



LI ZIXIANG

Licenciado in Engenharia Informática

SMARTCTT GEOCODING

MESTRADO EM ENGENHARIA INFORMÁTICA

Universidade NOVA de Lisboa
Julho, 2022

SMARTCTT GEOCODING

LI ZIXIANG

Licenciado in Engenharia Informática

Orientadora: João Carlos Gomes Moura Pires
Professor, NOVA University Lisbon

Coorientadora: Fernando Pedro Reino da Silva Birra
Professor, NOVA University Lisbon

SmartCTT Geocoding

Copyright © Li Zixiang, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa.

A Faculdade de Ciências e Tecnologia e a Universidade NOVA de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

AGRADECIMENTOS

Gostaria de salientar e agradecer aqui a todos aqueles que de uma forma ou de outra que tornaram este projeto possível a sua realização.

Em primeiro lugar, às instituições pela oferta desta oportunidade. Agradeço a faculdade onde estudei, a NOVA School of Science and Technology e a empresa CTT que propuseram este desafio.

Aos meus orientadores, o professor João Moura Pires e o professor Fernando Birra. Deram-me muito apoio e atenção durante o período da realização. Agradeço imenso pelos seus conselhos e sugestões para a concretização deste trabalho.

Aos membros da equipa do CTT pelo seu apoio e fornecimento de material para o desenvolvimento do sistema.

Aos membros do Júri que aprovaram o trabalho que realizei.

E no fim, aos meus famílias e amigos, obrigado por compreender e encorajar-me durante este período.

*“You cannot teach a man anything; you can only help him
discover it in himself.” (Galileo)*

RESUMO

Os serviços oferecidos pela empresa CTT, necessitam de uma informação tão completa quanto possível das coordenadas das sequências de portas assim como uma estimativa do tempo de entrega associado a cada porta. Apesar de os CTT dispor uma enorme base de endereços com número de porta, devidamente geolocalizados, ela não é completa, como consequência, este dificulta o processo da definição e das otimizações das rotas dos carteiros. Pretende-se então, o desenvolvimento de uma metodologia de geocodificação que se possa caracterizar todos os pontos para uma determinada região, de modo a mais tarde, este possa ser aplicada a diferentes regiões do país.

A geocodificação é um processo composto por uma serie de etapas, incluindo a fase da normalização, a fase da correspondência e a fase de localização, com o objetivo final de estimar a posição geográfica apartir de uma dada informação textual. É provável a existência de erros na estimação, podemos sempre utilizar um indicador de geocodificação para validar os resultados obtidos, de modo maximizando a precisão.

O presente documento apresenta um estudo sobre as técnicas e as tecnologias relacionados na área de geocodificação e também, uma possível resolução do desafio proposta pela empresa CTT, sendo o objetivo de completar ao máximo possível a informações geográficas na base de dados da empresa. De modo a verificar a qualidade e precisão da solução desenvolvida neste projeto, são utilizadas diversas APIs, sendo as mesmas serviços de geocodificação como Google ou Bing, que retornam resultados que irão ser utilizados para um efeito de comparação. Um outro critério de verificação é recorrer a técnica de Ground truth, testando a solução com a informação disponível na base de dados do CTT.

Palavras-chave: Geocodificação, Geolocalização, Georreferenciação, SIG, Endereço

ABSTRACT

In the services offered by the company CTT, its necessary to provide as complete information as possible about the coordinates of the door sequences, such as an estimate of the delivery time associated with each door. Although CTT provides an huge base of addresses with door numbers, properly geolocated, its not complete, as a consequence, the process of altering and optimizing the postman's routes is difficult. Its intended, then, the development of a geocoding methodology in which it can characterize all points for a specific region, so that later, this can be applied to different regions of the country.

Geocoding is a process composed of a series of steps, including a normalization phase, a correspondence phase and a localization phase, with the ultimate goal of estimating a geographical position from a given textual information. There're likely to have application errors, always use a geocoding indicator to validate the results obtained in order to maximize accuracy.

This document presents a study on techniques and related technologies in the area of geocoding and also, a possible solution to challenge proposed by the company CYY, with the objective of completing as much geographic information as possible in the company's database. In order to verify the quality and accuracy of the solution developed in this project several APIs are used, the same geocoding services as Google or Bing, which return results that will be used for a comparison effect. Another verification criterion is to perform the ground truth technique, testing a solution with information available in the CTT database.

Keywords: Geocoding, Geolocation, Georeferencing, GIS, Address

ÍNDICE

Índice de Figuras	x
Índice de Tabelas	xii
1 Introduction	1
1.1 Motivação	1
1.2 Descrição e Contexto	2
1.3 Contribuição Prevista	4
1.4 Organização do Documento	4
2 Estado de Arte	6
2.1 Endereços	7
2.1.1 Endereços de acordo com CTT	8
2.1.2 Abreviaturas no endereço	8
2.2 Dados na Geocodificação	9
2.2.1 Dados de Entrada	9
2.2.2 Dados de Referência	10
2.2.3 Dados de Saida	11
2.3 Algoritmos de Geocodificação	11
2.3.1 Componente de atribuição	11
2.3.2 Componente de correspondência	12
2.3.3 Componente de localização	15
2.4 Indicador de Certeza de Geocodificação	17
2.4.1 Indicador de certeza de análise	18
2.4.2 Indicador de certeza de correspondência	18
2.4.3 Indicador de certeza de Localização	18
2.5 APIs de Geocodificação	19
2.6 Conclusão	21
3 Abordagem	22

3.1	Dados	22
3.1.1	Dados do Cartaxo	23
3.1.2	Normalização dos Dados	26
3.2	Estudo dos Apis	27
3.3	Conclusão	29
3.4	Problemas e Considerações	29
3.5	Estrutura Geral da Solução	31
4	Implementação	33
4.1	Fase de pré-processamento	33
4.1.1	Normalização dos dados	33
4.1.2	Projeção dos Dados	34
4.1.3	Deteção da paridade	35
4.2	Fase da localização	36
4.2.1	Escolha da referência	36
4.2.2	Estimação das coordenadas	37
4.3	Melhoria do Algoritmo	40
4.4	Introdução da régua dos carteiros no geocoding	43
5	Avaliação de Resultados	45
5.1	Criterios de Avaliação	45
5.2	Geocoding com dados Cartaxo	46
5.2.1	Comparação dos resultados com GoogleAPI	48
5.3	Geocoding com a régua do Carteiro	49
5.3.1	Validação de algumas resultados c/ Google Street View	49
6	Conclusão e Trabalhos Futuros	50
6.1	Conclusão	50
6.2	Trabalho Futuro	50
	Bibliografia	52

ÍNDICE DE FIGURAS

1.1	Ideia abstracta sobre geocodificação	4
2.1	Um exemplo simples de geocodificação reversa	7
2.2	Um exemplo comum de endereço postal em Portugal	8
2.3	Modelo em geral de um endereço português	12
2.4	Modelo em geral de um endereço americano	12
2.5	Exemplo do processo de atribuição	12
2.6	Geocodificação com interpolação linear	16
2.7	Geocodificação com interpolação linear e deslocamento	16
2.8	Situação real de posicionamento das casas numa rua	17
2.9	Fórmula para calcular o GCI	18
2.10	Fórmula para calcular o LCI com dados lineares	19
2.11	Fórmula para calcular o LCI com dados poligonais	19
2.12	Tipos de determinações para resultado do <i>Google</i>	20
2.13	Tipos de determinações para resultado do <i>Bing</i>	20
2.14	Tipos de determinações para resultado do <i>MapQuest</i>	21
2.15	Tipos de determinações para resultado do <i>Here</i>	21
3.1	Cartaxo	23
3.2	Distribuição no Cartaxo por CP7	24
3.3	Segmentos que descrevem as ruas no Cartaxo	25
3.4	Visualização gráfica dos rotas dos carteiros parcialmente	25
3.5	Endereços com a mesma designação mas diferentes apresentações	26
3.6	Resultados obtidos dos APIs após a execução de 100 endereços	27
3.7	Resultados obtidos dos APIs após a execução de 1000 endereços	27
3.8	Tipo de determinação do resultado	28
3.9	Distância entre o resultado do tipo Rooftop e real	28
3.10	Distância entre o resultado de outros tipos e real	29
3.11	Exemplos de portas com numeração não regular	30
3.12	Rua com segmento incompleto	30

3.13	Esquema para solução proposta	31
4.1	Exemplo de classificação de componentes	34
4.2	Exemplo de projeção dos endereços na artéria	35
4.3	Exemplos de artérias com/sem a noção de paridade	35
4.4	Exemplo de determinação de paridade de uma artéria	36
4.5	Exemplo de Candidatos vizinhos para a porta número 26	37
4.6	Exemplo de endereços com a sua posição relativa	38
4.7	Exemplo de dados para extrapolação	39
4.8	Reprojeção com o deslocamento	40
4.9	Exemplo de portas do tipo Lote	41
4.10	Exemplo de grafo	41
4.11	Exemplo de construção do troço	42
4.12	Exemplo de candidatos vizinhos pertencerem troços diferentes	43
4.13	Ordem da entrega pré-definido de uma rota de carteiro	44
5.1	Exemplo de como medir o erro	46
5.2	Classificação das portas dos dados do Cartaxo	46
5.3	Resultado obtido com endereços do cartaxo para as portas numéricas	47
5.4	Resultado para as portas não numéricas	48
5.5	Comparação de resultado entre SmartCTT e GoogleApi	48
5.6	Exemplo de um resultado valido	49

ÍNDICE DE TABELAS

3.1	Classificação dos componentes sobre a porta de um endereço	26
-----	--	----

INTRODUCTION

Este capítulo apresenta a motivação e o contexto que leva o desenvolvimento desta dissertação. Introduzindo alguns conceitos básicos e mencionando objetivos e contribuições na sua realização. E no final, uma apresentação sobre a estrutura geral do documento.

1.1 Motivação

Atualmente, os serviços baseados em localização são cada vez mais utilizados, definir um destino no GPS (Sistema de posicionamento global) ou visualizar o mapa no seu *smartphone*, todos eles são exemplos de aplicações de SIG (Sistema de Informação Geográfica).

O SIG é um tipo especial de tecnologia da informação que pode nos ajudar a entender e se relacionar com “o que”, “quando”, “como” e “porque”[2]. Com a evolução das tecnologias, a importância de SIG torna-se cada vez mais elevada, um dos maiores problemas presentes neste tipo de sistema é a obtenção dos dados de recurso. Muitas vezes, as informações geográficas são incompletas e imprecisas. A medida em que o SIG é cada vez mais utilizado, maior é a tendência e a necessidade de obter informação geográfica completa.

A geocodificação e geocodificação reversa são componentes fundamentais do SIG, estes permitem-nos obter as informações sobre a nossa localização, seja ela a posição exata ou relativa. Entrar em mais detalhe, o funcionamento do sistema GPS é caracterizado através destes elementos do SIG.

Apesar da existência de diversas empresas que fornecem serviços para geocodificação, muitas vezes, os resultados obtidos não são confiáveis. Para garantir a maior controle das operações executadas, seria um motivo forte a construção do nosso próprio algoritmo de geocodificação.

Desta forma, o estudo que leva esta dissertação permite compreender as principais

questões envolvidas na área relacionada, e seria uma excelente oportunidade de poder lidar com eles.

1.2 Descrição e Contexto

Esta dissertação está integrada num projeto em colaboração direta entre o Departamento de Informática da Faculdade de Ciências e Tecnologias da Universidade Nova de Lisboa e a empresa CTT.

CTT (Correios, Telégrafo e Telefone) Correios de Portugal, S.A. é uma empresa portuguesa fundada em 1520, essencialmente focada em negócio dos correios. Esta é uma das maiores e mais conhecidas empresas que fornecem o serviço de correio do país.

Tendo em conta o funcionamento deste tipo de empresa, o trabalho essencial seria os seus trabalhadores, neste caso, os carteiros, a fazer os seus percursos da entrega, e levar as cartas ou encomendas a casa dos seus clientes. E para cada carteiro, é atribuído uma lista de cartas ou encomendas a ser entregue durante a sua vida diária. E para evitar o desperdício de recurso, seria extremamente importante que a sua rota seja otimizada.

Para otimização da rota, é responsável pelo *RouteSmart*, onde este é o *software* utilizada pela empresa CTT, para a definição e otimização das rotas dos carteiros. E tendo a parametrização do software, seria necessária a introdução dos endereços e respetivas coordenadas geográficas.

Os CTTs possuem duas bases de dados sobre endereços dos seus clientes para diferentes aplicações nomeadamente, SigPostal (Plataforma de Gestão da Informação Postal) e GIB (Base agregadora de dados geográficas resultante da operação GEO10), entre elas, há uma relação unívoca entre 70% das portas dos dois sistemas. Para um dado endereço, seria possível encontrar as suas coordenadas registadas recorrendo estas duas bases de dados, sendo numa fase posteriori, utilizar no *software RouteSmart*.

Apesar dos CTTs serem a empresa que possuem a maior base de dados sobre endereço do país, as informações presentes não são completas. Existem uma percentagem significativa de portas onde apresentam informações que o caracteriza, mas as suas coordenadas geográficas são desconhecidas. E ainda, algumas portas com conhecimento das coordenadas mas a sua caracterização é desconhecida.

Para o contexto do problema, a dissertação terá como objetivo principal estimular as informações geográficas em falta para as portas existentes na base de dados do CTT, recorrendo a análise das coordenadas existentes, extraindo as características da zona respectiva, implementando as técnicas de geocodificação.

Geocodificação é um componente fundamental da análise espacial e que pode ser aplicado em diversos domínios e áreas científicas [7]. Tomado literalmente, geocodificação significa codificar uma informação geográfica [7]. É o processo em que permite transformar uma informação textual em coordenadas geográficas, i.e., a capacidade de converter um endereço ou nome de um local em coordenadas de latitude e longitude. Talvez não seja realista de o computador encontrar a localização apenas com uma descrição, mas sim, com um conjunto de referências que consiga relacionar. Portanto, podemos também dizer que a geocodificação é um processo que procura uma relação entre onde pretendo saber e o que já sei sobre o mapa.

Para manter o conteúdo de localização atualizada na base de dados geográficas, requer a inserção de dados manualmente, mas não seria nada prático a inserção de todos os pontos existentes, não só exige um elevado custo para a inserção manual, mas também exige uma base de dados de grandes dimensões. Com o uso de método de geocodificação, este permite minimizar o trabalho realizado na inserção, possibilitando a deteção e eliminação de erros numa fase inicial. Outro aspeto importante seria, este permite uma otimização em armazenamento, quando se trata em dados geográficos, geralmente possui grandes dimensões, com a ajuda da geocodificação, diminuindo o volume, guardando apenas os dados mais revelantes, e estimulando os outros casos em falta.

De um ponto de vista mais geral, a geocodificação pode ser definida em a relação entre quatro elementos, como ilustrado na figura 1.1. Sendo os elementos:

- **Os dados de entrada (*Input*)** – dados não espaciais.
- **Os dados de referência** – base de dados previamente conhecidos.
- **Algoritmos** – relaciona entre os dados de entrada e os dados de referência.
- **Os dados de saída (*output*)** – dados espaciais resultantes.

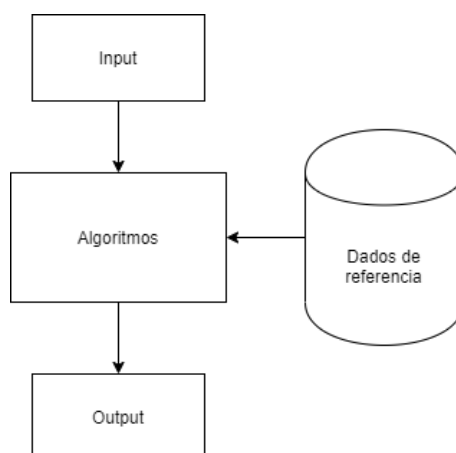


Figura 1.1: Ideia abstracta sobre geocodificação

O projeto baseia-se no desenvolvimento de um modelo para estimular coordenadas geográficas, e que pretende-se que o resultado final desta aplicação seja integrado no *software RouteSmart*.

