo das suas diferentes inicializações e medidas de distância. Começámos com uma qualidade de 0,35 inicializado com os centros aleatórios e medida de distância Euclidiana, chegando a um resultado de 0,49 pontos quando escolhemos as modas como inicialização dos centros e distância Euclidiana ou Cityblock (soma do módulo das diferenças). Embora os melhores resultados tenham sido obtidos quando mudámos o algoritmo para *Hierachical Clustering* (HC) obtendo 0,59 pontos.

Depois de analisados os grupos encontrados pelo algoritmo *Hierachical Clustering*, foi possível perceber que estes eram mais ou menos equivalentes aos seguintes perfis de utilizador:

1. Utilizadores *Ocasionais*

Este grupo apresentava uma utilização muito baixa e horários de utilização muito dispersos.

2. *Trabalhadores*

Este grupo apresentava muitas validações, horários muito regulares (principalmente de manhã, almoço e fim de tarde), paragens de autocarro frequentes.

3. *Crianças*

Este grupo apresentava menos validações que o anterior, horários regulares (principalmente de manhã e fim de tarde), paragem de autocarro única ou dupla.

4. *Estudantes*

Este grupo apresentava muitas validações, horários regulares embora se dispersem mais pelo dia que o grupo *Trabalhadores* e paragens de autocarro frequentes.

Achámos bastante interessante que, ao observar o campo *layout do cartão*, existirem também quatro grupos, neste caso definidos como: Títulos Recarregáveis, Títulos Passe, Título Crianças e Título Jovens. Embora estes dados não correspondam a um perfil, podem aproximar-se razoavelmente destes. Assim decidimos utilizar este campo para medir a qualidade dos clusters.

A distribuição dos grupos está apresentada a seguir na Figura 7.

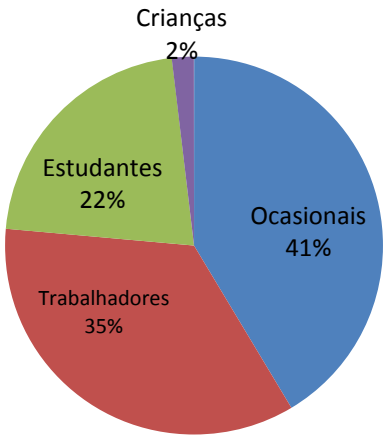

Figura 7 - Resultados Clusters

4. Dados Geográficos

Os Pontos de Interesse (ou POIs) correspondem a informação geográfica que pode ser associada às paragens de autocarro por proximidade. Desta forma, podem contribuir para caracterizar os potenciais utentes dessas paragens. Com o objetivo de classificar os utilizadores através de informação geográfica, assumimos que os utilizadores se deslocam de e para os pontos de interesse na proximidade de cada paragem de autocarro. Assim, podemos associar os perfis de utilizador diretamente aos perfis de paragem de autocarro derivados dos POIs.

4.1. Sistemas de Informação Geográfica

Como sistema de informação geográfica (em Inglês GIS) decidimos utilizar o sistema Open StreeMaps (OSM) que inclui uma base de dados *Open Source* e também por ser uma das base de dados geográficas mais completas e mais rica em termos de descrição de pontos de interesse. Inicialmente foi ponderado o uso do próprio sistema GIS do projeto Tice.Mobilidade embora quando se iniciaram as experiências se tenha constatado que a plataforma ainda não estava suficientemente populada para ser utilizada no projeto. Um dos problemas do OSM é que, por ser uma plataforma colaborativa, pode ter erros associados. Isto poderia ser um problema se estivéssemos a criar uma aplicação que interferisse diretamente com as pessoas. Uma vez que o objetivo é uma classificação de utilizadores, esta questão não se coloca, podendo, na pior das situações, aumentar o erro associado à classificação. Foi também ponderada a utilização do Google Maps (sem os problemas referidos), mas as restrições impostas por este serviço fizeram-nos optar pelo OSM. O facto do acesso aos dados do OSM ser facultado de uma forma muito simples e sem restrições foi um fator, portanto, decisivo para a sua escolha.

4.2. Organização de Categorias de POIs

Depois de escolhido o sistema de informação geográfica, começamos por analisar as categorias (ou TAGs) utilizadas pelo OSM para classificar os POIs [49] ajudando a decidir quais os perfis de utilizador possíveis de identificar nestes dados. De seguida, as categorias foram organizadas de acordo com a influência que estas têm para os perfis de utilizador. As categorias do OSM propostas para melhor caracterizar os utilizadores estão descritas na Tabela 2.

Ao explorar as categorias apercebemo-nos que nem todas elas têm a mesma frequência de pontos de interesse por área, o que pode influenciar uma categorização por área. Por exemplo, existem 126 *restaurantes* na zona do Porto mas só existem 12 *bibliotecas*, o que implica dar um peso superior a bibliotecas que a restaurantes. Na pontuação de uma paragem que tenha na sua abrangência 4 *cafés* (perfil *Lazer*) e 1 *biblioteca* (perfil *Estudante*), o perfil *Lazer* não deve ser favorecido em relação ao perfil *Estudante* por haver um maior número de cafés que bibliotecas. Uma vez que existem pontos de interesse com uma densidade muito mais elevada que outros é necessário criar um sistema de influência para cada categoria de ponto de interesse.

Ainda assim, certos pontos de interesse têm associada uma massa populacional muito mais elevada que outros. Por exemplo uma *universidade* tem necessariamente mais pessoas associadas que um *pub* ou *loja*. Para isso decidiu-se calcular a massa populacional para cada categoria. Surgiu a necessidade de utilizar então o *Foursquare* (aplicação de *check in* onde as pessoas dizem voluntariamente a sua localização) e medir o número de utilizadores para cada categoria, Para calcular o número de pessoas associadas (quarta coluna da Tabela 2) foi utilizada a API do *Foursquare*, [50] e calculada a média de utilizadores para cada categoria. Para cada categoria foram escolhidos 30 POIs num raio de 2km do centro do Porto (limites da API do Foursquare).

Por fim a influência de um dado POI na pontuação da paragem deve ainda decrescer com a distância do POI à paragem de autocarro, uma vez que a probabilidade de um utilizador se deslocar entre o POI e a paragem também decresce [51].

Para a contagem do número de POIs foi utilizada uma aplicação Java [52] disponibilizada pelo OSM que facilitou a recolha dos dados. Delimitando uma dada área geográfica, esta aplicação faz múltiplas chamadas à API do OSM permitindo uma recolha facilitada dos dados necessários. Nesse sentido resolvemos contar o número de pontos de interesse associados a cada categoria.

Na Tabela 2 apresentam-se as categorias recolhidas do OSM (2ª coluna) associadas aos perfis de utilizador / perfil de paragem. Uma vez que é a área que pretendemos representar é o centro urbano do Porto, apresenta-se também na tabela o número de pontos de interesse de cada categoria e o número de utilizadores.

Perfil	Categorias (Tags OSM)	Número de POIs (Porto)	Massa Populacional (Porto)
Lazer	Sport	173	924
	Shop	242	463
	Leisure	388	587
	Café	130	773
	Restaurant	126	665
	Theatre	8	121
	Marketplace	4	362
	Bar	34	653
	Pub	129	653
	Nightclub	14	653
	Cinema	4	425
Turista	Tourism	170	962
	Natural	1086	962
	Historic	68	962
	Hotel	76	856
	commercial	4	463
	Chapel	3	30
	Church	4	30
	place_of_worship	64	30
	beach_resort	0	1
Trabalhador	Office	50	352
	public_building	97	1
	Bank	53	254
	courthouse	2	1
	commercial	4	1
	Landuse	1	1
	industrial	13	1
	Warehouse	11	1
Estudante	Library	12	683
	University	23	742
	Dormitory	1	362
Criança	School	71	254
	College	2	342
	Kindergarten	2	1

Tabela 2 – Divisão de Categorias de Pontos de interesse por Perfil de Utilizador e seus POIs e utilizadores associados.

4.3. Método de Recolha de Informação

Tal como anteriormente referido, de modo a recolher os POIS do OSM usando a API disponível, começámos por filtrar os pontos de interesse na proximidade das paragens de autocarro. Uma vez que a API do OSM apenas permite recolher POIs limitados por duas coordenadas (canto inferior esquerdo e canto superior direito da Figura 8) que definem um retângulo, foi necessário fazer alguns cálculos para obter os POIs dentro da área desejada. De acordo com "Transit Capacity and Quality Service Manual Study" [51], 400 metros é a distância que as pessoas estão dispostas a percorrer para apanhar uma paragem de autocarro. Assim os pontos de interesse utilizados para a caracterização de uma paragem de autocarro têm que estar dentro de um círculo de 400 metros de raio centrado na própria paragem de autocarro. Os limites dessa caixa foram calculados utilizando as seguintes fórmulas em (2), (3), (4) e (5):

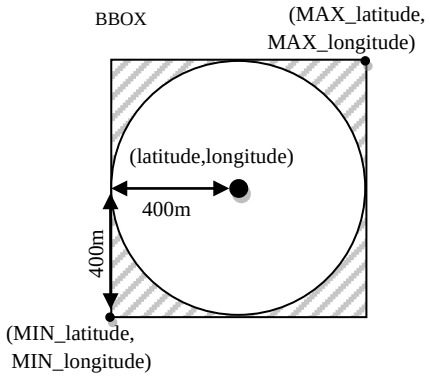

Figura 8 - Caixa de recolha OSM e círculo de interesse

$$\mathrm{MIN}_{\mathrm{latitude}} = \mathrm{latitude} - \mathrm{h1} \quad ; \quad \mathrm{MIN}_{\mathrm{longitude}} = \mathrm{longitude} - \mathrm{h2} \qquad \textbf{(2) e (3)}$$

$$\mathrm{MAX}_{\mathrm{longitude}} = \mathrm{latitude} + \mathrm{h1} \quad ; \quad \mathrm{MAX}_{\mathrm{longitude}} = \mathrm{longitude} + \mathrm{h2} \qquad \textbf{(4) e (5)}$$

onde h1 e h2 correspondem, em graus, a um deslocamento vertical e horizontal, respetivamente, de 400 metros, à latitude e longitude do Porto. Dos POIs devolvidos pelo OSM, são apenas tomados os que se encontram no interior de um círculo de 400 metros de raio, excluindo os pontos em excesso que estão a sombreado na Figura 8.

4.4. Método de Classificação

Para classificar os utilizadores segundo dados geográficos é necessário em primeiro lugar atribuir às paragens de autocarro uma pontuação para cada perfil de utilizador. Para isso identificam-se os POIs próximos dessa paragem de autocarro (num raio de 400m). Como cada categoria de POI está associado a um perfil de utilizador segundo a Tabela 2. A pontuação para cada paragem de autocarro, *BS_score*, consiste em acumular a influência de cada POI no seu perfil. Como foi referido anteriormente na secção 4.2, existem três fatores a considerar: a frequência dos POIs relativamente à sua categoria, f_{cat}, a massa populacional da categoria a que se associa um peso

empírico, w_{cat}, e a distância à paragem com peso $wDist = 1 - dist(x)/400$. Para cada perfil de utilizador, *p*, é calculada a pontuação para a paragem *x*, tendo em conta os *n* POIs das categorias associadas a esse perfil. Tal é indicado em (6) :

$$BS\_score(x,p) = \sum_{i=\text{POI}\,1}^{\text{POIn}} \frac{1}{f_{\text{cat}}(i)} \cdot w_{\text{cat}}(i) \cdot wDist(i) \quad \textbf{(6)}$$

Depois de obtida a pontuação para as paragens de autocarro, um dado utilizador, *u*, é classificado num dos perfis da Tabela 2, tendo em conta as cinco paragens de autocarro que ele mais frequenta. A identificação dos utilizadores diz respeito aos dados de bilhética. Para cada perfil de utilizador são acumuladas as pontuações *BS_score* multiplicadas pela frequência da sua utilização, *freq*(*u*,*x*), tal como se indica em (7).

$$Profile\_geo\_score(u,p) = \sum_{x=\text{BusStop}\_1}^{\text{BusStop}\_5} freq(u,x) \cdot BS\_score(x,p) \quad \textbf{(7)}$$

O perfil de utilizador escolhido de acordo com os dados geográficos corresponde ao perfil que obtém a máxima pontuação de *Profile_geo_score*, como podemos ver em (8):

$$Profile_{geo} = \arg\max_{p}(Profile\_geo\_score(u,p)) \quad \textbf{(8)}$$

5. Solução Integrada

Neste capítulo as soluções correspondentes as dados de bilhética e aos dados geográficos são integradas de forma a criar um algoritmo final de classificação de perfis de utilizador. Esta escolha deve-se às vantagens relativas de cada uma das soluções anteriores, nomeadamente as características de utilização e o conhecimento associado à informação geográfica. Os perfis de utilizador derivados pelas duas análises dos capítulos anteriores são redefinidos em conformidade com a solução integrada que se pretende.

