



MARIA SOFIA SERRA DO NASCIMENTO

Licenciada em Engenharia Informática

PLATAFORMA DE MAPAS TEMÁTICOS PARA DADOS ESPACIOTEMPORAIS

CASO DE ESTUDO PARA OS DADOS RSB

MESTRADO EM ENGENHARIA INFORMÁTICA

Universidade NOVA de Lisboa
Dezembro, 2022

PLATAFORMA DE MAPAS TEMÁTICOS PARA DADOS ESPACIOTEMPORAIS

CASO DE ESTUDO PARA OS DADOS RSB

MARIA SOFIA SERRA DO NASCIMENTO

Licenciada em Engenharia Informática

Orientador: João Carlos Gomes Moura Pires

Professor Associado, NOVA University Lisbon

Coorientador: Fernando Pedro Reino da Silva Birra

Professor Auxiliar, NOVA University Lisbon

Júri

Presidente: Vasco Miguel Moreira Amaral

Professor Associado, FCT-NOVA

Arguente: Nuno Miguel Soares Datia

Professor Adjunto, ISEL

Vogal: João Carlos Gomes Moura Pires

Professor Associado, NOVA University Lisbon

Plataforma de Mapas Temáticos para dados Espaciotemporais

Copyright © Maria Sofia Serra do Nascimento, 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.

Por resiliência.

AGRADECIMENTOS

À Faculdade de Ciências e Tecnologia, à universidade NOVA e ao NOVA Lincs. pela paciência e disponibilidade dos orientadores e contributos académicos destes. Aos professores de mestrado de MIEI por serem um exemplo, tais como os de Dresden UT onde passei um semestre de Erasmus e tirei *Data Visualization* e me inspiraram a seguir esta área. Por fim, e não menos importantes, aos meus amigos, Carolina Peres, Daniel Flamino e Diogo Serrano por me ajudarem a continuar em momentos menos bons.

RESUMO

Antes de haver computadores, as visualizações já eram usadas para representar dados espaciotemporais. O popular gráfico da campanha russa de Napoleão, feito por Charles Joseph Minard em 1961, é bastante conhecido por ser uma excelente visualização e exemplo para este tipo de dados.

A visualização de dados espaciotemporais é complexa. Este tipo de dados contém tanto referências espaciais como temporais, sendo particularmente desafiante por ser composto por mais que quatro dimensões, que torna complicado de representar eficientemente numa tela 2D. Para além disto, o conceito de tempo não é intuitivo para o ser humano, dado que não somos capazes de nos movimentar nele como fazemos no espaço. Movemos-nos unicamente no presente e para a frente, na direção do futuro. Porém, esta não é a única dificuldade de trabalhar com o tempo, também as suas diferentes granularidades, o tipo de calendários e os ciclos devem ser considerados.

Apesar destas dificuldades, existem soluções para este tipo de dados. Contudo, a maioria destas resultam apenas de dados com características específicas. Para além desta restrição, outra razão pela qual as visualizações podem não ser apropriadas é por se encontrarem antiquadas. Ora não permitem interações, ora usam animações para representar a mudança do tempo, que se sabe ser uma técnica que sofre de cegueira à mudança, entre outros.

Esta dissertação pretende focar-se em dados espaciotemporais que usem geografia de polígonos ou eventos espaciotemporais, e desenvolver nessa área novas propostas de visualizações interativas. Começando pelo processo de analisar a taxonomia dos dados, seguindo pelo estudo das tarefas e questões inerentes, e terminando com o levantamento das soluções já presentes. Após esse estudo, o desenho de novas propostas começa, onde mais tarde serão implementadas e testadas com outros utilizadores.

Palavras-chave: Visualização de Dados, Mapas Temáticos, Dados Espaciotemporais, Visualização Analítica, Visualização Interativa.

ABSTRACT

Before the existence of computers, visualizations were already used to represent spatiotemporal data. The popular graph of Napoleon’s Russian campaign made by Charles Joseph Minard in 1961, is quite well known for being an excellent visualization and example of this type of data.

The visualization of spatiotemporal data is complex. This type of data contains both spatial and temporal references, being particularly challenging for being composed of four or more dimensions, which becomes complicated to represent efficiently on a 2D canvas. Furthermore, the concept of time isn’t intuitive to human beings, given that we aren’t capable of moving in it as we do in space. We move uniquely in the present and forward, in the direction of the future. However, this isn’t the only difficulty of working with time: the different granularities, the types of calendars, and cycles all should be considered as well.

Despite these difficulties, there are solutions for this type of data. However, most of these work only with data with specific characteristics. In addition to this restriction, another reason why visualizations may not be appropriate is that they are outdated. Sometimes they do not allow interactions, sometimes they use animations to represent the change of time, which is known to be a technique that suffers from blindness to change, among others.

This dissertation intends to focus on spatiotemporal data that use polygon geography or spatiotemporal events, and develop new proposals for interactive visualizations in this area. Starting with the process of analyzing the taxonomy of the data, continuing with the study of the inherent tasks and issues, and ending with the survey of the solutions already present. After this study, the design of new proposals begins, and they will later be implemented and tested with other users.

Keywords: Data Visualizaton, Thematic Maps, Spatiotemporal data, Analytic Visualization, Survey.

ÍNDICE

Índice de Figuras	ix
1 Introdução	1
1.1 Contexto e Motivação	1
1.2 Descrição do Problema	3
1.3 Contributos	4
1.4 Organização do Documento	5
2 Estado da Arte	6
2.1 Conceitos Básicos de Visualização de Dados	6
2.1.1 Suporte Interativo	8
2.1.2 Suporte Analítico	11
2.2 Dados Espaciotemporais	13
2.2.1 Características de dados Temporais	13
2.2.2 Características de dados Espaciais	16
2.2.3 Exemplos de Visualizações de dados Espaciais	20
2.3 Levantamento de Visualizações de dados Espaciotemporais	22
2.3.1 Pencil Icons	23
2.3.2 Helix Icons	23
2.3.3 Data Vases	24
2.3.4 Wakame	26
2.3.5 Icons on Maps	26
2.3.6 Value Flow Map	27
2.3.7 Mapas de Anéis	28
2.3.8 Time-Oriented Polygons on Maps	29
2.3.9 VIS-STAMP	30
2.3.10 Cubo Espaço-Temporal	31
2.3.11 GeoTime	32

3	Trabalho Realizado	34
3.1	Abordagem	34
3.2	Tecnologias Utilizadas	39
3.2.1	Tableau Prep	40
3.2.2	Tableau Desktop	41
3.2.3	Tableau Online	43
3.3	Processamento de Dados	43
3.3.1	Dados iniciais	44
3.3.2	Limpeza e Transformação dos Dados	45
3.3.3	Análise Exploratória dos Dados	50
3.4	Dashboards Desenvolvidos	57
3.4.1	D1 - Análise geral da distribuição temporal e espacial das ocorrências RSB	58
3.4.2	D2 - Análise de Padrões Cíclicos	61
3.4.3	D3 - Tendências de Freguesias	64
3.4.4	D4 - Tendências ao longo dos Anos	66
3.4.5	D5 - Análise de Acidentes Graves	67
3.5	Conclusões	70
4	Avaliação do Trabalho Realizado	72
4.1	Metodologia de Avaliação	72
4.2	Participantes	75
4.3	Resultados	76
4.4	Conclusões	78
5	Conclusões e Trabalho Futuro	79
	Bibliografia	81
	Apêndices	
A	Propostas de Abordagens	83
A.1	Acidentes Rodoviários e Incêndios Recorrentes	84
A.2	Acidentes Rodoviários com Encarcerados	85
A.3	Serviço de Abertura de Porta	85
A.4	Incêndios e VE37 – Veículo com Escada Giratória de 37 metros	86
A.5	Queda de Árvores	86
A.6	Cheias Rápidas	86
B	Questionário com soluções	88
C	Interaction Visualization Modeling Language	92

ÍNDICE DE FIGURAS

2.1	Variáveis visuais sugeridas por Bertin em 1983. Figura retirada de [14].	18
2.2	Ranking de tarefas percetuais. Figura retirada de [14].	19
2.3	Ranking da precisão de cada variável visual para as tarefas percetuais quantitativas, como verificado empiricamente por Cleveland and McGill. Figura retirada de [14].	20
2.4	Mapa de pontos da morada de pacientes a sofrerem de cólera em Londres no ano 1854, feito por John Snow.	21
2.5	Mapa de símbolos graduados que mostra o total de mortes por doenças cardiovasculares por município em Michigan, entre os anos 2004 e 2008. Figura retirada de [15].	22
2.6	Mapa coropleto a média anual de mortes por causas cardiovasculares nos Estados Unidos, por município nos anos 2015 a 2017. Figura retiradas de [15].	23
2.7	Mapa <i>isopleth</i> com a quantidade de precipitação máxima anual num período de 5 dias consecutivos na Coreia do Sul, entre 1981 e 2010.	24
2.8	Técnicas de visualização que fazem uso de extrusões gráficas de pontos espaciais. Figuras retiradas de [2].	25
2.9	Dois exemplos de gráficos com vasos de dados, que mostram número de crimes violentos e a taxa de desemprego na Carolina do Norte. Figura retirada de [2].	25
2.10	Gráfico Wakame e a interface do sistema. Figura retirada de [2].	26
2.11	Exemplo de um gráfico de ícones em mapas que para efeitos demonstrativos usa vários tipos de gráficos para as de miniaturas de visualizações. Figura retirada de [2].	27
2.12	Exemplo de um mapa de <i>Value Flow</i> sobre crimes nos Estados Unidos. Figura retirada de [2].	28
2.13	Dois exemplos de mapas de anéis. No gráfico da esquerda são alertas em códigos postais na figura da direita, e nível de atividade humana durante um dia no da direita. Figura retirada de [2].	29

2.14	Três variantes de Polígonos Orientados Temporalmente em Mapas da esquerda para a direita: cunhas, anéis e fatias. Representam a evolução da população de estudantes no ensino secundário. Figura retirada de [2].	30
2.15	VIS-STAMP que mostra dados sobre crimes nos Estados Unidos. Figura retirada de [2].	31
2.16	Cubo espaço-temporal que mostra dados climatéricos. Figura retirada de [2].	32
2.17	Mapa GeoTime que mostra as trajetórias de táxis envolvidos num atropelamento e consequente (delito de) fuga. Figura retirada de [2].	33
3.1	Fluxo da metodologia CRISP-DM [18].	35
3.2	<i>Flow</i> de limpeza usado no Tableau Prep para o ficheiro sobre as viaturas. . .	46
3.3	<i>Flow</i> de limpeza usado no Tableau Prep para o ficheiro sobre os locais e aonde junta com o ficheiro das viaturas no Join.	46
3.4	<i>Flow</i> de limpeza usado no Tableau Prep para juntar as correções aos dados feitas manualmente.	50
3.5	Distribuição da quantidade de ocorrências por ano. Pode-se observar que há alguma variância entre os anos.	51
3.6	Distribuição da média, dos cinco anos, de ocorrências por mês. Pode-se observar que há alguma variância entre os meses.	51
3.7	Distribuição da média, dos cinco anos, de ocorrências por dia da semana. Pode-se observar que há alguma variância entre a média de ocorrências dependendo do dia.	52
3.8	Distribuição da média, dos cinco anos, de ocorrências por hora. Pode-se observar que há alguma variância entre a média de ocorrências dependendo da hora do dia.	52
3.9	Frequência global da duração total das ocorrências (em minutos) com a quantidade de ocorrências. O azul-petróleo realça onde há menos de 10 ocorrências.	53
3.10	Duração total média por ano.	53
3.11	Tempo médio em Rota por ano.	53
3.12	Porcentagem das ocorrências que ocorreram, por trimestres, que pertencem à natureza de primeiro nível.	54
3.13	Quantidade de ocorrências que ocorreram, por trimestres, que pertencem à natureza de primeiro nível.	54
3.14	Caixa de bigodes da duração total de cada ocorrência, em minutos, com a quantidade de ocorrências.	55
3.15	Gráfico da participação por viaturas em relação ao total de todos os anos das ocorrências do RSB.	56
3.16	Mapa coropleto de Lisboa com a quantidade de ocorrências nos cinco anos do <i>dataset</i> por freguesia.	56
3.17	Mapa de Densidade com as ocorrências nos cinco anos dos dados. Áreas mais claras têm uma maior densidade de ocorrências nesse local.	57

3.18	Mapa coropleto de Lisboa com a duração média das ocorrências nos cinco anos do <i>dataset</i> por freguesia.	58
3.19	Captura de ecrã do <i>dashboard</i> D1, intitulado Análise geral da distribuição temporal e espacial das ocorrências RSB.	59
3.20	Captura de ecrã do <i>dashboard</i> D1, com granularidade da janela maior que 4 meses.	60
3.21	Granularidades do mapa de hexágonos: Scale 700 (Pequeno), Scale 400 (Médio) e Scale 100 (Grande) respetivamente.	60
3.22	Análise ao dia atípico de 19 de janeiro de 2013 utilizando o D01.	61
3.23	<i>Tooltip</i> do mapa de <i>hexabins</i> do <i>dashboard</i> D01.	62
3.24	Captura de ecrã do <i>dashboard</i> D2, intitulado Análise da Ciclicidade Semanal.	63
3.25	Informação apresentada na <i>tooltip</i> do gráfico de barras do D02.	63
3.26	Exemplo de uma análise conduzida com o <i>Dashboard</i> 02, a comparar as diferenças entre os verões (esquerda) e os invernos (direita).	64
3.27	Captura de ecrã do <i>dashboard</i> D3, intitulado Tendências de Freguesias.	65
3.28	Exemplo da seleção de uma freguesia no mapa do <i>dashboard</i> D3.	66
3.29	Captura de ecrã do <i>dashboard</i> 04, intitulado Tendências ao longo dos Anos.	67
3.30	Captura de ecrã do <i>dashboard</i> 05, intitulado Análise de Acidentes Graves. Está dividido em duas visualizações, A e B.	68
3.31	Exemplificação da função de <i>highlight</i> no <i>Dashboard</i> 05.	68
3.32	Gráfico de densidade com as ocorrências do tipo Acidente com Encarcerados.	69
4.1	Distribuição das Idades e da sua experiência com Visualização de Dados.	75
4.2	Gráfico com a taxa de sucesso de cada questão do questionário.	76
4.3	Distribuição do tempo médio e número de ações média por cada questão.	77
4.4	Distribuição das médias de tempo e número de ações por utilizador. Cada ponto é um utilizador.	77
C.1	Diagrama de <i>tooltips</i> do dashboard 01, com duas <i>tooltips</i>	93
C.2	Diagrama de interações do dashboard 01, com quatro interações.	94

INTRODUÇÃO

Eventos espaciotemporais são um tipo de dados compostos por uma componente espacial, uma componente temporal e têm uma duração nesse tempo [2]. Este tipo de eventos e dados existem todos à nossa volta. Um exemplo seria o exames numa faculdade, os quais têm uma localização, a sala, um fator temporal, a data e hora, e uma duração máxima de tempo.

Este tipo de dados não é novo, mas são pouco explorados devido à sua complexidade. Tanto a dimensão espacial como temporal são pouco intuitivas de se explorar numa tela 2D, portanto é de esperar que a sua combinação piore a sua visualização e consequente interpretação dos dados. Há também que considerar que apenas nesta era moderna é que se encontram este tipo de dados com mais frequência e em conjuntos de dados exponencialmente maiores, devido à facilidade que é hoje em dia recolher dados das mais variadas maneiras. Para além do crescimento da quantidade de dados espaciotemporais, este aumento de dados também cria grande diversidade de dados espaciotemporais e com esses igualmente novos desafios de visualização. Existem poucas técnicas de visualização que resolvam este problema, pois há novas dificuldades a processar conjuntos de dados tão grandes. Como qualquer outra área, a sua análise e compreensão deste problema poderão melhorar a área a que a temática de dados e as suas respetivas fontes e ofícios.

1.1 Contexto e Motivação

Visualizações são uma ferramenta de transmissão de ideias que estão cá há milhares de anos [3]. Funcionam onde palavras pode não ser suficientes, eficientes ou até uma barreira.

O aumento da procura do mercado tem criado mais e maiores mercados de trabalhos, que significam investimentos tanto no equipamento como nas equipas. Em particular, uma das ferramentas que tem sido desenvolvida com a computadorização tem sido visualização de dados. Seja para explorar, confirmar hipóteses ou apresentar ideias, é inegável que passou a ser fundamental uma boa comunicação. Mais do que isso, a eficiência, adequação e expressividade são considerados critérios fundamentais que garantem posições de relevância a visualizações durante muito anos, enquanto a outras fazem com que sejam

procuradas melhores alternativas.

A visualização de dados espaciotemporais é complexa, tanto na sua construção como na sua interpretação. Este tipo de dados contém tanto referências espaciais como temporais, sendo particularmente desafiante por ser composto por quatro ou mais dimensões. Esta é a chamada maldição da dimensionalidade, que torna complicado de representar tantas dimensões eficientemente numa tela 2D. Em alguns casos, opta-se por visualizar este tipo de dados no que são chamados mapas temáticos. Estes mapas, em geral, envolvem visualizações dentro de visualizações, ou seja, uma visualização temporal em cima do ponto geográfico correspondente.