1.3 Contribuição Prevista

As informações geográficas são extremamente importantes para as empresas, seria muito vantajoso conhecer as posições geográficas dos seus clientes de modo a aplicar estratégias de *marketing*. No contexto do problema, a empresa CTT paga uma taxa anual significativa devida a falta de registos na sua base de dados. Para os casos desconhecidos, estes não se conseguem ser tratados de uma forma automatizada, seriam tratados de uma forma especial, requer a necessidade da intervenção humana, originando elevados custos. Espera-se que o resultado deste projeto consiga trazer benefícios para empresa, de modo reduzir o custo gasto nos processos relacionados.

Outra contribuição prevista, seria no desenvolvimento de geocodificação no país, enriquecendo o conhecimento sobre a distribuições de localização geográfica em Portugal.

1.4 Organização do Documento

O presente documento está estruturado em quatro capítulos, organizados da seguinte forma:

1. Introdução

Este capítulo corresponde a uma breve descrição do problema, incluindo a sua motivação e as contribuições do resultantes esperadas no final.

2. Estado da arte

Este capítulo descreve o estado de arte relativamente às técnicas e tecnologias disponíveis atualmente para este tipo de problema.

3. **Abordagem**

Este capítulo apresenta uma pequena ilustração sobre a solução trabalhada com algumas discussões sobre as suas componentes.

4. **Implementação**

Este capítulo explica com detalhe como são implementados os algoritmos usados na solução proposta.

5. **Avaliação de Resultados**

Este capítulo demonstra os resultados obtidos e uma pequena discussão sobre como os resultados são avaliados.

6. **Conclusão e Trabalhos Futuros**

Este capítulo apresenta uma conclusão sobre todo o trabalho escrito neste documento, e uma pequena discussão sobre trabalhos futuros.

ESTADO DE ARTE

Na nossa vida diária, se houver alguém que viessem nos perguntar sobre um dado endereço, conseguimos indicar a sua localização ou zona aproximada se já possuímos algum conhecimento previa sobre o mesmo. Sendo esta ideia aplicando nas máquinas, designando por geocodificação, um processo em que transforma um endereço na sua representação reconhecível pelos humanos numa sistema de coordenadas, onde podemos construir uma visualização sobre a sua posição no mapa, que mais tarde podem ser utilizados como por exemplo para definir as rotas de uma viagem. Ao explorarmos as posições no mapa, podemos também querer inverter o processo, uma vez tendo as coordenadas, queremos saber qual endereço correspondente, e este processo é designado por geocodificação reversa.

Geocodificação reversa como o nome indica, é o inverso do processo de geocodificação, este é o processo em que transforma as coordenadas geográficas em um endereço postal ou nome de um local. Este processo permite-nos a possibilidade de reconhecer onde estamos, e o que está a nossa volta. É de esperar que não é trivial para o ser humano conseguir interpretar as coordenadas geográficas, mas sim, um conjunto de referências ou endereços de designação. Quando estamos a usar um dispositivo ou aplicação baseado em serviços de localização, as coordenadas são obtidas pelo GPS, e este é transformado em um endereço de modo possa ser interpretável para os seus utilizadores [12].

De um ponto de vista mais abstrata, o processo de geocodificação reversa funciona da mesma forma que o processo de geocodificação, sendo os mesmos componentes, dados de entrada, dados de referência, algoritmos e dados de saída. Um exemplo de funcionamento do processo pode ser ilustrado na figura 2.1, tendo o objetivo de determinar o endereço postal onde se encontra localizado.

Para um dado input é inserido para o nosso dispositivo, os mesmos são comparados com um conjunto de dados de referência, e são determinados por métodos matemático de modo a obter um conjunto de dados mais próximos, que de seguida, são utilizados para a computação do resultado final. A etapa de localização é efetuada conforme o tipo dos dados, o resultante tanto pode ser um ponto mais próximo que esteja contido na base de dados ou um ponto estimulado pelo uso de método de interpolação linear, é possível

também recorrer a combinação entre os dois tipos de dados, aproveitando a natureza de cada um [13]. Após as fases das computações, um endereço postal é devolvido.

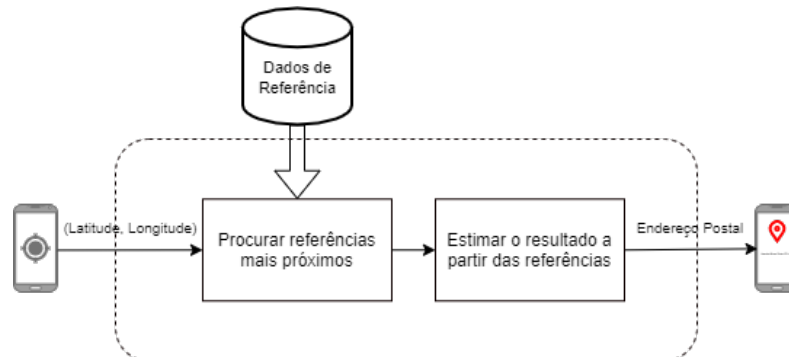


Figura 2.1: Um exemplo simples de geocodificação reversa

Para um melhor entendimento sobre o problema em causa, foram efetuados estudos sobre as técnicas e tecnologias existentes para atacar este tipo de problema. Neste capítulo serão apresentados os resultados das pesquisas e das leituras sobre Estado de Arte relacionado.

2.1 Endereços

A forma mais comum de expressar a nossa posição é através do uso dos endereços, este é construído por uma sequência de caracteres que é capaz de identificar uma determinada localização. Ao contrário do sistema de coordenadas de latitude e de longitude, este não apresenta uma forma universal, podem ser encontradas formatos diferentes em lugares diferentes.

Os endereços fornecem informações que permitem localizar fisicamente um edifício, e são frequentemente encontradas no processo de geocodificação. Um endereço por si é apenas uma expressão textual, podem ser interpretados pelos humanos caso possuam algum conhecimento sobre ele. Para poderem ser interpretados pelas máquinas, como por exemplo, uso de GPS, exige a necessidade de transformar-lho num objeto espacial, com coordenadas, de modo possa permitir efetuadas análises ou estudos mais sofisticados.

Os endereços podem ser **diretos** ou **indiretos**, onde os diretos são aqueles que apresentam uma descrição estruturada que nos permite expressar onde este encontra-se localizado como por exemplo o endereço postal, enquanto os indiretos se referem apenas um número identificativo do local.

Para os endereços diretos, estes ainda podem ser subdivididos em **absolutos** ou **relativos**, onde os absolutos identificam um local de acordo com uma referência fixa e os relativos identificam o posicionamento de um local em relação ao outro.

2.1.1 Endereços de acordo com CTT

De acordo com CTT, o endereço é considerado como um conjunto de informação precisas e completas em que permite o encaminhamento de correspondência e a sua entrega ao destinatário sem procura nem equívoco [3]. Este é constituído pela presença dos elementos como Lugar, Artéria, Local, Identificador de Porta, Identificador de Alojamento e Código Postal. Onde representam:

- **Lugar** - Um conjunto de edifícios contíguos ou próximos a que corresponde uma designação.
- **Artérias** - Os elementos em que caracterizam o espaço urbana, como Rua, Praça, Levada e etc.
- **Local** - Elemento toponímico complementar, identifica a localização geográfica de artérias ou partes de artérias.
- **Porta** - Um edifício com a finalidade de abrigar atividades humanas, podem conter um ou mais alojamentos residenciais ou funcionais.
- **Alojamento** - Instalações ocupadas num edifício por uma entidade para fins habitacionais ou funcionais.
- **Código Postal** - O código postal é um código desenvolvido pelo CTT com a intenção de facilitar o tratamento e da distribuição dos objetos para um endereço postal. O código postal também pode ser designado por CP7 (CP4 + CP3).

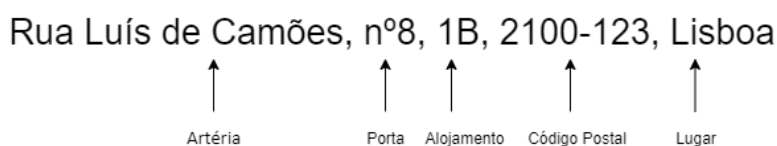


Figura 2.2: Um exemplo comum de endereço postal em Portugal

2.1.2 Abreviaturas no endereço

No preenchimento de endereços em formulários, muitas vezes são abreviadas palavras para agilizar o processo e este também ajuda na gestão do espaço do campo da grelha. Do ponto visto de geocodificação, as abreviaturas são consideradas um dos maiores problemas existentes, existe a necessidade de normalizar os endereços para que estes possam ser utilizados como dados do processo. Para um endereço português, os tipos e títulos de artérias são frequentemente abreviados, como por exemplo: R (Rua), AV (Avenida), PROF DR (Professor Doutor), D (Dom) e também para os tipos de porta e de alojamento como: LT (Lote), VV (Vivenda), ESQ (Esquerdo), DTO (Direito). Os CTT classificam as

abreviaturas em normalizadas e não normalizadas, existe uma lista de abreviaturas recomendadas pelo CTT e ainda um conjunto de regras devem ser respeitadas no uso das abreviaturas [3]:

1. O designativo de tipo de artéria pode ser abreviado;
2. As partículas (de, da, das, do, dos) podem ser suprimidas;
3. O título, cargo ou profissão no nome da artéria pode ser abreviado e, em caso de extrema necessidade, mesmo suprimido;
4. Como terminadores de abreviaturas não devem ser utilizados pontos. O separador de palavras deve ser o espaço;
5. Os nomes de artérias e localidades quando são constituídos por uma única palavra não devem ser abreviados;
6. Nos nomes de artérias e localidades quando são formados por três ou mais palavras não se deve abreviar nem a última, nem a primeira palavra (caso não se trate de um título). Das palavras intermédias, caso existam e quando necessário, pode ser inscrita apenas a letra inicial.

2.2 Dados na Geocodificação

No mundo de computação, os elementos básicos e essenciais são os dados, estes permitem-nos no efetuar análises, ajudar a definir estratégias e tomar decisões. E para o processo de geocodificação, tudo anda a volta dos dados, sem os dados, nada é concretizável, A geocodificação na realidade, não é mais do que encontrar e estabelecer uma relação entre os dados. Para este sistema, os dados podem ser classificados em três tipos: de entrada, de referência e de saída.

2.2.1 Dados de Entrada

Os dados de entrada são os dados do problema, consiste em textos que descreve a localização e estes serão transformados em dados espaciais no final do processo de geocodificação. Estes podem ser **absolutos** ou **relativos**. Onde os dados absolutos são aqueles que indicam uma localização absoluta, este corresponde à localização de um lugar em função de um sistema de referência, como endereço postal, por exemplo. Os dados absolutos por si já são capazes de produzir um resultado de localização geográfica, enquanto para os dados relativos, ainda não são suficientes. Os dados relativos são aqueles que indicam uma localização relativa, como por exemplo cruzamento da rua X com rua Y, estes indicam a posição de uma localização em relação a outro.

Os dados de entrada podem ser apresentados em diversas formas e formatos, muitas vezes, exige a necessidade de um tratamento previo para que possa ser iniciado o processo de geocodificação. Como o caso das abreviaturas, estes podem ser frequentemente encontrados nos endereços e que podem ser apresentados em formas diferentes, como por exemplo, apartamento pode ser abreviado em **ap** ou **apart**, perante tal situação, requer um tratamento em que possa uniformizar as abreviaturas. O importante aqui serão os dados encontrarem-se no mesmo formato de modo a que se possa prosseguir para os passos seguintes.

2.2.2 Dados de Referência

Os dados de referência são um conjunto de dados de localização previamente conhecidos e validados, esses dados geralmente são colecionados e inseridos manualmente e normalmente são dados integrados em SIG, além de conter a descrição textual sobre a localização e também da sua posição geográfica. Os dois tipos mais comuns são dados **vetoriais** e dados **rasterizados**, sendo que o anterior é o mais frequentemente usado para o processo de geocodificação devido a sua estrutura permitem ser facilmente manipulados, enquanto para os dados rasterizados, estes são baseados em pixels e são mais difíceis de trabalhar, apenas são usadas em contextos específicos como por exemplo correção ou confirmação de dados.

Para os dados vetoriais, estes ainda podem ser classificados em três tipos diferentes, podem ser **pontos**, **linhas** ou **polígonos** e são caracterizados por:

- **Ponto** – Um objeto representado por um único vértice, possui uma posição no espaço, mas não apresentam comprimento ou área. Estes representam o nível mais baixo de complexidade geográfica, a sua usabilidade em termo de potencial de interpolação é quase inexistente uma vez que estes não possuem atributos adequados para derivar novos objetos geográficos. Geralmente este descreve um endereço absoluto.
- **Linha** – Um objeto representado no mínimo dois vértices conectados e que possui um comprimento, também pode conter outros atributos que são explorados no processo de geocodificação, como por exemplo a direção. Estes podem ser linhas retas ou curvas, e apresentam alto potencial de interpolação. Geralmente este descreve uma rua.
- **Polígono** – Um objeto representado no mínimo três vértices conectados e que forma uma figura fechada com lados, este possui uma área. Pode representar tanto como casas ou até mesmo cidades e países.

Existe uma relação direta entre a qualidade de dados de referência e o resultado da geocodificação, é extremamente importante a validação desses dados, pois esta tem efeitos diretos no resultado da geocodificação. A ideia da geocodificação é na verdade,

estabelecer uma relação entre os dados de entrada e os dados de referência, encontrar uma associação entre os dois tipos de dados. Este é um dos componentes que decide o sucesso e o insucesso do processo de geocodificação. Como descrito nos dados de entrada, este também exige a necessidade de tratamentos prévios para que possam ser utilizados.

2.2.3 Dados de Saída

Os dados de saída são os resultados obtidos através dos dados de entrada com a execução dos processos de geocodificação. Consiste em dados espaciais de um dado texto descritivo não espacial, geralmente representadas por coordenadas de latitude e de longitude. Este representa a relação entre os dados de entrada e os dados de referência. Existe a possibilidade de não produzir resultados ou resultados errados, quando não existir ou não for possível encontrar uma associação entre dados de entrada e de referência.

2.3 Algoritmos de Geocodificação

Quando estamos a falar da geocodificação, não existe apenas um algoritmo, mas sim um conjunto de algoritmos. Cada etapa do processo de geocodificação utiliza algoritmos diferentes, e este é constituído por três etapas. Sendo a etapa inicial, a fase de definição, onde são estabelecidos um conjunto de regras para o sistema, de modo a que possa dar o sentido das palavras das descrições textuais. A etapa intermediária, a fase de correspondência, onde encontra e estabelece uma correspondência entre os dados. E a etapa final, a fase de localização, onde encontra o posicionamento geográfico dos dados de entrada.

2.3.1 Componente de atribuição

Os dados textuais para as máquinas são apenas um conjunto de caracteres que não possui nenhum significado, para podermos usar os dados, precisamos de definir um conjunto de regras em que possa ser interpretado pela máquina e de tal modo consiga representar ou valorizar mais do que conjunto de caracteres.

Perante o primeiro passo de geocodificação, na receção de um endereço, a máquina não consegue arranjar uma associação imediata, só mesmo para os casos em que esteja escrito de forma igual nos conteúdos de base de dados de referência, mas isso poucas vezes acontece. Precisamos de encontrar e de definir o que representa o conteúdo da informação textual. Sendo o caso, o termo normalização seria utilizado e aplicada não só para dados de entrada, mas também os dados de referência..

Antes de iniciar o processo, devíamos estabelecer um modelo geral em que possa indicar o que representa no conteúdo do endereço, isso é, encontrar um padrão para os endereços. Este modelo pode ser muito diferente de país para o país, visto que cada país possui o seu próprio formato na organização dos conteúdos presentes no endereço. Sendo o caso de endereço português, geralmente é apresentada um modelo como ilustrado na figura 2.3 que difere num modelo para o endereço americano figura 2.4.

<Tipo de Artéria> <Nome da Artéria> <Número de Polícia> <Código Postal>
<Município>

Figura 2.3: Modelo em geral de um endereço português

<Número de Polícia> <Nome da Artéria> <Município> <Abreviatura do Estado>
<Código Postal>

Figura 2.4: Modelo em geral de um endereço americano

Após a definição do modelo, efetuamos o *parsing* do endereço, isto é, dividir o texto do endereço em vários átomos, de modo a que consiga corresponder ao nosso modelo estabelecido previamente. Um exemplo de *parsing* pode ser encontrado na figura 2.5. E é nesta fase, que existe a possibilidade de perda de dados, isso acontece devido ao estarmos a efetuar a normalização dos dados, alguns dados seriam descartados, no exemplo figura 2.5, o endereço inicial, “Rua do Canto do Muro”, após o processo de normalização, a preposição “do” foi descartada, ficando apenas com Rua, e Canto do Muro.