5.1. Perfis de Utilizador Integrados

Depois do estudo efetuado aos *Smart Cards* e feita a análise aos pontos de interesse foram derivados os perfis de utilizador apresentados na Tabela 3.

É de ter em conta que os perfis de utilizador presentes nesta tabela tiveram em conta a análise do estado da arte e questionários realizados no âmbito dos transportes públicos [23][24][25][26]. Observou-se que na sua maioria, os perfis mais comuns identificados eram os *Trabalhadores* e *Estudante*s, uma vez que estes têm, em geral, padrões muito regulares na sua deslocação nos transportes públicos.

Estes são os perfis principais que conseguimos identificar à partida, uma vez que assumem as características relevantes que podem ser induzidas através dos padrões de mobilidade presentes na base de dados de validações e base de dados geográfica. Os perfis de utilizador mais difíceis de identificar são as *Crianças* e os *Idosos* uma vez que são os que apresentam padrões de mobilidade pouco regulares.

A partir da análise aos *Smart Cards* chegámos à conclusão que era possível de identificar 4 perfis de utilizador: *Estudantes*, *Crianças*, *Trabalhadores* e utilizadores *Ocasionais* uma vez que a análise detalhada dos cluster produzidos pelos algoritmos de agrupamento estavam bem correlacionados com as características destes perfis de utilizador. Como foi antes referido, o perfil *Ocasional* foi derivado por regras.

A partir da análise aos pontos de interesse esperam-se conseguir identificar todas as classes de acordo com as TAGs associadas, com exceção da classe *Idosos* e utilizadores *Ocasionais* uma vez que não apresentam TAGs associadas.

Perfil de Utilizador	Características principais	Categorias dos POIs
Estudantes Universitários	Distância percorrida menor que 5km; Uso essencialmente nos períodos da manhã ou alternado (manhã-almoço-tarde): Associados a cartões mensais; Utilização pelo menos 1 a 2 vezes por dia.	Library; University; Dormitory; Bar; Pub; Nightclub.
Crianças	Distância percorrida menor que 5km; Uso essencialmente nos períodos da manhã ou alternado (manhã-almoço-tarde): Associados a cartões mensais; Descontos nas viagens.	School; College; and Kindergarten
Ocasionais	Sem perfil de distância; Rara utilização; Uso maioritariamente matinal.	No POIs TAGS assigned
Turistas	Cartões semanais; Uso intensivo num curto período espaço de tempo	Tourism; Natural; Historic; National_park; protected_area; Hotel; Commercial Building,; Cathedral; Chapel; Church; Place_of_worship,; Beach_resort.
Idosos	Cartão especial; Uma a duas viagens por dia; Descontos nas viagens	No POIs TAGS assigned.
Trabalhadores	Viagens regulares; Paragens regulares; Horários rígidos; Uso intensivo.	Office; Public_building; Bank; Courthouse; Commercial; Landuse; Industrial; Warehouse.
Lazer	Uso nos fim-de-semana e feriados; Baixa utilização; Pouco regulares.	Sport; Shop; Leisure; Café; Restaurant; Theatre; Marketplace; Cinema.

Tabela 3 - Características dos perfis de utilizador e classes de POI

5.2. Algoritmo Geral de Classificação

Então, para o modelo final de classificação de perfis de utilizador de transporte público, pretendeu-se fundir a informação proveniente das técnicas de agrupamento, sistema de regras discutidos no Capítulo 3 e com a técnica de classificação de base geográfica discutida no Capítulo anterior. Podemos ver na Figura 9 um diagrama onde a arquitetura geral do algoritmo do sistema de classificação de perfis de utilizador de TP é apresentada.

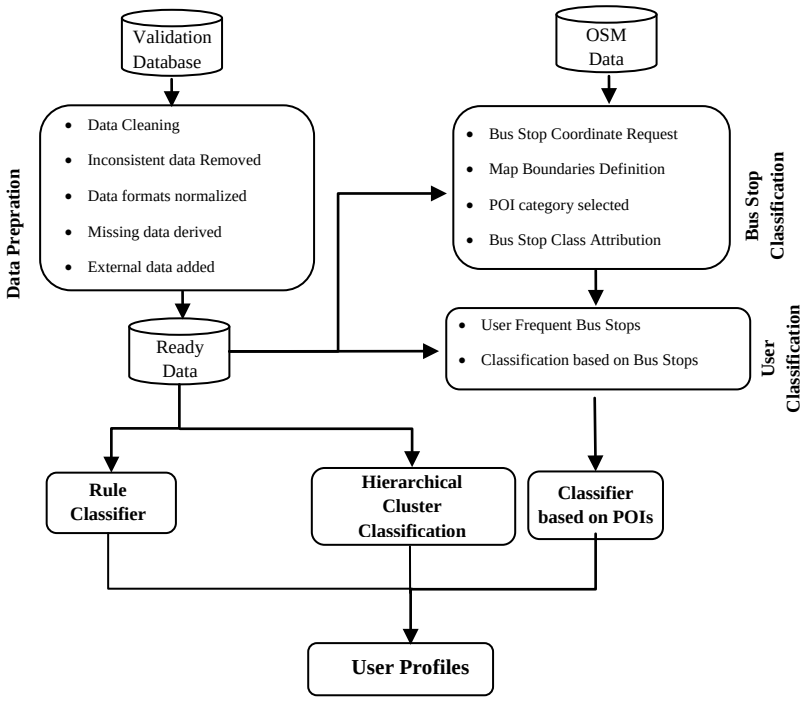

Figura 9 - Arquitetura do Algoritmo de classificação

Num primeiro passo, foi decidido que os utilizadores *Ocasionais* deveriam ser agrupados através de um classificador simples, baseado em regras, uma vez que a identificação desta classe é relativamente fácil de derivar através dos padrões de utilização. Esta categorização do perfil *Ocasional* tem a vantagem de ser fiável, simples e de separar utilizadores ocasionais dos restantes.

Para juntar os resultados dos dois restantes classificadores, foi decidido reajustar a pontuação dada pelo classificador com base em POIs, em função do cluster a que ele foi associado, uma vez que o resultado do *clustering* não tem uma pontuação associada. Interessa ter em conta que a 3 dos clusters encontrados foi atribuído um perfil coincidente com os perfis da Tabela 2. A nova pontuação consiste em aumentar a pesagem associada ao perfil atribuído pelo cluster *c* em 50% do máximo dos pesos dos perfis *p*, como podemos ver em (9):

$$Profile\_new\_score(u,c) = Profile\_geo\_score(u,c) + \frac{1}{2}\max_p(Profile\_geo\_score(u,p)) \quad \textbf{(9)}$$

A justificação desta abordagem prende-se com o facto de, em caso de desacordo entre os classificadores, se tentar valorizar o perfil encontrado dado pela análise dos padrões de utilização. Assim, os padrões de utilização ficam com um peso significativo no classificador. No caso especial de indecisão por parte dos dados geográficos os dados sobre a utilização do transporte prevalecem. Quando o perfil da pontuação máxima coincide com o do cluster, a pontuação nesse perfil fica ainda mais reforçada.

A classificação final segue a mesma expressão (8), agora para a nova pontuação, como podemos ver em (10).

$$Profile_{global} = \arg\max_p(Profile\_new\_score(u,p)) \quad \textbf{(10)}$$

6. Resultados

6.1. Estatísticas Gerais

Para além das ferramentas de análise já descritas anteriormente, foram ainda utilizadas outras ferramentas entre elas HTLM e *Javascript* associado a biblioteca gmaps.js que amplifica o potencial do Google Maps de uma maneira simplificada (A Figura 10 e Figura 11 são exemplos da aplicação desta biblioteca). Através destas ferramentas é possível adicionar e manipular formas geométricas (como polígonos e círculos) ao Google Maps.

Na Figura 10, podemos observar quais as paragens de autocarro onde os utentes mais validam os seus títulos. O centro de cada círculo representa uma paragem de autocarro e o tamanho do círculo é proporcional à afluência da respetiva paragem.

Figura 10 - Paragens com maior afluência

Observa-se que existe uma grande afluência nas paragens perto do centro da cidade do Porto, como era de esperar tratando-se de um centro urbano. Existem também algumas paragens com muita afluência na periferia, o que pode ser explicado pelo facto de existirem menos paragens nestas zonas, levando os utentes das redondezas a deslocarem-se para estas. No Grande Porto entram diariamente 90.276 pessoas que se deslocam a por razões de trabalho, estudo ou outras. Estas deslocações acontecem principalmente devido aos grandes pólos de emprego ou ensino situados nestas áreas [53].

Embora estes resultados pareçam triviais são muito importantes uma vez que têm muita informação relevante para a melhoria dos sistemas de transporte. Por exemplo, tem interesse saber quais as paragens que precisam de mais manutenção, ou aquelas nas quais se deve investir mais (por exemplo nem todas as paragens têm cabine), ou onde um poster publicitário deve ser mais caro.

Na Figura 11, podemos observar as viagens mais comuns. Cada linha representa uma origem e um destino mais realizados pelos utilizadores (O calculo desta linha origem destino é explicado ao

pormenor na secção seguinte, Figura 29). As linhas têm uma transparência diferente proporcional ao número de utilizadores que são representados por essa linha como viagem (O-D) mais frequente.

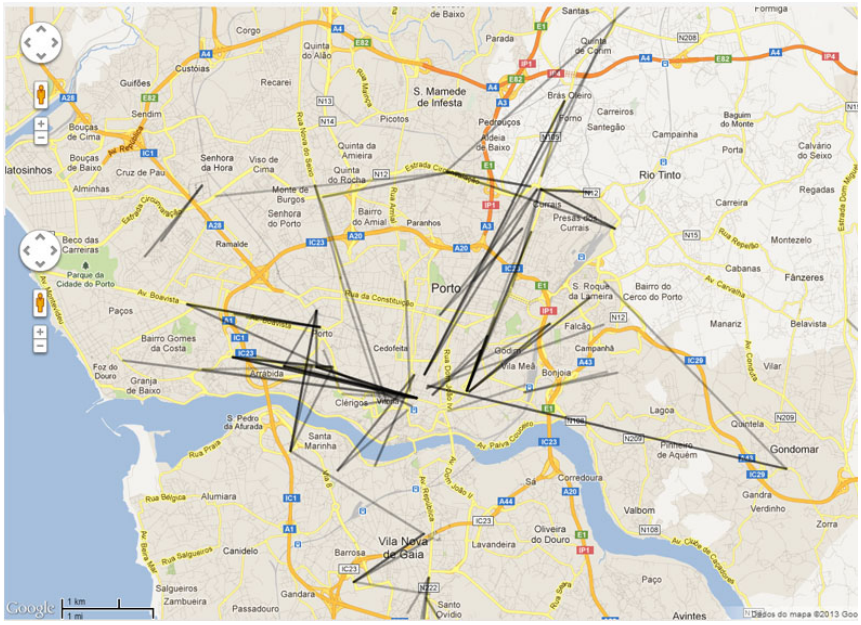

Figura 11 - O-D mais comuns

Realça-se que, como seria de esperar com a discussão anterior, a maior parte das viagens mais comuns são realizadas entre o centro urbano da cidade e a periferia da cidade. Mais uma vez, estas viagens mostram quais as viagens que as empresas de transporte se devem focar para melhorar o seu sistema.