Para além disto, o conceito de tempo não é intuitivo para o ser humano, dado que não somos capazes de nos movimentar nele como fazemos no espaço. Movemos-nos unicamente no presente e para a frente, na direção do futuro [2]. Porém, esta não é a única dificuldade de trabalhar com o tempo: as diferentes granularidades do tempo, tipo de calendários e ciclos também devem ser considerados.

Apesar de sermos confrontados com dados espaciotemporais quase todos os dias, as visualizações destas ainda estão pouco desenvolvidas devido às dificuldades já mencionadas. Porém, estamos numa altura oportuna para estudar esta questão, explorando e propondo as melhores visualizações.

Atualmente já foram desenvolvidas visualizações para dados espaciotemporais. Nem todas estas visualizações servem, nem deviam, para todo o tipo de casos [2][4]. Contudo, é um campo ainda em evolução. Há técnicas obsoletas que não satisfazem as características mencionadas na secção 1.1, visto que tratam o tempo como um parâmetro meramente numérico, ignorando as propriedades especiais deste. Outro problema comum é negligenciarem a exploração interativa, ou a pesquisa em séries temporais, que são essenciais para boas análises baseadas em visualizações. Ou até mesmo a utilização de animações ou *sliders* para representar a mudança do tempo em mapas temáticos com dados espaciotemporais que, tal como a memória visual do ser humano, é falível, sofrendo de cegueira à mudança.

A qualidade de uma visualização de dados é subjetiva. Neste sentido o processo de visualização está dependente de dois aspetos: o tipo de dados a ser analisado e as questões, sobre os mesmos dados, a que queremos responder com a visualização. Assim, resume-se a duas perguntas “O que queremos apresentar?” e “Porquê queremos apresentar?” [2]. Pode-se concluir que a escolha da visualização apropriada para um certo caso é altamente dependente do contexto no qual é aplicada.

Quando se fala em dados espaciotemporais existem pelo menos duas categorias de dados: as variáveis espaciais, as variáveis temporais e as variáveis alfanuméricas dos dados [2], que são a informação relativa à temática que queremos analisar no contexto das duas. Eventualmente pode haver sobreposição da temática e esta ser também uma variável temporal ou espacial, e não alfanumérica como mencionado, querendo dizer que as categorias utilizadas seriam apenas duas, mas há sempre uma componente temática seja ela qual a categoria a que pertença. Imaginando todas as combinações possíveis

destas categorias, por exemplo, para a parte espacial pode ser regiões, linhas ou pontos, e exemplos para a parte temporal podem ser intervalos de tempo, eventos ou outras maneiras de medir o tempo, dependendo do(s) mapeamento(s) usado(s) nos dados. Cada uma dessas combinações é um desafio diferente com soluções igualmente únicas. O que se pretende destacar é que cada combinação destas categorias de dados é no seu conjunto um tipo de dados único. Isto dá-se porque a mistura das categorias gera questões únicas à combinação criada e às características da temática. Por exemplo, analisar viagens de táxis não é igual a analisar votos numa eleição, porque as suas categorias de dados são diferentes - na parte espacial são linhas da trajetória *versus* regiões, e parte temporal são intervalos de tempo *versus* ponto único no ano -, logo as suas soluções também o serão. Para questões diferentes são adequadas visualizações de dados diferentes, visto que cada uma se especializa num tipo ou secção de dados para responder às questões em causa. É neste aspeto que esta tese se pretende focar.

1.2 Descrição do Problema

Como referido anteriormente, não existem muitas soluções para dados espaciotemporais devido à sua complexidade, mas especialmente devido à sua diversidade. Este facto também constitui a razão pela qual as soluções que existem não são únicas, ou seja, não se aplicam universalmente aos variados tipos de dados espaciotemporais. Os mapas temáticos são um tipo de visualização de dados que fazem uso de um mapa e variáveis visuais sobrepostas sobre esse mapa como fundo, variáveis essas que podem ser usadas para representar o tempo, ou aplicando outro tipo de técnicas, que é o que iremos explorar.

O tema desta dissertação centra-se em encontrar soluções para visualizar dados de eventos espaciotemporais, dos quais devemos agradecer ao Regimento Sapadores de Bombeiros (RSB) por nos terem fornecido os seus dados para instanciar os *dashboards* com casos significativos. Nesta dissertação iremos investigar soluções específicas para dados de eventos localizados em pontos.

Portanto, o objetivo desta dissertação é criar mapas temáticos interativos para exploração de dados com eventos espaciotemporais. Estes mapas devem cumprir os objetivos de explorar a distribuição e tendências ao longo da dimensão temporal e espacial. Para avaliar a sua eficiência, será identificado um conjunto de questões sobre os dados, que será utilizada para definir funcionalidades de certas ferramentas (dividindo o domínio do problema em partes). Mais tarde, essas ferramentas serão diretamente avaliadas contra essas funcionalidades e objetivos, para que a qualidade e eficiência destas ferramentas possa ser medida e considerada.

Além do mais, como objetivo secundário, tendo em conta a parceria com o RSB, tentaremos devolver uma análise dos dados disponibilizados. Os dados que nos foram disponibilizados passarão por uma análise de qualidade, que revelará erros que podem ser corrigidos na base de dados, dando origem a um conjunto de dados transformado, limpo e consolidado. Espera-se que as ferramentas desenvolvidas forneçam análises interessantes

comparativamente às conseguidas com ferramentas já existentes e, além disso, esperamos ainda que estas se tornem importantes para análises futuras.

1.3 Contributos

Tendo em consideração, como referido anteriormente, que o objetivo desta dissertação é explorar métodos para trabalhar com dados espaciotemporais, os contributos desta dissertação foram os seguintes:

- Relatório de qualidade de dados: Como a temática de visualização de dados é baseada em dados, esta dissertação segue esse esquema, e como tal teve que haver uma exploração e limpeza dos dados utilizados no curso da elaboração. Este processo resulta num relatório de qualidade de dados, incluído em apêndice, que beneficia a entidade que partilhou os dados. Os dados referidos no relatório pertencem ao RSB.
- Dados RSB processados: É o produto final da exploração e limpeza de dados, ou seja, os dados que após a escrita do relatório de qualidade, foram processados. Este é subconjunto dos dados originais, com as devidas limpezas de dados erróneos e processamento para que se possa reduzir o trabalho em etapas seguintes, como dividir colunas com múltiplas informações (por exemplo, classe e designação da viatura, em que podemos ter interesse em agrupar as ocorrências pelo tipo de viatura necessária para socorrer a uma natureza). Mais tarde, estes dados foram partilhados com a entidade que os disponibilizou, tendo em conta o especial interesse na correção dos erros encontrados.
- Ferramentas para análise e exploração de dados espaciotemporais: Este é o principal objetivo, e, portanto, também se espera que seja o seu maior contributo. O desenvolvimento de ferramentas que permitam então a análise, exploração ou deteção de padrões em dados espaciotemporais. Para este efeito foram criados vários *dashboards* para responderem a uma ou mais questões especificadas, relevantes no contexto de dados espaciotemporais. Estes podem, então, no futuro, serem usados como exemplos para outros estudantes da temática de dados espaciotemporais.
- Linguagem gráfica para auxiliar o processo de *design* e documentação de *dashboards*: Na elaboração desta dissertação sentiu-se dificuldade a desenvolver ideias de *dashboards*, e mais tarde documentá-los por escrito. Para ajudar a combater este problema eu participei com um grupo de estudantes no desenvolvimento de uma linguagem de diagramas para *dashboards* a que chamamos *Interaction Visualization Modeling Language* (IVML) que continua em desenvolvimento.

1.4 Organização do Documento

Este documento está organizado em quatro capítulos, sendo um deles o capítulo que se encontra esta secção, a Introdução com o objetivo de introduzir o tema e contexto envolvente. Os restantes focam-se noutros aspetos diferentes da temática, que são os seguintes:

- **Estado da Arte:** Capítulo 2, como diz o título do capítulo, é onde está resumido o estado da arte, a informação contextual do presente sobre a matéria da temática. Esta sendo focada na visualização de dados, especificamente envolvendo dados espacial, temporais e uma combinação de ambas.
- **Trabalho Realizado:** Capítulo 3, é descrito o trabalho feito na realização desta dissertação, começando pela abordagem e metodologia, quais as ferramentas utilizadas, os processos feitos nos dados disponíveis e terminando nos resultados destes passos, uma apresentação dos *dashboards* executados.
- **Avaliação do Trabalho Realizado:** Capítulo 4, contem os resultados e conclusões da fase de avaliação das ferramentas criadas durante esta dissertação, para julgar a sua eficácia.
- **Conclusões e Trabalho Futuro:** Capítulo 5, a última divisão desta organização e que finaliza também o documento desta dissertação. Faz um resumo do trabalho concluído, o impacto e quais as conclusões retiradas, e ainda como se pode dar seguimento a este projeto em trabalhos futuros.

ESTADO DA ARTE

O termo visualização é uma definição muito abrangente. A definição que interessa para o contexto deste trabalho refere-se à arte de criar um modelo ou imagem de algo na mente de outra pessoa [5], usando métodos computadorizados. Para este efeito, tem-se feito estudos sobre a percepção humana [6] e foram ainda estudadas diferentes maneiras de se transmitir informação para que esta seja feita de forma eficiente .

Este capítulo pretende estabelecer os básicos e os fundamentos sobre os quais o trabalho assume ser verdade. Primeiro iremos rever as noções básicas de visualizações de dados, questionando como são criadas novas visualizações. Depois avançaremos para quais as técnicas que contribuem para boas visualizações enquanto ferramentas, revendo os métodos de interação e análise de dados em contexto dos dados espaciotemporais. A seguir, expandir-nos-emos sobre as características desses mesmos dados, olhando primeiro para a componente temporal, investigando a fundo todos os aspetos temporais estudados, e de seguida para a componente espacial, onde faremos algo semelhante, mas com foco no mapas temáticos. Abordaremos também assim quais são as variáveis visuais pertinentes utilizadas neste tipo de mapas. Na última secção veremos o culminar de toda esta informação nos exemplos de visualizações utilizadas para ver dados espaciotemporais, mais concretamente um caso complexo e específico que demonstra que nem todas as visualizações se representam da mesma maneira.

2.1 Conceitos Básicos de Visualização de Dados

Para que se possa realmente chegar ao que este campo de visualização de dados pretende, alguns fundamentos básicos têm que ser estabelecidos. Estes fundamentos codificam o universo de visualização e permitem a construção sobre os mesmos.

Existem três critérios que uma boa visualização deve atingir: expressividade, eficiência e adequação [2]. Expressividade refere-se à parte dos dados que devem ser mostrados, sendo estes completos ou até mesmo um subconjunto dos dados que estamos a tratar. Eficiência é o que o nome indica: existência de uma fácil interpretação dos dados, ou seja, boa transmissão do modelo mental que estamos a tentar passar. Isto pode envolver outros

critérios, como por exemplo a forma como se tira partido das capacidades do sistema visual humano ou ainda se a informação visualizadas é interpretada e reconhecida de forma intuitiva. Por último, adequação é a avaliação do custo-benefício, sendo que esse custo é normalmente interpretado pela rapidez da computação da visualização, ainda pelo quanto e como o espaço é utilizado.

O processo de visualização tem que ter em consideração dois aspetos: os dados e a tarefa. Em termos leigos, reduz-se em três questões: O que queremos apresentar? (*What?*); Porque queremos apresentar? (*Why?*); e Como é que queremos apresentar? (*How?*). Isto traduz-se na especificação dos dados, na especificação da tarefa e nas necessidades da ferramenta de visualização. O que se pretende realçar é que os dados e as suas questões influenciam quais as visualizações mais apropriadas para maximizarem os critérios de expressividade, eficiência e adequação.

Podemos dividir as visualizações pelos objetivos ou tarefas que pretendem realizar. A identificação destes objetivos permite desenvolver melhores ferramentas, visto que requerem representações visuais totalmente diferentes [2]. Segundo o livro *Interaction in data Visualization* de Ward et al., a um nível abstrato existem os seguintes objetivos [7]:

- análise exploratória, onde se procura compreender os dados e desenvolver hipóteses;
- análise de confirmação, onde se procura confirmar hipóteses;
- apresentação de resultados da análise, onde se pretende transmitir informação previamente descoberta a outras entidades.

Estes objetivos formam o chamado processo de visualização. Este documento focar-se-á na análise exploratória e na de confirmação.

O processo de visualização que estamos a considerar é o Dos Santos e Brodlie (2004) [8], baseado na versão original de Haber e McNabb (1990), cuja principal mudança é a introdução de uma análise dos dados antes e durante a filtragem dos mesmos. O pipeline é o seguinte:

1. Análise dos Dados, onde se tenta ver os dados de várias perspetivas e visualizações. Deve-se tentar perceber os dados e as suas distribuições nesta fase;
2. Filtragem é a fase onde se selecionam porções dos dados que são interessantes, costuma ser uma fase centrada no utilizador;
3. Mapeamento é o passo do processo de visualização no qual se decide quais são as variáveis visuais que se vão mapear as colunas de valores no dados;
4. Renderização é a etapa final do pipeline, onde se vê o resultado das escolhas dos passos anteriores e consequentemente a visualização dos dados.

2.1.1 Suporte Interativo

A introdução de interação numa visualização é um método que permite ao modelo de visualização ser mais flexível e mostrar diferentes perspectivas dos dados, ou subconjuntos deste, mudando certos aspetos [9] para que seja demonstrada uma região dos dados que nos interesse, por exemplo um período temporal específico. Este processo tem a característica de ser iterativo, onde o computador gera o *feedback* para refletir a mudança que resultou da interação do utilizador. Assim, o utilizador pode ir recursivamente focando-se em subconjuntos de dados que lhe interessam. A principal vantagem desta abordagem é que permite a exploração dos dados, um passo importante para a análise dos mesmos. Não apenas isso, mas também permite mostrar mais informação, ou seja, mais dimensões, usando técnicas que permitem mostrar essa informação apenas em níveis de detalhe onde a sua existência é pertinente e aumenta a claridade dos dados.

Iremos falar bastante do trabalho de Shneiderman nesta secção. Começando pela taxonomia de tarefas por dados de Shneiderman, publicada em 1996, que inclui sete tarefas de alto nível sobre interação e as suas tarefas visuais:

1. Overview: ganhar uma visão geral de todo o conjunto de dados;
2. Zoom - aumenta a área visual dados de interesse;
3. Filtragem - filtrar informação desinteressante;
4. Details-on-Demand - seleccionar dados de interesse e receber detalhes adicionais quando necessário;
5. Relacionar - ver relações entre elementos dos dados;
6. História - manter o histórico de ações para suportar desfazer e refazer;
7. Extrair - permitir a extração de dados e pesquisa de parâmetros.

Mais recentemente foi publicado por Yi et al.'s (2007) um estudo mais profundo sobre o que motivava os utilizadores a interagir, e introduziu uma lista de categorias que descrevem as motivações a alto nível que os utilizadores tem para interagir:

- **Selecionar** – Marcar algo como interessante

O utilizador quer marcar ou realçar os dados que acha relevantes ao encontrá-los na visualização. Isto pode ser implementado através de etiquetas temporárias ou memorizar permanentemente resultados de análises importantes.

- **Explorar** – Mostrar outra coisa

Em conjuntos de dados multivariados de grande volume não é boa prática mostrar as demais dimensões. Em vez disso, mostram-se apenas sub-regiões dos dados para os tornar mais digeríveis aos olhos do utilizador, como por exemplo, ir por períodos de tempo sequencialmente até ter uma ideia geral dos dados.

- **Reconfigurar** – Mostrar numa configuração diferente

Existe uma necessidade do utilizador permitir explorar os dados de diferentes ângulos e parâmetros. Isto é evidente quando se pensa nos arranjos temporais que podem ser lineares ou cíclicos. A ferramenta deve permitir que os dados sejam vistos de diferentes maneiras para que sejam investigadas as características dos mesmos.

- **Mapeamento** – Mostrar uma representação diferente

Semelhante ao que foi dito anteriormente sobre os arranjos temporais, o mapeamento dos dados para as variáveis visuais tem um grande impacto na representação visual. Como os dados e as tarefas são muito versáteis, os utilizadores precisam de ser capazes de adaptar o mapeamento conforme as suas necessidades. Isto pode ser apenas para melhorar a eficiência de certas tarefas ou para confirmar a hipótese gerada noutra representação visual.

- **Abstrair/Elaborar** – Mostrar mais ou menos detalhe

Durante a análise visual, os utilizadores necessitam de diferentes níveis de observação. Se por um lado pode ser importante visualizar certas pormenores em detalhe, por outro, uma representação esquemática pode ser suficiente. A estrutura hierárquica do nível de granularidade do tempo, onde abstrações de alto nível fornecem *overviews* agregadas e os níveis inferiores mantêm os detalhes correspondentes, são uma combinação natural para conduzir uma busca interativa de informações em dados orientados ao tempo.