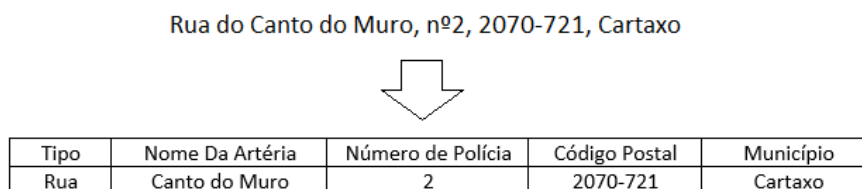


Figura 2.5: Exemplo do processo de atribuição

2.3.2 Componente de correspondência

Após a organização dos dados, entra em ação a fase de correspondência, é nesta fase em que se escolhe um conjunto de dados de referência para que possa ser utilizado para continuar o processo de geocodificação. Requer estabelecimento de um conjunto de regras de como os dados são comparados. Na fase anterior, estabelecemos um conjunto de atributos, e para esta, comparamos as informações que são guardados nos atributos. Há muitas escolhas de implementação para esta fase, sendo que podemos decidir como são comparados os caracteres ou valores numéricos. E ainda podemos classificar os algoritmos de correspondência em duas categorias principais: **determinísticos** ou **probabilísticos**.

Os métodos determinísticos são aqueles em que se baseiam em um conjunto de regras e que são processados numa sequência específica, as regras geralmente são operações binárias que indicam se existe uma correspondência para um determinado atributo. A

vantagem principal de usar este tipo de método é que estes são fáceis de implementar. Deve-se ter cuidado na definição de regras, visto que quando a regra é demasiado restritiva, pode originar resultados falsos negativos. Como por exemplo, a regra de correspondência mais simples possível segundo o autor [6]: “Todos os atributos de um endereço da entrada corresponder aos mesmos atributos de um dado de referência.”

Como o resultado desta regra, ou ocorre uma correspondência perfeita, ou não existe uma correspondência, a razão deve-se a esta regra ser muito restritiva, na maioria dos casos em que não se consegue corresponder um dado de referência mesmo que na verdade ele existe. Um exemplo facilmente observável é quando existe falta de informação num dos atributos de endereço da entrada e que para o mesmo atributo encontra-se a informação disponível em dado de referência, perante esta situação a regra falha, uma vez que não se consegue corresponder para aquele atributo.

Os métodos probabilísticos são aqueles que utilizam um esquema computacional para determinar a verossimilhança ou a probabilidade que um atributo consegue corresponder, e classificando para cada elemento analisado. Este provém da teoria de probabilidade e de decisão e são utilizados no processo de geocodificação desde início [6]. A implementação deste envolve processos matemáticos e são mais complicados de implementar.

Antes de considerar o método probabilístico é de relembrar que a probabilidade de um dado acontecimento A , $P(A)$, pode ser calculado pela lei de Laplace 2.1, e que este ainda pode ser calculado pela sua inversa, onde $P(A)$ é equivalente a $(1 - \text{a probabilidade de não ocorrer } A)$, i.e. $P(A) = 1 - P(\neg A)$.

$$P(X) = \frac{\text{Nmero de casos favorveis}}{\text{Nmero de casos possveis}} \quad (2.1)$$

E ainda a relembrar que a probabilidade condicionada, que é a probabilidade de um dado evento ocorre sabendo mais informação sobre esse facto, sendo a probabilidade do acontecimento B sabendo que tinha ocorrido A é dada pela equação 2.2

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \quad (2.2)$$

Na correspondência probabilística, as probabilidades da correspondência indicam o grau de confiança em que um atributo tem a correspondência baseado em um determinado critério de avaliação, onde esta toma valores entre 0 e 1, de onde 0 representa que não existe a probabilidade de que estes corresponderem, enquanto 1 representa 100 % de probabilidade. Devemos também definir um limite de confiança para decidir como são classificados os resultados obtidos, sendo os valores acima do limite são aceites e rejeitamos quando estes são abaixo do limite estabelecido. Para uso desta metodologia, ainda devemos aceitar um nível de risco em que os resultados podem ser errados.

Existem muitas formas de proceder a correspondência probabilística, uma das formas mais comuns é chamado **Peso dos Atributos**, segundo [6], este método consiste em determinar uma pontuação composta através da soma ou subtração do valor que esta associado em cada atributo e decidir com este se existe ou não a correspondência.

Para cada atributo de um dado endereço, são associados dois valores para o peso, onde estes representam o nível da importância do atributo e que são resultados pela combinação entre a probabilidade de correspondência e a probabilidade de não correspondência. A probabilidade de correspondência é a probabilidade de dois atributos corresponderem, m , dado que dois registros correspondem, M . Denotado matematicamente pela probabilidade condicionada $P(m | M)$, este pode ser calculado pelas estáticas sobre uma pequena amostra do conjunto de dados no qual um dado de entrada e um dado de referência realmente correspondem. Também podemos calcular este tendo em conta as taxas de erro, α , onde este representa dois atributos não corresponderem mesmo sabendo que os dois registros correspondem, ou seja, $P(m | M) = 1 - \alpha$. Na literatura existente, as notações da probabilidade geralmente são descartadas e $P(m | M)$ é apenas apresentado por m .

A probabilidade de não correspondência é a probabilidade de dois atributos corresponderem, m , dado que dois registros não correspondem. Denotado matematicamente pela probabilidade condicionada $P(m | \neg M)$, este pode ser calculado pela estatística sobre uma pequena amostra do conjunto de dados no qual o dado de entrada e o dado de referência não coincidem. E na literatura, este é apresentado por u .

A partir das duas probabilidades podem ser calculados os índices de frequência para concordância, f_a , e para discordância, f_d , 2.3 e estes são usados para computar o peso positivo e o peso negativo do atributo 2.4.

$$f_a = \frac{m}{u}, f_d = \frac{1 - m}{1 - u} \quad (2.3)$$

$$w_a = \log_2(f_a), w_d = \log_2(f_d) \quad (2.4)$$

Os pesos para cada atributo de dados de referência são calculados a priori. Durante a execução da correspondência, apenas é computado a pontuação composta do dado de entrada, são comparados os atributos de dado de entrada e de dado de referência, quando há uma correspondência, é somado o valor de w_a , caso contrário, é subtraído o valor de w_d . O resultado final é comparado com o limite de confiança estabelecido, para poder concluir se estes correspondem ou não.

Segundo o autor [11], existe três problemas principais para esta fase e devem ser tratados, sendo o caso de existência de abreviatura, o caso da existência de erros de ortográficos no endereço e o caso em que existe várias correspondências, o mesmo autor também apresenta uma solução para cada um deles.

Para o caso de abreviatura, este pode ser resolvido com um pré-processamento dos dados, estabelecendo uma uniformidade no sistema. Substituindo as ocorrências de abreviatura com uma lista das abreviaturas comuns. Como por exemplo, a substituição de abreviatura “Av.” por Avenida.

Para o caso de erros ortográficos no endereço, a resolução deve-se à escolha de algoritmo de comparação. Na comparação do texto, podem ser comparações exatas, ou comparações aproximadas. Com os algoritmos de comparação de texto aproximada, os

textos podem ser considerados equivalentes mesmo que não sejam escritos na mesma forma.

Para o caso da existência de várias correspondências, usamos uma informação adicional, de modo a remover alguma das correspondências, como por exemplo, para o caso de existir duas ruas com nomes iguais, podemos considerar o atributo código postal para remover a informação errada.

2.3.3 Componente de localização

Depois da escolha de dados de recurso para a correspondência, entra em questão de onde se posicionam exatamente os dados de entrada. Dependente do tipo de dados que são usados como referência, diferentes técnicas são escolhidas e aplicadas.

Quando são dados baseados em pontos, cada um possui o par de coordenadas onde representa a localização geográfica de um endereço. Para um dado endereço, se este está contida na nossa base de referências, então é devolvida a sua posição geográfica. Caso contrário, é georreferenciado no endereço com o número de polícia mais próximo. Este pode obter resultados mais precisos, mas também é a técnica com a taxa de correspondência mais baixa, pois requer uma base de dados extremamente completa para que consiga obter resultados com efeitos desejados. Este é o tipo de dados menos usado para geocodificação, visto que exige um elevado custo para obtenção dos dados.

Os dados também podem ser apresentados na forma linear, quando se trata de dados representados com dois vértices, como por exemplo, o ponto inicial e final de uma rua. Para este caso, utiliza-se o método de interpolação linear para a decisão da localização final.

O método interpolação linear trabalha com a base das linhas de segmentos utilizando as suas propriedades de geometria espacial determinando a posição de um dado ponto. Este método requer o conhecimento sobre dado de referência, a geometria, o intervalo de número de portas válidos e a sua posição geográfica da rua (linha de segmento). Para um dado endereço de entrada, assumindo que ele conseguiu encontrar uma correspondência na fase anterior, é utilizado o seu atributo o número de porta, e é calculada a proporção da distância em função do comprimento total e o intervalo de número de portas válidos da rua referenciada. Segundo o autor [6], para uma rua simples e contínua, caracterizado por dois vértices, a distância entre o início da rua e a posição do resultado, d , é calculado como uma proporção do comprimento total da rua, l , o número de endereço de entrada, a , e o tamanho de intervalo de número de portas, r , onde este último é equivalente a metade do valor entre o número da porta do início, r_s , e do fim da rua, r_e (uma rua normalmente tem dois lados). 2.5, 2.6

$$r = \frac{|r_s - r_e|}{2} \quad (2.5)$$

$$d = l \frac{a}{r} \quad (2.6)$$

E ainda usando a distância calculada, combinando com as coordenadas do vértice inicial e do vértice final da rua, é possível determinar a posição exata de onde este é localizado. O autor[6] refere que o uso de cálculo esférica é o caso mais comum, e este é a escolha mais precisa, como as equações 2.7 uma vez que para curtas distância como os segmentos da rua mais comuns, o erro introduzido é insignificante.

$$x_a = d(x_e - x_s), y_a = d(y_e - y_s) \quad (2.7)$$

O resultado final obtido é apenas uma aproximação, sendo como exemplo o endereço “Rua de Camões nº5”, este é estimado a posição da porta número 5 na rua “Rua de Camões”. Tendo em conta as propriedades de dado de referência, este provavelmente esta apontado para o meio da rua, ilustrado na figura 2.6. Visto que este não considera o lado da rua em que iria se posicionar, perante esta situação, podemos resolver com parâmetros adicionais, uma distância perpendicular entre a posição das casas e a linha central da rua [17]. De modo aproximar mais da posição real. Conforme o esquema da numeração da porta usada, diferentes decisões são tomadas, como por exemplo, em Portugal, o sistema mais utilizado é o esquema europeu, visualizando numa direção em que o número de porta aumenta, lado com números ímpares geralmente encontra-se do lado esquerdo. 2.7

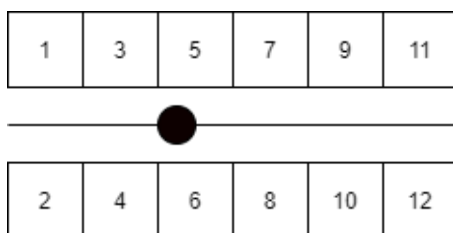


Figura 2.6: Geocodificação com interpolação linear

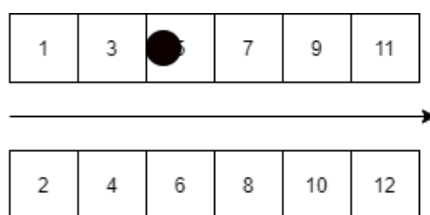


Figura 2.7: Geocodificação com interpolação linear e deslocamento

Estas abordagem apenas são aplicadas em casos mais simples, uma vez este assume que uma rua seja uma reta, completamente preenchida, e que as casas estão distribuídas uniformemente e ainda que tenham todos a mesma dimensão, mas muitas vezes, isso não é verdade. As casas nem sempre se posicionam logo no começo e no fim de uma rua, e também é frequentemente encontrada numa rua em que existe uma intervenção de uma outra rua, formando com ele uma esquina, e deixando um “buraco” entre as casas. 2.8 Para cada situação, exige dados adicionais ou técnicas mais sofisticadas para poder

obter maior precisão. Para o caso das dimensões, se tivermos o conhecimento sobre as dimensões das casas, podemos fazer uma soma de distância para determinar a posição geográfica. E para o problema das esquinas, seria ainda necessário ter o conhecimento sobre a localização da esquina, para poder dividir a linha da rua, em vários segmentos, estimulando assim, a posição do dado de entrada, com a soma de dimensão das casas.

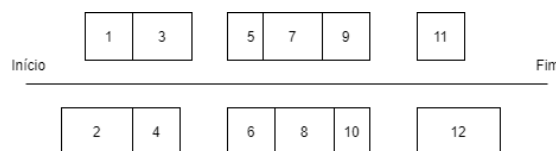


Figura 2.8: Situação real de posicionamento das casas numa rua

O outro caso possível é quando se tratam de dados poligonais, este forma um polígono com vários vértices, geralmente descreve uma estrutura dum construção ou de uma cidade, perante tal situação, existem duas possibilidades de escolha como resultado, uma seria considerar o centroide do polígono ou então, escolhe-se um ponto aleatório na área interior do mesmo. Segundo o [11], seria preferível o uso do ponto aleatório, pois este evita a sobreposição dos resultados, e também como não conseguimos determinar com uma melhor precisão, qualquer ponto dentro da área resultante tem a mesma probabilidade de ser a verdadeira correspondência.

Os três tipos de dados referidos podem coexistir ao mesmo tempo, sempre possível, a preferência entre as três, seria o de baseado em pontos, pois este apresenta a maior taxa de precisão, e de seguida o de linear, só quando não existir a possibilidade de usar essas duas técnicas, prossegue-se com o caso de georreferenciação com dados poligonais, este apenas nos dá uma ideia sobre onde está localizado, mas não o que realmente se encontra localizado.

2.4 Indicador de Certeza de Geocodificação

Nem todo o resultado obtido através do processo de geocodificação é considerado válido, apesar de não conseguirmos determinar se os resultados são realmente verdadeiros, podemos criar um critério para descartar alguns deles. Para o tal efeito pretendido, utiliza-se o método de indicador de certeza de geocodificação (GCI - *Geocoding Certainty Indicator*), este é o método proposto por [11], tem como objetivo de determinar a confiança que temos sobre o resultado de um dado input.

GCI pode ser usado como um filtro, permitindo que o utilizador descartar os resultados obtidos abaixo de um certo limite de confiança estabelecido. Pode ser utilizado como parâmetro de ponderação, para efeitos de análises e estatísticas, ele determina o nível de importância de um determinado ponto. E também como um critério de avaliação, ajuda nos a identificar os resultados mais prováveis.

O valor de GCI varia entre 0 e 1, onde 0 representa completamente desconfiados que o resultado seja válido, e 1 representa que estamos completamente confiantes. Este valor é calculado a partir de produto entre três componentes que podemos observar na figura 2.9, PCI (*Parsing Certainty Indicator*), MCI (*Matching Certainty Indicator*) e LCI (*Locating Certainty Indicator*), e para cada um deles, toma valores entre 0 e 1.

$$\text{GCI} = \text{PCI} \times \text{MCI} \times \text{LCI}$$

Figura 2.9: Fórmula para calcular o GCI

2.4.1 Indicador de certeza de análise

O PCI é um indicador que indica o grau da integridade dos componentes de um dado endereço. Durante a fase de atribuição, fazemos classificação do conteúdo de um endereço, de modo que seja distribuído pelo modelo previamente criado. Dependente do sistema de endereçamento local, algumas atribuídas são mais importantes para um sistema enquanto que a outra não exige a necessidade. No caso em Portugal, existem muitas ruas e vilas com o mesmo nome, talvez seria útil a presença do código postal para diferenciar entre elas, enquanto para o México, um endereço comum contém a presença do município, a presença de código postal para este caso é mais irrelevante [11]. Variando do sistema para o sistema, os componentes podem tomar valores diferentes para o peso, e o valor de PCI é definido em função desta variante.