Segundo o estudo realizado no Censos 2011 sobre os movimentos pendulares (interações regionais), "no Grande Porto é visível a interação com diversas NUTS III vizinhas[5]", sendo principalmente nesta zona e também na zona da Grande Lisboa onde se notam mais estes movimentos pendulares (ver Figura 12). Segundo este mesmo estudo estes movimentos têm vindo a aumentar desde 2001 quer dentro da mesma região, quer para regiões diferentes [53].

[5] As NUTS III da região Norte são compostas pelas seguintes Sub-Regiões: Alto Trás-os-Montes, Ave, Cávado, Douro, Entre Douro e Vouga, Grande Porto, Minho-Lima e Tâmega [55]

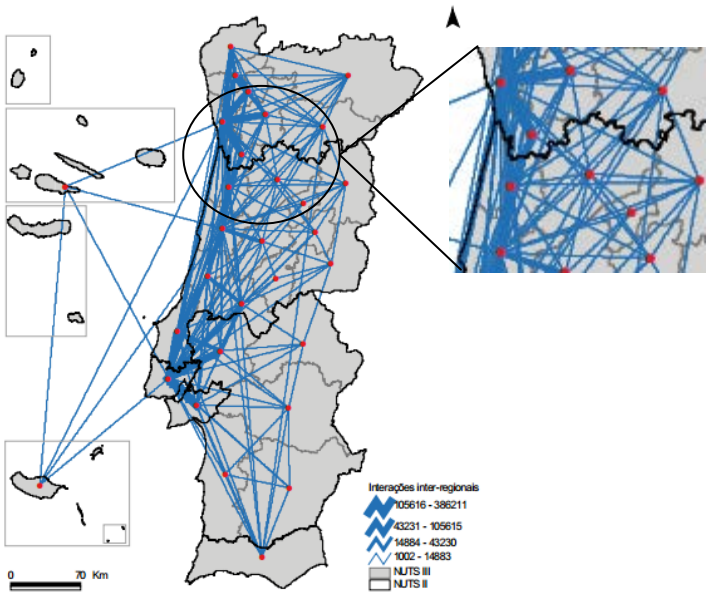

Figura 12 - Censos 2011 Movimentos Pendulares [53]

Os resultados obtidos no estudo dos Censos 2011 mostram que no Grande Porto é visível a interação com diversas NUTs III vizinhas como Minho - Lima, Cávado, Ave e Tâmega [53]. Estes dados validam de algum modo, ou estão em concordância com o estudo realizado por nós. É possível observar, nos nossos resultados, que algumas das viagens mais comuns e paragens mais utilizadas correspondem ao centro do Porto / Regiões adjacentes referidas no estudo (Figura 10 e Figura 11).

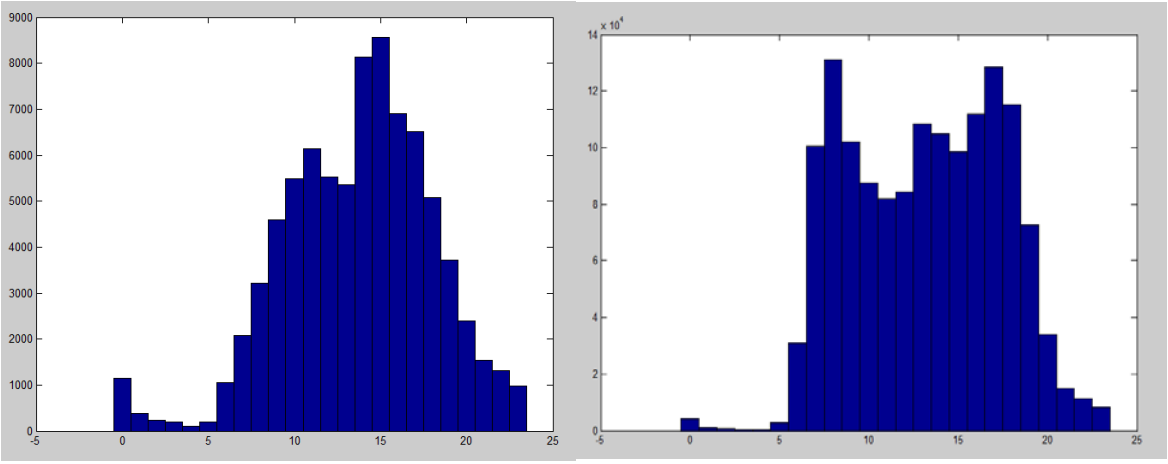

Figura 13 –Viagens em função da hora do dia (C1) ao domingo (esq.) e nos dias úteis (dir.).

Relativamente aos dados do conjunto C1, podemos observar na Figura 13 (esq.) que o horário de utilização mais intenso ao domingo é por volta das 15:00, onde o histograma atinge o seu máximo. Por se tratar de um dia de descanso, a utilização nos períodos da manhã é mínima. Já nos dias úteis, Figura 13 (dir.), existem 3 picos de utilização observáveis. O primeiro no período da manhã e pelas 8:00 da manhã. O seguinte pela hora do almoço por volta das 13:00 e um último pico às 17:00. É

muito provável que estes picos sejam influenciados pelos horários de trabalho dos utilizadores. Esta afirmação ganha alguma consistência quando comparamos a utilização dos dias úteis com a utilização no domingo.

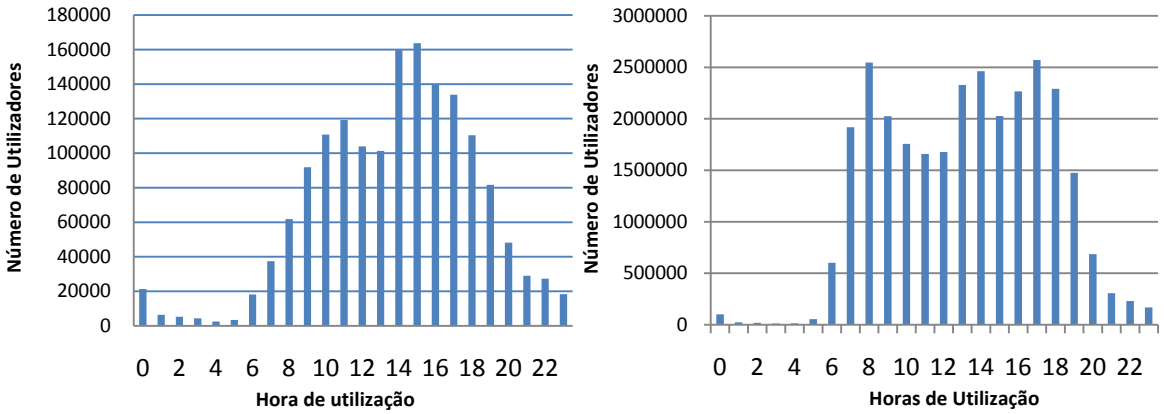

Figura 14 - Número de Viagens em fução da hora (C2) domingo (esq.) e dias uteis (dir.)

Comparando com os resultados mais atuais (Figura 14) já com a análise à base de dados completa (conjunto C2) verificamos que as tendências verificadas na primeira análise se mantêm quando a janela temporal foi alargada. As diferenças são mínimas e quando existem verificam uma oscilação pouco significativa.

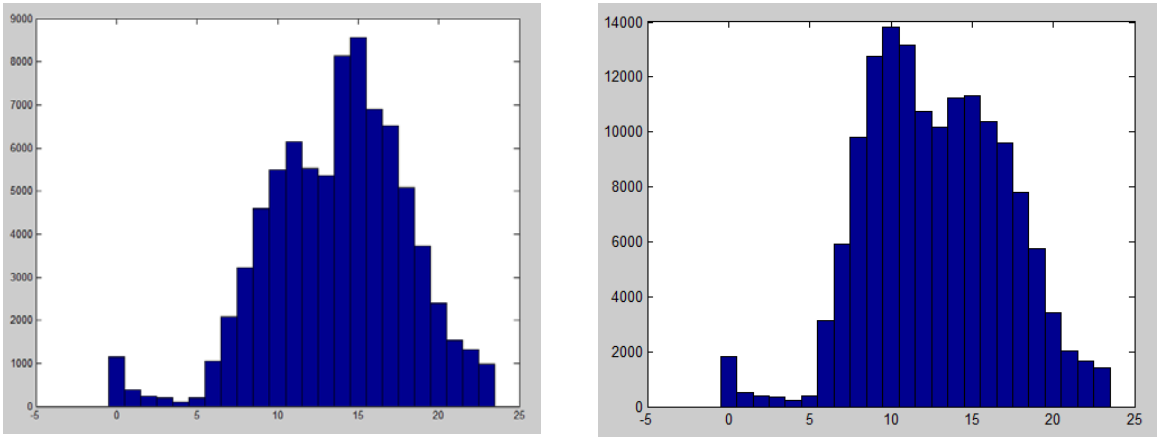

Figura 15 - Número de Viagens em função da hora (C1) ao domingo (esq); e ao sábado (dir.)

Se observarmos agora a Figura 15, relativa ao conjunto C1, podemos verificar que o sábado tem uma grande influência na quantidade de viagens efetuadas ao fim de semana. A resposta mais provável pode ser a de que muitos utilizadores trabalham também ao sábado. No entanto, outra resposta possível pode encontrar-se em utilização dos TP para compras e ou lazer. Isto porque pode-se observar que os picos neste histograma relativos a sábado estão bem definidos: O primeiro entre as 10:00 e as 11:00 e um outro entre as 14:00 e as 15:00. Uma vez que o sábado é composto por um misto de trabalho e lazer (de acordo com a cultura Portuguesa) por parte dos utilizadores, em termos de análise de resultados indicará que este dia pode não ser um dia bom para caracterizar os utilizadores relativamente ao perfil concreto de *Trabalhadores*. Por outro lado, isto prova que, tal como anteriormente referido, há necessidade de várias separações temporais: sazonais e semanais.

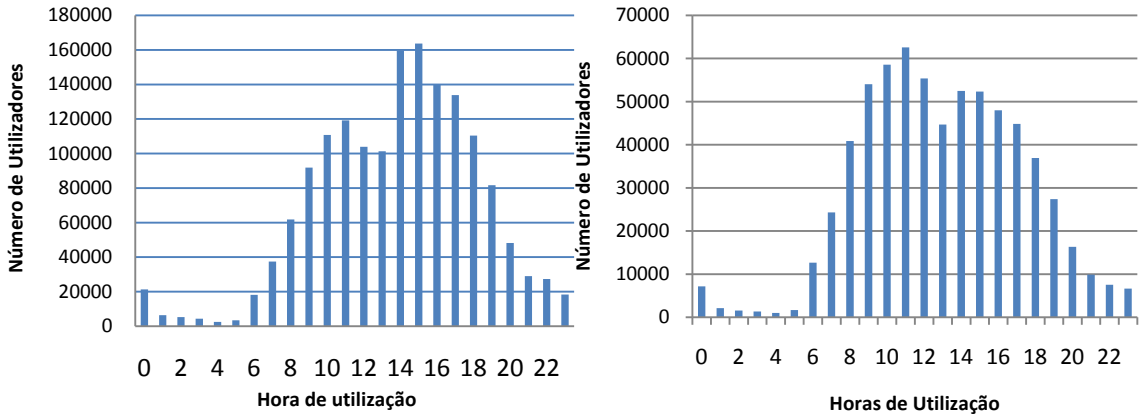

Figura 16 - Número de Viagens em função da hora (C2) ao domingo (esq) e ao sábado (dir.)