- **Filtrar** – Mostrar condicionalmente

Quando os utilizadores pesquisam por informação ou avaliam hipóteses nos dados, faz sentido restringir a visualização para mostrar apenas os elementos dos dados que verificam condições impostas pelo critério de pesquisa ou as restrições da hipótese. Interativamente filtrar ou atenuar os dados irrelevantes para "limpar a vista" permite aos utilizadores focarem-se na tarefa em questão.

- **Associar** – Mostrar elementos relacionados

Quando os utilizadores fazem descobertas potencialmente interessantes no dados, normalmente perguntam-se se existem outras descobertas semelhantes ou relacionadas noutra(s) parte(s) do conjunto de dados. Portanto, os utilizadores querem interativamente encontrar, comparar e avaliar outras semelhanças ou relações, como por exemplo, se uma tendência descoberta numa estação do ano está presente noutras variáveis ou se esta é repetida na mesma altura noutros anos.

- **Anular/Repetir Ação** – Deixar voltar ao estado anterior

Os utilizadores irão precisar de navegar no tempo e em diferentes granularidades, tentar diferentes arranjos e mapeamentos visuais e experimentar condições de

filtragens e de diferentes limites de similaridade. Para compensar o sistema de tentativa e erro que um processo de exploração acarreta é necessário um mecanismo de histórico que permita fazer e refazer operações cujo resultado não funcionou como o esperado.

- **Alterar a configuração** – Ajustar a interface

Os utilizadores também podem querer adaptar o sistema da ferramenta ao seu método ou restrições de trabalho. Isto pode acontecer ao ajustar a posição das diferentes interfaces ou fazendo a gestão de recursos como a quantidade de memória a ser usada.

Todas estas intenções em conjunto descrevem a sinergia entre o humano e as capacidades da máquina. Estas abordagens serão novamente discutidas na secção referente ao *survey* de visualizações de dados espaciotemporais. Existem tarefas que são consideradas fundamentais, como marcar itens interessantes, outras que nem são consideradas de todo. Isto deve-se provavelmente ao esforço de implementação que seria considerar todas as intenções. Contudo, isto não quer dizer que todas não sejam igualmente importantes e técnicas equivalentes devem ser fornecidas.

Existem vários métodos básicos estabelecidos que permitem controlar a visualização para além do que nos é apresentado nas interfaces gráficas banais. Segundo o livro de *Visualization of Time-Oriented Data*, existem três métodos chave para implementar interação em visualizações [2]:

- **Manipulação Direta** é um conceito introduzido por Shneidermann em 1983. Foi a solução encontrada para quando os controlos da interface gráfica davam *feedback* visual numa área diferente do ecrã da que se estava a interagir, ou que não garantia que as mudanças fossem claras. Este método descreve então o controlo da interação diretamente na visualização sem desvio [2]. Um exemplo comum é permitir o uso da roda do meio do rato para fazer zoom e ver mais detalhes. Outro exemplo mais específico é a escolha de um intervalo de tempo na TimeWheel. Existem várias aplicações para este conceito.
- **Brushing & Linking** elabora sobre a ideia de manipulação direta. *Brushing* é o nome da técnica que permite aos utilizadores escolher elementos de um conjunto de dados. Frequentemente é implementada com um *point and click* para escolher pontos individuais. O passo seguinte a essa seleção é o destaque dos pontos escolhidos aos restantes dados. Para a técnica de *Linking* é necessário estarem presentes várias visualizações ou vistas dos mesmos dados na tela. Esta técnica refere-se a fazer *brushing*, ou seja, selecionando os dados numa vista é possível destacá-los em todas as outras (daí o nome da técnica, porque conecta as vistas) sendo esta a vantagem desta técnica.

- **Queries Dinâmicas** é o conceito de procurar informação visual, filtrando dados irrelevantes. Isto é uma vantagem em séries temporais, porque normalmente existe um grande volume de dados, portanto dinamicamente omitir dados remetentes a tarefas específicas é útil [2]. Em relação à omissão de dados, pode ser feita de duas maneiras. Os dados irrelevantes são visualizados em menor detalhe, ou então não são apresentados de todo, tornando-se invisíveis (relevante para quando existem demasiados dados para se visualizar na tela). Os sistemas de filtragem normalmente têm mecanismos dedicados e adaptados consoante as características dos dados, como controlos deslizantes ¹ de limite ou intervalo que são eficientes a filtrar por período de tempo, ou qualquer outra variável numérica.

Esta secção focou-se primeiro em dar uma visão geral de quais são os interesses do utilizador num panorama interativo e de seguida enumerar vários métodos clássicos básicos e complexos de implementar e integrar essas intenções nas visualizações. Como dito anteriormente, existem casos em que não é possível implementar todas as técnicas, mas todas estas contribuem para uma melhor experiência de interação com os dados.

2.1.2 Suporte Analítico

A interação serve para ajudar o utilizador a explorar os dados, fazendo partido da eficaz perceção humana a reconhecer padrões. Este método serve, exceto quando o tamanho dos dados é demasiado grande para ser visto todo de uma vez. É aqui que o suporte analítico se encaixa, analisando dados quando os humanos não são capazes. Os computadores são mais rápidos e ainda analisam o problema nas vertentes numéricas, cálculos simbólicos, lógica e pesquisa, nas quais são melhores que os humanos [2].

Estes processos são os conhecidos por mineração de dados e descoberta de conhecimento, dois temas em expansão no campo da aprendizagem automática. Estes resumem-se a extrair estruturas informativas de grandes volumes de dados. As seguintes categorias de tarefas são as relevantes quando aplicadas a dados temporais:

- **Classificação** é quando se dá um algoritmo um conjunto preestabelecido de classes para que ele possa atribuir estas etiquetas a um conjunto de dados, conforme o que aprendeu no seu treino. Este pode ser aplicado a reconhecimento da fala, ou até mesmo para gestos, sendo estes apenas alguns dos exemplos de aplicações.
- **Agrupamento**² refere-se ao método que tenta agrupar um conjunto de dados conforme a sua semelhança, sendo este um critério a ser escolhido. Pode ser conforme a distância entre pontos, por exemplo, ou por alguma dimensão que seja tipológica, ou temporal, ou de várias outras maneiras. Um exemplo são as análises financeiras para perceber quais são as ações com tendências semelhantes pelo tempo.

¹Termo original a partir do inglês *slider*.

²Termo original do inglês *clustering*.

- **Search & Retrieval** é uma tarefa que começa por procurar pesquisas específicas, definidas *a priori*, em conjuntos de dados possivelmente de grande volume. Isto é o chamado query-por-exemplo. Este método pode ser usado para localizar correspondências exatas ou aproximadas (fuzzy), tais como encontrar utilizadores com padrões de pesquisa *online* semelhantes pelo tempo.
- **Descoberta de Padrões** é semelhante ao *Search & Retrieval*, mas não requer uma pesquisa *a priori* porque faz esta descoberta automaticamente. Ou seja, olha para as pesquisa e retira padrões que conseguir identificar seja ele sequencial, periódico, entre outros. Por exemplo, para se notar se um anúncio irá funcionar ou não numa certa época.
- **Previsão** é importante na análise temporal. Aprender com o passado, mas mais do que isso, tentar inferir qual é o comportamento do futuro baseado nos dados recolhidos no passado. Modelos de regressão automática, modelos não estacionário e estacionários, etc.

Reconhecer a tarefa analítica que pretendemos cumprir é o primeiro passo a escolher a ferramenta adequada para obter os resultados pretendidos e que técnicas integrar na análise.

Neste tópico não podemos deixar de falar sobre a possibilidade de abstrações temporal dos dados, um conceito introduzido em 1985 por Clancey. O objetivo da abstração de dados é suprimir os dados irrelevantes de maneira a expor os conceitos chave nos dados. Ou seja, tentar usar valores, classes ou conceitos em vez de apresentar os dados todos. É particularmente útil para lidar com volumes grandes de dados e de grande complexidade [2]. Para que a abstração seja atingida, normalmente exige uns quantos passos de métodos diferentes, como selecionar os dados relevantes, filtrar os irrelevantes, fazer cálculos, ordenar e agrupar [2]. Existem três tipos de de abstração de métodos simples, segundo *Combi et al. (2010)* [10]:

- **Abstração Qualitativa** significa converter expressões numéricas em expressões qualitativas. Por exemplo, o caso de chamar de 'poucas' quando a quantidade de viaturas mobilizadas pelos bombeiros para um local é de uma a duas.
- **Abstração de Generalização** é o método de tentar agrupar instâncias em classes mais abstratas (que representem parte dos dados). Por exemplo, um "fogo rural"(ou seja, com origem doméstica ou industrial) ou um "fogo florestal" ambos pertencem à classe conceptual de "fogo", que necessitam de carros cisterna.
- **Abstração de Definição** é semelhante à abstração de generalização, mas aplicado em diferentes conceitos de taxonomias.

Para além deste métodos, também foram apresentados uns métodos mais complexos. Estes consideram uma ou mais variáveis, em conjunto, que especificamente integram com

a dimensão do tempo num raciocínio temporal. Existem mais quatro tipos de métodos complexos: abstração de Fundição (ou de estado); abstração de Persistência; abstração de Tendência (ou gradiente, ou taxa); abstração Periódica. Deve-se ter em conta todas estas opções quando se está em processo de análise, sendo portanto um conceito vital nesta área [10].

2.2 Dados Espaciotemporais

Dados espaciotemporais são a junção de dados temporais com dados geográficos, também chamados de temporais, com uma terceira componente alfanumérica que define que tipo de dados estamos a analisar. Quando as visualizações envolvem mapas geográficos para representar a parte espacial chama-se, a essa categoria de visualizações, mapas temáticos.

Para construir visualizações para dados específicos é útil estudar esses dados para que se entenda quais são as suas estruturas inerentes. O estudo das combinações de possibilidades facilita o desenho de propostas de visualizações.

2.2.1 Características de dados Temporais

Quando se fala em dados temporais referimos-nos a dados que tem valores associados a primitivas temporais. Existem vários modelos para dados temporais, uns que se baseiam em dados contínuos, outros em discretos. Iremos falar brevemente sobre esses neste capítulo. Começando pela modelação do tempo em si que pode ter as seguintes estruturas:

- **Escala:** ordinal, discreta ou contínua. Na categoria ordinal, apenas relações relativas de tempo são consideradas, como por exemplo "antes" ou "depois" de um evento (não sendo obrigatoriamente fornecido uma data para esse mesmo evento também). No domínio de dados discretos são consideradas distância temporais, baseados na unidade de tempo mais pequena possível. É o caso normalmente dos segundos ou milissegundos em sistemas UNIX. Os modelos temporais contínuos são caracterizados por um possível mapeamento a números reais, como no caso de entre dois pontos temporais, ou a existência de outro ponto no tempo. Este modelo também é conhecido por tempo denso.
- **Escopo:** baseado em pontos ou intervalos. Tempo baseado em pontos são momentos discretos com extensão temporal igual a zero. Ao invés do tempo baseado em intervalos, onde são consideradas regiões entre dois pontos no tempo, ou seja, extensão temporal maior que zero. Se considerarmos o exemplo de uma data, como 30 de março de 2020, em termos do modelo de tempo esta data poderá referir-se às 00:00:00 de 30 de março ou, no caso do modelo baseado em intervalos, a região de tempo com início nas 00:00:00 de 30 de março até 23:59:59 do mesmo dia.
- **Arranjo:** linear ou cíclico. O tempo como o costumamos experienciar é linear. Ou seja, cada momento tem um único predecessor e sucessor. É o exemplo dos anos,

cujo o número vai crescendo assim que passa o anterior, mas aplicável a outras granularidades de tempo. No entanto, como sabemos, a periodicidade é muito comum. Este é o exemplo dos meses ou das estações do ano (que se repetem anualmente).

- **Ponto de Vista:** ordenado, ramificado ou múltiplas perspectivas. Este aspeto é relacionado com a vista de tempo que são modeladas. O domínio ordenado considera que eventos se passam uns após os outros. Isto, num nível mais detalhado, pode ser totalmente ordenado, ou parcialmente. A diferença é que enquanto o domínio totalmente ordenado considera que os eventos acontecem um de cada vez, no parcialmente ordenado estes eventos podem sobrepor-se ou acontecer em simultâneo.

Para além das características do tempo há que considerar outras perspectivas de o ver. Em geral, é útil ver os dados a vários níveis de detalhe. Falaremos então de abstrações temporais importantes para a sua análise:

- **Granularidades e Calendário:** nenhuma, única ou múltiplas. Um calendário é um sistema de múltiplas granularidades em malha. Granularidades descrevem mapeamentos de valores temporais para grandes ou pequenas unidades conceptuais. Isto é o caso de minutos, horas, dias, semanas, meses, etc. Se ambas a granularidade de um calendário forem considerados, diz-se ser múltiplas granularidades. Se todos os valores forem dados numa única medida temporal, como por exemplo horas, diz-se ser granularidade única. É de notar que as granularidades temporais têm a peculiaridade de não manterem as igualdades presentes num grão grosso não se mantêm num grão mais fino e vice-versa (e.g. dois dias podem pertencer à mesma semana, mas não ao mesmo ano).
- **Primitivas Temporais:** instante, intervalos ou período³. Estes conceitos relacionam dados com o tempo. As primitivas podem ser classificadas em duas categorias: ancoradas (ou absolutas) e primitivas desancoradas (relativas). Instantes e intervalos pertencem ao grupo das ancoradas, porque são localizadas num ponto fixo do domínio temporal. Os períodos são relativos, visto que não têm posição no tempo fixa. A maior importância deste aspeto são as todas as relações e combinações possíveis destas primitivas, que são úteis no estudo de dados.
- **Determinação:** determinadas ou indeterminadas. A incerteza é outro aspeto em ser em consideração. Há sempre algum grau de incerteza nos dados capturados, visto que nenhum é capaz de tirar informação completa ou exata sobre especificações do tempo, ou se as primitivas forem convertidas de uma granularidade para outra. Para se lidar com este aspeto, foi introduzido a incerteza que captura se os dados são determinados ou indeterminados. Indeterminados quer dizer que não se tem ideia de quando a informação foi capturada, ou de quando foi exatamente capturada. Isto

³Tradução do inglês. Original: *span*

é o caso de dados como quando é que a terra foi formada, ou de planos futuros (e.g. o projeto terminará daqui a duas a três semanas).

Um conjunto de dados não é só formado com dados temporais. Normalmente estes são acompanhados com outras dimensões com outras informações alfanuméricas (com significados semânticos diversos). Falaremos então agora de algumas características desses dados no contexto temporal:

- **Escala:** esta pode ser quantitativa ou qualitativa. Dados quantitativos referem-se a dados com métricas numéricas, contínuas ou discretas. As escalas qualitativas são também discretas por natureza e podem ser ordenadas, chamadas de ordinais (por exemplo a escala de Likert), ou desordenadas, ou seja nominais.
- **Estrutura de Referência:** abstrata ou espacial. Faz também sentido separar os dados abstratos dos espaciais, ou seja, os que respondem à pergunta "Onde?", dos restantes. Isto tem a ver com a peculiaridade de dados espaciais serem fáceis de identificar e terem uma área de estudo bastante desenvolvida. Os dados espaciais são melhor visualizados mapeamento as coordenadas ao próprio mapa. A categoria de dados abstratos pode ser uma infinidade de possibilidades (médica, engenharia, organização, etc) e não é fornecida um mapeamento espacial à priori.
- **Tipo de Dados:** eventos ou estados. É um critério que se refere a se estamos a lidar com eventos temporais, ou estados. Eventos são marcadores de mudanças de estado, por exemplo uma carro a partir do estacionamento. Estados referem-se a fases de continuidade entre eventos, como por exemplo o carro estar em andamento. São conceitos semelhantes, mas devia ser claro se os dados se referem a um ou a outro, ou se até uma combinação dos dois, quando estão a ser visualizados.
- **Número de variáveis:** uni-variadas ou multi-variadas. É um critério que conta o número de variáveis que dependem do tempo, visto que para diferentes números de dimensões, existem variadas visualizações apropriadas (com avanços mais significativos nas visualizações para dados uni-variados). Se cada primitiva temporal estiver associada a apenas um único valor de dados, é uni-variada. Senão, a primitiva está relacionado com vários valores dos dados.

Aspetos relacionados com variáveis temporalmente dependentes não foram ainda extensivamente estudadas, mas existem duas definições importantes: o tempo interno refere-se a dimensões cujo dados são inerentes ao modelo de dados, e o tempo externo quando é extrínseco ao modelo de dados.

- **Dados Estáticos Não-Temporais** - Se ambos os tempos internos e externos forem compostos de um único elemento temporal e os dados forem completamente independentes do tempo.

- **Dados Estáticos Temporais** - Se o tempo interno contem mais que uma primitiva temporal, enquanto o externo só contem uma, pode-se considerar que os dados são dependentes do tempo.
- **Dados Dinâmicos Não-Temporais** - Se o tempo interno conter apenas uma primitiva temporal e o externo for composto de múltiplas, então os dados dependem do tempo externo. Noutros termos, os dados podem mudar durante o tempo.
- **Dados Dinâmicos Temporais** - Se ambos os tempos interno e externo forem compostos de múltiplas primitivas temporais, então os dados podem ser considerados bi-temporalmente dependentes. Ou seja, os dados contem variáveis dependentes do tempo (interno) e do estado atual das mudanças de dados no tempo (externo).