2.4.2 Indicador de certeza de correspondência

MCI é um indicador que determina a qualidade de correspondência. Durante a fase de correspondência, vários atributos são comparados, para os mesmos, atribuindo um peso de relevância a cada um deles. Sendo a classificação deste indicador que é dada pela soma dos pesos dos atributos, apenas consideramos os quais que tenham uma correspondência. E o autor[11] também indica que no uso das comparações de texto aproximada, devemos notificar o utilizador sobre a presença de diferença de texto durante a comparação, diminuindo o valor de MCI por um fator de 0.05. Esta penalização é aplicada em cada atributo.

2.4.3 Indicador de certeza de Localização

LCI é um indicador que determina a confiança que temos sobre a determinação das coordenadas geográficas. Durante a fase de localização, três tipos de técnicas são utilizadas para diferentes tipos de dados, e para a cada técnica possui o seu próprio critério de avaliação. No caso em que os dados são pontos, quando se trata do input pertencer no conjunto de dados de referência, o valor de LCI é 1.

Para o linear, este é determinado pela equação ilustrada na figura 2.10, onde max e o min corresponde a porta com a numeração maior e menor do mesmo lado (par ou ímpar) da artéria correspondente.

$$LCI = 1 - \frac{2}{(\max - \min)}$$

Figura 2.10: Fórmula para calcular o LCI com dados lineares

Para o polígono, quando é escolhido um ponto aleatório para o resultado, este é determinado pela equação ilustrada na figura 2.11, onde RA representa a área do polígono onde o ponto é escolhido, e RM, a área da vila ou município.

$$LCI = 1 - \frac{RA}{MA}$$

Figura 2.11: Fórmula para calcular o LCI com dados poligonais

2.5 APIs de Geocodificação

O uso de geocodificação traz diversas vantagens e benefícios, e existe uma grande tendência de procura no mercado. Atualmente existem muitas empresas que fornecem serviços de geocodificação, seja ele livres, pagos ou pagos parcialmente, cada um possui a sua estratégia de determinação, e seria possível de observar diferentes resultados entre eles. Os serviços de geocodificação são fáceis de utilizar, conseguimos obter resultados imediatos sem nos preocupar como este é calculado, nem a preocupação de obtenção e atualização da base de dados de referência. Mas por outro lado, perdemos o controle do processo de geocodificação, este não permite definir os nossos critérios de avaliação nas etapas da computação, e também desconhecemos a qualidade dos resultados obtidos. Nesta seção, iria ser mencionados alguns APIs atualmente existentes e ainda, apresentada uma pequena análise sobre eles. Os APIs foram estudados são:

1. *Google Geocoding API* [8]
2. *Bing Maps locations API* [1]
3. *MapQuest Geocoding API* [14]
4. *HERE Geocoding API* [10]

Diferentes APIs produzem resultados diferentes, uma vez em que cada um tem as suas próprias base de referências e as suas estratégias de determinação, para um melhor

entendimento sobre eles, foram examinadas as suas documentações e também a realização de alguns testes. Tanto como *Google geocoding API* e os outros restantes possuem os seus termos e limitações de uso, e podem ser gratuitos até certas condições.

Para os quatro APIs estudados, nas suas respostas, além das coordenadas geográficas, também são incluídas informações sobre cada componente de endereço em causa, bem como o endereço na forma normalizada de acordo com eles. Quando ocorre uma situação de indecisão, todos eles devolvem uma lista de respostas por uma ordem decrescente de probabilidade de ser hipótese correta.

Todos eles ainda descrevem na sua resposta, como este foi determinado. Apesar do resultado da posição serem semelhantes, ainda existe uma diferença do algoritmo usado por cada um deles. No caso de *Google* 2.12, os tipos de localização podem ser *Rooftop* (exata), *Interpolação linear*, *Ponto centro da geometria* e *Aproximada*, mas este não descreve a informação sobre a correspondência do endereço com a sua base de referência. Enquanto para o *Bing* [1], tem mais dois campos adicionais, *Confiança* e *Código de correspondência*, que caracteriza a qualidade do resultado na fase de correspondência. Em que a *Confiança* descreve no caso de *Here* 2.15, possui menos estratégias de localização, mas são mostradas as pontuações obtidas para cada atributo do endereço, e para *MapQuest* 2.15, apenas descrevem a qualidade, e não é mostrado como este foi localizado.



Figura 2.12: Tipos de determinações para resultado do *Google*

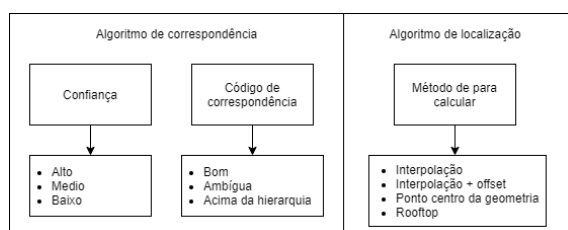


Figura 2.13: Tipos de determinações para resultado do *Bing*

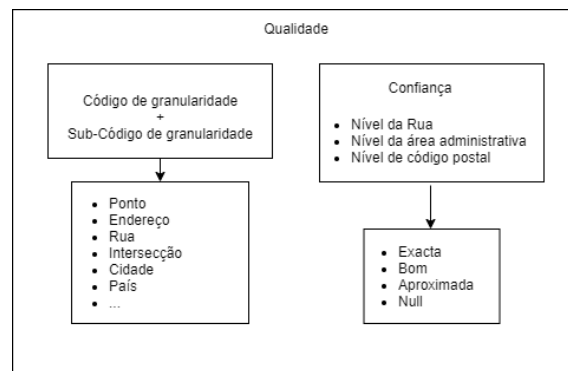


Figura 2.14: Tipos de determinações para resultado do *MapQuest*

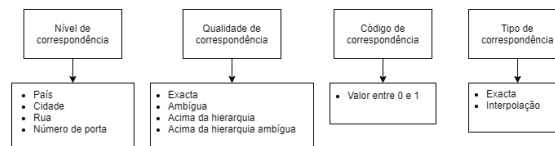


Figura 2.15: Tipos de determinações para resultado do *Here*

2.6 Conclusão

Com base de conhecimento descrito neste capítulo, podemos concluir que o processo de Geocodificação é composto por um conjunto de etapas, onde envolve diversos tipos de dados e algoritmos. Sendo importante a validação de cada elemento composto, uma vez em que estes afetará diretamente a qualidade dos resultados.

ABORDAGEM

Este capítulo irá apresentar ideia sobre como vai ser resolvido o problema em causa, identificando os requisitos e discutindo a modelação do sistema.

Relembrando os objetivos descritos no capítulo 1, os principais objetivos neste trabalho seria desenvolver uma metodologia que estima as coordenadas dos endereços de uma forma eficiente tendo em atenção a precisão dos resultados.

Tendo como base de conhecimento sobre a geolocalização dos endereços, os dados fornecidos pelos CTTs, à partida, assumimos que estes são válidos e confiáveis. Desta forma, decidimos como uma primeira solução, estudar o comportamento dos serviços existentes no mercado. Iremos demonstrar, neste capítulo, os resultados obtidos pela execução dos APIs do Google, Bing, MapQuest e Here, dos mesmos foram referidos com mais detalhe no capítulo anteriores.

Depois de uma pequena discussão sobre os resultados, ainda neste capítulo, iremos apresentar a estrutura geral da nossa própria solução, e uma explicação breve sobre cada um dos componentes.

3.1 Dados

Todo o processo de geocodificação, anda à volta dos dados. A validade dos dados é extremamente importante, uma vez em que estes irão afetar diretamente os resultados. No projeto presente, os dados são fornecidos pela empresa CTT. À partida, assume-se que são válidos e confiáveis.

Numa etapa inicial deste projeto, pretende-se o desenvolvimento da metodologia de geocodificação para uma região específica, de modo a que, mais tarde, esta possa ser aplicada para todo o país. A geocodificação é extremamente dependente dos dados do problema, é importante a região escolhida apresentar diversas características de modo a que possa extrair uma solução que consiga generalizar para todos os casos. Tendo em conta a este critério, os CTT propuseram para esta primeira fase, a cidade do Cartaxo, onde este apresenta características fortes e relevantes para o processo de geocodificação, contendo uma parte mais urbana e uma parte rural.

Cartaxo, é uma cidade portuguesa do distrito de Santarém, é a sede de um município com 158,17 km² e de acordo com PORDATA [16], este contava 23 843 habitantes até 31 de Dezembro de 2019. Esta cidade é ainda subdividida em 6 freguesias: Cartaxo e Vale da Pinta, Ereira e Lapa, Pontével, Valada, Vale da Pedra, Vila Chã de Ourique (ver figura 3.1).

Do ponto vista da carta geográfica, entre todo o conselho, 11 892 ha da sua área está reservado para cultura agrícola, 2 475 ha para área florestal e 1 461 ha para área social ou inculta [5]. Uma das maiores fontes de economia da cidade deve se à produção do vinho, este também é considerado um dos maiores a nível do nacional, sendo que grande parte da área agrícola estão reservados para vinha [5].



Figura 3.1: Cartaxo

Fonte: Wikipédia

3.1.1 Dados do Cartaxo

Para o contexto do problema numa fase inicial, os dados que serão trabalhados, essencialmente, são disponibilizados pelos CTTs. Sobre tudo estes dados são extremamente confiáveis e não exigem a necessidade de serem pré-processados, uma vez em que os mesmos já se encontram normalizados segundo as regras dos CTTs. E para uma fase mais avançada, o sistema trabalhado irá integrar os dados que provém de uma fonte exterior como por exemplo os dados disponibilizados por OpenStreetMap [15], de modo a manter os dados de referência mais completo possível.

Toda informação disponibilizada pelos CTTs, relevante para a resolução do problema, está essencialmente dispersa em três ficheiros, sendo eles: um ficheiro que possui informações sobre os endereços do Cartaxo, um ficheiro sobre os segmentos que descreve as

artérias do Cartaxo e um ficheiro que descreve as rotas dos Carteiros no Cartaxo. Sendo os dois primeiros ficheiros será a nossa fonte principal para construção da base de dados de referência.

Para os dados sobre os endereços do Cartaxo, temos informação sobre a sua descrição e as suas coordenadas geográficas, ainda possui alguns atributos de identificação criados pelos CTTs. Estes dados encontram-se na forma vetorial com a implantação pontual. Para os dados fornecidos, um total de 13 830 portas registadas e ainda entre eles, 2 887 portas sem a identificação. Os registos estão distribuídos entre 482 CP7s distintos, podemos ver a sua distribuição na figura 3.2. Numa primeira vista ao análise de dados, facilmente observamos dois tipos de distribuição, sendo uma delas encontra-se na freguesia da Valada, e que as portas se encontram distribuído de uma forma mais uniforme e linear, enquanto para a freguesia da Pontével, as portas encontram-se distribuídos numa forma mais aleatório. Também podemos observar que na freguesia de Vila Chã de Ourique apresentam ambos das características.

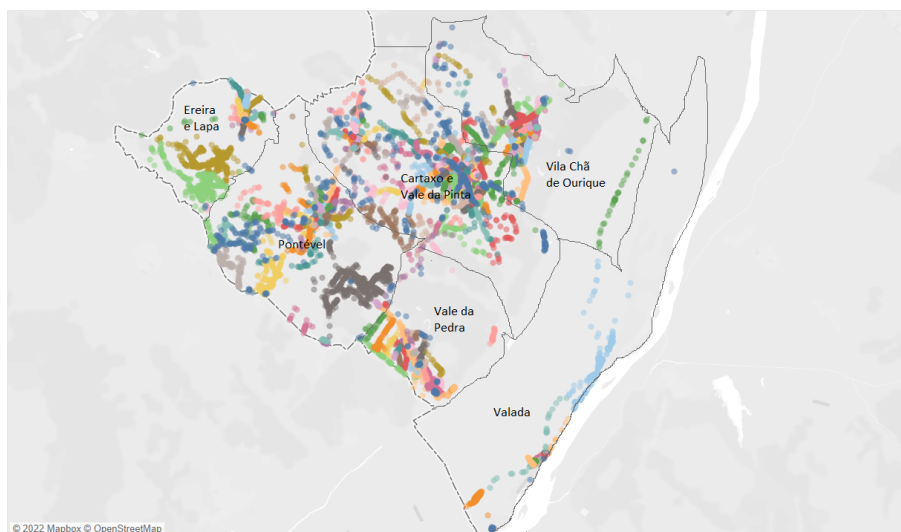


Figura 3.2: Distribuição no Cartaxo por CP7

Para os dados sobre os segmentos que descrevem as artérias do Cartaxo, cada segmento é denotado por uma identificação interna dos CTTs e corresponde uma e só uma artéria, mas nem todas as artérias são identificadas. Este encontra-se em forma vetorial com a implantação linear. Para os dados fornecidos, um total de 2 326 registos, distribuído em 732 Artérias diferentes, ilustrado na figura 3.3.

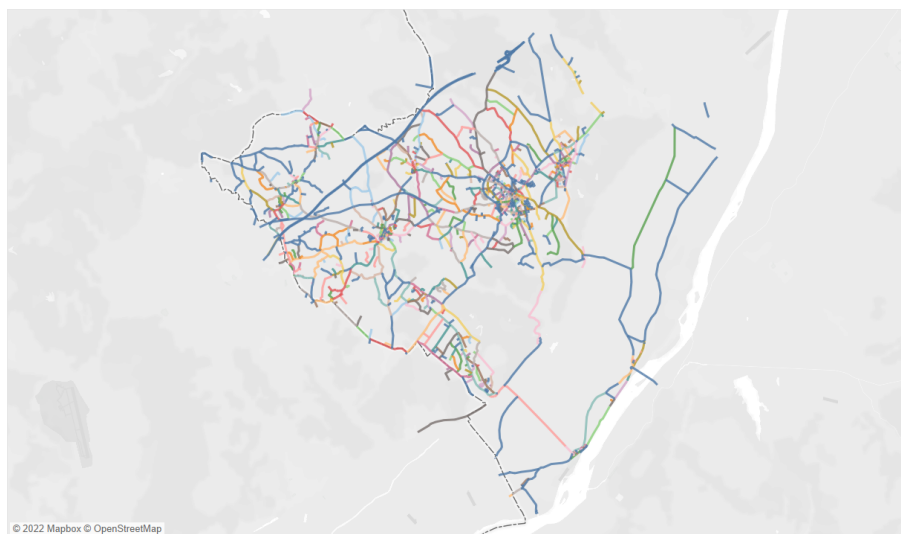


Figura 3.3: Segmentos que descrevem as ruas no Cartaxo

E para o ficheiro sobre as rotas dos carteiros, temos a designação dos endereços, mas não temos a sua informação geográfica. Temos ainda informação sobre a rota a que ele pertence e a sua ordem de entrega, a chamar atenção aqui que a ordem não está otimizada, ou seja, esta ordem pode não ser respeitada pelos carteiros. O ficheiro possui 14 450 endereços e estes distribuem-se em um total de 21 rotas diferentes. Os registos possuem o mesmo atribuído de identificação como no ficheiro dos endereços do cartaxo, e podemos associá-las diretamente de modo a obter algumas coordenadas. Houve uma intersecção de 9 487 endereços entre os dois ficheiros, e restou um total 4 963 endereços com coordenadas desconhecidas. Sendo o nosso objetivo da solução inicial será focado em determinar estes endereços que não foram associados.

Apesar de os dados não serem completos, ainda podemos agrupar-lhos por rota, e conectando entre os eles seguindo a sua ordem de entrega, (saltando pontos em falta), conseguimos obter uma visualização sobre as rotas. Podemos observar na figura 3.4.