Mais uma vez, Figura 16, embora existam pequenas diferenças, a análise feita ao conjunto C1 prevalece observando-se as mesmas tendências reforçando assim as nossas conclusões.

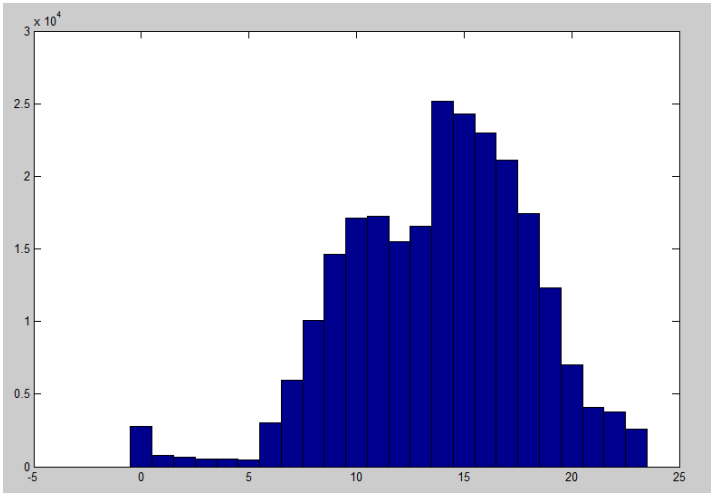

Figura 17 - Número de Viagens aos Feriados em função da hora do dia

A Figura 17 representa o histograma tomado em 2 feriados e pode observar-se que tem uma dispersão de utilização muito idêntica ao domingo embora o número de validações seja muito superior.

Segundo os estudos baseados em questionários [23][24][25][26], a utilização principal dos transportes públicos ocorre maioritariamente como forma de deslocação para o emprego, o que nos levou a estudar se a classe trabalhadora se fazia notar na utilização semanal dos transportes públicos.

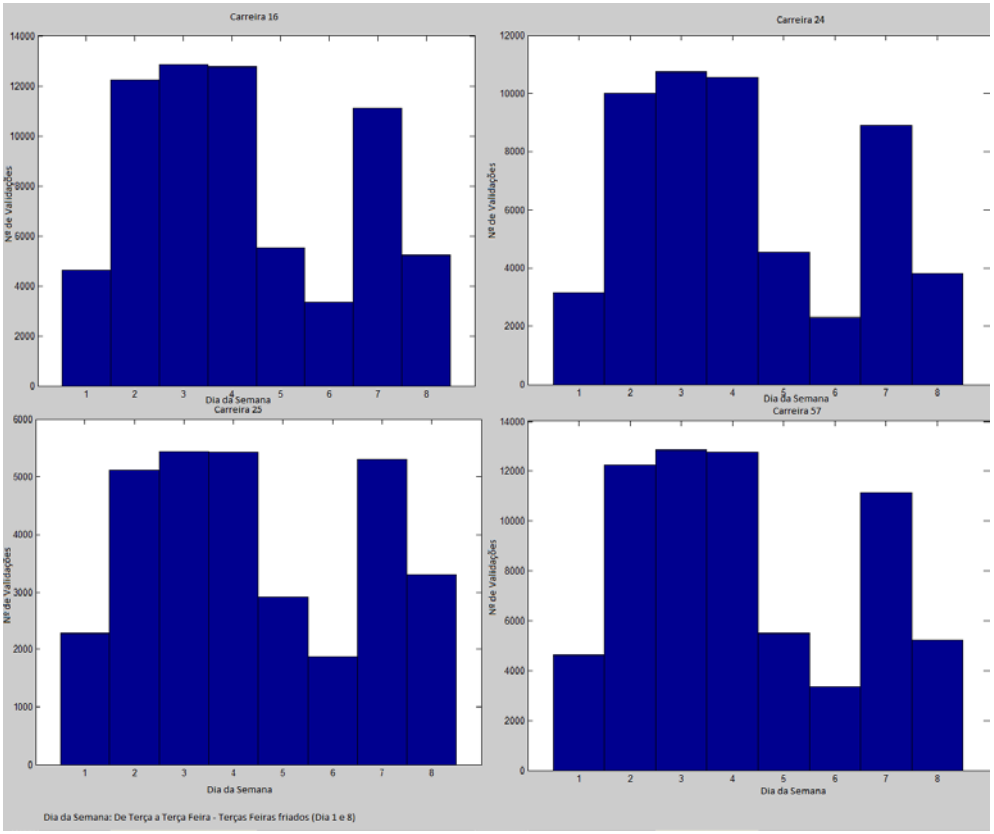

Figura 18 - Utilização de 4 carreiras no Conjunto C1: dias 1 a 8 de dezembro

É possível observar este facto através da Figura 18 (relativamente ao conjunto C1) que esta classe se faz notar nas estatísticas semanais, uma vez que a maior utilização ocorre nos dias úteis. O dia com menos utilização é o domingo (dia 6) seguido dos dias de feriado (dia 1 e dia 8), e por fim o sábado. Os restantes dias úteis tem uma utilização constante nos dias analisados.

Constatou-se ainda de um ponto de vista mais global, através da análise do conjunto C2, que estas tendências se mantiveram iguais quando estendidas para os 5 meses de análise como se pode ver na Figura 19.

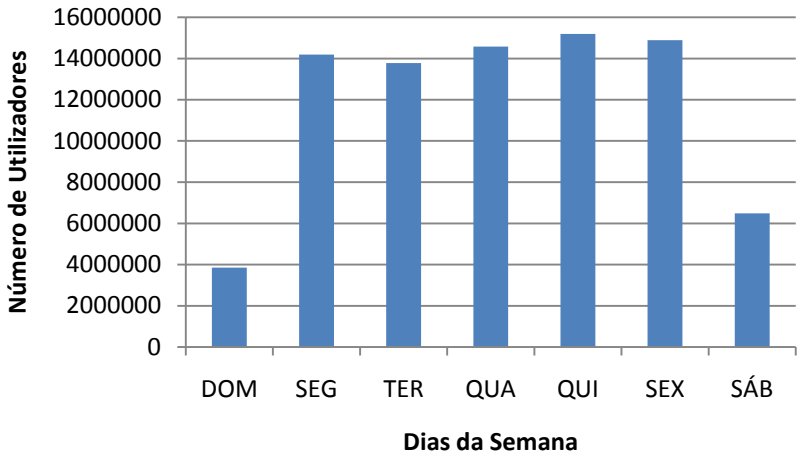

Figura 19 - Utilização por dia de Semana (C2)

Podemos observar (Figura 19) que o domingo continua a ser o dia com menos utilização seguido do sábado, estando os restantes dias nivelados com a mesma utilização.

Depois de uma análise mais geral sobre a utilização dos transportes públicos decidimos fazer uma estatística mais focada nos utilizadores e o potencial perfil de utilizador a que estes podem pertencer.

6.2. Estatísticas de Utilização

Com o fim de perceber quais os perfis mais adequados aos dados de utilização, foi feito um estudo preparatório onde foram analisadas várias características. Chamamos de *subperfil* às categorias analisadas. Por exemplo a característica de utilização do serviço (número de viagens ocorridas) consistiu em dividir os utilizadores em 5 fatias de regularidade de utilização (por exemplo, Muito Regular, Regular ou Pouco Regular; ver Figura 20).

No estudo efetuado aos dados disponibilizados foram especificados 6 subperfis que passamos a descrever.

Subperfil por Regularidade de Utilização

Este subperfil por utilização (Figura 21 e Figura 20) pretende fazer a estatística do número de viagens que cada utilizador faz durante uma semana. As divisões foram feitas com base na média de viagens por dia. Por exemplo, na categoria *regular*, os utilizadores fazem em média pelo menos uma viagem por dia útil, o utilizador *bastante regular* pelo menos duas viagens por dia e assim consecutivamente. Podemos observar que a categoria mais comum é a categoria *Pouco Regular* ao qual pertencem 52% de todos os utilizadores. A figura indica as divisões consideradas.

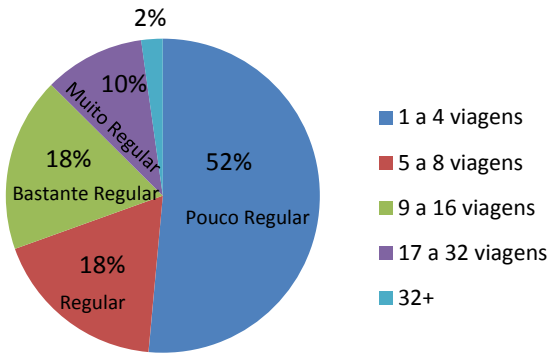

Figura 20 - Subperfil por Utilização (C1)

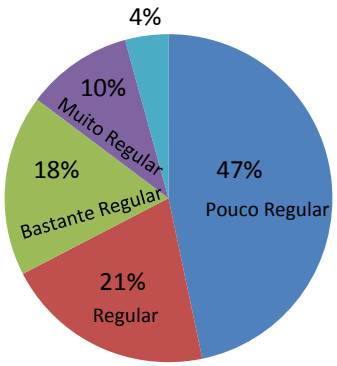

Figura 21 - Subperfil por Utilização (C2)

Comparando os resultados da Figura 21 (agora com uma janela temporal expandida – conjunto C2) com os resultados obtidos anteriormente (usando apenas o conjunto exploratório C1) foi possível de observar que este subperfil por utilização se mantém quase igual, sendo que a classe que mais variou foi o perfil *Pouco Regular* que diminuiu 5%. Esta diminuição faz sentido pois os utilizadores que no período curto apareciam poucas vezes nos registos agora passam a aparecer

mais justificando também o subperfil Regular ter aumentado 3%. As variações restantes não são significativas, mantendo-se a tendência anterior.

Subperfil Horário

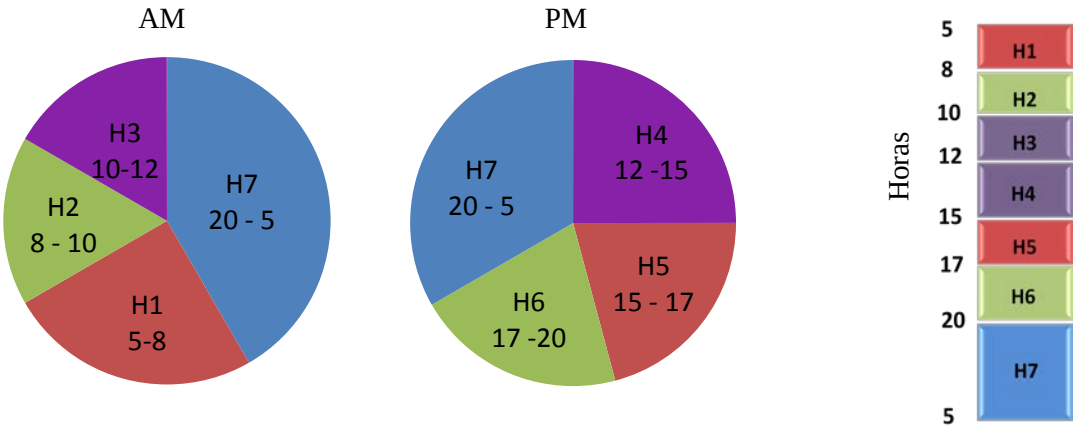

Figura 22 - Espaços Temporais

A semelhança do trabalho realizado por Agard et al [17], foi fragmentado um dia em diferentes porções baseadas nos horários mais prováveis para as pessoas utilizarem o serviço de transportes públicos como se pode observar na Figura 22. Por exemplo o espaço temporal H4 foi escolhido devido a representar a hora de almoço por parte dos trabalhadores do setor público e privado da maior parte das instituições. Assim como o espaço temporal H2 e H5 representam a hora respetiva de entrada e saída destas mesmas empresas.