2.2.2 Características de dados Espaciais

Estamos habituados a ver o espaço desde o local onde estamos, a vermos a topografia do terreno à nossa volta para nos podermos mover nela. Sabemos que desde da época pré-história usamos mapas, na altura pintavam em paredes onde se podiam ver os arredores dos nossos ancestrais, para nos ajudar a visualizar, explorar e planejar [11]. Nos últimos anos esta área tem continuado a crescer com a quantidade de dados espaciais disponíveis. A capacidade gráfica também tem aumentando, abrindo mais oportunidades para soluções diferentes de representações de mapas. O foco com esta tecnologia tem sido o realismo, quando na história anteriormente o foco eram abstrações, que tem vindo a evidenciar que um bom balanço entre os dois é necessário.

2.2.2.1 Representação do espaço

Os dados espaciais são referências de pontos ou regiões no mundo real. Podem também ser valores, medições, observações ou outras anotações humanas a que nos referimos a características do terreno [12]. Os cartógrafos costumam dividir o espaço de mapas em duas categorias [13]:

- **Mapas Gerais ou de Base:** o foco destes mapas enfatizar a localização de objetos espaciais do mundo real, tais como nos mapas topográficos, hidrográficos e físicos;
- **Mapas Temáticos:** o foco é evidenciar possíveis padrões dos atributos ligados ao espaço descrito no conjunto de dados. Um exemplo destes podem ser mapas económicos, políticos, climáticos ou de estradas, etc. Estes podem ter como fundo uma camada com um mapa geral.

Estas duas categorias não são mutuamente exclusivas, sendo que pode existir um mapa geral com padrões (por exemplo um mapa do evacuação de águas da chuva). Na secção 2.2.2.2 focar-nos-emos nos mapas temáticos devido à sua importância.

A representação gráfica espacial destes mapas pode ser total ou parcial da superfície terrestre. Estes tornam o que é uma superfície quase esférica, a Terra, numa representação plana da superfície. A essa transformação denomina-se de projeção cartográfica. Esta projeções podem distorcer a superfície, portanto são um aspeto a ter em conta ao trabalhar com cartografia.

2.2.2.2 Mapas Temáticos

Estes mapas, como o nome indica, são construídos com base na informação presente num mapa geral (podendo este não estar visível), sendo outros dados sobrepostos sobre esse mapa. Em geral, essa informação é baseada numa variável visual, ou seja, existe um elemento gráfico que representa um fenómeno (dos dados) no mapa. Estes dados são baseados num tópico real, como a distribuição de entregas de uma empresa ao longo de um ano, e é este aspeto dos mapas temáticos que lhe dá o nome. Demonstra a flexibilidade de tópicos que podem ser explorados espacialmente via esta técnica. Cada tópico costuma ser representado de maneira diferente, visto que é provável que tenham objetivos e restrições diferentes.

Como se pode ver na figura 2.1, existem várias opções de variáveis visuais que podem ser mais apropriadas que outras dependendo do tipo de dados que representam e a sua importância. Estes mapas podem ser classificados pelo número de variáveis visuais utilizadas: univariado; bivariado e multivariado. Mapas temáticos podem ser usados de três maneiras básicas [12]:

- Dar informação específica sobre localizações específicas;
- Dar informações gerais sobre padrões espaciais;
- Comparar padrões entre dois ou mais mapas.

É devido a esta flexibilidade que os mapas temáticos mostram o seu forte, podendo adaptar-se a vários níveis de complexidade, como é o caso da diversidade dos dados espaciotemporais e a diversidade dos seus temas.

2.2.2.3 Variáveis Visuais

O criador de variáveis visuais é um cartografo francês, Jacques Bertin's, que publicou um trabalho chamado "Semiologie graphique" em 1967 com a análise detalhada da representação gráfica de dados de duas dimensões. Semiologia é o estudo de símbolos com o intuito de traduzir informação para símbolos. Isto quer dizer que o objetivo de uma variável visual é codificar um atributo com variados valores em uma propriedade visual, como o tamanho de um símbolo. Desde então Bertin está desatualizado, visto que se focou em gráficos que podiam ser impressos em papel, sem animações, navegação ou gráficos 3D. É de notar que Bertin obviamente não criou esta semiologia exclusivamente

para uso em mapas, portanto nem todos os seus aspetos funcionam com esta categorias de visualizações. O cartografo distingue três tipos de marcas diferentes:

- **Ponto** (localização): um ponto representa a localização num plano que não tem teoricamente comprimento ou área e é independente do tamanho e das características da marca usada para o tornar visível no mapa. Pontos podem variar em todas as variáveis visuais.
- **Linha** (limite, trajetória, ligação): uma linha representa um fenómeno no plano que tem comprimento mensurável, mas não têm área. É independente da largura e das características da marca usada para o tornar visível no mapa.
- **Região** (cidade, pais): uma área represente algo no plano que tem tamanho mensurável. A marca visível é uma representação totalmente dependente da área que representa. É de notar que a área pode mudar de posição, mas não de tamanho, forma ou orientação sem mudar o significado da região.

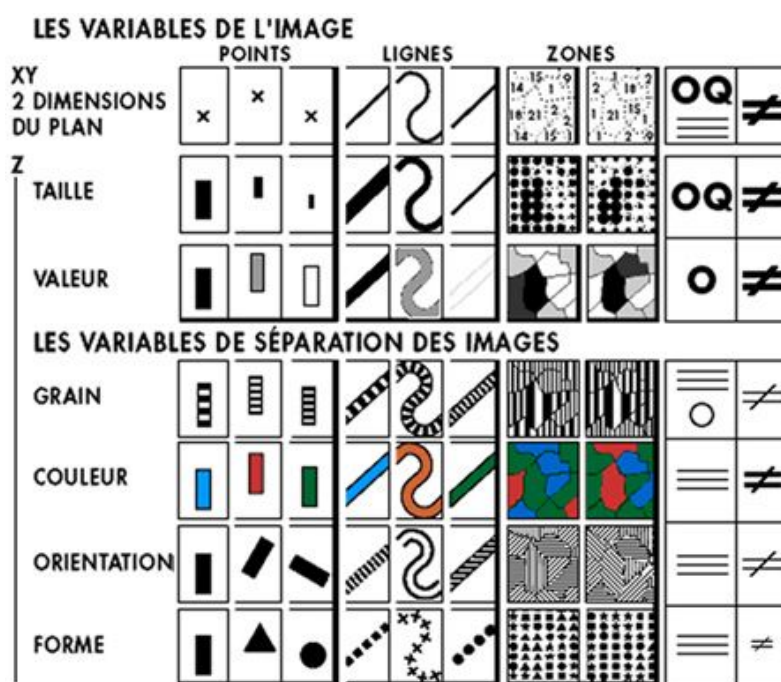


Figura 2.1: Variáveis visuais sugeridas por Bertin em 1983. Figura retirada de [14].

Na figura 2.1 são apresentadas seis das oito **variáveis visuais** inicialmente propostas: posição, tamanho, saturação, orientação, cor e por fim textura. Existem quatro categorias perceptuais que influenciam o mapeamento. Os dados podem ter categorias (associatividade e seletividade), ordem e quantitatividade. Nos categóricos podemos querer mostrar uma classe, que é a associatividade, ou evidenciar uma subclasse, que é a seletividade. Por exemplo, num conjunto de dados sobre peixes, mostrar a totalidade do conjunto e

todas as localizações das espécies de peixes encontradas para a associatividade, ou só as localizações dos salmões na seletividade.

Isto é importante porque para cada **categoria percetual** existe uma variável visual geralmente mais adequada. Poderá haver exceções à regra. A variável posição é a melhor para todos os objetivos, contudo esta pode já estar utilizada. Nesse caso, veremos então quais as indicações gerais. Para a associatividade as formas, cor e posição são apropriadas. Para a seletividade é recomendado cor e tamanho. Para ordem e quantitatividade é recomendado o tamanho. Existem duas métricas, segundo o ranking de Mackinlay's [14], para avaliar a eficácia de uma codificação de variáveis visuais: expressividade e eficácia. Um mede o entendimento e o outro mede o tempo que demora o utilizador a perceber a visualização. Na figura 2.2 pode-se ver uma lista detalhada das combinações recomendadas para maximizar a eficácia, ordenadas por decrescente. As tarefas a cinzento nessa figura não são relevantes para os tipos de dados relevantes.

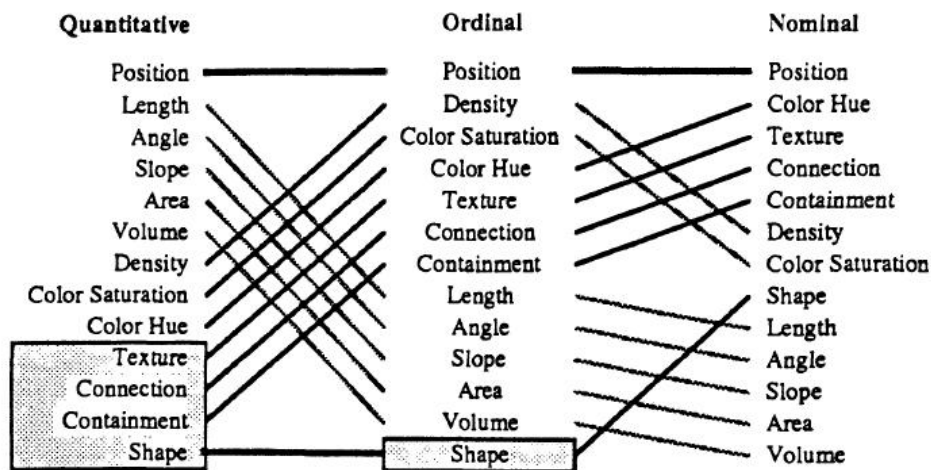


Figura 2.2: Ranking de tarefas percetuais. Figura retirada de [14].

Por último, vamos abordar brevemente sobre as cores e a importância da escolha de paletas. As capacidades da variável visual cor é a **distinção** para agrupar elementos, para **realçar** e marcar elementos ou área, **replicar** fenómenos natural (temperatura, quente a vermelho e frio a azul) e ainda **expressão do humor** (associação cultural de cor a certos estados de espírito).

Para uma compreensão rápida de todas as cores de uma paleta sugere-se que se use apenas cinco a oito cores no máximo. Quando estamos a desenhar uma paleta de cores, por exemplo para um mapa coropleto, devemos tentar maximizar a distinção (distância percetual de cores), ponderar também a maximização de cores com nomes únicos para facilitar a discussão e ainda considerar de existem objetivos de associatividade ou seletividade (ou se as cores sugerem erroneamente tais agrupamentos). Para dados nominais, as cores devem usar a mesma luminosidade. Para dados quantitativos ordinais são usadas escalas de cores, sendo que para criar estas escalas preserva-se a ordem e

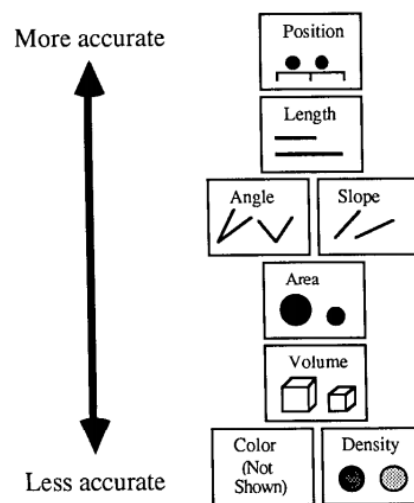


Figura 2.3: Ranking da precisão de cada variável visual para as tarefas perceptuais quantitativas, como verificado empiricamente por Cleveland and McGill. Figura retirada de [14].

as diferenças de cor observadas devem ser proporcionais aos dados originais. Para este último objetivo é usado a interpolação de Beziér, ou correção de luminosidade (disponível em ferramentas online como o Color Brewer ou o Chroma.js). Têm que ter cor inicial e final óbvias, sem ciclos. As diferenças de tonalidade devem ser reservadas para áreas críticas, que tomam proveito de serem um processo pré atencional. Também se deve ter em consideração daltonismo.

2.2.3 Exemplos de Visualizações de dados Espaciais

Nesta secção pretendemos mostrar alguns dos tipos básicos de dados espaciais e como são representados em mapas. Para isso foram selecionados um exemplo de mapa temático que use uma das marcas referidas por Bertin, ou seja, pontos, linhas ou regiões.

- **Mapa de Pontos:** Esta técnica usa pontos para mostrar a presença de um atributo ou ocorrência, mostrando um padrão espacial e densidade relativa. Os pontos podem representar uma ou mais ocorrências. Apropriado para contagem de dados, ou para mostrar vários conjuntos de dados (utilizando formas ou cores para diferenciar) [15]. Na figura 2.4 encontra-se um exemplo feito por John Snow, em 1854, cujo cada ponto representa a residência de um paciente infetado com cólera. Este mapa permitiu observar que existe uma maior densidade à volta de uma bomba de água, indicando assim que este era o poço contaminado e a fonte da epidemia;
- **Mapa de Símbolos Graduados:** este tipo de mapa usa símbolos que variam em tamanho para mostrar os valores quantitativos dos dados. Requer dados com pontos/localizações. É um mapa adequado para usar quando há muita variação e um grande domínio de dados. O objetivo é mostrar a magnitude de fenómenos nos

dados [15]. Na figura 2.5 encontra-se um exemplo que mostra o total de mortes por doenças cardiovasculares por município em Michigan entre os anos 2004 a 2008. Este permite comparar superficialmente a quantidade de incidentes entre os municípios e ainda notar quais os que quebram o padrão normal e tem um excesso de incidentes, em relação ao que percebemos ser a média. Por curiosidade chamo à atenção o facto que os aumentos nesta figura coincidem com aumentos da população residente no município;

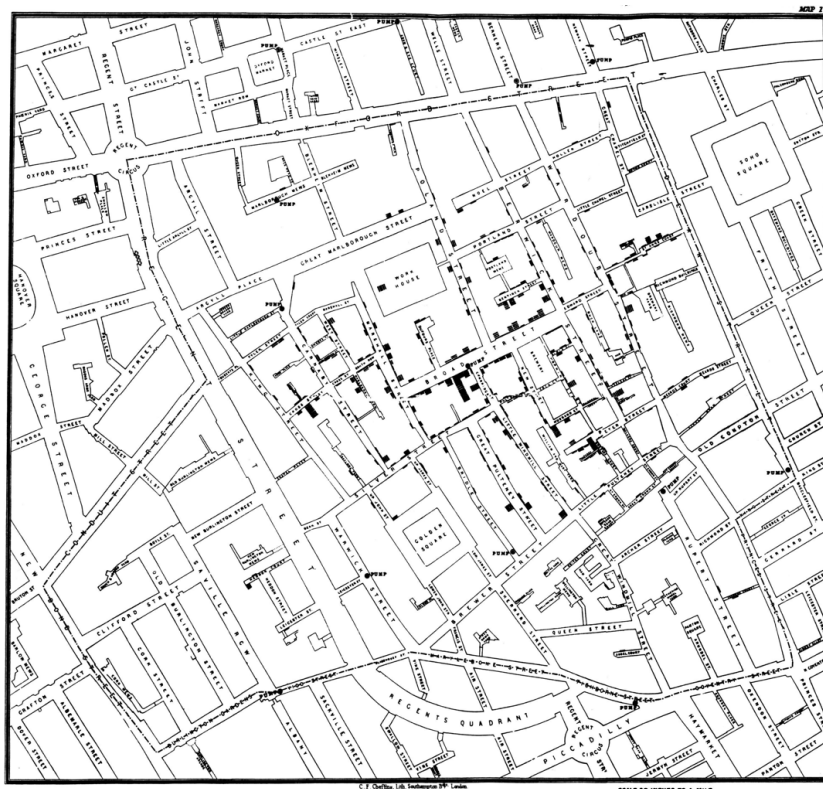


Figura 2.4: Mapa de pontos da morada de pacientes a sofrerem de cólera em Londres no ano 1854, feito por John Snow.