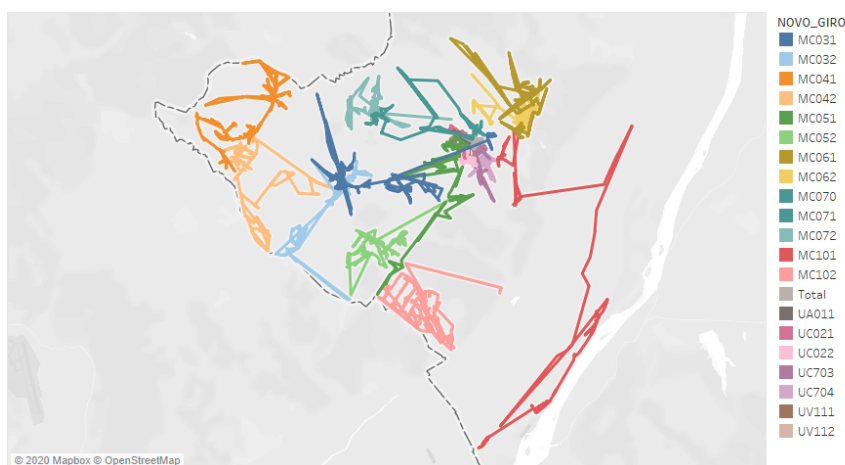


Figura 3.4: Visualização gráfica dos rotas dos carteiros parcialmente

3.1.2 Normalização dos Dados

Para um dado endereço inserido na máquina, ela apenas reconhece como uma *String*, numa fase a priori, devemos definir o significado de cada componente de modo a server para as fases seguintes. Segundo os CTTs, uma porta pode ser decomposto em tipo, número e sufixo. Podemos observar na tabela 3.1, para a porta "5", o seu tipo é subentendido, que é considerado como uma porta numérica, e que não possui sufixos. Ao contrário da porta "Lote 35 B", que é uma porta do tipo Lote, a sua numeração é 35 e tem como sufixo a letra B. É importante ter esta classificação visto que para os processos seguinte, iremos recorrer às portas como dados de referência, é importante escolhermos as portas corretas para a validade dos resultados. A mesma ideia aplica-se também para as artérias.

Tabela 3.1: Classificação dos componentes sobre a porta de um endereço

Porta	Tipo	Número	Preposição	Sufixo
5		5		
116B		116		B
Vivenda Família Sousa	Vivenda			Família Sousa
Lote 35B	Lote	35		B
Casal do João	Casal		do	João

Apesar dos dados serem fornecidos pelos CTTs, a representação dos endereços não se encontra uniforme, podemos observar na figura 3.5 em que foi verificado nos dados, para a mesma designação, este pode estar ou não em forma abreviado. De modo a garantir consistência dos endereços, decidimos abreviar tudo quando possível seguindo a lista de abreviaturas recomendadas pelos CTTs mencionados na secção 2.1.2. Também é observável na figura 3.5, para uma mesma designação, a utilização indiscriminada de maiúsculas, minúsculas e/ou com acentuação.

Qta do Sonho	VV Vilela
Qta Laranjeiras Família G..	VV VINTÉM
Qta Santa Helena	VV Vintém
Qta Vida Nova	VV Violeta
Quinta Abeguaría	VV Zézé
Quinta Adélias	Vv. H. Silva
Quinta Africano	
Quinta Água	VV Sérgio
	VV Sergio

Figura 3.5: Endereços com a mesma designação mas diferentes apresentações

3.2 Estudo dos Apis

De modo a avaliar a qualidade dos resultados obtidos pelas API, foram extraídos ao acaso um total de 100 endereços nos dados fornecidos pelos CTTs. A figura 3.6 demonstra os resultados obtidos comparados com a distância em relação a posição real utilizando a fórmula de haversine. Podemos observar que o *API da Google* possui resultados ligeiramente melhores. A maioria dos resultados encontra-se numa distância aceitável, i.e. se considerarmos o resultado dentro de um raio de 50m de distancia comparado com o alvo. Mesmo assim, existe em média de 30 casos para cada API, cujo resultado se encontra a mais de 200 metros da posição real, e ainda, uma média de 15 casos em que os resultados se encontram a uma distância superior a 1 km. A razão principal de tal acontecimento deve-se ao facto do problema da correspondência, uma vez que este pode fazer um mapeamento com informação desatualizada ou mesmo erros nos dados.

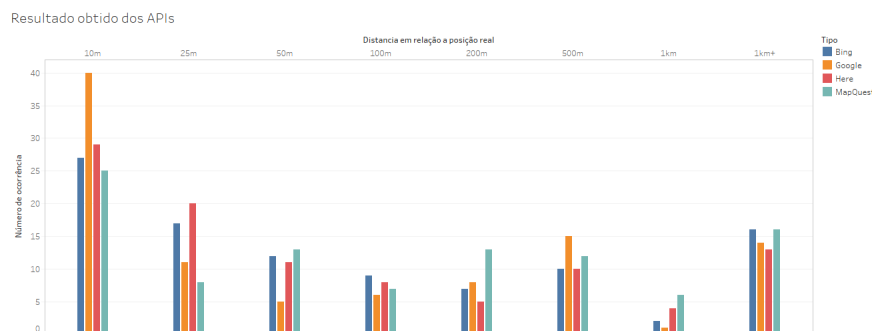


Figura 3.6: Resultados obtidos dos APIs após a execução de 100 endereços

Para um estudo mais profundo sobre os comportamentos das API, ainda foram realizados testes com um total de 1000 endereços escolhidos aleatoriamente. Podemos observar na figura 3.7, apesar da existencia de situações em que não se conseguiu produzir resultados após a inserção do endereço, o aspeto geral sobre os comportamentos das API não diferiu muito comparado com os testes realizados apenas com 100 endereços.



Figura 3.7: Resultados obtidos dos APIs após a execução de 1000 endereços

Pelo motivo de este ser o tipo de serviço atualmente mais popular [4] e de facto observou-se que este produziu resultados ligeiramente melhores comparado com os outros, decidiu-se efetuar um estudo mais profundo sobre a API Google[8]. Como foi referido anteriormente, quatro tipos de determinação da posição são utilizados por esta

API, para os 100 testes realizados. Podemos observar na figura 3.8, 63% dos casos são do tipo Rooftop, 21% para Interpolação linear, 8% para Ponto centro da geometria e para Aproximada, como nesta bateria de teste, para todos os endereços foram encontradas respostas, não foi possível ilustrar caso em que não se encontrou resultado.

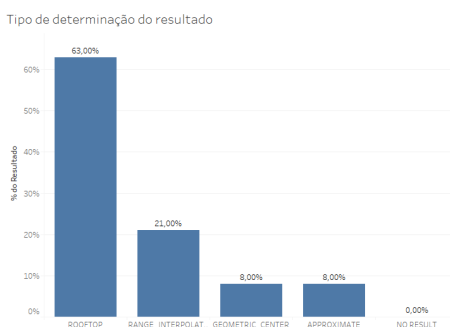


Figura 3.8: Tipo de determinação do resultado

Para o caso em que o resultado seja do tipo Rooftop, mais de 50% está a menos de 10 metros e que ainda existem 19.05% dos casos em que se encontra uma distância superior a 200 metros. A razão deve-se ao problema de encontrar a correspondência, visto que estas são posições exatas. Enquanto que os restantes tipos, podemos ver na figura 3.10, a maioria dos resultados não correspondem a posição real, uma vez que, os métodos de interpolação linear e aproximação são extremamente dependentes das características da localização em causa e para os pontos centros da geometria, a dimensão da geometria em causa.

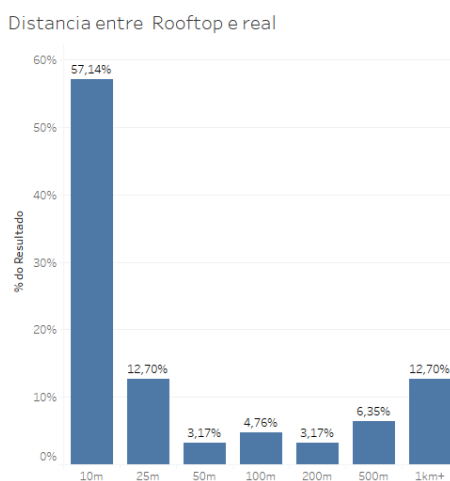


Figura 3.9: Distância entre o resultado do tipo Rooftop e real

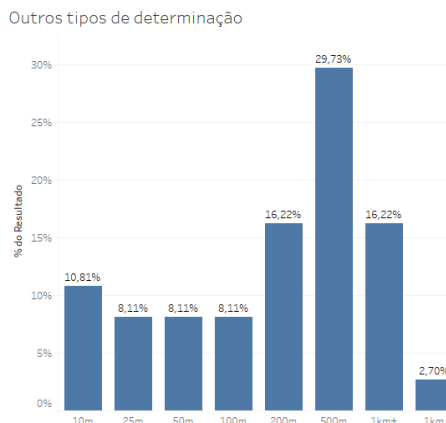


Figura 3.10: Distância entre o resultado de outros tipos e real

Com base da análise e dos resultados obtidos, podemos concluir que a exatidão dos resultados é aumentada se for a nível de Rooftop, onde possui os melhores resultados, mas mesmo assim, o que é produzido nem sempre é válido e verdadeiro, e que para que os outros casos de determinação, devemos colocar em dúvida de confiança sobre a qualidade do resultado.

3.3 Conclusão

Observando as análises efetuadas no ponto 3.2, para além das limitações dos API, nem todos os resultados se podem aproveitar. Muitas vezes o resultado leva-nos a um local completamente diferente, uma vez que desconhecemos as suas bases de dados e algoritmos usados para a produção do produto. O ideal seria construir o nosso próprio sistema de geocodificação, de modo a garantirmos o maior controlo possível e tendo sempre em consideração a qualidade dos resultados, visto que para o contexto do problema, é extremamente importante o posicionamento correto das portas a fim de otimizar as rotas dos carteiros.

3.4 Problemas e Considerações

Levando em consideração os objetivos previamente estabelecidos e as informações obtidos através dos estudos realizados no estado da arte, e das análises efetuadas, alguns problemas foram encontrados e explorados nos dados disponibilizados. Sobretudo, a deteção dos problemas existentes afeta diretamente o design da proposta de solução. Um dos problemas frequentemente encontrado foi a ordem da numeração das portas não regulares numa dada rua, como podemos observar na figura 3.11. Isto exige a necessidade de identificar os casos possíveis. Nem sempre acontece a verificação da paridade entre os lados da rua, sendo ele apenas números pares de um lado e ímpares de outro como habitualmente. É importante determinarmos se as portas tem ou não as portas distribuídas

desta forma, visto que devemos escolher corretamente os candidatos de endereços com a numeração da porta mais próximos para que possa efetuar um procedimento de interpolação apenas no segmento que descreve adequadamente. Outra grande categoria de irregularidade sobre a numeração da porta foi a existência de representação da porta por um valor não numérico, geralmente, por um nome de designação, e também da existência de portas sem identificação. Para esta situação, não seria possível efetuar a interpolação apenas com o conhecimento sobre o segmento e o posicionamento sobre os endereços vizinhos, requer alguma informação adicional para a continuação do procedimento.



Figura 3.11: Exemplos de portas com numeração não regular

Outro problema frequente foi a falta de informação sobre os dados de referência. Detetaram-se algumas ruas sem o segmento descritivo, mas existindo endereços com coordenadas registadas na mesma rua, ilustrado na figura 3.12. Para o procedimento de geocodificação, é extremamente importante possuímos os dados mais completo possível, devemos sempre ter a possibilidade de receber as novas fontes de dados para enriquecer a nossa base de referência.



Figura 3.12: Rua com segmento incompleto

3.5 Estrutura Geral da Solução

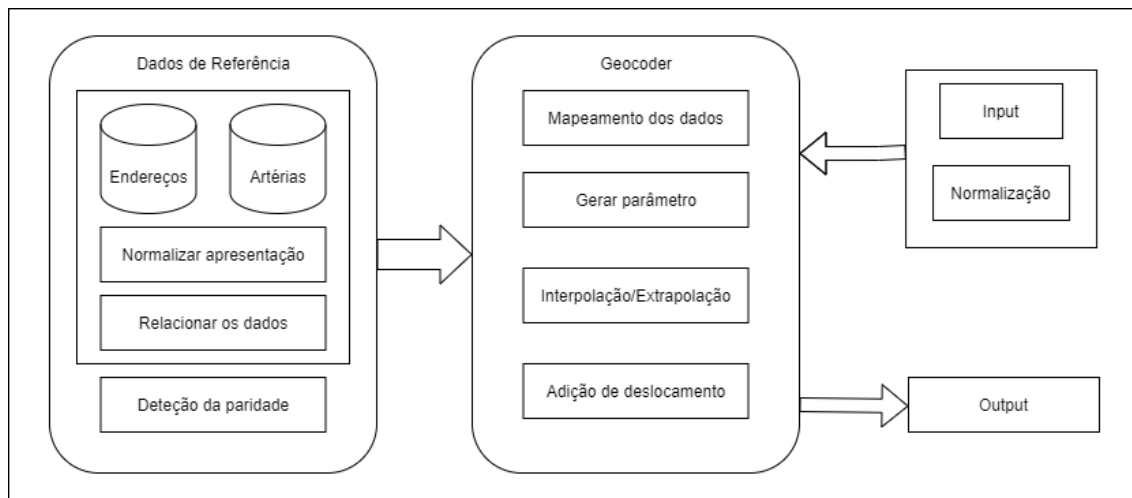


Figura 3.13: Esquema para solução proposta

Tendo em conta os problemas referidos, foi desenhado o esquema do trabalho para a resolução do problema, ilustrado na figura 3.13. É de notar que para o esquema funcionar corretamente, deve-se considerar que os dados de referência sejam sempre confiáveis, de preferência de valores coletados e por inserção manual, e mais uma vez, a chamar atenção pelo facto da qualidade dos dados de referência está diretamente proporcionada com a qualidade dos resultados. Sendo a base de dados, os dados fornecidos pelos CTTs e/ou utilizar outras fontes confiáveis de modo completar o máximo possível. E também de observar que nem todos endereços teriam a possibilidade de serem transformado com apenas o processo automático, a tendência e o desafio seria tornar o processo mais automaticamente possível, evitando a necessidade da intervenção humana.

De acordo com o contexto do problema, os dados de entrada e as bases de dados são disponibilizados pelos CTTs, mas mesmo assim, a representação dos endereços não segue o mesmo padrão, exige a necessidade de um tratamento prévia para garantir a possibilidade de se relacionar. De um ponto vista de eficiência, ainda nesta fase de pré-processamento, nos dados de referência, optamos por relacionar entre eles, de modo obter a posição do endereço projetado na faixa da artéria para preservar a geometria na fase de interpolação/extrapolação, uma vez os mesmos candidatos de referência são utilizados N vezes, este processo repete-se N vezes. Pela mesma razão, a deteção da paridade também ocorre nesta fase.

Após os dados serem pré-processados, o procedimento de geocodificação inicia-se com o mapeamento dos dados de referência, no caso em que não encontrar nenhuma referência utilizável, o programa deixa os dados de input numa lista e estes terão de ser resolvidos por intervenção humana. Caso contrário, dependendo dos resultados obtidos pela fase da correspondência, diferentes parâmetros são gerados pelo sistema. Se estamos a falar

de uma rua com a noção de paridade nos lados, a forma de calcular a sua localização iria ser diferente do que para uma rua sem paridade, ou mesmo a forma de determinar um endereço cujo o número da porta está entre as outras portas é diferente do que determinar uma porta que esteja na extremidade da rua. Os parâmetros aqui consistem em quais são os candidatos de referência e as suas características, e com eles, decide como o processo vai prosseguir.

Depois do sistema inserir os parâmetros, o resultado de uma interpolação ou extrapolação é devolvido, este resultado indica o ponto estimado se encontra o endereço projetado na faixa da artéria, para obtenção da posição real, projetamos o ponto para fora da faixa numa direção perpendicular e a uma distancia de deslocamento estimado.

No capítulo seguinte será apresentado com mais detalhe a forma de implementação de cada uma das componenetes, a descrição sobre a normalização dos dados, o pré-processamento nos dados de referência, e os conjuntos algoritmos utilizados para o processo de geocodificação.

IMPLEMENTAÇÃO

Este capítulo descreve, numa perspetiva de implementação, todos os componentes referidos no capítulo anterior. Até agora tudo foi apresentado de uma forma abstrata, com a informação presente neste capítulo, o leitor irá conseguir compreender melhor a solução proposta.