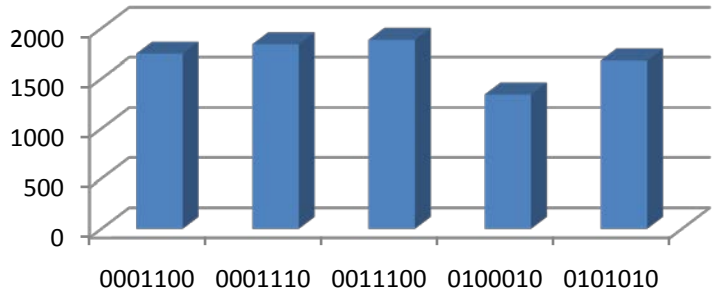

Figura 23 - Perfil Horário

Na realidade, este subperfil foi criado com objetivo de identificar padrões de utilização dos utilizadores. Em que períodos os utilizadores utilizam os transportes públicos? Utilizam sempre nos mesmos horários? Para responder a estas perguntas decidimos criar um mecanismo que permitisse responder a estas questões. A este mecanismo chamámos de *máscaras*, que pretendem representar a altura do dia em que os utilizadores mais utilizam os transportes públicos.

Na Figura 23, podemos observar os 5 conjuntos de máscaras horárias mais frequentes entre todos os utilizadores. Uma máscara é construída a partir dos espaços temporais referidos na Figura 22. Assim, existindo 7 espaços temporais, foi criada uma máscara com 7 bits, representando cada bit um espaço temporal.

Exemplificando, um trabalhador que tem um horário muito regular, imaginando que vai todos os dias no período H2 para o trabalho, no período H4 vai almoçar a casa e regressa ao trabalho, e por fim sai do trabalho no período H6, teria assim a máscara "0101010" associada (na Figura 24 é possível perceber melhor esta atribuição).

H1	H2	H3	H4	H5	H6	H7
0	1	0	1	0	1	0

Figura 24 - Exemplo de Atribuição de Mascara

Na Figura 23, pode observar-se que esta máscara é uma das mais frequentes do conjunto de dados analisados, representando cerca de 1.700 utilizadores. Por fim, uma vez que existem muitas máscaras, é necessário agrupá-las. As máscaras foram agrupadas de acordo com o espaço temporal que ocupavam (de manhã, almoço, tarde) e foi ainda criado um grupo misto aos quais os utentes regulares devem pertencer, uma vez que o horário de utilização é muito disperso.

Subperfil de Redução de Custo

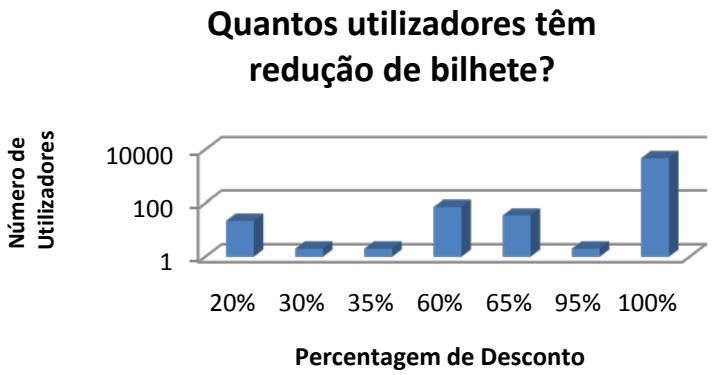

Figura 25 - Redução de Bilhete

Uma vez que poucos utilizadores têm acesso a desconto, e os que têm quase todos usufruem de um desconto de 100% (como se pode verificar na Figura 25) chegou-se a conclusão que apenas fazia sentido haver duas classes no perfil de redução de custo.

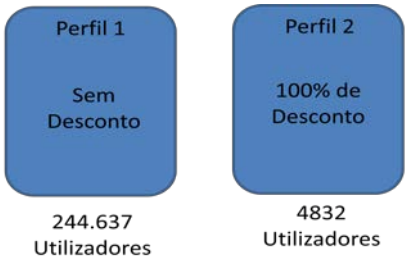

Figura 26 - Subperfil de Redução de Custo

Assim existem dois perfis de Redução de Custo (Figura 26), o *sem desconto* que pretende representar um utilizador normal e o *100% desconto* que pretende representar um utilizador privilegiado (por exemplo um trabalhador da câmara municipal, crianças ou um pensionista).

Subperfil deTipo de Utilização

É ainda possível identificar o tipo de utilização associado a um utente. Na podemos ver que através do número de viagens utilizadas numa validação é possível saber se um utilizador possui um título passe (Perfil 1, na), um título com viagens limitadas (Perfil 2, na) ou até se o validador é partilhado por mais que um utilizador (Perfil 3, na).

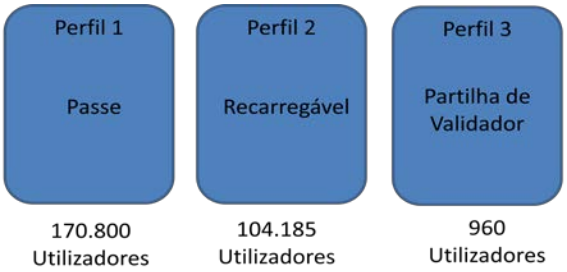

Figura 27 - SubPerfil por tipo de utilização

Subperfil de Viagens Disponíveis

Figura 28 - SubPerfil por viagens disponíveis

As divisões deste subperfil (Figura 28) foram escolhidos tendo em conta que o número de viagens que um utilizador pode comprar de uma vez é de 2, 5 ou 10 viagens. Pode observar-se que a maior parte dos utilizadores tem poucas viagens disponíveis no validador. Como se pode observar pelo subperfil de *Regularidade de Utilização*, cerca de metade dos utilizadores pertencem ao à subdivisão *Pouco Regular*, o que justifica as poucas viagens disponíveis por parte dos utilizadores.

Subperfil de Distância Percorrida

É necessário relembrar que apenas é conhecida a paragem onde é feita a validação do título (a paragem de origem) e não a paragem de destino. Para calcular um perfil de distância percorrida foram escolhidas as paragens de validação mais comuns que cada utente utiliza e calculada a distância entre essas paragens (Ver exemplo na Figura 29). Para os utentes regulares é possível então saber com algum rigor qual a viagem mais comum.

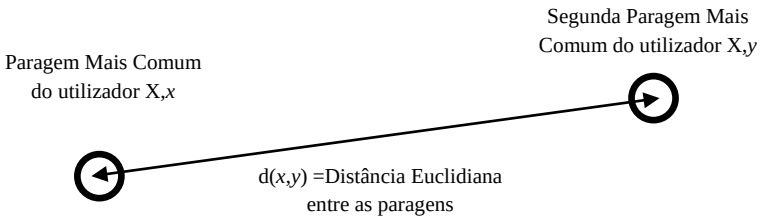

Figura 29 - Calculo de distância percorrida

Dos cerca de 250.000 utilizadores, 70.000 (28%) não puderam ser classificados ou porque não tinham viagens suficientes para caracterizar as paragens origem-destino (O-D), ou porque as paragens de O-D não tinham informação georreferenciada associada.

A distância entre as paragens mais frequentes dos utilizadores varia entre poucos metros, no caso dos utilizadores que validam sempre os seus títulos em paragens muito perto umas das outras e 70 km, o caso dos utilizadores que percorrem distâncias muito elevadas.

Para definir um subperfil de distância percorrida, esta classe foi dividida da seguinte forma apresentada na Figura 31:

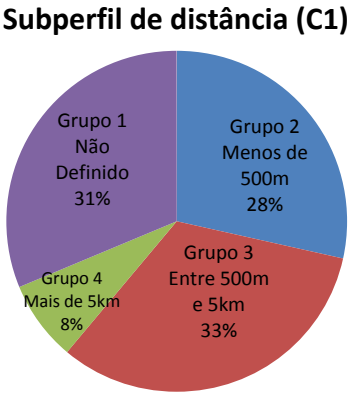

Figura 31 -
Subperfil por Distância Percorrida (C1)

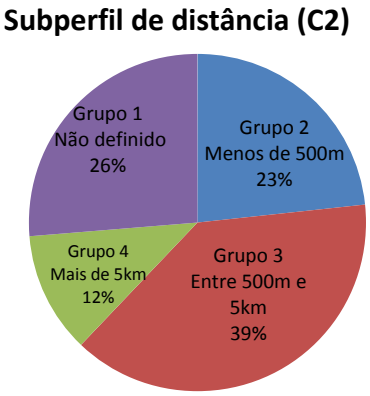

Figura 30 –
Subperfil por Distância Percorrida (C2)

O grupo 1 (“*Não Definido*”) diz respeito aos utilizadores que não têm um subperfil de distância atribuído. O grupo 2 (“*Menos de 500 m*”), diz respeito aos utilizadores que utilizam frequentemente paragens de autocarro muito próximas umas das outras, por exemplo um utilizador que para sair de casa utiliza um transporte público e para regressar a casa utiliza outro metodo de transporte como boleia, a pé ou metro. O grupo 3 (“*Entre 500m e 5 km*”) diz respeito aos utentes que têm um padrão regular e que utilizam os transportes públicos numa distância relativamente curta. O grupo 4 (“*Mais de 5 km*”) aos utentes que fazem viagens maiores, provavelmente utentes que moram na periferia da cidade e utilizam o transporte público para se deslocar ao centro urbano ou vice-versa.

Ao observar os novos resultados obtidos sobre o conjunto completo de dados (Figura 30), podemos observar que houve, tal como seria de esperar ao aumentar a quantidade e tempo de dados, uma alteração aos subperfis de Distância iniciais. Houve um decréscimo de 5% no Grupo 1 justificado

por haver mais dados, logo relativamente menos utilizadores sem informação suficiente para criar estas distâncias. O Grupo 2 teve também um decréscimo de 5%, o que se deverá justificar com o fato de o conjunto C1 ter a particularidade de referir os feriados de dezembro. O Grupo 3 teve um aumento de 6% passando a ser o grupo dominante. Por se tratar de uma distância esperada para a maioria dos utilizadores não surpreende este resultado. O Grupo 4 que teve um aumento de 4 %, manifestando-se maior quantidade de utentes com percursos maiores na sua rotina. Pensamos que no geral, os resultados fazem sentido, visto que a diminuição do Grupo 1 era inevitável pelo aumento de informação, logo estes utilizadores foram ocupar os restantes grupos do gráfico.

Para além destes subperfis terem sido utilizados pelas técnicas de *clustering*, toda esta análise sobre a base de dados de validações, incluindo as estatísticas gerais, como as estatísticas mais enquadradas com o utilizador, ajudaram a definir os perfis de utilizador que se pretendiam identificar, sendo então crucial para esta dissertação.