- **Mapa Coropleto:** Este mapa temático requer regiões geográficas, onde as transforma em variáveis visual através do uso da cor, tonalidade ou padrão em relação a um valor (como uma escala por exemplo) [15]. Este tipo de mapa é um dos mais usados para dados espaciais e existem várias variantes do mesmo. Um caso de uso popular é a visualização do alinhamento político das regiões de um país, através da contagem de votos da última eleição do país. Como no exemplo, como se pode ver na figura 2.6, mostra a média anual de mortes por causas cardiovasculares em adultos com mais de 35 anos nos Estados Unidos, ao nível dos municípios nos anos 2015 a 2017. Este mapa mostra bem, primeiro onde estão exceções visuais de cores mais escuras, *outliers* nos municípios à sua volta, e em segundo lugar permite visualizar onde se concentram os municípios este problema;

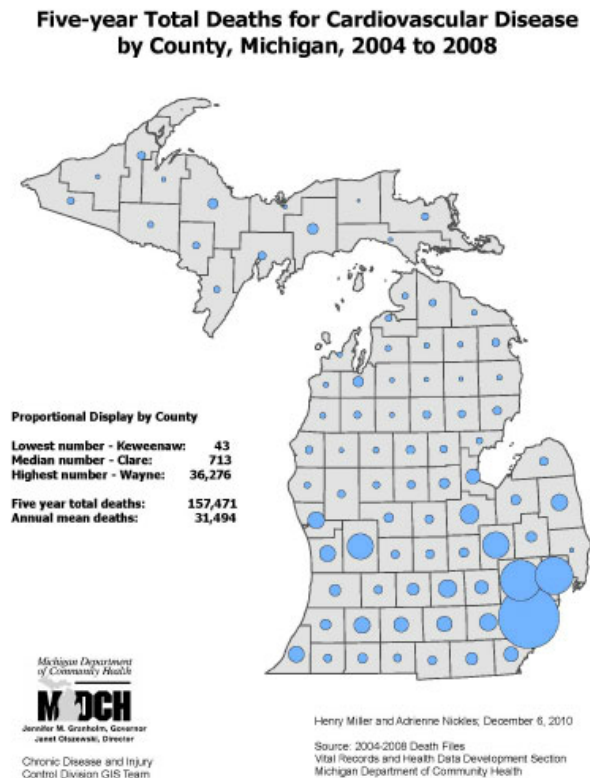


Figura 2.5: Mapa de símbolos graduados que mostra o total de mortes por doenças cardiovasculares por município em Michigan, entre os anos 2004 e 2008. Figura retirada de [15].

- **Mapa Isopleth:** Um mapa de contornos que descreve informações contínuas, como por exemplo a precipitação ou outros dados meteorológicos. Os dados são representados usando linhas que conectam pontos de igual valor numérico que criam a abstração de regiões entre os intervalos das linhas, e consequentemente os valores que as definem. Na figura 2.6 encontra-se um exemplo com a precipitação na Coreia do Sul, onde podemos observar que as chuvas são mais ou maiores na zona norte, decresce no centro do país e volta a aumentar um pouco na zona sul.

2.3 Levantamento de Visualizações de dados Espaciotemporais

No intuito de perceber melhor o estado da arte atual é importante estudar algumas das visualizações e mapas temáticos já existentes. Estes indicam concretamente que técnicas resultam e para que áreas disciplinares são mais utilizadas. Também modelam a expectativa do utilizador, visto que quando mais comuns algumas técnicas são, melhores estas conseguem interpretar novas visualizações que se baseiam nos mesmos princípios. De

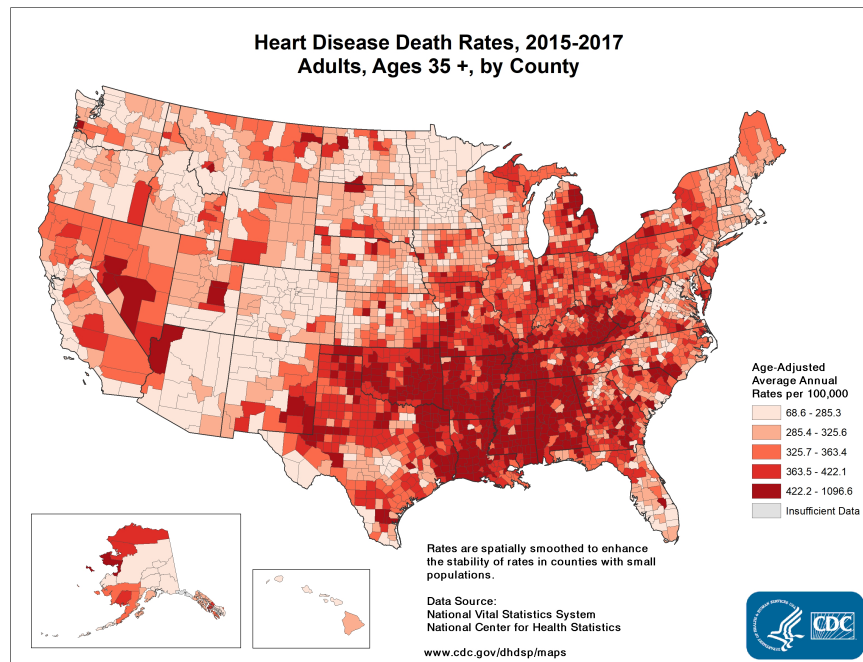


Figura 2.6: Mapa coropleto a média anual de mortes por causas cardiovasculares nos Estados Unidos, por município nos anos 2015 a 2017. Figura retiradas de [15].

seguida, será apresentada uma lista selecionada de mapas temáticos aplicáveis a dados espaciotemporais considerados importantes.

2.3.1 Pencil Icons

Esta técnica foi desenvolvida por Tominski et al. 2005 para visualizar dados espaciotemporais multivariados. Podemos ver na figura 2.8 (a) um exemplo desta técnica. É chamado assim devido à semelhança dos lápis prismáticos às formas usadas para representar os dados, que usam com um bico cônico como pino para marcar um lugar, e um tubo hexagonal com várias marcas para mostrar as variáveis temporais, sendo que o eixo temporal estende-se ao longo do comprimento do lápis, para na direção do topo do lápis. Estes lápis são posicionados em localizações onde existem dados, mostrando a informação disponíveis, que comunica logo a dependência espacial dos dados. São aplicados às visualizações das faces um mapeamento de cores para mostrar os diferentes valores. Para lidar com omissão e oclusão de dados, técnicas de interação e navegação são necessárias, podendo-se rodar um lápis (que afeta todos os outros da mesma maneira através de *linking*) ou o ambiente.

2.3.2 Helix Icons

Ícones em Hélice é semelhante aos *Pencil Icons*, descrita por Tominski et al. em 2005, mas focada em enfatizar a natureza cíclica de dados espaço-temporais. Podemos ver na figura 2.8 (b) um exemplo desta técnica. Num mapa são posicionadas, na sua localização espacial

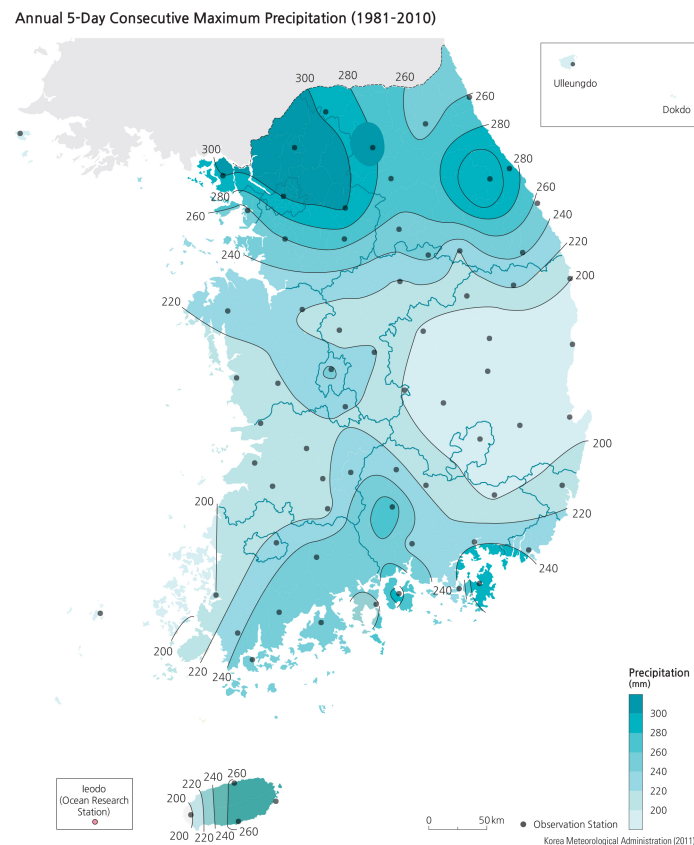


Figura 2.7: Mapa *isopleth* com a quantidade de precipitação máxima anual num período de 5 dias consecutivos na Coreia do Sul, entre 1981 e 2010.

correspondente, uma hélice cilíndrica que é composta e desenhada a partir de várias fita. Cada fita representa uma variável temporal, e cada segmento da fita representa um instante ou intervalo de tempo usando mapeamento de cores. A fita pode ser subdividida em sub-fitas mais estreitas para mostrar mais variáveis temporais ao mesmo tempo (cada fita representa unicamente uma variável). Esta técnica sobre a normal obstrução que visualizações 3D sofrem, portanto, interação e navegação é necessária. Existem também técnicas interativas para ajudar os utilizadores a escolherem o número apropriado de segmentos por ciclo, para que os padrões periódicos sejam mais evidentes. Esta técnica comunica logo a dependência espacial dos dados.

2.3.3 Data Vases

Os vasos de dados (*Data Vases*) é uma técnica multivariada focada em mostrar a progressão temporal de variáveis. Quando originalmente descrita por Thakur and Rhune (2009), foi sugerida duas alternativas: 2D e 3D. Para construir um vaso 2D, imaginemos um gráfico de linhas com a região preenchida que é depois espelhado no eixo temporal, criando uma forma simétrica. É esta simetria criada, pelas séries temporais, que têm a aparência de vasos de olaria tradicionais os quais deram assim o nome à técnica de visualização.

2.3. LEVANTAMENTO DE VISUALIZAÇÕES DE DADOS ESPACIOTEMPORAIS

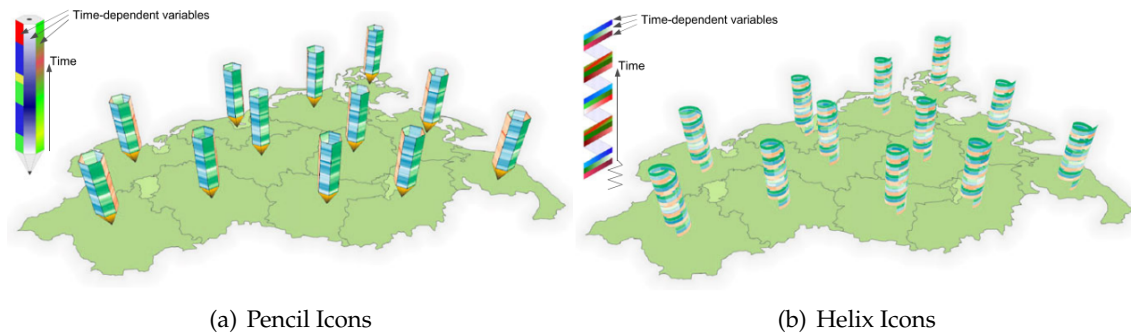


Figura 2.8: Técnicas de visualização que fazem uso de extrusões gráficas de pontos espaciais. Figuras retiradas de [2].

Estas séries temporais podem depois ser atribuídas uma escala de cores e mapeadas a outras variáveis do conjunto de dados, mostrando assim mais uma dimensão dos dados. Se este *dataset* for espaciotemporal, esse mapeamento pode ser ligado ao componente geográfico composto por discos que representar uma primitiva temporal e as suas variáveis associadas são mapeadas ao tamanho e cor desse mesmo disco. Se o conjunto de dados for grande, a interação torna-se vital nesta técnica, tal como filtragem ou seleção. Na figura 2.9 podemos ver dois exemplos com dados diferentes. É possível observar no topo a versão 2D, onde se pode referir ao mapa para associar as cores à localizações a que os vasos se referem, ou a alternativa 3D embutida num cubo espaciotemporal.

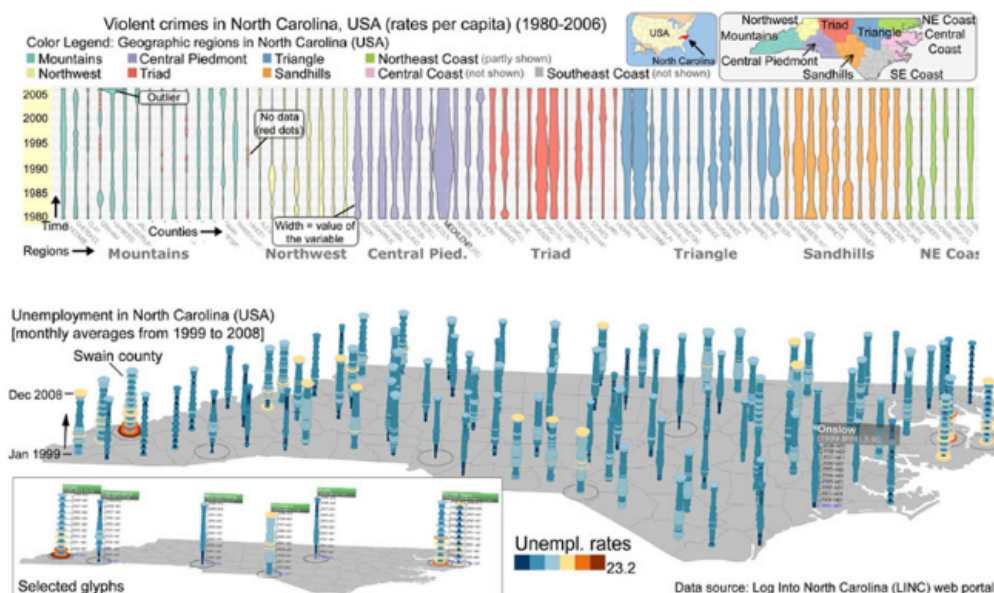


Figura 2.9: Dois exemplos de gráficos com vasos de dados, que mostram número de crimes violentos e a taxa de desemprego na Carolina do Norte. Figura retirada de [2].

2.3.4 Wakame

O Wakame é também um conjunto de técnicas de visualização multivariada [16]. Foi inventada por Forlines e Wittenburg em 2010, descreve um sistema interativo para visualizar especificamente dados espaciotemporais. Os aspetos temporais são mapeados para glifos multivariados chamados de Wakame (nome derivado de uma alga marinha comestível com o mesmo nome, comum nas costas rochosas japonesas). Podemos imaginar estas torres como um *radar chart* que foi extrudado para três dimensões, mais o eixo temporal que cria a sua altura, que permitem mostrar tendências temporais nos dados nas três variáveis escolhidas. Estas podem ser sobrepostas num mapa para facilitar a compreensão espacial.

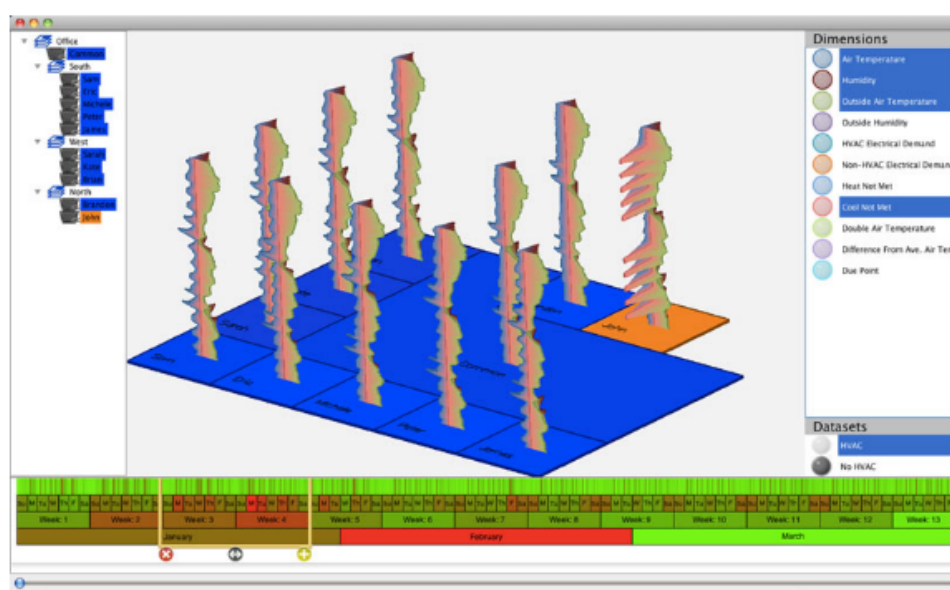


Figura 2.10: Gráfico Wakame e a interface do sistema. Figura retirada de [2].

Embora o nome da técnica tenha sido inspirado na geometria do artefacto visual no centro do sistema, Wakame é tudo o que vemos na figura 2.10, incluindo barra lateral com a hierarquia dos dados, a lista de atributos dos dados (seleccionável) e na parte inferior uma linha temporal navegável. O mais notável nesta visualização é o seu poder de interação e animações. A câmara é inteligente, ajudando os utilizadores a mostrar perspectivas cuja probabilidade de mostrarem aspetos interessantes é maior. O eixo temporal é interativo, dividido pelas granularidades temporais e mostra logo os valores com uma escala de cores que expõe o quão diferente cada primitiva temporal é das suas vizinhas.