4.1 Fase de pré-processamento

Antes de implementarmos o funcionamento do sistema de geocodificação, vamos preparar os dados de modo a que sistema seja eficaz durante a sua execução. Os dados que estamos aqui a falar são os dados fornecidos pelos CTTs que descrevem os endereços e as artérias do Cartaxo (Referido na secção 3.1.1). Estes são a nossa fonte principal de bases de dados de referência.

4.1.1 Normalização dos dados

De modo a garantirmos a uniformidade do sistema, sendo o primeiro passo, vamos normalizar toda a sua forma da apresentação. Um endereço por si é apenas uma expressão textual, de modo a possibilitar a interpretação com a máquina, dividimos a descrição do endereço em vários componentes. O importante aqui seria a máquina reconhecer o significado de cada um deles, como por exemplo, distinguir qual o número de polícia num dado endereço. Para os dados fornecidos, um endereço já se distingue em componente de artéria, nome de porta, código postal e localidade. Mas, mesmo assim, ainda temos de entrar em mais detalhe no ponto vista de geocodificação do nosso sistema. É importante reconhecermos no componente nome de porta, o que representa o seu tipo, número e sufixo. Dividimos a sua descrição em vários componentes separados com o “espaço”, e começamos por remover as proposições existentes, visto que estas não têm um valor significativo para o nosso contexto. Para informação restante, abreviamos, tanto quanto possível, usando uma lista de constantes pré-definidas no sistema. Esta descreve as abreviaturas recomendadas pelos CTTs (2.1.2). Quando a informação é mapeada com sucesso, o componente é substituído pela sua abreviatura. Após a abreviação, recorreremos a uma

outra lista de constantes para classificar o seu tipo, conhecer qual o seu tipo de porta ou artéria, como por exemplo, reconhecer se o endereço em causa é uma avenida ou uma rua, se é uma porta numérico ou é uma porta do tipo Lote. Esta segunda lista é constituída pela observação sobre os tipos mais frequentes encontrados nos dados. Ambas as listas têm a possibilidade de serem atualizados. Podemos observar um exemplo de classificação de componentes na figura 4.1.

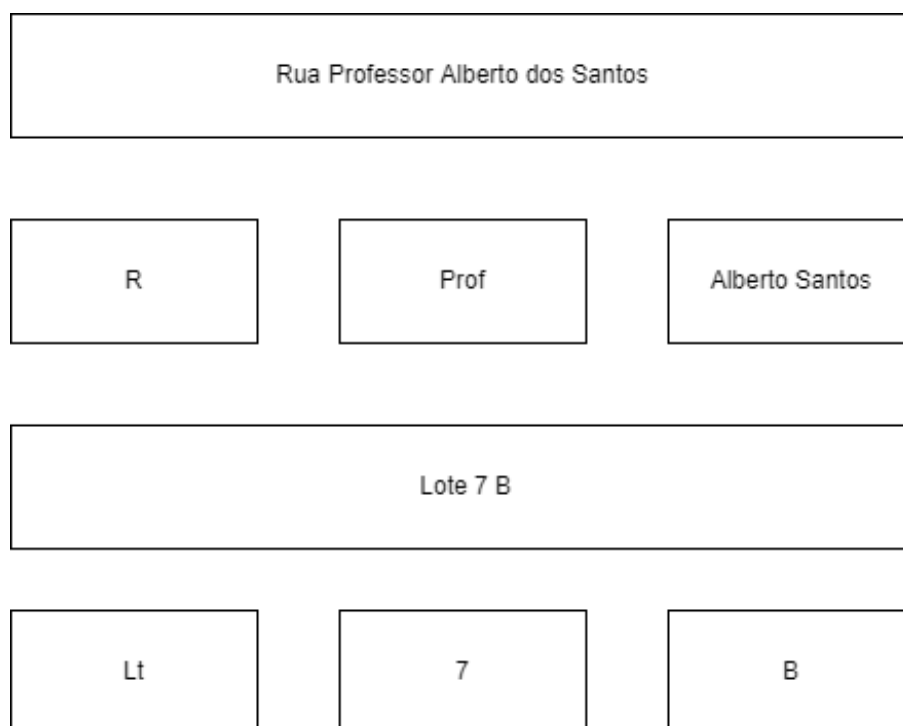


Figura 4.1: Exemplo de classificação de componentes

4.1.2 Projeção dos Dados

Relembrando a ideia do processo da geocodificação, iniciamos com a procura dos dados de referência, após obtenção dos mesmos, prosseguimos para a fase da localização. Mas então se apenas utilizarmos endereços como referência, iremos estar a perder a geometria formada pelas casas na zona correspondente, e se utilizarmos apenas as artérias, iremos perder a precisão e os detalhes oferecidos pelos endereços. A solução aqui seria misturarmos os dois tipos de dados, projetando os pontos dos endereços nos segmentos das artérias. Recordar o que foi referido na secção 2.2.2, os endereços correspondem a dados em forma de ponto e as artérias são dados em forma de linha. A nossa estratégia de implementação aqui foi converter os pontos de endereço em vértices inseridos na artéria sem estragar a sua forma geométrica, isto é, projetamos o ponto do endereço numa direção perpendicular a linha da artéria, preservando o deslocamento do ponto à faixa da artéria e a direção em relação ao segmento (se este se encontra à esquerda ou à direita da artéria), uma vez que estes são atributos fundamentais para recuperar o ponto original. Também

é importante mantermos os vértices originais e a ordem da construção dos segmentos (a ordem em que a linha é desenhada), de modo a manter a estrutura dos segmentos originais.

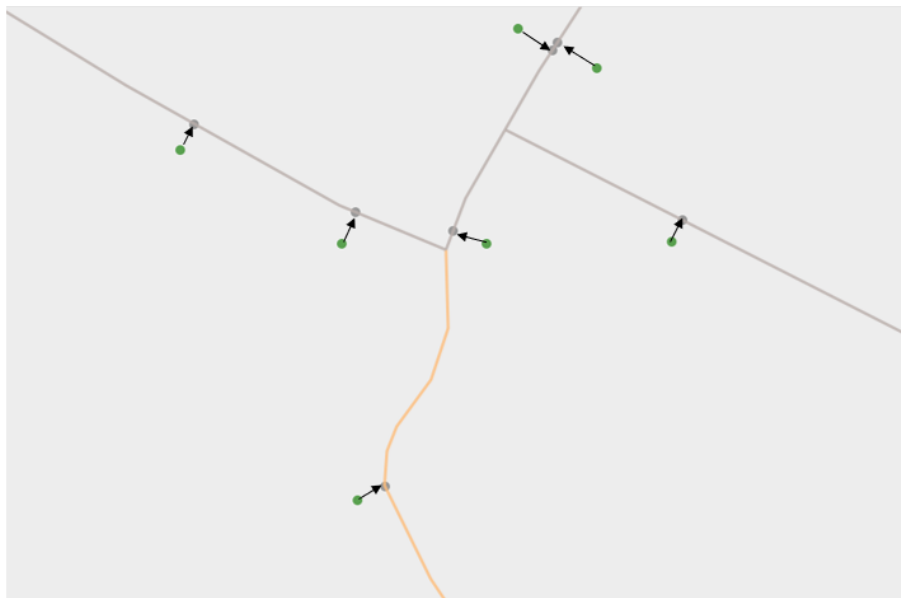


Figura 4.2: Exemplo de projeção dos endereços na artéria

4.1.3 Detecção da paridade

No mundo real, a numeração das portas não se distribuem da mesma forma em todas as artérias. É frequente encontrar numa artéria que tenha portas com numeração par de um lado e ímpar do outro, mas nem sempre se aplica. Esta noção de paridade é importante visto que afeta a distribuição geográfica das portas, observável na figura 4.3, em que temos aqui um exemplo de uma artéria sem paridade, e facilmente concluímos que a sua distribuição é diferente do que uma artéria com paridade.

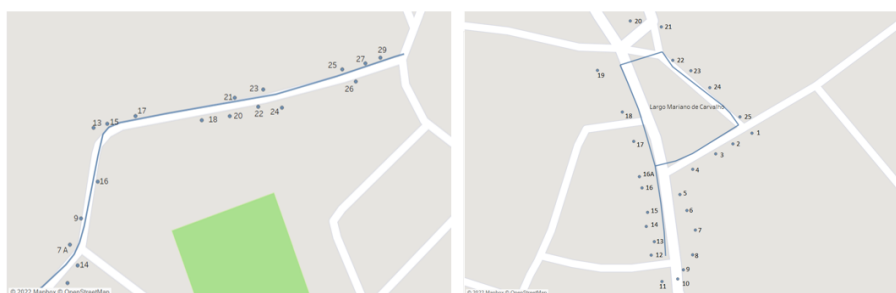


Figura 4.3: Exemplos de artérias com/sem a noção de paridade

Ao longo da dissertação, recorrendo os dados fornecidos, foram estudadas diferentes técnicas para determinação desta noção. E aqui vamos deixar a explicação do algoritmo com que se obteve o melhor resultado.

Numa artéria com paridade, sabemos que as portas estão divididas pelo eixo da artéria, com pares dum lado, ímpares do outro e, na fase da projeção dos dados, temos o conhecimento sobre a sua posição em relação ao eixo da artéria e o deslocamento. A nossa estratégia aqui é utilizar o deslocamento considerando um lado com valor de deslocamento positivo e outro lado negativo. Podemos observar na figura 4.4, onde cada ponto corresponde a um endereço já apresentados com o número da porta e o valor de deslocamento (já com o fator da paridade). A partir do princípio que as portas com a mesma paridade estão isolados pelo eixo da via, o sinal da soma de todos os valores do deslocamento com numeração par vai ser diferente da soma para os ímpares (se for o caso de existir a noção da paridade na artéria em causa).

Pegando no exemplo da figura 4.4, vamos obter um valor de -4 ao somar as portas pares e um valor de 23.4 para as portas ímpares. Como os valores obtidos têm sinais diferentes, classificamos que a artéria em causa tem a noção da paridade. No entanto, este algoritmo é apenas aplicado quando existem portas numéricas na artéria, caso contrario, não conseguimos determinar a noção da paridade.

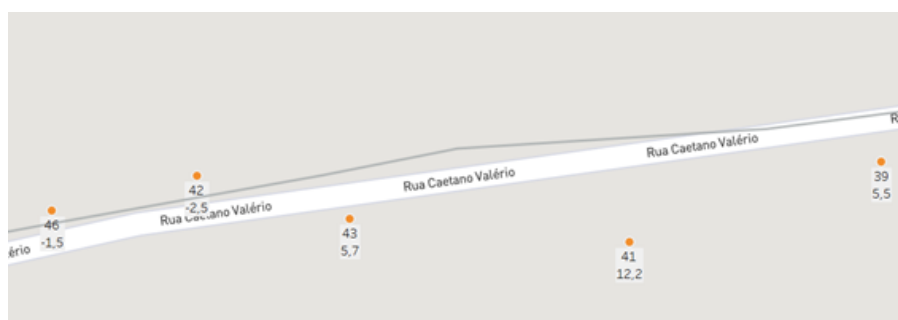


Figura 4.4: Exemplo de determinação de paridade de uma artéria

4.2 Fase da localização

Depois de organizarmos os dados na fase pré-processamento, entra agora em funcionamento o nosso sistema de geocodificação, tendo como objetivo, converter uma descrição textual em coordenadas geograficas. A técnica implementada será descrita nas próximas subsecções.

4.2.1 Escolha da referência

O processo de geocodificação inicia-se com o mapeamento dos dados de referência, a ideia aqui seria encontrar os dados mais detalhados e especificos possíveis, começando por encontrar todos os pontos que pertencem à artéria envolvida e, de seguida, determinar quais são os pontos de referências mais convenientes para determinar a localização de um dado input. Chamamos a estes pontos, **Candidatos vizinhos**, e ainda para estes candidatos, distinguimo-los como minorante e majorante. Por exemplo, na figura 4.5,

para a localizar a porta 26, iremos considerar como minorante a porta 24 e majorante a porta 28. O número de candidatos pode variar, uma vez, consoante estamos perante um caso de interpolação (2 candidatos) ou extrapolação (1 candidato). Quando nenhum candidato é encontrado, o sistema não consegue produzir resultados devido a falta de dados de referência.



Figura 4.5: Exemplo de Candidatos vizinhos para a porta número 26

4.2.2 Estimação das coordenadas

Uma vez tendo a informação sobre os dados de referência para o nosso alvo, definimos os mesmos como parâmetros para iniciar o processo da localização. Dependendo do conteúdo dos nossos parâmetros, prosseguimos para **interpolação** ou **extrapolação**.

Quando estamos perante a existência dum minorante e dum majorante, pelo senso comum, o nosso alvo deverá estar entre eles. Assim sendo, efetuamos uma interpolação ao longo da artéria associada de modo a não perder a sua geometria. Começamos por estimar quantas casas possíveis poderão existir entre os candidatos vizinhos (baseado em numeração real ou ordem alfabética). Entre minorante/majorante e o nosso alvo, o número de portas existentes é dado pela diferença entre as suas numerações. No casos da existencia da noção da paridade, o valor passa a ser metade. Mas muitas vezes, as portas numéricas, possuem sufixos, como por exemplo porta 7b, com sufixo letra b. Se estimarmos quantas portas existem entre porta 7 e porta 7b, apenas considerando a sua numeração, iríamos perder casos no meio, visto que, pelo senso comum, provavelmente existe uma porta 7a entre eles. Para o nosso sistema, tentamos também estimar o intervalo de portas determinados pela numeração conhecida diferenciada pelos sufixos. Este processo considera o maior sufixo conhecido como sendo o último, isto é, tendo a porta 7b apenas conseguimos estimar a existência da porta 7a, mas não conseguimos dizer algo sobre existencia da porta 7c. Uma vez tendo o conhecimento sobre o intervalo das portas possíveis, determinamos a posição relativa entre o minorante e majorante.

Na fase de pré-processamento, projetamos os endereços na linha que define artéria, conseguimos conhecer as suas posições relativas em relação a artéria, e agora nesta fase, utilizamos as posições relativas dos vizinhos de referência, para determinar a posição do nosso alvo projetada na linha. Sendo a posição relativa na artéria é dado pelo intervalo $[0, 1]$, com 0 início da artéria e 1 o final. A posição relativa do nosso alvo é calculada pela fórmula 4.1. Dependente do sentido da artéria, uma vez que o minorante pode estar posicionado no final da artéria e não no lado inicial, a posição relativa (PR) é dada pela soma/subtração entre a posição relativa do minorante (PRm) e a proporção necessária de minorante até ao nosso alvo (pmI). Onde a proporção é determinado pela fórmula 4.2, que é o produto de numero de intervalos entre as portas existentes de minorante até ao nosso Input (imI) com a proporção estimado para cada intervalo, sendo esta proporção é dada pela quociente da distancia entre posição relativa do Majorante (PRM) e do minorante (PRm) com o total de numero de intervalos entre as portas existentes de minorante até ao Majorante (TimM). .

$$PR = PRm \pm pmI \quad (4.1)$$

$$pmI = imI \cdot \frac{|PRM - PRm|}{TimM} \quad (4.2)$$

Podemos observar um exemplo na figura 4.6, onde cada ponto corresponde a posição projetada na linha da artéria, e é denotado por número da porta e a sua posição relativa (o valor dentro de parênteses). Temos aqui uma artéria com a noção da paridade e tem o mesmo sentido de ordenação da numeração da porta. Sendo o nosso alvo a porta 5, e os valores de posição relativa para minorante (porta 1) e majorante (porta 7) são 0.0222 e 0.4271. O total de intervalos entre as portas de minorante até majorante é 3, e de minorante para o nosso alvo é 2, substituindo na fórmula, iremos obter como resultado 0,2921, onde este se encontra um pouco afrente da porta 4.