6.3. Resultados Gerais de Classificação de Perfis de Utilizador

Depois de aplicado o algoritmo descrito no capítulo 5 foram obtidos os seguintes resultados que passamos a analisar:

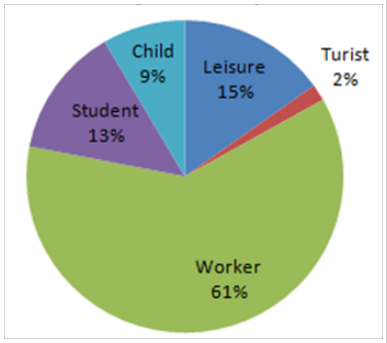

Figura 32 - Perfis de Utilizador (POI)

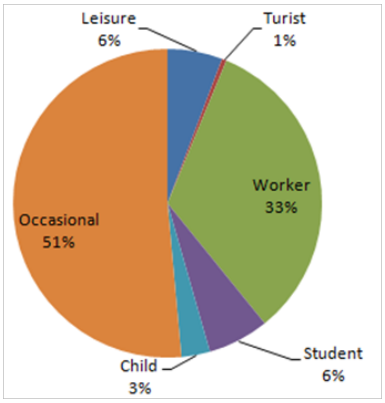

Figura 33 - Perfis de Utilizador Finais

Na Figura 32, que se refere à classificação obtida com base nos pontos de interesse, observámos que o perfil *Trabalhadores* é o dominante, observado em 61% dos utilizadores. O perfil com maior expressividade seguinte é o perfil de utilizador que usa os TP por motivos de *Lazer*, que representa 15% dos utilizadores seguida do perfil *Estudantes* e *Crianças* com respetivamente 13% e 9%. Uma vez que a cidade do Porto é uma das principais cidades de Portugal com um significativo polo de estudantes, facilmente se percebe este resultado. Podemos ainda referir o facto de o perfil *Turista* ter uma expressão mínima de 2%. Pensamos que isto se deve ao facto de ser um perfil difícil de identificar devido à falta de padrões de mobilidade desta classe (período curto de utilização; compra direta de bilhete ao condutor, etc.).

Na Figura 33 pode observar-se o resultado final depois de recalculados os perfis de utilizador com a informação tanto de utilização como geográfica. Analisando este gráfico podemos constatar o que

o perfil de utilizador Ocasional representa mais de metade dos utilizadores (51%) concordando com a análise feita na subsecção 6.2. Nota-se que a maioria dos utilizadores continua a usar os TP como transporte para o trabalho (33%). Por outro lado, 6% dos utilizadores utiliza TP para se deslocar com objetivos de lazer (locais de desporto, cafés, pubs, restaurantes, teatros, entre outros). É de notar que tal como antevisto, a classe *Idoso* não foi categorizada devido a inexistência de padrões de mobilidade suficientes associados a esta classe para identificação. Por outro lado a classe *Criança* acabou por ter resultados surpreendentes, em especial na classificação com a base nos sistemas de GIS.

6.4. Validação do Algoritmo

A validação é necessariamente limitada por não termos disponível uma *ground-truth* absoluta para comparação. Naturalmente que tivemos de recorrer às ferramentas possíveis, ou seja, fazer uma validação parcial comparando os resultados obtidos com os estudos previamente consultados na literatura.

Os resultados apresentados na secção 6.5 estão concordantes com os trabalhos consultados [25], onde os *Trabalhadores* representam 56% dos utilizadores dos TP. A classe estudantil, no nosso estudo dividida em *Estudantes universitários* e *Crianças*, tem em [25] uma representação de 31%, o que mais uma vez é bastante próxima quando comparada com os nossos resultados (*Estudantes Universitários* + *Crianças* =22%). Por fim, 3% dos utilizadores pertencem à classe *Turista*, resultados estes também muito próximos da nossa análise.

Relativamente aos motivos de viagem apresentados no inquérito de mobilidade da população residente na área do Porto [54], onde cerca de 55% das pessoas se deslocam por motivos de trabalho, cerca de 18% por motivos de deslocação académicos, aproximadamente 10% deslocam-se por motivos de lazer ou compras. Os restantes 15% são classificados como outros motivos. Estes resultados, embora não tenham em conta a classe Turista, apresentam valores também muito próximos dos concluídos. Uma vez que este inquérito é referente à região estudada, ainda mais valor acrescentam aos nossos resultados. Esta aproximação dos resultados é muito importante uma vez que mostra que os perfis de utilizador obtidos pelo algoritmo se assemelham à realidade.

Num segundo nível de validação foram comparados o tipo de cartões dos utilizadores com a nossa classificação atribuída aos utilizadores (assumindo que esta categoria está associada aos perfis de utilizador). Existem 4 tipos de cartão (*Trabalhadores*, *Estudantes*, *Crianças* e utilizadores *Ocasionais*) que coincidiram em 62% com os perfis traçados pelo nosso algoritmo. O perfil de utilizador *Lazer* não foi possível de validar uma vez que não existe uma representação deste perfil nos estudos ou dados utilizados. A classe *Idosos* foi excluída no início uma vez que a inexistência de padrões de utilização desta classe dificultou a sua identificação. Para esta classe prevemos que um conjunto de dados mais completo ajude na sua identificação.

Acreditamos que a proximidade dos resultados com as estatísticas da literatura assim como a taxa de acerto dos algoritmos de *clustering* é entusiasmante para uma primeira experiência e marca um passo importante no nosso trabalho. O passo seguinte para conseguir uma melhoria dos resultados passa pela utilização de uma base de dados de melhor qualidade que nos permita uma classificação dos utilizadores mais robusta. Claro que o acesso a dados mais pessoais ajudaria imenso na

validação. A própria melhoria do algoritmo de classificação está também dependente da própria qualidade dos dados.

6.5. Análise aos resultados

Com os perfis de utilizador atribuídos podemos ver como estes são distribuídos pelas paragens de autocarro, em especial por aquelas que verificámos serem mais utilizadas na cidade do Porto (Figura 34).

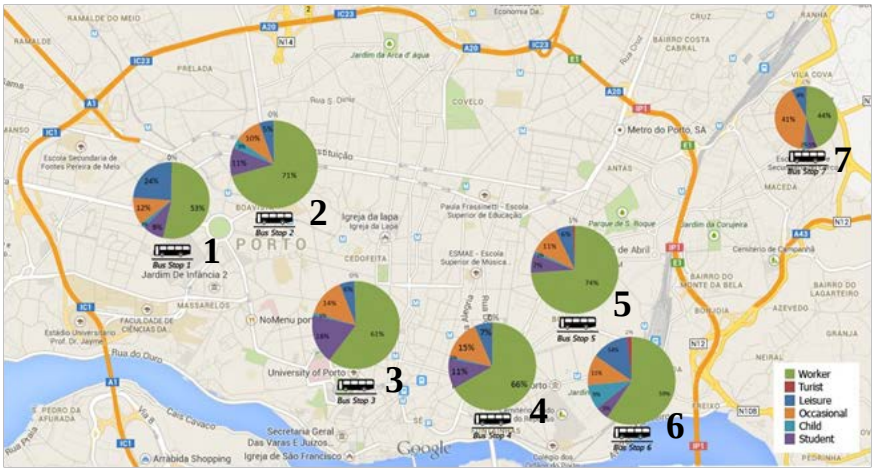

Figura 34- Distribuição dos perfis de utilizador pelas paragens de autocarro

Como era esperado, as paragens mais usadas pertencem ao centro urbano da cidade uma vez que a maioria das pessoas utiliza os TP para se deslocar para o emprego ou estabelecimento de ensino. Isto explica também a taxa elevada da classe de *Trabalhadores* nos resultados. Obviamente que, perto da Universidade do Porto, as paragens 3, 4 e 2, por esta ordem, mostram uma maior quantidade de utentes do perfil *Estudantes*, uma vez que são esta classe é mais esperada perto destes locais. De facto, a paragem 3 mostra que cerca de 1/3 dos utentes que a utilizam são *Estudantes*. Na paragem de autocarro 1 é possível ver que o perfil de *Lazer* tem aqui uma expressão significativa quando comparada com as restantes, uma vez que esta área é o coração da cidade onde existem diversos tipos de lazer como restaurantes, bares, jardins e lojas. Podemos ainda observar que na paragem de autocarro 7 que está mais longe do centro urbano, são os utilizadores *Ocasionais* que se expressam mais significativamente. Pensamos que as pessoas nestas áreas possivelmente utilizarão outro tipo de TP (por exemplo comboio) de forma regular.

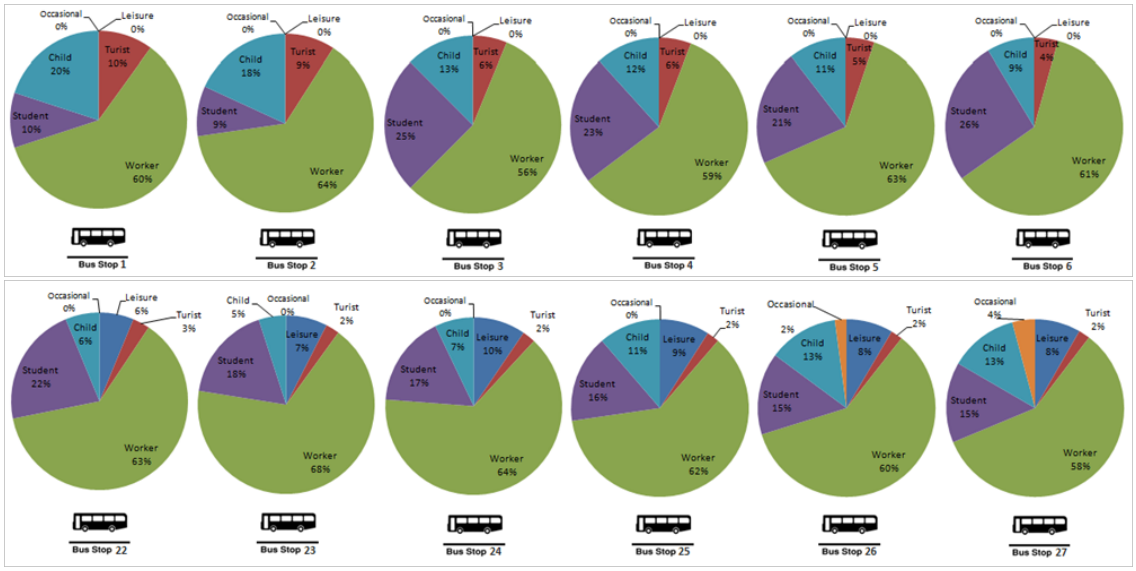

Figura 35 - Perfis de utilizador
ao longo das paragens de autocarro

Pareceu interessante também explorar um exemplo de como se comportaria o modelo na realidade. Para isso escolhemos uma das viagens de autocarro da base de dados de validações e iniciamos a sua caracterização em termos de perfis traçados pelo modelo desenvolvido. A rota escolhida foi uma das que apresentava maior número de validações, deslocando dos arredores do Porto para o centro do Porto, a partir das 8 horas da manhã. Na Figura 35 podemos observar que os perfis de utilizador se adaptam assim que novos utilizadores entram no veículo. Foram escolhidos apenas dois intervalos de paragens de autocarro (1 a 6 e 22 a 27) devido à enorme quantidade de paragens presentes. Nesta viagem o perfil mais representado é o perfil *Trabalhador* (*Worker*) e no caso de se pretender aplicar este conhecimento para um sistema de divulgação de conteúdos esta classe teria que ser o público-alvo. As duas primeiras paragens de autocarro têm uma percentagem da classe *Crianças* (~20%) relativamente elevada o que poderia implicar um aumento no conteúdo dirigido a esta classe. Nas paragens (3 a 6) observa-se uma mudança na distribuição dos perfis de utilizador. A classe *Estudantes* (*Student*) inicia um crescimento de 9% até 25% (na paragem 3) mantendo-se nas três paragens seguinte, uma vez que há universidade perto destas paragens.

Um dos problemas desta análise é não termos qualquer indicação explícita da saída de um utilizador do veículo, o que leva a que os perfis de utilizador, no final da viagem, e sendo cumulativos tenham um erro associado (a validação apenas acontece no início da viagem no caso dos autocarros). Uma possível solução passa por prever quais as paragens de autocarro que cada classe tem maior probabilidade de sair, ou até quais as saídas mais comuns para cada utilizador.Uma possível análise interessante pode contemplar o estudo exaustivo dos ‘ids’ de um conjunto dos *Smart Cards* de modo a verificar, nos mesmos dias para classes rotineiras (trabalhadores, por exemplo), onde estes são validados, tentando inferir destinos e, assim, perceber as saídas.