2.3.5 Icons on Maps

Adaptar técnicas existentes e integrar noutras é válido no campo de visualização de dados e isso claro que se estende na área de dados espaciotemporais. Ícones em mapas é uma solução que usa mapas para representar o contexto espacial e coloca imagens reduzidas de

visualizações 2D, os chamados ícones, em partes desse mapa para representar a progressão temporal associada à parte espacial respetiva, que tem que ser fixa e constantes nos dados. Essa componente espacial podem ser regiões, pontos ou outro. A figura 2.11 mostra um exemplo que, para efeitos demonstrativos, usa dois ícones diferentes, os gráficos *ThemeRiver* à esquerda e gráficos *TimeWheel* à direita. A parte norte da figura mostra o processo de otimização de posicionamento de figuras, representado num gráfico de conflitos no qual as linhas vermelhas são potenciais problemas de oclusão devido a sobreposição. Uma variante desta técnica é *Value Flow Map*, criada pelos Andrienko e Andrienko (2004), visto que usa ícones, contudo limita-se apenas a gráficos de silhueta (*Silhouette Graph*) que mostram a variação de valores por tendências positivas ou negativas.

Este tipo de visualização acarreta um problema. Demasiados ícones no mapa criam oclusão, cuja solução passa por otimização do posicionamento dos ícones para evitar oclusão global ou local. O mesmo pode ser mencionado para figuras demasiado grandes na visualização que pode ser necessário devido a não se notar detalhes nos ícones, e a solução é de novo um equilíbrio na granularidade, ou comprometer e analisar o mapa por secções aumentadas.

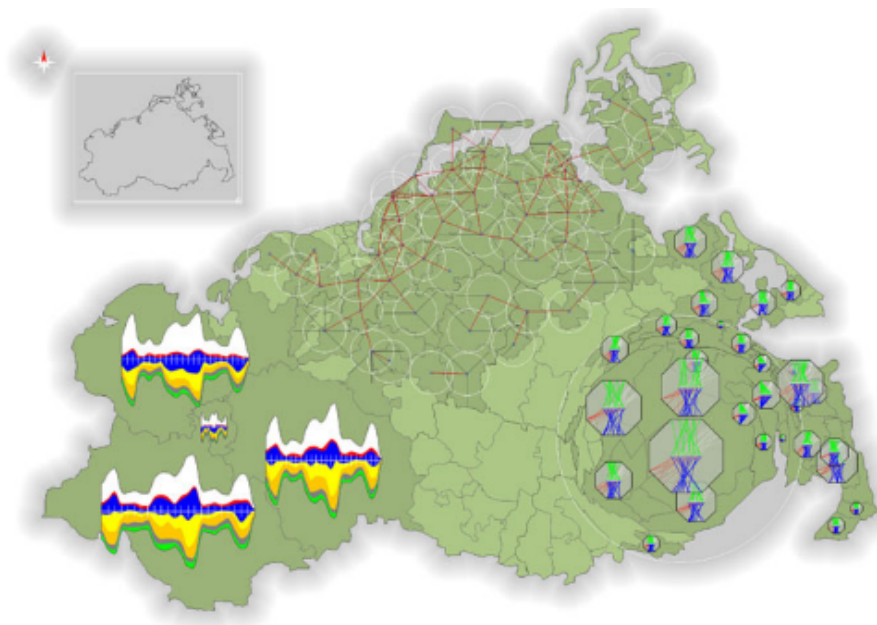


Figura 2.11: Exemplo de um gráfico de ícones em mapas que para efeitos demonstrativos usa vários tipos de gráficos para as de miniaturas de visualizações. Figura retirada de [2].

2.3.6 Value Flow Map

Esta técnica desenvolvida pelos Andrienko e Andrienko (2004) tem como objetivo visualizar a variação de dados espaciais. Na figura 2.12 pode-se ver que o mapa mostra uma miniatura de um gráfico de silhueta para cada área cartográfica com o intuito de representar o comportamento temporal de uma variável nessa área. Aplica-se regularização

temporal (*temporal smoothing*) para que o valor base, onde está posicionado o eixo dos gráficos, seja a média dos valores do intervalo selecionado. É assim possível descartar pequenas flutuações nos dados e apenas as tendências mais acentuadas são visíveis. O nome *Value Flow Map* deriva das transformações de dados que podem ser aplicadas, nomeadamente uma relacionada com o desvio padrão, que efetivamente mostra a diferença dos valores em relação à média em valores positivas ou negativas. Isto torna a evolução dos dados no espaço e no tempo mais clara no gráfico de *silhouette*. Esta clareza é necessária para exploração e análise de dados, permitindo uma boa comparação da evolução temporal em diferentes localizações (tanto semelhanças como *outliers*).

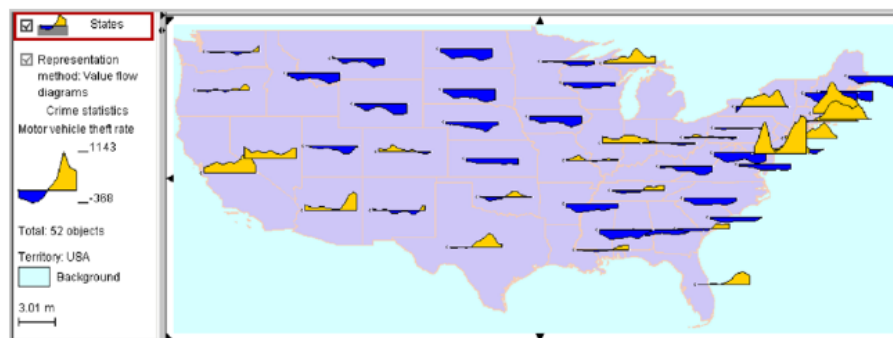


Figura 2.12: Exemplo de um mapa de *Value Flow* sobre crimes nos Estados Unidos. Figura retirada de [2].

2.3.7 Mapas de Anéis

Este mapa de anéis (originalmente *Ring Maps*), como o nome indica, é composto por vários anéis, cada um maior que o da camada anterior, subdividido num número igual de segmentos. Há várias maneiras de utilizar os anéis e os seus segmentos, assim como o espaço no centro destes. Para se visualizar os dados é utilizado um mapeamento de cores, contudo existem outras variáveis visuais que podem ser utilizadas (para outros atributos) como se pode ver na parte direita da figura 2.13 com o relevo. Podemos observar na figura 2.13 dois exemplos. O mapa do lado esquerdo mostra os alertas relativos à saúde (verde, amarelo, laranja e vermelho, por gravidade crescente respetivamente) para alguns códigos postais durante um período de 24 horas. A figura da direita mostra o nível de atividade humana ao longo de um dia, em que cada anel representa uma das 96 atividades listadas no conjunto de dados. Podemos notar imediatamente alguns padrões e a maior vantagem desta visualização é que destaca os padrões cíclicos (que vimos na secção 2.2.1 ser uma forte componente de dados temporais).

Para o caso específico de dados espaciotemporais, o centro do mapa pode ser utilizado para uma nova vista onde se representa os segmentos de um particular intervalo angular da área do mapa, como se pode ver no exemplo da figura 2.13, do lado esquerdo, onde diferentes ângulos mostram dados para diferentes códigos postais. Uma série temporal

para cada região pode ser representada pelos anéis, usando a primeira série para no anel mais central ou interior, e a última série no anel mais exterior.

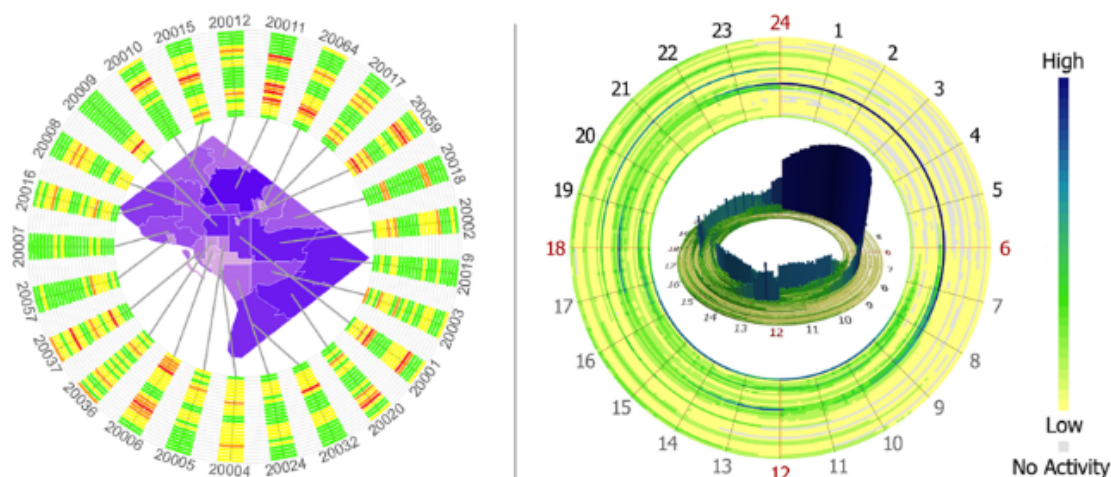


Figura 2.13: Dois exemplos de mapas de anéis. No gráfico da esquerda são alertas em códigos postais na figura da direita, e nível de atividade humana durante um dia no da direita. Figura retirada de [2].

2.3.8 Time-Oriented Polygons on Maps

Shanbhag et al. (2005) é o autor deste método, como o nome indica, é uma técnica para mapas divididos em regiões (polígonos). Existe três variantes desta técnica que se focam nas principais maneiras de representar mudanças em conjuntos de dados espaciotemporais: cunhas, anéis ou fatias. Para todas estas, os dados são representados usando componente de cor (tonalidade, saturação e brilho). Na disposição em cunha a área do polígono é particionada pela ordem dos ponteiros dos relógios em secções radiais iguais ao número de pontos temporais. Na disposição em anel será igual, com o anel mais central sendo o ponto temporal mais recente. E por fim, nas fatias o polígono é dividido verticalmente, e os pontos temporais são ordenados da esquerda para a direita pela progressão temporal.

Estas visualizações foram úteis na repartição de distritos escolares, onde as variáveis eram o ano escolar, estudantes em carência de refeições grátis e notas dos testes, para analisar a alocação de recursos. Podemos ver na figura 2.14 três exemplos desta técnica, todas sobre um conjunto de dados que se refere à população de estudantes no ensino secundário nos anos 2005, 2006 e 2007. Demonstra várias variantes de técnicas para representar os dados nos polígonos.

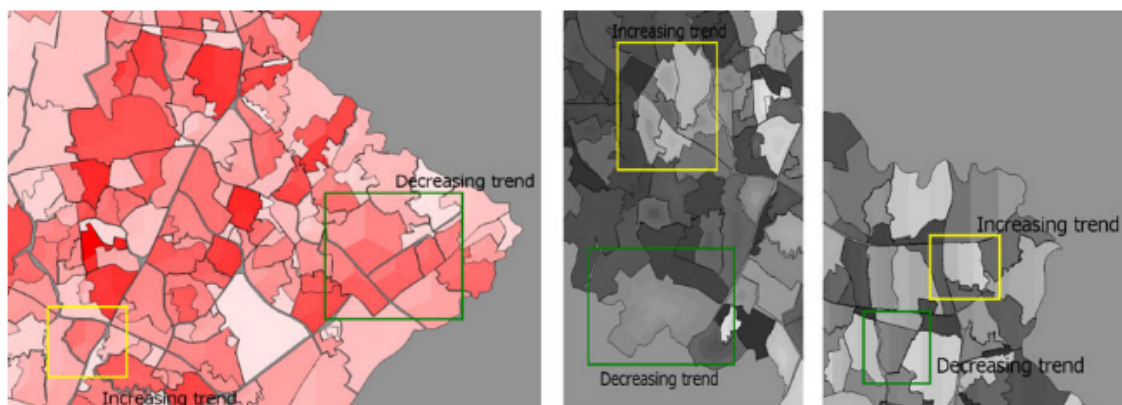


Figura 2.14: Três variantes de Polígonos Orientados Temporalmente em Mapas da esquerda para a direita: cunhas, anéis e fatias. Representam a evolução da população de estudantes no ensino secundário. Figura retirada de [2].

2.3.9 VIS-STAMP

O VIS-STAMP é das poucas técnicas aqui referidas que não é apenas uma visualização de dados, mas sim um conjunto delas, ou seja, torna-a particularmente interessante por ser um *dashboard*. A diversidade de vistas complementa bem a complexidade multifacetada de dados espaciotemporais. O responsável por esta ferramenta é Guo et al. (2006). Esta ferramenta integra métodos computacionais, visuais e cartográficos para análise e exploração dos dados.

Como podemos ver na figura 2.15, na parte inferior esquerda, esta técnica faz uso de um *Self-Organizing Map* (SOM) que é usado para agrupamento multivariado, ordenação e atribuição de cores. O *dashboard* incorpora também, como se pode ver na figura 2.15, uma vista com uma matriz (canto superior esquerdo), vista com mapa (canto superior direito) e um gráfico de coordenadas paralelas (inferior direito), para além da vista para o SOM (inferior esquerdo) anteriormente referida. A matriz representa as regiões nas linhas e o tempo nas colunas. No exemplo da figura 2.15, as regiões são os estados, permitindo assim a observação da evolução temporal de cada região. A vista do SOM permite a visualização e manipulação do mesmo. O gráfico de coordenadas paralelas mostra todos os atributos dos dados. A representação do mapa, da componente espacial, segue uma abordagem de *small multiples* para mostrar uma série temporal que usam as cores como variável visual. A paleta de cores utilizada é a mesma para todas as vistas, mantendo assim a consistência e possibilitando comparações entre as mesmas. Esta técnica tem muitas outras técnicas automáticas e manipulação interativas (como ordenação e classificação) que facilitam a sua utilização e, por consequência, a análise de dados.

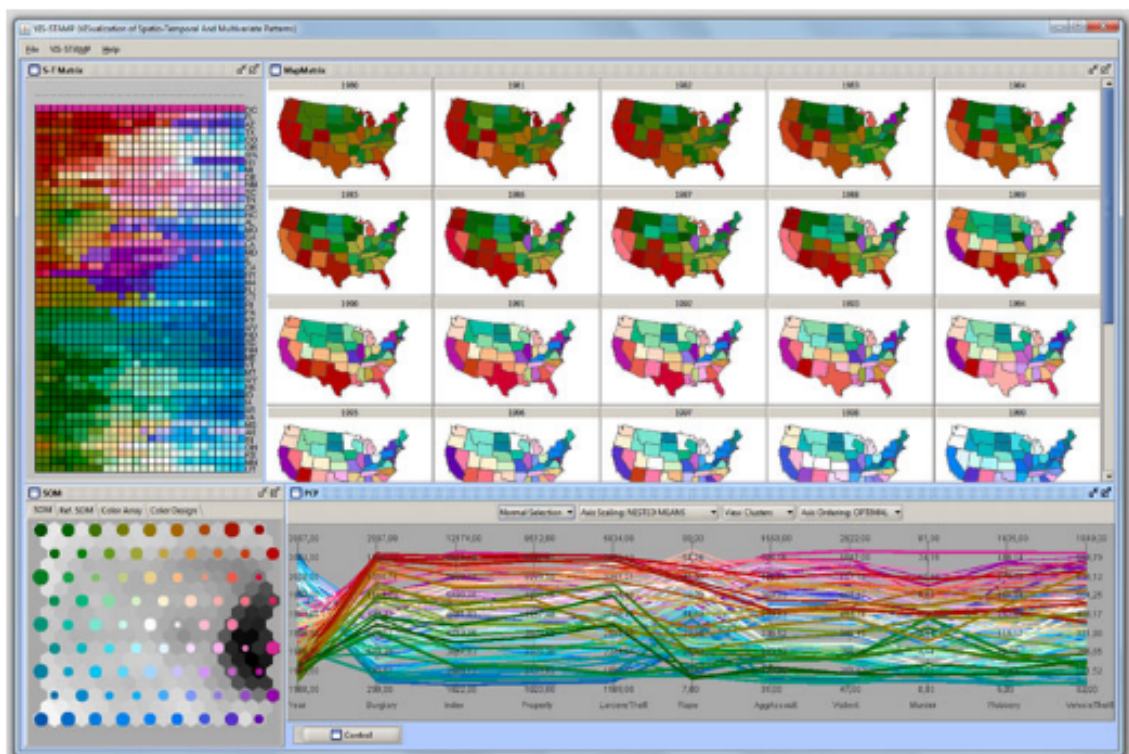


Figura 2.15: VIS-STAMP que mostra dados sobre crimes nos Estados Unidos. Figura retirada de [2].

2.3.10 Cubo Espaço-Temporal

O cubo espaciotemporal[17] (originalmente *Space-Time Cube*) é um conceito com importância histórica para a área de visualização de dados que, tal como o nome indica, combina dados espaciais aos temporais. A ideia básica resume-se a mapear duas dimensões espaciais (um mapa plano) a dois eixos de um cubo virtual de três dimensões a um terceiro eixo mapeado ao tempo. Este modelo normalmente requer interação para permitir aos utilizadores observarem os dados das várias perspetivas necessárias para verem os dados [2].

Existem várias técnicas de codificação para as variáveis visuais que se enquadram dentro deste conceito, que deram origem a variantes do cubo. Pode-se representar objetos no cubo através de marcas dos pontos de interesse, ou construir trajetórias que ilustram os caminhos desses objetos, como no Space-Time Path. Dados adicionais costumam ser mapeados ao tamanho e cor dessas marcas [2]. Uma outra técnica é posicionar várias camadas ao longo do eixo tempo, cada uma representando os dados de um tempo específico.