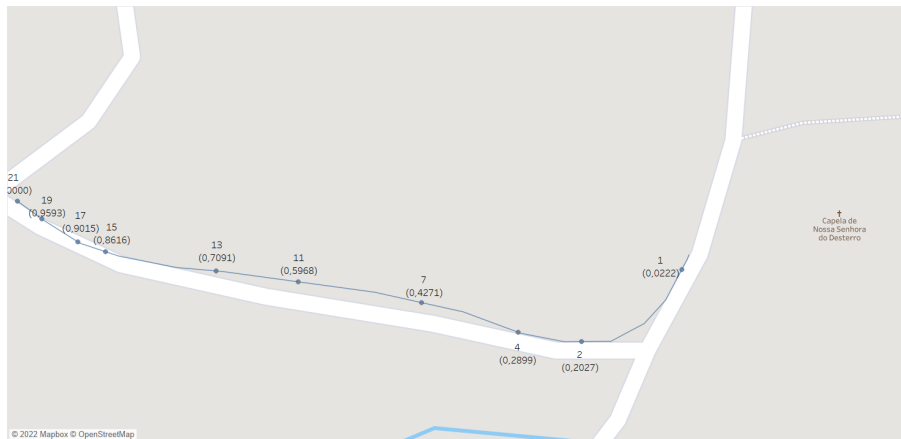


Figura 4.6: Exemplo de endereços com a sua posição relativa

Outro progresso é seguir com uma extrapolação. Perante este caso, apenas temos um candidato vizinho, tanto podendo ser minorante como majorante. Diferente no processo de interpolação, a proporção é calculada pela fórmula 4.3, onde é utilizada a distância relativa média entre duas casas até ao N vizinhos existentes nos dados, sendo o N um valor escolhido pelo utilizador. Podemos ver um exemplo na figura 4.7, tendo nos dados de referência a posição relativa (valor dentro de parênteses) e a distancia entre o vizinho (valor dentro de parênteses retos, e este é computado na fase de pré-processamento). Queremos determinar a porta 22, com N igual a 2. O candidato para referência neste caso seria a porta 20, e o sistema vai somar a distância entre a porta 16 e 18, a distância entre porta 18 e 20, e no final, divide por N para obter a média. Sendo a media obtida toma o valor de 0.0471. Devido ao sentido da artéria (contra a ordenação da numeração de portas), o resultado vai ser dado pela subtração entre a posição relativa da porta 20 e a proporção obtida, que toma o valor 0.1775. No caso em que a posição resultante não esta contida na artéria (o valor não pertencer ao intervalo [0,1]), este é colocado nos extremos da artéria (ser valor 0 ou ser valor 1).

$$pmIe = imI \cdot \frac{d1 + d2 + d... + dN}{N} \quad (4.3)$$



Figura 4.7: Exemplo de dados para extrapolação

Depois de efetuar uma interpolação ou extrapolação, o resultado encontra-se na faixa da artéria, ou seja, o resultado obtido é um ponto projetado na linha da artéria. Para o contexto desta dissertação, a posição projetada teria mais utilidade para a empresa visto que o objetivo do projeto é completar toda informação geográfica possível para poder definir as rotas para os carteiros e que as rotas seriam caminhar pelas artérias. Mas de qualquer forma, na nossa solução, também estimamos a posição não projetada. A partir dos candidatos vizinhos, extraímos os seus valores do deslocamento e o lado em relação

ao eixo da artéria, desfazemos a projeção com a média do valor dos deslocamentos, numa direção perpendicular para o mesmo lado. Como ilustrado na figura 4.8.

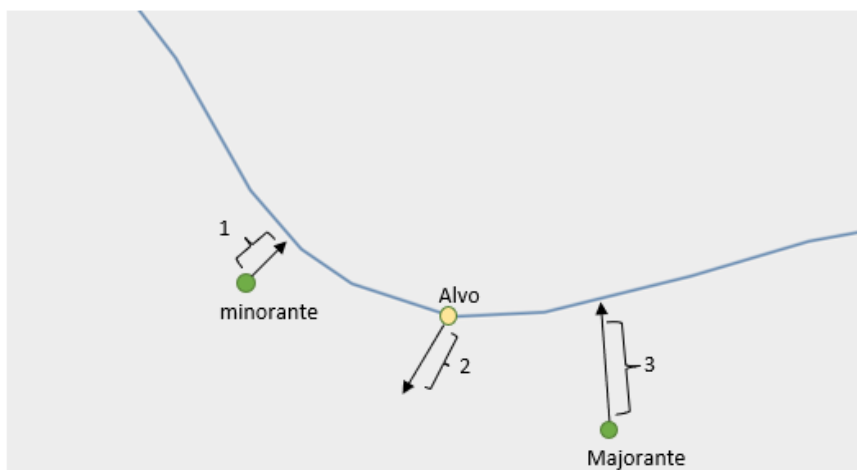


Figura 4.8: Reprojeção com o deslocamento

4.3 Melhoria do Algoritmo

Apesar de já conseguirmos produzir alguns resultados com o que foi apresentado nas secções anteriores, ainda existem muitas limitações. Uma vez que o processo apenas funciona para as portas numéricas. Quanto aos outros tipos de porta, ficamos mais limitados. Como por exemplo, portas representadas por nomes. Estes não possuem numeração para efetuar interpolação ou extrapolação. Para podermos estimar estas portas, utilizamos a ordem presente no ficheiro das réguas dos carteiros, esta ordem nem sempre é respeitada, mas podemos ter uma ideia de onde se posiciona o endereço.

Outra limitação encontrada foi para os endereços do tipo Lote. Estes possuem o valor numerico para representação da porta, mas o posicionamento não é regular. Podemos observar na figura 4.9 em que a relação entre a numeração e o posicionamento das portas não é contínua. Pertante tal situação, o nosso sistema muitas vezes produz os resultados errados. Como se pode ver na figura 4.9, se o nosso alvo for a porta LT 7, o sistema iria escolher os candidatos vizinhos porta 6 e porta 9, efetuando uma interpolação entre eles, e o resultado é colocado ao pé da porta LT 11, mas se observarmos a imagem, a porta LT 7 provavelmente seria resultante pela extrapolação com majorante LT 6 ou extrapolação com minorante LT 9. Visto que, o agrupamento das portas se dividem em varias secções. Para lidarmos com a esta situação, nesta secção, vamos introduzir novos algoritmos para o sistema de modo garantir resultados melhores.

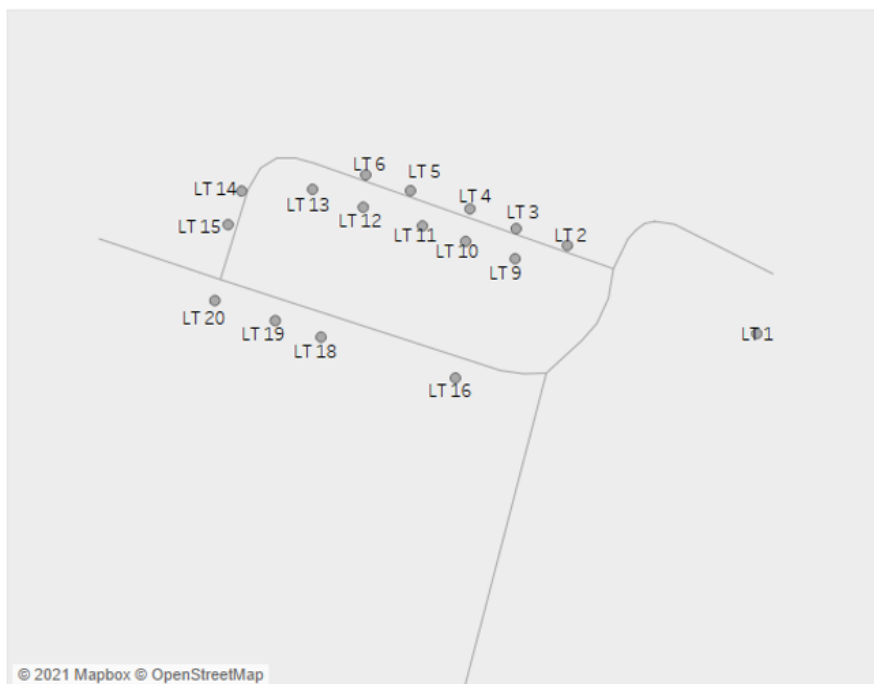


Figura 4.9: Exemplo de portas do tipo Lote

Podemos conhecer as secções das artérias a partir dos dados disponibilizados, mas para os endereços, este processo não é assim tão trivial. A ideia aqui seria construir um grafo de modo permitir conhecer a geometria formada pelas portas. Sendo os nós as extremidades dos segmentos das artérias ou as portas projetadas, e os arcos a ligação entre nós consecutivos. Podemos observar um exemplo na figura 4.10.

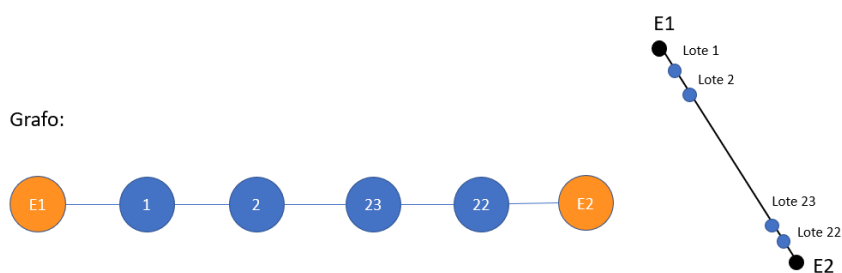


Figura 4.10: Exemplo de grafo

Depois da construção dos grafos, vamos caminhar pelos nós, de modo extrairmos a geometria formada pelas portas. Introduzimos aqui um novo termo, **Troço**, uma sequência ordenada de endereços, projetados no eixo da via, aos quais corresponde um trajeto passando uma única vez em cada ponto, ou seja, a partir de um troço, conseguimos descrever geometria de um conjunto de portas que pertencem a mesma secção. Para o exemplo de

grafo apresentado 4.10, obtivemos o resultado como na figura 4.11. O processo inicia-se com a produção de uma lista de portas presentes na artéria e agrupa-las por ordem crescente. No exemplo mencionado, o ponto inicia-se na porta 1, e verifica na lista a porta que vem a seguir, que é a porta 2, existe um caminho, caminhamos para a porta 2. O próximo seria a porta 22, encontramos a porta 23 no caminho, o sistema regista e continua. Estamos agora na porta 22, vamos caminhar para a porta 23, o sistema notifica que já passamos por este nó, a cadeia é quebrado, e criamos um novo troço.

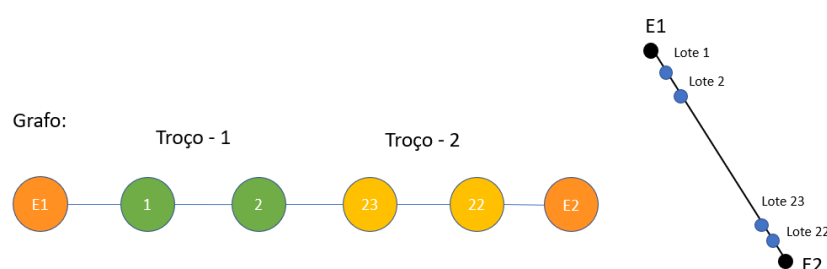


Figura 4.11: Exemplo de construção do troço

Após a criação dos troços, estes seriam integrados na nossa base de referência. Cada endereço é denotado por uma identificação, e este indica a qual dos troços pertence. Uma vez iniciado o processo de geocodificação, o sistema seleciona as referências vizinhas com base na numeração da porta e também considera os troços correspondentes. A ideia aqui é tentar encontrar a cadeia das portas mais conveniente para que a interpolação produza resultados que façam sentido, principalmente para as portas do tipo Lote, onde o agrupamento das portas da mesma rua se subdivide em várias secções.

Na fase de seleção de referência, podemos encontrar situações em que dois troços diferentes sejam escolhidos, isto acontece no caso em que os candidatos vizinhos pertencem a troços diferentes. Como se pode observar na figura 4.12, sendo o nosso alvo a porta 7, os candidatos vizinhos seriam a porta 6 e a porta 9, mas estes encontram-se em troços diferentes. Pelo senso comum, esta porta tanto pode ser colocada a esquerda da porta 6 ou mesmo a direita do 9. Dependendo do troço escolhido (azul ou laranja), diferentes resultados irão ser produzidos. Designamos por troço *up*, se este contiver as portas numéricas maiores do que o nosso alvo, e o troço *down* o caso contrário. Perante nesta situação, o sistema irá devolver os dois resultados, e o utilizador decide qual seria mais conveniente.

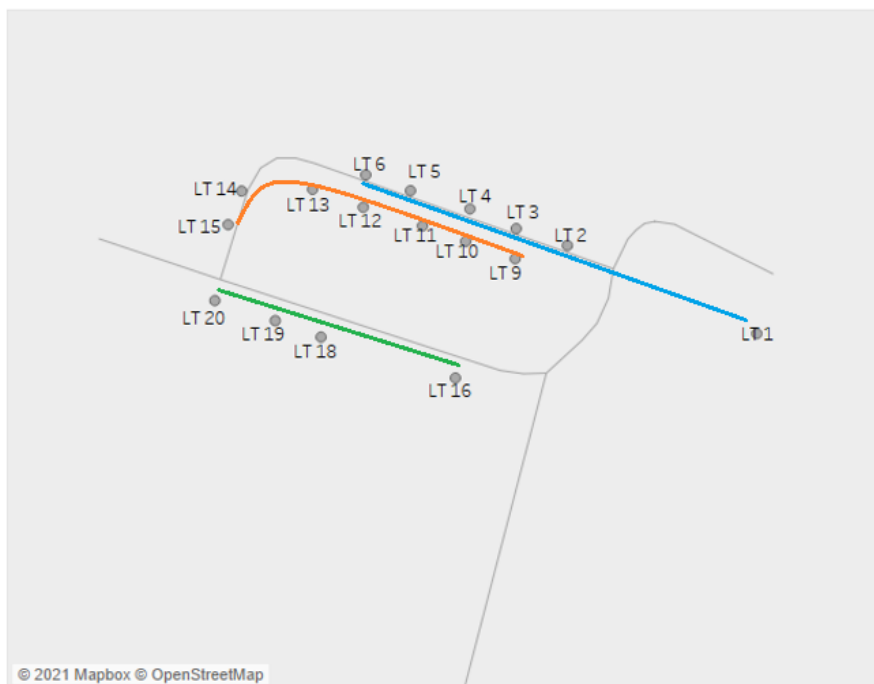


Figura 4.12: Exemplo de candidatos vizinhos pertencerem troços diferentes

4.4 Introdução da régua dos carteiros no geocoding

Até agora, toda a informação para o nosso sistema de geocodificação é baseada em endereços presentes no Cartaxo, ou seja, os endereços de que temos o conhecimento sobre a sua localização geográfica. Com toda informação disponibilizada, além dos dados geográficos ainda podemos utilizar outras informações úteis para estimar as posições. Introduzimos aqui, a geocodificação com a régua dos carteiros.

Relembrando a informação contida na régua, as quais possam ser úteis para geocodificação, como atributo a ordem da entrega na rota, apesar de este não corresponder a ordem otimizada, mas garante-nos que o mesmo endereço está entre algures na zona. Esta característica também permite a possibilidade de estimar as portas com designação não numérica. Visualizando a figura 4.13, onde os endereços estão identificados por rota que o pertence e a respetiva ordem de entrega. Podemos observar que a ordenação da rota não esta completa, pelo senso comum, podemos concluir que as portas em falta deve estar algures entre as portas com ordens presentes.



Figura 4.13: Ordem da entrega pré-definido de uma rota de carteiro

E ainda, lembrando que a régua para o contexto do problema, servirá de input, visto que o desafio deste projeto será estimar as coordenadas dos endereços presentes neste ficheiro, outro ponto podemos aproveitar aqui para o nosso sistema de geocodificação é o conhecimento sobre a quantidade de endereços existentes. Para uma dada artéria, se temos o conhecimento sobre a localização de dois pontos, e ao conhecermos a quantidade de sucessor e predecessores entre eles, podemos então atribuir uma distância proporcionada para os endereços desconhecidos, evitando adicionar / faltar uma distância extra para portas inexistentes / existentes.

AVALIAÇÃO DE RESULTADOS

Este capítulo irá apresentar os testes realizados para a validar o sistema construído durante a fase do desenvolvimento e os resultados provenientes deles.

5.1 Critérios de Avaliação

Antes de apresentar os resultados obtidos, iremos introduzir nesta secção como vamos avaliar a nossa solução. Segundo o contexto do problema, o objetivo principal deste projeto é estimar as coordenadas de modo a construir as rotas do carteiro de uma forma eficiente e otimizada, ou seja, um carteiro, irá fazer o seu percurso ao longo da linha que descreve a artéria e o importante aqui seria detectarmos o sentido da sua caminhada. Portanto, na medição do erro, vamos optar em medir o nosso resultado projetado com a posição real projetado, ilustrado na figura 5.1.