De seguida é apresentada uma análise aos perfis característicos utilizando algumas linhas (carreiras) específicas:

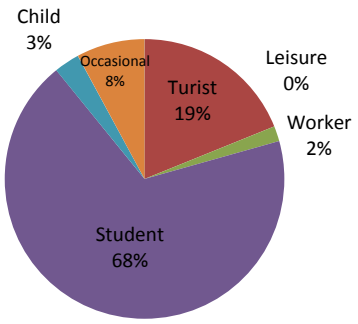

Figura 37 - Distribuição dos perfis de utilizador na linha 50

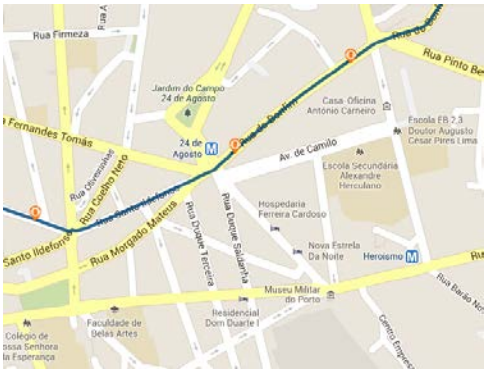

Figura 36 - Zona característica da linha 50

Na linha 50 podemos observar pelo gráfico circular da Figura 37 que o perfil maioritário é o *Estudante* (*Student*) uma vez que esta linha serve uma quantidade de escolas e universidades bastante elevado como pode ser observado na Figura 36 (nesta zona foram identificadas pelo menos 8 instituições escolares). Portanto deve tratar-se de uma linha que serve essencialmente *Estudantes* na sua rotina casa-escola/escola-casa.

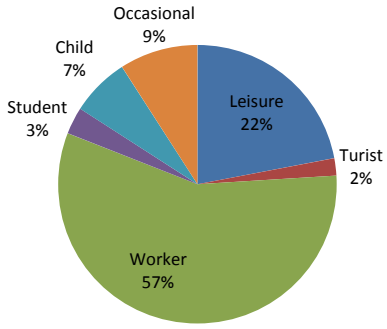

Figura 38 - Distribuição dos perfis de utilizador na linha 53

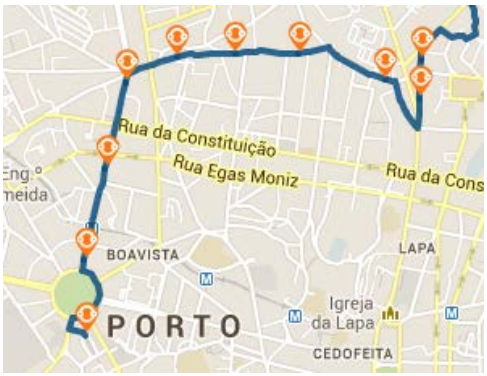

Figura 39 - Zona característica da linha 53

Noutro exemplo, na linha 53, pode observar-se a distribuição dos perfis maioritários de utilizadores na Figura 38, onde se destaca o perfil *Trabalhadores* (*Workers*) que se deslocam desde a zona de Rio Tinto (periferia do Porto) até à rotunda do Boavista, como pode ser observada no mapa da Figura 39. Esta zona pode também explicar o motivo do perfil *Lazer* (Leisure) se manifestar significativamente na distribuição da figura, uma vez que nesta zona existem bastantes locais de lazer, como restaurantes, bares, jardins e lojas, como vimos já anteriormente.

Numa análise ligeiramente diferente, decidimos ainda observar como os perfis de utilizador evoluem durante os períodos do dia. Para isso começamos por apresentar a distribuição geral dos perfis de utilizador para a linha em estudo, linha 51.

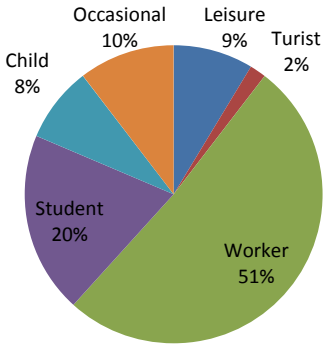

Figura 40 - Distribuição dos
perfis de utilizador na linha 51

Analisando esta distribuição, podemos ver pela Figura 40, que os perfis maioritários que utilizam esta linha são *Trabalhadores* (*Worker*) e *Estudantes* (*Student*) representando cerca de 70% dos perfis de utilizador. Entre os restantes perfis de utilizador, as *Crianças* (*Child*), os utilizadores *Ocasionais* (*Occasional*) e o perfil *Lazer* (*Leisure*) apresentam uma distribuição idêntica (cerca de 10%) ficando o perfil *Turista* (Turist) em minoria (2%).

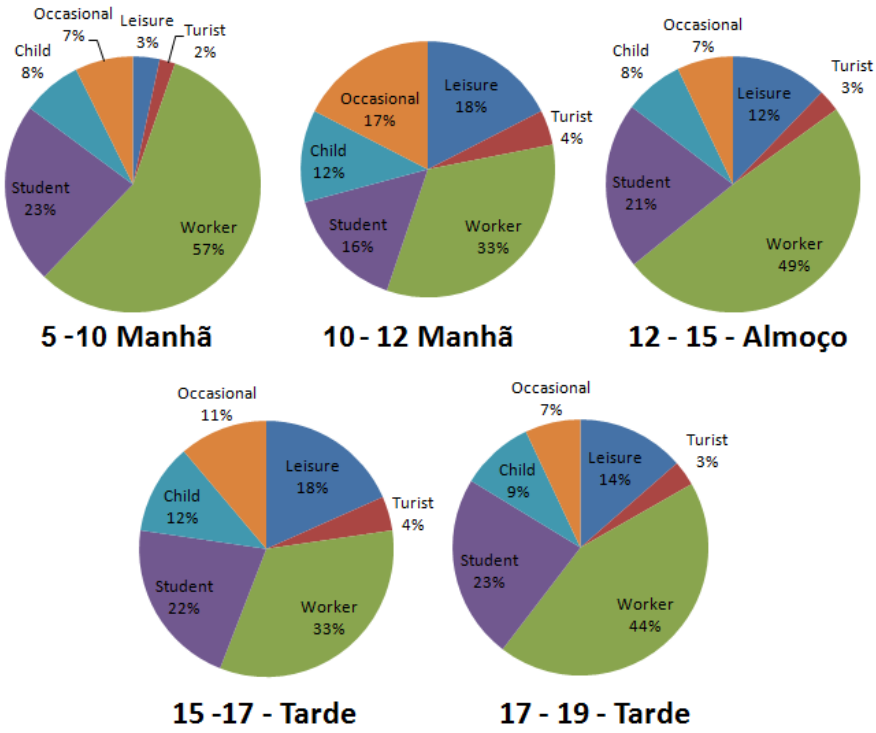

Figura 41 - Distribuição dos
perfis de utilizador por hora da linha 51

Relativamente à distribuição horária, podemos observar pela Figura 41 que existe uma percentagem maior de *Trabalhadores* e *Estudantes* (*Worker*, *Student*) nos períodos do início da manhã, hora de almoço e fim da tarde e que estará possivelmente ligada com a rotina destes perfis de utilizador (Ida para o Emprego/Escola – Hora de almoço - Regresso). Apesar disso a diferença não é tão acentuada no perfil de *Estudantes* (*Students*) uma vez que estes têm um horário mais flexível que o perfil *Trabalhador*.

7. Conclusões e Próximos Passos

A nossa proposta era construir uma nova abordagem para criar um gerador de perfis de utilizador nos transportes públicos aplicando análise de dados para extrair características relevantes dos utilizadores destes transportes. Para isso decidimos usar dados de bilhética de sistemas de transportes públicos (*Smart Cards*) com a caracterização de espaços associados as paragens de autocarro através de pontos de interesse.

Como os nossos resultados mostram (secção 6.7) conseguimos uma acuidade de 62% nas técnicas de *clustering* assumindo que o tipo dos cartões corresponde a um grupo correto (o que pode levar a um resultado ainda melhor, pois nem sempre isto acontece). Quando comparámos os nossos resultados com os da literatura relacionada podemos observar que estes estão muito próximos daqueles que estão publicados. Os resultados usados para validação são fruto de contacto real com os utilizadores o que torna os nossos resultados animadores, uma vez que estes resultaram apenas de técnicas de análise de dados.

Com esta exploração iniciática, é fácil concluir que a análise de dados nos transportes públicos é uma área de grande interesse e, como seria de esperar, muito importante para melhorar os espaços urbanos e a rede de transportes. Apesar disso, existe ainda muito trabalho a realizar para que se possam aplicar estas técnicas num contexto real. A importância que tem no sentido de melhorar a nossa sociedade torna esta área muito motivadora.

As principais dificuldades na realização desta dissertação consistiram no facto de a fonte de dados disponibilizada ser incompleta, faltando muitos detalhes à fonte de dados como, por exemplo, as coordenadas geográficas das paragens de validação, faixa etárias dos utilizadores, género do utilizador. Por outro lado, o facto de muitos dos dados apresentarem incoerências como viagens mais rápidas do que é possível, viagens em que a mesma paragem é usada para todas as validações, muitos valores em falta (*NULL*), diminui a qualidade e quantidade de dados fiáveis. Estas falhas fizeram com que as análises em questão fossem menos eficientes fazendo com que os valores tivessem quer ser ponderados com recurso a fontes externas ter que ser utilizadas. Para além disso o facto de se tratar de uma base de dados de dimensão elevada faz com que os algoritmos tenham tempo de execução elevados. Numa situação real a quantidade de dados a processar é muito menor do que analisada nesta dissertação, sendo o processamento por exemplo diário ou semanal, em vez de dados de 5 meses de validações.

Assim ao longo desta dissertação foram identificadas algumas tarefas a desenvolver futuramente no decorrer deste projeto:

1. **Criar um mecanismo de previsão de utilizadores com base nos dados históricos**.
   Como a previsão dos utilizadores que estão no autocarro e a escolha de conteúdo pode não ser feita em tempo real é necessário criar um mecanismo de previsão de utilizadores para uma dada linha, local, hora entre outras características.
2. **Melhoria do algoritmo de classificação.**
   Várias melhorias podem ser feitas relativamente ao algoritmo de classificação. Por exemplo, podem ser adicionadas novas fontes de dados para a caracterização dos utilizadores como novos sistemas GIS mais completos e com melhor desempenho. A melhoria da informação relativa aos dados dos *Smart Cards* tem que ser melhorada uma

vez que os dados atuais estão muito longe de ser os ideais. Informação mais detalhada e específica para a caracterização de perfis é necessária.

3. **Utilização de Eventos Públicos para prever desvios à normalidade**
   Apesar de nesta dissertação não termos arranjado uma fonte de dados suficientemente boa para a caracterização dos utilizadores através de eventos públicos pensamos que esta área é muito importante para a previsão de desvios à regularidade de perfis. Por exemplo um jogo de futebol ou um concerto de música podem facilmente alterar radicalmente os perfis de utilizador previstos para uma dada altura e lugar.

Outros aspetos seriam interessantes como por exemplo prever as paragens de saída dos utilizadores com base nas paragens de validação mais utilizadas.

8. Referências

[1] J. P. Rodrigue, "The Geography Of Transport Systems - Chapter 6 Urban Transport Problems", rodrigue2009geography , 2012

[2] Éva Molnár, Unece. "Consolidated Resolution on Road Traffic",

[3] Proposta do TICE
http://ebookbrowse.com/proposta-tice-mobilidade-onestoptransport-vp-04-2009-doc-d135897200 (acedida em 22-01-2013)

[4] Inquérito à Mobilidade da População Residente – 2000
(acedida em 24-09-2013)

[5] Tice.Mobilidade
http://www.tice.mobilidade.ipn.pt (acedida em 06-12-2012)

[6] Programa de acção TICE MOBILIDADE
http://www.tice.pt/documentos/ProgramaAc%C3%A7ao_TICE.pdf (acedida em 19-09-2012)

[7] Américo Henrique Pires da Costa. "Transportes Públicos", Comissão de Coordenação e Desenvolvimento Regional do Norte, 2008.