Podemos ver na figura 2.16 um exemplo que mostra duas dimensões espaciais (latitude e longitude) no eixo do y e no eixo do z. Várias camadas foram mapeadas por cores e estão incorporadas no cubo para visualizar os dados climáticos.

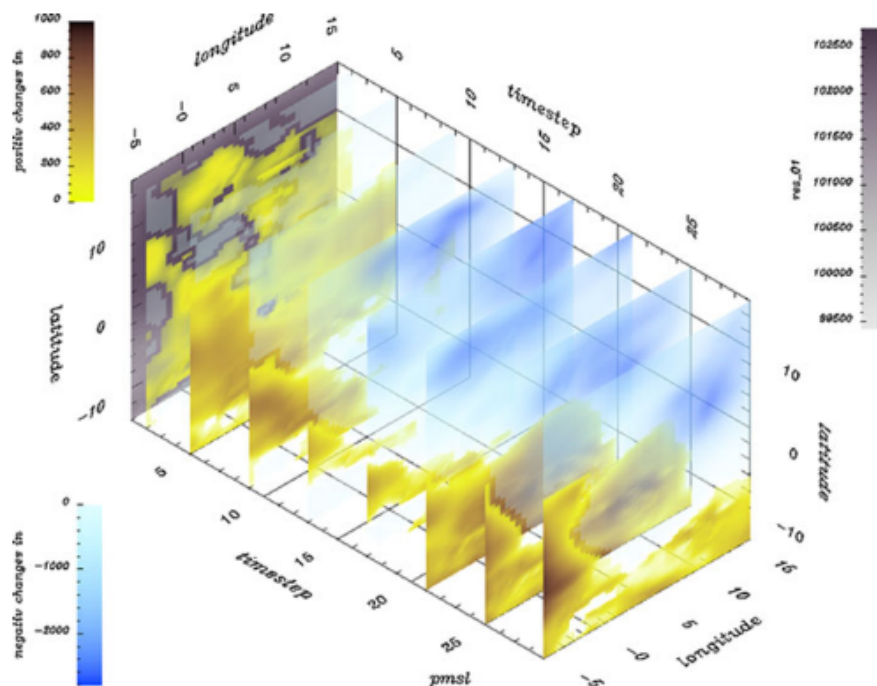


Figura 2.16: Cubo espaço-temporal que mostra dados climáticos. Figura retirada de [2].

2.3.11 GeoTime

GeoTime é um sistema da autoria de Kapler e Wright (2005) que se desenhado para permitir visualizar conjuntos de dados no seu contexto espacial e temporal. É um alternativa, capaz de uma análise dinâmica e interativa, do conceito do cubo espaciotemporal. Como podemos ver no exemplo da figura 2.17, e tal como no cubo espaciotemporal, o eixo temporal também está posicionado verticalmente na direção da extrusão do mapa, ou seja, quanto mais acima a marca, mais recente o evento. São posicionados marcas e, entre as marcas pertinentes ao mesmo objeto ou entidade nos dados, são desenhadas trajetórias.

Podem ser selecionados intervalos temporais pelo utilizador e os eventos são animados ao longo do eixo temporal. Existem outras projeções alternativas que podem focar-se mais noutros aspetos espaciais ou temporais. O mais notável nesta ferramenta é o poder de anotação, narrativa (*storytelling*) e reconhecimento de padrões nos atributos. O processo de anotação e narrativa é automático, e pode ser melhorado pelo utilizador.

2.3. LEVANTAMENTO DE VISUALIZAÇÕES DE DADOS ESPACIOTEMPORAIS

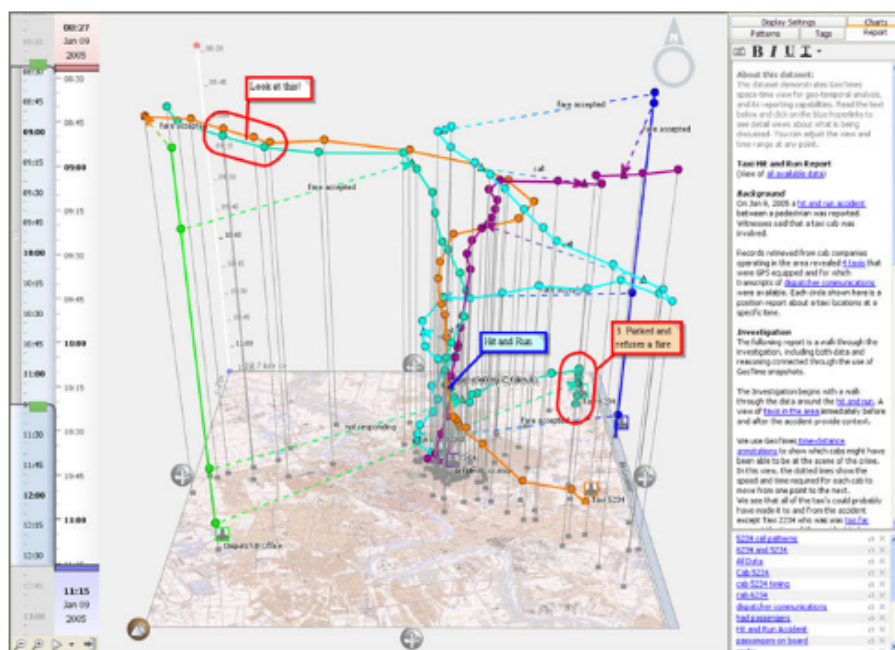


Figura 2.17: Mapa GeoTime que mostra as trajetórias de táxis envolvidos num atropelamento e consequente (delito de) fuga. Figura retirada de [2].

TRABALHO REALIZADO

Neste capítulo serão descritos os passos e o resultante trabalho que foi desenvolvido na produção desta dissertação. A primeira secção descreve a metodologia teórica seguida na abordagem ao tema, assim como a sua aplicação e efeito nas diferentes fases no progresso desta dissertação. Após uma explicação das ferramentas usadas para dar contexto ao trabalho realizado e familiarização do ambiente de desenvolvimento, tal como uma exposição dos conjuntos de dados que foram alvo de análise na secção 3.3, é então explicado o trabalho realizado e o seu impacto na secção 3.4.

3.1 Abordagem

Foi definido inicialmente que a realização desta dissertação ir-se-ia inspirar no processo *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*). Este foi desenhado originalmente para mineração de dados, sendo contudo possível de adaptar a outras áreas. Este processo está dividido em seis passos, ligadas entre si, que podem ser repetidos de maneira iterativa. Na figura 3.1 pode-se ver um diagrama que representa o fluxo que compõe o CRISP-DM, e descreve-se em seguida as suas fases:

Entender o negócio - No início do CRISP-DM começa-se por entender o contexto de onde foram retirados os dados, o seu tema e o negócio a que pertencem. Os dados com que trabalhamos pertencem ao RSB, portanto esta fase envolveu conhecer o tipo de ocorrências, quais as diferenças entre as respostas e tentar perceber com os clientes se havia alguma preocupação ou perguntas importantes.

Entender os dados - O passo seguinte é perceber os dados em si, pois são apenas uma representação simplificada da realidade. Aqui é feita uma familiarização e exploração com o intuito de perceber a representação dos dados e atributos no negócio. Etapas:

1. Reunir os dados iniciais
2. Descrição dos dados
3. Exploração dos dados
4. Verificar a qualidade dos dados

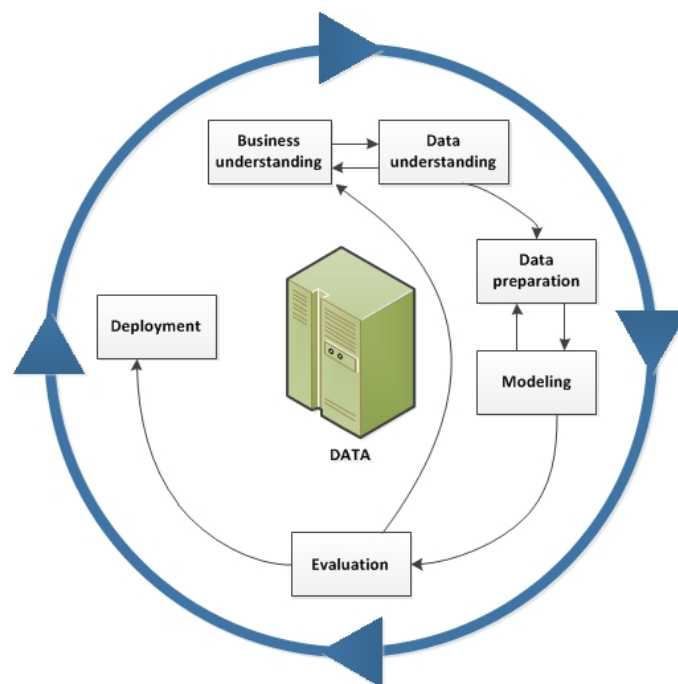


Figura 3.1: Fluxo da metodologia CRISP-DM [18].

Esta fase é onde se produz o relatório de qualidade de dados.

Preparação dos dados - Em termos de mineração de dados este é um dos passos mais vitais, visto que sem formatação apropriada a grande quantidade de dados pode influenciar exponencialmente os tempos de funcionamento de um algoritmo. Em visualização de dados também é importante para reorganizar dados alfanuméricos que um humano interpreta facilmente, não esquecendo que para serem usados em algoritmos precisam de ser reestruturados. Além disso, é nesta fase que são escolhido os atributos não triviais, os dados, é feita a limpeza necessária, integração com outros dados, etc.

Modelação - Neste passo é selecionado uma técnica de modelação, sendo instanciados vários modelos com descrições de como podem ser testados. O objetivo é criar vários modelos com diferentes parâmetro e então classificar para escolher o mais apropriado. Este passo têm muita interação com o anterior, normalmente para adaptar a novas exigências de alguns modelos.

Avaliação - Sendo a penúltima etapa, tem como objetivo testar o modelo da etapa anterior, de maneira a avaliar a sua aptidão para a próxima fase.

Distribuição - Última etapa do processo com o objetivo de distribuir, publicar ou partilhar o trabalho concluído, a quem interessar.

Esta dissertação baseou-se no processo do CRISP-DM. Isto não significa que o processo tenha sido seguido inteiramente, mas sim que os seus passos foram usados como guia, sujeito a mudanças para que se adapte ao contexto e tipo de trabalho. Explicarei de seguida como foi então aplicada esta metodologia.

Embora o objetivo desta dissertação seja criar *dashboards* interativos gerais que funcionem com todo o tipo de dados espaciotemporais, para se trabalhar e avaliar durante o processo é preciso instanciar em dados reais espaciotemporais. Neste caso foram os dados fornecidos pelos RSB. É nos dados que começa o entendimento do negócio, ou seja, entender o contexto dos dados para que estes façam sentido ao desenvolvedor e para que isso se possa traduzir no trabalho. O RSB ajudou-nos nesta parte disponibilizando o seu tempo para reuniões e responderem a e-mails para esclarecer as dúvidas que ocorreram na leitura dos dados. Logo, primeiro explorou-se o conjunto de dados para se obter uma visão de familiarização dos dados, seguido de pesquisas sobre definições próprias do tema. Uma pesquisa exaustiva por atributos para perceber o que cada coluna significa no mundo real e qual o impacto que essa variável tem no negócio.

Para se entender o resto da dissertação vou primeiro explicar qual a linguagem e de maneira simplificada como funcionam os RSB. O Regimento Sapadores Bombeiros chama a todos os eventos que socorre (com ou sem urgência) ocorrências. Estas ocorrências têm vários dados registados, um deles sendo o tipo de evento, por exemplo se é incêndio ou acidente rodoviário, a que chamam Natureza. Esta natureza têm um código associado, mas estes detalhes são melhores explicados na secção 3.3. Cada ocorrência regista o dia e a que hora ocorreu a chamada, quando saíram viaturas do quartel, e outros eventos pertinentes. Também está registado várias partes componentes espaciais, como a morada, e as coordenadas geográficas. O ficheiro de viaturas também regista os veículos que foram a uma ocorrência (referida por ID), e está organizado por nome da viatura (que costuma ser uma mistura da sigla do tipo de viatura, e um número que indica qual a quantidade que o RSB tem ou teve daquela viatura, por exemplo VOPE-1).

Após a familiarização do contexto dos dados foi iniciado o processo de compreensão dos mesmos. Dado o acesso a um conjunto de dados dos RSB de um outro tema, e as novas reuniões com a entidade, foi-nos fornecido um novo conjunto de dados com outro tempo e colunas, nomeadamente o tipo e quantidade de viaturas que assistiram uma ocorrência. Para além destes, também tivemos acesso a vários outros dados, como meteorologia, *geojson* com a geometria das freguesias (pré e pós reorganização de 2013), um *geojson* com as áreas de influencia de cada quartel de bombeiros e informação sobre trânsito com dados do Waze que foram utilizados pontualmente para responder a curiosidades. Como o foco desta tese era a temática dos RSB, logo no início terá sido feita apenas a exploração dos dados relativos à temática, sendo os restantes explorados, entendidos e limpos conforme a necessidade. Este processo foi híbrido entre o Tableau Desktop, para se poderem esboçar ideias básicas e ter uma apreciação fundamental do conteúdo dos dados, e o Tableau Prep onde foram corrigindo os erros encontrados nos dados, ou anotado modificações úteis para se poder apresentar dados de uma maneira que necessitasse um novo formato no conjunto de dados. Este passo foi fechado com uma análise integral e minuciosa dos dados no Tableau Prep que resultou no relatório de qualidade de dados que pode ser consultado no 3.3.

O relatório de dados foi escrito simultaneamente à preparação dos dados. Ou seja,

quando foram notados problemas com a qualidade, estes foram corrigidos de imediato. O objetivo é aumentar a qualidade dos dados e consequentemente das análises feitas com estes. Para além de limpar dados problemáticos, ou erros do sistema, ou ortografia, este processo também modifica os conjuntos de dados e adapta-os de maneira a que resultem em conjuntos de dados mais robustos, e os tornem mais flexíveis e compatíveis com os programas que os vão analisar. Este processo é descrito em detalhe na secção 3.3.

Neste ponto do processo, com familiarização do negócio, dos dados, e confiança na qualidade dos dados pode-se então iniciar o processo de análise dos dados através da criação de ferramentas por modelação. Neste tema, o termo modelação é usado muito levemente, devido ao trabalho ser muito prático e se apoiar em iterações fundamentadas em *feedback*. Há pouco uso para modelação ou planeamento prévio excessivo, considerando que na iteração seguinte se torna obsoleto. É assim preferível manter o ambiente flexível.

No entanto, é aceite que para se desenvolver software é necessário analisar o domínio do problema que os utilizadores querem que a ferramenta ajude a navegar [2]. Existe portanto um mínimo de modelação por ferramenta, especificamente para se definir os objetivos e requisitos para que as iterações possam seguir nessa direção apenas. Em visualização de dados é comum este domínio de problemas serem as questões, em texto informal, a que se querem saber as respostas sobre os dados. E explorar a vertente das questões analíticas foi uma das partes importantes desta dissertação, visto que iriam guiar o trabalho a realizar. Como queríamos que estas ferramentas funcionem com outros dados, outras temáticas, as questões que se geraram são gerais. Estas podem ser lidas na seguinte lista:

Q1.G Quais as tendências ao longo dos anos?

Q1.1. Quais os valores absolutos?

Q1.2. Quais os valores relativos ao total de pedidos?

Q1.3. Qual os valores relativos entre sub-naturezas?

Q1.4. Qual a média móvel de 7, 14 e 21 dias por ano? Existem diferenças entre os anos?

Q2.G Existe sazonalidade? Padrão repetidos? Nas seguintes diferentes granularidades:

Q2.1. Meses

Q2.1.a. Existe um mês mais acentuado que os outros?

Q2.4. Dia da Semana

Q2.5. Horas do Dia

Q2.5.a. Fim de semana

Q2.5.b. Dia normal

Q3.G É semelhante entre as freguesias? Existem comportamentos diferenciados?

Q3.1. Comparação entre as freguesias com as questões da Q1

Q3.2. Comparação entre as freguesias com as questões da Q2

Q4.G Existem *hotspots* na cidade para este de chamadas?

Q4.1. Qual a densidade espacial Global das chamadas?

Q4.2. Qual a densidade espacial Anual de cada ano dessas chamadas?

Q4.2.a. Como se compara essas densidades entre os anos? Existem diferenças?
Quais?

Q5.G Há padrões nos Tempos de Resposta?

Q6.G Há padrões nos Tempos Totais de resolução da ocorrência?

Q7.G Existem *hotspots* no uso de veículos? Aonde são usados mais de um tipo de viatura?

Q8.G Que impacto tem a Meteorologia?

Como refiro anteriormente, os RSB deram-nos a oportunidade de participar em reuniões e deste modo partilhar questões reais e pertinentes ao contexto específico, que reflexem problemas reais que eles sentiam. Estas reuniões permitiram também esclarecimento de dúvidas e sugestões de ramificações da análise com interesse para o mundo real, que podíamos tentar incorporar na investigação com a confiança da veracidade da sua importância. Numa das reuniões, tornou-se pertinente o pedido de mais dados e ainda de quais as questões que gostariam de ver resolvidas, algo que se pode ler no Apêndice A. Todas as questões são instâncias de pedidos de análise focadas em certas naturezas como se pode ver na tabela 3.1.