Para a medição da distância entre o ponto estimado e o ponto real utilizamos a fórmula de haversine, onde este permite calcular a distância mais curta entre dois pontos com coordenadas latitudes e longitudes numa superfície esférica. Como alternativa, a proporção na linha da artéria resultando entre os dois pontos pode ser utilizada juntamente com o comprimento da artéria para obter a distância. Devido à geometria da artéria, ambos dos métodos podem apresentar erros. Recorrendo à fórmula de haversine, a distância real pode ser maior do que obtida, visto que o método ignora a geometria da artéria. E em relação ao segundo método, quando mais complexo for a artéria maior seria a sua complexidade e podem surgir outros erros, devido a este motivo, para os resultados ilustrados nas secções seguintes, as distâncias são calculadas pela fórmula de haversine [18].

E ainda, uma vez que o objetivo final deste projeto é para ajudar a definir as rotas otimizadas para os carteiros, consideramos que, um raio até 50 metros seria aceitável. Por tanto, uma distância até 50 metros, consideramos que é um resultado válido e correto.



Figura 5.1: Exemplo de como medir o erro

5.2 Geocoding com dados Cartaxo

Numa primeira fase, para garantirmos a validade do nosso sistema, decidimos estimar com a nossa solução a todos os endereços presentes no ficheiro do Cartaxo, isto é, para um total de 13 830 portas, usamos tanto como referência e também como input. É de claro quando é usado para efeito de input, este é retirado do conjunto de referência. Uma vez consideramos que os dados fornecidos pelos CTTs são fiáveis e válidos, podemos utiliza-las como base de *Groundtruth*.

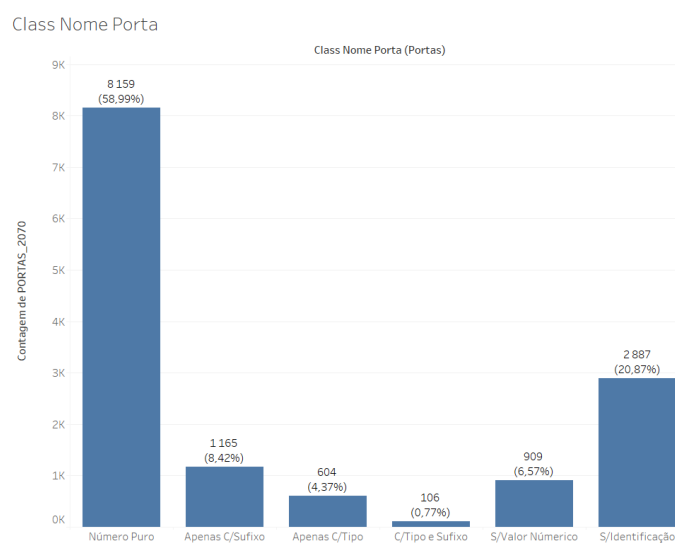


Figura 5.2: Classificação das portas dos dados do Cartaxo

A figura 5.2 apresenta a distribuição do número da porta de acordo com a classificação do tipo de porta. Tal como se observa na figura, temos 20,87% das portas não tem identificação, e serão descartadas visto que não possuem características para o processo de geocodificação, sendo as restantes, um total de 10 943 portas contribuem para o nosso input. E ainda no mesmo conjunto, temos 909 portas cuja a sua designação não possuem qualquer identificação numérica em que possa determinar a sua posição relativa, neste caso, iremos usar geocodificação com a régua dos carteriros.

Para as portas que possuem identificação numérica, isto é, para além das portas do tipo numérico, também incluem como as portas do tipo Lote e Bloco. Podemos observar na figura 5.3, para os resultados obtidos, num total de 10 034 portas, atingimos 90.44% dos endereços válidos e corretos segundo ao nosso critério. 4.15% dos casos se encontram numa distancia a cima de 50 metros, e a maioria deles encontram-se numa zona mais rural, como consequência, os resultados encontram-se mais afastados devido a distribuição dos dados de referência. E ainda 5.41% dos casos como *"not found"*, onde estes não apresentam dados de referência suficientes para efetuar a geocodificação.

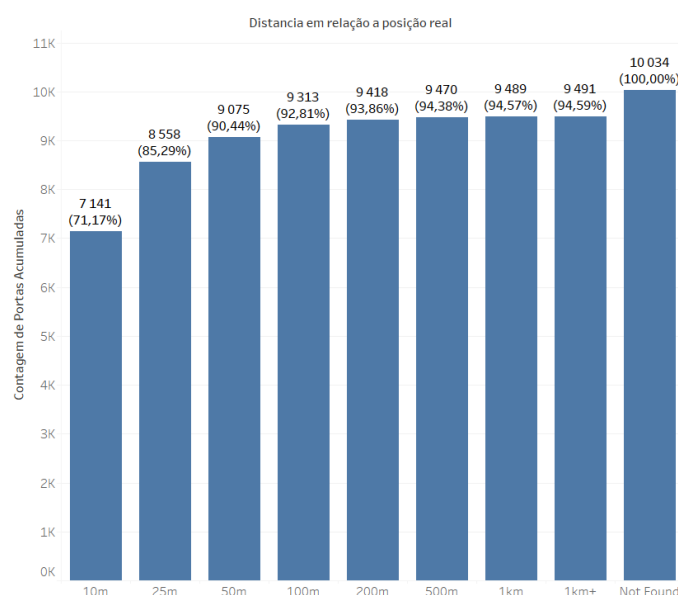


Figura 5.3: Resultado obtido com endereços do cartaxo para as portas numéricas

E para as portas não numéricas, conseguimos obter alguns resultados, mas não tanto como as portas numéricas, ilustrando na figura 5.4, 46.64% dos casos não se obteve resultado, uma vez que a maioria deles não possuem a identificação na rota dos carteiros e uma pequena parte com inexistencia de dados de referência. E ainda observando na figura, a qualidade dos resultados não se encontram num estado desejável. O motivo deve-se ao uso da geocodificação com a régua dos carteiros, em que um dos fatores de computação foi o uso da ordem de entrega na rota, e este nem sempre representa um caminho otimizado.

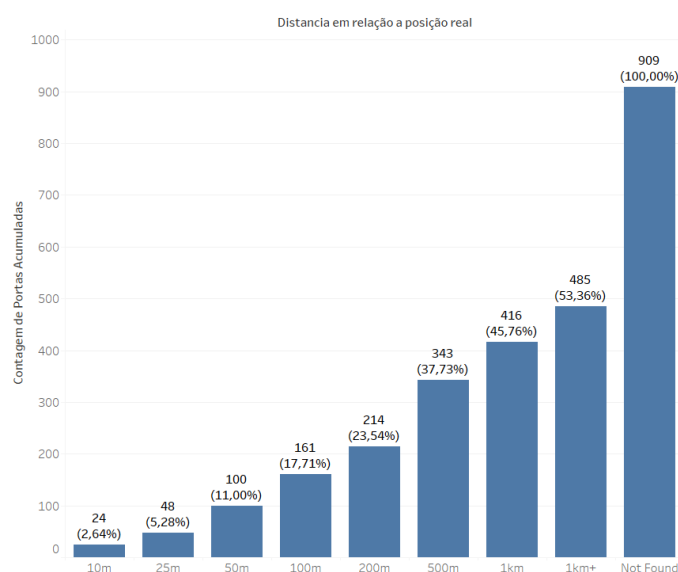


Figura 5.4: Resultado para as portas não numéricas

5.2.1 Comparação dos resultados com GoogleAPI

Para estudarmos a vantagem em utilizar a nossa solução resolvido com os APIs existentes, decidimos ainda, comparar os nosso resultado com GoogleApi, uma vez este foi considerado com os melhores resultados na secção 3.7. Usando o mesmo conjunto de exemplos, obtivemos o resultado ilustrado na figura 5.5. Podemos observar que neste conjunto de exemplos, conseguimos obter resultados semelhantes ao API da Google. Até uma distância de 25 metros, a nossa solução apresenta melhores resultados, mas existem uma quantidade de casos significativos que não se obteve uma resposta, enquanto isto não acontece para o API da Google.

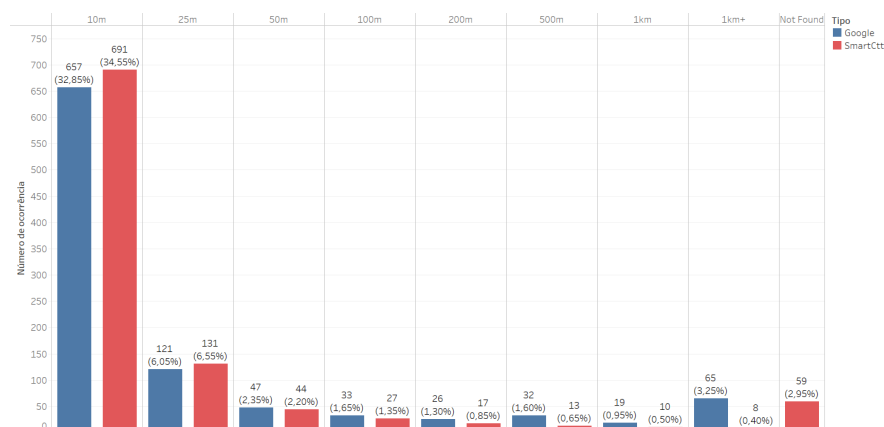


Figura 5.5: Comparação de resultado entre SmartCTT e GoogleApi

5.3 Geocoding com a régua do Carteiro

Nesta secção, iremos demonstrar os resultados obtidos com os endereços presentes no ficheiro das régua dos carteiros. Os endereços em causa aqui são os endereços que não resultam da interseção com o ficheiro do Cartaxo, estes já foram ilustrados na secção 3.1.1. Aqui não temos as coordenadas base para validar os resultados e para ter uma ideia sobre a qualidade. Infelizmente nem todos os endereços se obtiveram resultados. Sendo como input, os 4 963 registos não intersectados, conseguimos estimar as coordenadas para um total de 4 302 portas. Entre os restantes 661 portas, 183 casos não possuem identificação da porta e 478 não foram produzidos resultados devido a falta de dados de referência.

5.3.1 Validação de algumas resultados c/ Google Street View

Uma vez não tendo as coordenadas que representa a posição real, decidimos visualizar as casas com as imagens foto aéreas, de modo a concluir a validade teórica do nosso sistema. Para o efeito, utilizamos *Google Street view* do [9], visualizamos a olho, se os resultados estimados encontram-se perto da posição real. Para cada endereço, registamos como válido e correto se encontra dentro de raio de 50 metros, como válido se na mesma artéria ou artérias adjacentes, e caso contrário, consideramos como *not found*. Podemos ver na figura 5.6. Um exemplo de um resultado válido.



Figura 5.6: Exemplo de um resultado válido

CONCLUSÃO E TRABALHOS FUTUROS

Este capítulo apresenta as conclusões finais sobre a contribuição desta dissertação na sua área envolvida e as direções possíveis para trabalhos futuros.

6.1 Conclusão

Com todo o trabalho realizado pensamos em que este projeto atingiu aos requerimentos mínimos, conseguimos implementar um sistema sólida para estimar as coordenadas dos endereços. A informação presente neste documento juntamente com o código desenvolvido pensamos que já se deixa claro a ideia de como foi desenvolvido o nosso sistema.

Para o desenvolvimento do projeto, dividiu-se em duas fases, sendo a primeira, a nossa solução básica, que determinamos a paridade da artéria e de seguida, aplicamos uma interpolação ou extrapolação com os candidatos vizinhos escolhidos. Consideramos a segunda fase, um metodologia mais sofisticado de modo a melhorar a qualidade dos resultados com a introdução da noção dos troços.

Relembrando os objetivos definidos na secção, a ideia será completar máximo possível a informação em falta nas bases de dados dos CTTs. Para uma grande parte dos endereços desconhecidos, conseguimos dar um palpite nas suas coordenadas. Consideramos que os resultados obtidos encontram-se dentro do que era esperado.

6.2 Trabalho Futuro

A solução apresentada como base, pensamos ainda existe um grande potencial na sua exploração para realizar nos trabalhos futuros. Uma vez que todo projeto foi desenvolvido com os dados fornecidos pelos CTTs, um dos possíveis ramos seria integrar os dados que provem nas outras fontes, como por exemplo, os dados de Openstreetmaps, que existem atributos como direção e sentido que possam ser úteis para investigação, ou mesmo com imagens de satélite que possa ajudar a tomar algumas decisões de modo a garantir o máximo possível a qualidade dos resultados. Devemos considerar a importância na contribuição de enriquecer para os dados de referência.

Outro aspeto a trabalhar seria a geocodificação reversa, na observação do ficheiro dos endereços do Cartaxo, observou-se ainda uma quantidade de endereços consideráveis que conhecemos as coordenadas, mas não possuem identificação. A geocodificação reversa neste aspeto, pode ajudar a dar uma identidade a estes pontos “desconhecidos”, uma vez estes podem ser o nosso alvo que provem no ficheiro das réguas (o nosso ficheiro input).

E por fim, também um aspeto a melhorar no projeto seria a performance e complexidade, em todo o projeto focamos nos aspetos dos resultados obtido e ainda existem aspetos a melhorar para acelerar o funcionamento do sistema.

BIBLIOGRAFIA

- [1] *Bing Map Location API*. URL: <https://docs.microsoft.com/en-us/bingmaps/rest-services/locations/find-a-location-by-address> (ver pp. 19, 20).
- [2] J. Campbell e M. E. Shin. *Essentials of Geographic Information Systems*. 2011 (ver p. 1).
- [3] CTT. *Manual de Endereçamento*. Rel. téc. 2014, pp. 1–21 (ver pp. 8, 9).
- [4] *Datanyze*. URL: <https://www.datanyze.com/market-share/mapping-and-gis--121/google-maps-api-market-share> (ver p. 27).
- [5] *Ecos do Ribatejo*. URL: <http://www.ribatejo.com/ecos/economia/primario.html#Cartaxo> (ver p. 23).
- [6] B. D. W. Goldberg. *A Geocoding Best Practices Guide*. 2008 (ver pp. 13, 15, 16).
- [7] D. W. Goldberg, J. P. Wilson e C. A. Knoblock. “From Text to Geographic Coordinates : The Current State of Geocoding”. Em: (2006), pp. 33–46 (ver p. 3).
- [8] *Google Geocoding API*. URL: <https://developers.google.com/maps/documentation/geocoding/intro> (ver pp. 19, 27).
- [9] *GoogleMaps*. URL: <https://www.google.pt/maps> (ver p. 49).
- [10] *Here Geocoding API*. URL: <https://developer.here.com/products/geocoding-and-search> (ver p. 19).
- [11] C. A. D. Jr e F. T. Fonseca. “Assessing the Certainty of Locations Produced by an Address Geocoding System”. Em: May 2014 (2007). DOI: [10.1007/s10707-006-0015-7](https://doi.org/10.1007/s10707-006-0015-7) (ver pp. 14, 17, 18).
- [12] S. F. M. Lambert M. Surhone Miriam T. Timplendon. *Reverse Geocoding*. 2010 (ver p. 6).
- [13] D. J. Logan et al. *Reverse Geocoding System using combined Street Segment and Point Datasets*. 2010 (ver p. 7).
- [14] *MapQuest Geocoding API*. URL: <https://developer.mapquest.com/documentation/geocoding-api/> (ver p. 19).

- [15] *OpenStreetMap*. URL: <https://www.openstreetmap.org> (ver p. 23).
- [16] *População residente Cartaxo*. URL: <https://www.pordata.pt/Municipios/Popula%C3%A7%C3%A3o+residente++estimativas+a+31+de+Dezembro-120> (ver p. 23).
- [17] J. Ratcliffe. “On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units”. Em: *International Journal of Geographical Information Science* 15 (jul. de 2001), pp. 473–485. DOI: [10.1080/13658810110047221](https://doi.org/10.1080/13658810110047221) (ver p. 16).
- [18] *StringFixer Haversine Formula*. URL: https://stringfixer.com/pt/Haversine_formula (ver p. 45).