[8] Gabriela Beirão, J.A. Sarsfield Cabral. "Understanding attitudes towards public transport and private car: A qualitative study", beirao2007understanding, 2007

[9] Bernhard Kölmel, SpirosAlexakis. "Location Base Advertisement",
M-BUSINESS 2002 the First International Conference on Mobile Business, kolmel2002location, 2002

[10] GediminasAdomavicius, Alexander Tuzhilin. "User Profiling in Personalization applications through rule discovery and validation", 1999

[11] Bracha Shapira, Uri Hanani, Adi Raveh, Peretz Shoval, "Information Filtering: A New Two-Phase Model Using Stereotypic User Profiling", 1997 Kluwer Academic Publishers,1997

[12] Neuhold, J. Erich, "Personalization and User profiling & Recommender Systems", Proceedings of the WI/IM Information management Proseminar(2003), neuhold2003personalization, 2003

[13] S. J. Soltysiak and I B Crabtree. "Automatic learning of user profiles — towards the personalisation of agent services", soltysiak1998automatic, 1998

[14] Jeyanthi Hall, Michel Barbeau, EvangelosKranakis. "Anomaly-based Intrusion Detection Using Mobility Profiles of Public Transportation Users" , hall2005anomaly, 2005

[15] J. Balasubramaniyan, J. Garcia-Fernandez, D. Isacoff, E. Spafford, andD. Zamboni, "An architechture for intrusion detection using autonomous agents," COAST Laboratory Purdue University, Tech. Rep., 1998.

[16] P. Porras and P. Neumann, "EMERALD: Event monitoring enabling responses

to anomalous live disturbances," in Proceedings of the Twentieth
National Information Systems Security Conference, 1997

[17] Bruno Agard, Catherine Morency, Martin Trépanier. "Mining Public Transport User Behaviour from Smart Card Data", 12th IFAC Symposium on Information Control Problems in Manufacturing-INCOM., agard2006mining, 2006

[18] Neal Lathia, Licia Capra. "Mining Mobility Data to Minimise Travellers' Spending on Public Transport", lathia2011mining, 2011

[19] G. Tseytin, M. Hofmann, M. O'Mahony, D. Lyons . "Tracing Individual Public Transport Customers from an Anonymous Transaction Database" Journal of Public Transportation, Vol. 9, No. 4, lyons2006tracing, 2006

[20] Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wen-Ning Kuo and Vincent S. Tseng. "Urban Point-of-Interest Recommendation by Mining User Check-in Behaviors", ying2012urban, 2012

[21] Geopulse-context, http://www.factual.com/data-apis/places/geopulse-context (acedida em 13-06-2013)

[22] J. Fink and A. Kobsa. "User Modeling for Personalized City Tour", fink2002user, 2002

[23] NCOSS - Council of Social Service of New South Wales. "Who uses Public Transport?", 2006.

[24] TransLink, Queensland Government. "TransLink Public Transport User Survey Summary", 2010.

[25] American Public Transportation Association. "A Profile of Public Transportation Passenger Demographics and Travel Characteristics Reported in On-Board Surveys", neff2007profile, 2007

[26] Lang Yang, Charisma F Choudhury, Moshe Ben-Akiva, JoãoAbreu e Silva, Diana Carvalho. "Stated Preference Survey for New Smart Transport Modes and Services", 2009

[27] Filipe Carrito, Cristóvão Silva. "Análise de não conformidade nos SMTUC", 2011 (acedida em 24-09-2013)

[28] Accenture. "Ticket to the Future: Smart card technology in public transportation",

[29] Wikipedia Inteligência Artificial http://en.wikipedia.org/wiki/Artificial_intelligence (acedida em 17-12-2012)

[30] Wikipedia Machine learning http://en.wikipedia.org/wiki/Machine_learning (acedida em 17-12-2012)

[31] Toby Segaran. "Programming Collective Intelligence", segaran2007programming , 2007

[32] Kamath, Chandrika, "Scientific Data Mining", 2009

[33] Jiang, Shan, Joseph Ferreira, and Marta C. González. "Clustering daily patterns of human activities in the city" Data Mining and Knowledge Discovery 25.3 (2012): 478-510.

[34] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. "Introduction to Data Mining", 2006

[35] David Blei. "Clustering", 2008

[36] Adigun Abimbola Adebisi, Omidiora Elijah Olusayo
and Olabiyisi Stephen Olatunde, “An Exploratory Study of K-Means and Expectation Maximization Algorithms”

[37] Fernando Bação, Victor Lobo. “Self-organizing Maps as Substitutes for *K-Means* Clustering”, 2005

[38] Kevin Pang, “Self-organizing Maps ”,http://www.cs.hmc.edu/~kpang/nn/finalpresentation.ppt (acedida em 22-03-2013)

[39] OsamaAbuAbbas. “Comparasions between data clustering algorithms”, 2007

[40] Philip T. Blythe.
“Improving public transport ticketing through smart cards”, blythe2004improving, 2003

[41] Lathia, Neal, and Licia Capra. "How smart is your smartcard? measuring travel behaviours, perceptions, and incentives." Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011

[42] Ingrid Zukerman And David W. Albrecht. “Predictive Statistical Models for User Modeling”, zukerman2001predictive , 11p, 2001
http://www.springerlink.com/content/x883127200727k8x/fulltext.pdf (acedida em 22-01-2013)

[43] Thomas S. Heydt-Benjamin, Hee-Jin Chae, Benessa Defend, and Kevin Fu. “Privacy for Public Transportation”, heydt2006privacy, 2006

[44] STCP. “Relatório de Gestão e Sustentabilidade”, 2011
http://www.stcp.pt/pdfs/RGestao%20e%20Sustent_2011.pdf (acedida em 31-10-2012)

[45] Microsoft SQL Server 2012. “Data Mining Concepts”, 2012
http://technet.microsoft.com/en-us/library/ms174949.aspx (acedida em 09-01-2013)

[46] Site STCP, http://www.stcp.pt (acedida em 25-08-2013)

[47] Tan, Steinbach, Kumar. “The K-Means algorithm”, 2002
http://www.cs.uvm.edu/~xwu/kdd/Slides/Kmeans-ICDM06.pdf (acedida em 26-09-2012)

[48] NMI algorithm - http://www.mathworks.com/matlabcentral/fileexchange/29047-normalized-mutual-information/content/nmi.m

[49] OSM Map Features DOI=http://wiki.openstreetmap.org/wiki/Map_Features
(acedida em 03-04-2013)
[50] API Foursquare DOI= https://api.foursquare.com
(acedida em 14-04-2013)

[51] Basics: walking distance to transit
http://www.humantransit.org/2011/04/basics-walking-distance-to-transit.html
(acedida em 26-03-2013)

[52] JOSM http://wiki.openstreetmap.org/wiki/JOSM
(acedida em 24-05-2013)

[53] IMTT. “Censos 2011 – Mobilidade de Transportes”,

http://www.imtt.pt/sites/IMTT/Portugues/Noticias/Documents/2012/Censos2011_Mobilidade_Transportes_Pags31-37.pdf
(acedida em 23-01-2013)

[54] “Inquérito à mobilidade da população residente: Cavado-Ave, Grande Porto, Vale do Sousa-Baixo Tâmega, Entre Douro e Vouga – 2000”, 2012
http://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=7250963&PUBLICACOESmodo=2

[55] Wikipedia. “Unidades Territoriais Estatísticas de Portugal“
http://pt.wikipedia.org/wiki/Unidades_Territoriais_Estat%C3%ADsticas_de_Portugal
(acedida em 23-01-2013)

[56] Facebook explorer
https://developers.facebook.com/tools/explorer/
(acedida em 23-03-2013)

9. Anexos

9.1. Outras Tabelas

Nome Cabeçalho	Significado & informação sobre a amostra
Id	Corresponde ao ID único da base de dados
Código	Outro ID único
sn_concentrador	Número de série do concentrador - 545 Concentradores
sn_validador	Número de série do validador - 1096 validadores
Idvalid	Sem significado atribuído - 83 ID's distintos
sw_version	Versão do software do validador - 2 versões
Veiculo	Identificador de veículo - 558 Veículos
location	Sem significado atribuído - 2 locations
pos_valid	Sem significado atribuído - 3 (1 2 10)
provider_code	Código de fornecedor - apenas um valor (3)
sn_h	Sem significado atribuído - 1079 sn's
sn_l	Sem significado atribuído - 249371 sn's
tipo_cartao	Tipos de cartão - (um dígito) (1 2 3 4 6 8)
layout_cartao	Aspeto do cartão –(0 71 83 87)
Empresa	Sem significado atribuído - quase sempre a 0 (0 1 2 3 4 6
Desconto	Tipo de desconto - 0 1 2 3 4 5 9 18 19 20 39 40
reduction_rate	Taxa de desconto - 0 20 30 35 60 65 95 100
data_model	Sem significado atribuído - 69 71 84
contract_index	Sem significado atribuído - 1 2 3 4 5
tarif_code	Código da tarifa - 1 2 6 7 9 12 19 20 31 34 35 36
provider_titulo	Quem forneceu o titulo - (0 3)
contract_ariff	Sem significado atribuído - 11 valores distintos
contract_sale_device	Sem significado atribuído - 283902 valores distintos
contract_sale_date	Sem significado atribuído
contract_sale_number_daily	Sem significado atribuído
Zonamento	Zonas que abrange? do cartão ou da carreira: 2 3 4
zone_mask	Sem significado atribuído - 621 mascaras
num_zonas	de quê? (1-32) cartão ou carreira: 1 2 3 4 5 6 7 8 9 10 11 12
Origem	Sem significado atribuído - tudo a 0
Destino	Sem significado atribuído - tudo a 0
counter_type	Sem significado atribuído –(0 4)
counter_value_before	Números de viagens antes de validar - 1626618 valores
nr_viagens	Viagens disponíveis - tem 0 ou um a menos que
duracao_viagem	Duração da viagem - unidades? de 0 a 65108.
linha_publico	nº da linha para o publico - 82 carreiras (string)
linha_interna	76 nºs de 1 a 127. Estranho serem menos que linha_publico
variante	Sem significado atribuído -1, 2
Sentido	Sentido em que vai o autocarro na linha - 1, 2 ou NULL

Tabela A 1 - Descrição dos Dados

Nome Cabeçalho	Significado & informação sobre a amostra
paragem_validacao	Paragem de validação - 2087 paragens diferentes.
zona_andante	Sem significado atribuído - 21 valores diferentes de 1 a 63
zona_monomodal	Sem significado atribuído - 1 2 3 4 5 7
zona_linha	Zona da linha –(1 2 3 4 5 6 7 8 9 10 11 15)
dh_ini_viagem	Hora e data do inicio da viagem, - nem sempre defnido.
data_hora_validacao	Hora em que foi validado o bilhete - alguns minutos depois de dh_ini_viagem
tipo_evento	Sem significado atribuído - valores 1 3 11
prev_data_hora_validacao	Sem significado atribuído - sempre NULL
prev_tipo_evento	Sem significado atribuído - 0 1 3
prev_provider_code	Sem significado atribuído - 0 2 3 5 6 9 11 12
Autenticacao	não existe autenticação - sempre NULL
sn_titulo	Número de série do título - identifica 249.475 passageiros distintos
Local	Sem significado atribuído - sempre 70.
Equipamento	Sem significado atribuído - 159 valores diferentes.
Zona	Zona em que validou - 38 zonas - string de 3 char alfanuméricos. Ex: '1 ', 'A ', 'C1 ', 'C10'
short_data_hora_validacao	Outra data de validação

Tabela A 2 - Descrição dos Dados (continuação)