Tabela 3.1: Lista de propostas de análise idealizadas pelo RSB.

Designação	Descrição da Proposta
P1.1	Onde se encontram zonas com Acidentes Rodoviários Recorrentes?
P1.2	Onde se encontram zonas com Incêndios Recorrentes?
P2	Quais são as zonas com mais Acidentes Rodoviários com Encarcerados?
P3	Houve um aumento das ocorrências da natureza "Serviço de Abertura de Porta" em 2020? Foi devido a pessoas idosas que se encontram mais isoladas?
P4	Quais são as zonas com mais Incêndios em edifícios com mais de 30 metros?
P5	Quando acontecem Queda de Árvores?
P6	Quando e onde acontecem Cheias Rápidas? Previsão deste tipo de eventos é possível?

Desta lista foram reduzidas algumas propostas de acordo com os dados disponíveis em comparação aos dados necessários para analisar certas perguntas e arranjar respostas robustas e satisfatórias. As restantes propostas não passaram nos critérios essenciais para

a sua análise, nomeadamente o maior problema foi a necessidade de dados adicionais que não estão presentes no conjunto de dados fornecidos, ou não existem em quantidade suficiente para serem analisadas. Para informação detalhada sobre cada razão de rejeição, aconselho a leitura do Apêndice A. Em suma, pode-se ver na tabela 3.2 quais as propostas que foram aceites.

Tabela 3.2: Lista de propostas de análise após exploração dos dados.

Designação	Estado
P1.1	Aceite, possível.
P1.2	Aceite, possível.
P2	Aceite, possível.
P3	Rejeitado, porque não temos acesso aos dados demográficos.
P4	Rejeitado, por insuficiência de ocorrências.
P5	Aceite, possível.
P6	Rejeitado, por insuficiência de ocorrências.

Ambas estas listas de questões e propostas serviram como objetivos e guia de eficiência de cada ferramentas, tanto para saber quando se dava por terminado o trabalho e mais tarde para avaliar se faz bem o trabalho para o qual foi desenhada, que é o que se faz na secção seguinte, a avaliação.

Em termos de desenvolvimento, a etapa final foi a avaliação experimental das ferramentas. Foi desenvolvido um ambiente de testes, nomeadamente um questionário e as suas devidas soluções. Este questionário foi apresentado a utilizadores e foi registado o tempo para atingirem cada resposta, tal como a quantidade de ações cumpridas até à apresentação da solução, e claro qual a solução. Com estes dados foi então possível avaliar a qualidade das ferramentas para cumprirem o objetivo para que foram desenvolvidas.

Por fim foi feita a organização e escrita deste trabalho e os seus resultados no documento da dissertação. Algumas das ferramentas foram também disponibilizadas ao RSB através do Tableau Online.

3.2 Tecnologias Utilizadas

No processo de investigação deste tema foram usadas ferramentas que tornam o trabalho possível, e noutras partes mais eficientes quando comparado com outras técnicas, e as usuais vantagens de ter um programa automatizado e otimizado para as tarefas que foram utilizados.

Nesta secção listo todas as ferramentas utilizadas, como foram utilizadas, quais as razões e como tornam melhor o trabalho ou tarefa para que foram escolhidas. A maioria destas ferramentas são da grande gama de produtos da empresa Tableau, que é uma empresa focada em ferramentas para visualização de dados. Descrevem-se como a melhor plataforma de analítica moderna, cuja função é ajudar a ver e entender melhor os dados.

Têm clientes conhecidos, como a Lenovo, Luftansa ou Hello Fresh, onde boas decisões fundamentadas em dados são importantes para o sucesso do negócio.

3.2.1 Tableau Prep

Começando pela ferramenta que foi utilizada na fase de preparação dos dados. A Tableau Prep é um programa que permite ligar a uma grande variedade de dados e fazer várias tarefas necessárias para se preparar os dados. O nome tem origem de *Preparation*, ou seja, preparação tal como o nome do processo que foi desenhado para fazer.

O Tableau Prep foi então utilizado no conjunto de dados fornecidos para as várias fases que se passam na preparação de dados: familiarização dos dados, limpeza de duplicados e dados inválidos, organizar e melhorar através da criação de novos atributos através de cálculos ou funções e para combinar com outros conjuntos de dados disponíveis. Neste último caso foram combinados dados com detalhes diferentes sobre as ocorrências, um sobre que veículos foram despachados e outros sobre a informação geral como localização, tipo e hora de chamada.

O programa chama à cadeia de modificações um fluxo, e esses fluxos são compostos por etapas. A quantidade, tipo e propósito de modificações por etapa cabe ao utilizador decidir. A vantagem de cada passo ser feito por um fluxo é que permite maior flexibilidade e organização para que possam ser exportados os dados em cada etapa de cada fluxo. Outra característica associada ao seu nome - fluxo- é o facto de se poder ramificar em várias opções para exportar diferentes dados. Como vários conjuntos de dados podem correr o mesmo fluxo, desde que tenham a mesma estrutura, isto é particularmente vantajoso para quando se quer limpar dados da mesma fonte e reduzir a quantidade de trabalho redundante. A experiência a trabalhar com esta ferramenta foi agradável pois conseguia visualizar o processo por etapas, com nomes reveladores das funções que efetuavam, tanto para organização mental, como para mais tarde, quando eventualmente se notou em etapas que podiam ser melhoradas ao aumentando a qualidade de certas visualizações, foi fácil fazer mudanças sem repetir passos, procurar os dados originais, ou ficar confuso com como fazer, pois é bastante intuitivo.

É mais rápido do que as alternativas como o *Microsoft Excel*, devido às suas funções e funcionalidades automáticas. Tais funcionalidades, como estarem sempre exibidas as distribuições dentro dos atributos, tornam o reconhecimento de erros e o impacto de funções mais rápido de se compreender. Outra vantagem é que este programa não muda o(s) ficheiro(s) originais, podendo o processo ser reutilizado para outros ficheiros semelhantes com a mesma estrutura, imaginando que são da mesma fonte, ou readaptados a mudanças sem grande problemas. Por fim, tem integrações com a ferramenta seguinte, o *Tableau Desktop*, onde se constroem as visualizações com os dados, que agiliza o processo.

Por fim, queria mencionar que este software não é grátis, tendo uma subscrição mensal de 70 USD por utilizador. São, no entanto, fornecidas licenças gratuitas renováveis de um ano a estudantes e investigadores académicos. Estas dão acesso a toda a gama de produtos

da Tableau.

3.2.2 Tableau Desktop

O Tableau Desktop é outro programa da empresa Tableau, que oferece uma variedade de ferramentas desenhadas para ajudar na leitura e visualização de dados. O Tableau Desktop permite criar visualizações interactivas quase instantaneamente, através de funcionalidades de *drag-and-drop* com os dados fornecidos. Ainda afirmam tirar o maior partido da capacidade humana natural de reconhecer padrões visuais rapidamente[19]. Estas visualizações podem ser usadas em *dashboards* ou *stories* (uma *feature* para apresentação mais elaboradas).

Um projeto no Tableau Desktop começa por se ligar os dados ao programa. Estes dados podem então ser unidos com outros dados numa base de dados usando *joins* ou *unions*. Tal como numa base de dados, os tipos por coluna de dados têm regras estritas, é crucial verificar os tipos atribuídos pelo Tableau automaticamente no caso que haja erro. É comum para identificadores numéricos serem confundidos com números ou outras dados semânticos que não se espera que um algoritmo perceba. A vantagem de ligar os dados é que se desbloqueia o potencial de todos os dados ao nosso dispor, tornando os conjuntos de dados combináveis para se usarem nas visualizações. Por exemplo, nos dados havia dois ficheiros que referiam uma ocorrência por identificador alfanuméricos, um listava os veículos usados na ocorrências (com outras informações relativos a esse veículo) e noutra a informação da ocorrência em si (a natureza, data e hora, etc). Estes dados foram *joined* pelo identificador da ocorrência, para que o Tableau pudesse referir os dados de um ficheiro a partir do outro.

As colunas de dados no Tableau estão divididas em dois grandes grupos no Tableau Desktop, que se chamam dimensões e medidas. Dimensões são colunas de dados qualitativas, enquanto medidas são colunas do tipo quantitativo. Outra diferenciação importante no Tableau são campos (como são representadas as colunas de dados) verdes ou azuis, que são *color-coded* para dizer se o campo é contínuo ou discreto respetivamente. Existem sete tipos de dados: Texto (string); Data; Data e Tempo; Valores Numéricos; Booleanos; Valores Geográficos. A classificação destes campos é importante para que o Tableau Desktop consiga identificar corretamente que tipo de funções e visualizações são possíveis com aquela tipologia de dados. Por exemplo, campos geográficos podem ser mostrados numa variedade de formas dependendo dos dados numa *layer* por cima de um mapa. As formas estão dependentes de novo do tipo de dados: um ficheiro geojson com *shapes* mostra as áreas descritas no ficheiro, enquanto uma linha com coordenadas geográficas é representado automaticamente por um ponto.

Após a fase inicial de ligar os dados e classificar os seus tipos, o programa principal do Tableau Desktop é a criação de visualizações, as quais são depois usadas para formar *dashboards*, ou seja, *dashboards* são um conjunto de visualizações num único painel. O

Desktop foi desenhado para que funções básicas como partilhar filtros de umas visualizações para outras no mesmo *dashboard* seja fácil, ou outro exemplo, expandir as opções de interações entre as visualizações usando ações, como dar *highlight* numa visualização, como seleccionar uma área no mapa e as restantes visualizações filtram para mostrar só informações pertinente aos dados seleccionados.

Há dois tipos de filtros, os que se usam em *dashboards* são temporários e reversíveis, para aumentarem a interatividade, as opções e flexibilidade dos dados que se estão a ver, assim como os filtros que se usam em visualizações, que são úteis para excluir dados não pertinentes, ou até para fazer zoom a certas secções dos dados. Imaginando duas visualizações a mostrar a quantidade total de ocorrências, os filtros na visualizações permitem trancar as visualizações a certos anos, como por exemplo 2020 e 2021 e, depois juntá-las no mesmo *dashboard* para comparar os dados. No entanto, esta é uma técnica usada em dados mais complexos, não devendo esquecer a importância da escala, que é algo que fica ao cuidado do colaborador. No *dashboard* pode haver ainda um filtro para ver a quantidade de ocorrências por natureza, que dá para mudar e reverter.

As visualizações podem depois ser agrupadas em páginas brancas a que são adicionados como elementos, sendo que ao conjunto de visualizações nessa página são o que chamamos *dashboards*. Podem ser reordenados como pretendido, serem adicionados caixas de texto, tanto para escrever títulos como para dar instruções de utilização. Podem ainda ser controlados quais os filtros que se podem controlar dessa página, mostrando legendas e controlando que visualizações filtram ou dão *highlight* a outras.

Este *software* não é grátis, tendo a mesma subscrição mensal que o Tableau Prep. No entanto, a mesma licença fornecida pela faculdade permitiu acesso gratuito a este produto.

A razão pela qual esta ferramenta terá sido a escolhida deve-se ao facto deste programa tornar substancialmente mais rápido a criação de visualizações e *dashboards* interativos, porque não é necessário implementar código. Para estender a funcionalidade do programa, existe a possibilidade de ainda assim usar código, ou especificamente cálculos como se veria num programa como o Excel, para criar novos campos de dados, contudo a maior partes destes campos são criados na parte de preparação de dados, o passo anterior. No que toca a outras ferramentas, como programar visualizações em D3 ou Python, o Tableau Desktop poupa imenso trabalho, pelo menos para os tipos de visualizações mais populares e simples (e não se sentiu falta de outras na duração desta elaboração).

O Tableau Desktop é uma ferramenta que permite criar visualizações, *dashboards* e apresentações. É um programa muito poderoso devido à sua flexibilidade, desde uma larga variedade de visualizações como a sua capacidade de ligação entre elementos só com *features* de *drag and drop*, fazendo com que o processo de criação seja rápido. Tal como qualquer ferramenta complexa, os diferentes graus de mestria incluem mais passos para visualizações e *dashboards* mais complicados, mas sempre mais rápido que outras opções no mercado.

3.2.3 Tableau Online

O Tableau Online é um produto também da Tableau, direcionado para se poder visualizar *dashboards* ou visualizações online, em qualquer *browser*, sem instalação necessária. Essas criações podem vir da ferramenta de criação de *dashboards* do Tableau Online, ou do programa local do Tableau, o Tableau Desktop. Isto é possível porque o Tableau Online é uma plataforma online hospedada na *cloud*, sendo assim possível o *upload* de um ficheiro e ser acessível de qualquer lugar, restrito apenas pelas escolhas de privacidade do utilizador.

Obviamente, isto torna-se vantajoso particularmente na partilha com outras pessoas, sendo a principal razão para a sua utilização, podendo obter *feedback* e se testarem as ferramentas sem a necessidade de estar localmente a presença de uma máquina com o Tableau Desktop instalado.

O processo de utilização é relativamente simples. Após se ter terminado o elemento pretendido no Tableau Desktop, pode-se exportar um ficheiro destas folhas, que empacota o trabalho com os respetivos dados, e dar upload no domínio oferecido pela faculdade. Certas subpartes do domínio podem ser partilhadas com outras contas após autorização do administrador, sendo essas divulgadas a elementos de interesse e apenas então necessário partilhar com a conta do utilizador pretendido, podendo ser da pasta, ou de um elemento desta.

3.3 Processamento de Dados

A análise da qualidade de dados é uma fase importante na exploração e conhecimento do domínio dos dados, assim como ao averiguar se é plausível a sua análise (dependendo da qualidade) e ao definir a abordagem de acordo com o tipo de dados. Uma boa qualidade de dados oferece uma maior confiança nas conclusões derivadas das fases anteriores. Esta seção é o relatório com a avaliação da qualidade dos dados das ocorrências do Regimento Sapadores de Bombeiros de Lisboa disponibilizados pelos mesmos. Esta fase é baseada na fase de compreensão do domínio de negócios e da fase de compreensão dos dados da metodologia CRIPS-DM [18]. As etapas desta fase, e deste relatório por sinal, são as descritas:

1. Dados iniciais;
2. Descrição dos dados;
3. Verificação da qualidade dos dados;
4. Estatística Descritiva.

A análise e resultados deste relatório poderão vir ser úteis para a fase seguinte de prospeção e preparação dos dados. Esta é a fase onde serão tomadas decisões sobre limpeza e tratamento dos dados, sendo que o caso mais frequente são falhas na consistência ou na qualidade dos dados.

Tabela 3.3: Quantidade de linhas por ficheiro e por tabela, que representam cada uma um ano de dados. Os números indicados não contam com o cabeçalho.

Folha	<i>DadosViaturasRSB_2016 a 2020.xlsx</i>	<i>DadosViaturasRSB_Locais_2016 a 2020.xlsx</i>
2016	22 737	18 224
2017	24 586	19 929
2018	25 889	21 133
2019	23 054	18 381
2020	19 583	15 828

Tabela 3.4: A tabela mostra uma lista e descrição das colunas de dados presentes no ficheiro Excel *DadosViaturasRSB_2016 a 2020.xlsx*. Este ficheiro contém informação referente à reação de uma viatura a uma ocorrência.

Nome da Coluna	Descrição
Ocorrência	Identificador interno da ocorrência
Viatura	Denominação da viatura utilizada
Natureza	Tipo de ocorrência
Freguesia	Freguesia onde ocorreu a ocorrência
Data	Data da ocorrência
Chamada	<i>Timestamp</i> da chamada que originou a ocorrência
Despacho	<i>Timestamp</i> do despacho no quartel
Saída	<i>Timestamp</i> da saída do complemento de Bombeiros
Chegada	<i>Timestamp</i> da chegada da viatura ao local da ocorrência
Retirada	Data e <i>timestamp</i> da retirada da viatura do local da ocorrência

3.3.1 Dados iniciais

Os dados referentes às ocorrências do Regimento de Sapadores de Bombeiros de Lisboa foram disponibilizados em formato Excel. Os dois ficheiros têm os seguintes nomes *DadosViaturasRSB_2016 a 2020.xlsx* e *DadosViaturasRSB_Locais_2016 a 2020.xlsx*. Ambos os ficheiros estão divididos em cinco folhas, cada uma representando um ano. Pode-se ver na tabela 3.3 o número de linhas por folha. É de notar que existem mais dados, tanto no total como por ano, no primeiro ficheiro que no segundo.

Começando pelo ficheiro *DadosViaturasRSB_2016 a 2020.xlsx*, cada folha apresenta o mesmo número de colunas, dez, e a mesma denominação. Dever-se-á conseguir juntar todas estas folhas. O segundo ficheiro, *DadosViaturasRSB_Locais_2016 a 2020.xlsx* apresenta o mesmo número de colunas em todas as folhas, cinco, e cabeçalhos idênticos. Dever-se-á também conseguir juntar todas estas folhas.

Para melhor compreensão dos dados poder-se-á ver a tabela 3.4 e a 3.5 que contém os metadados para os conjuntos de dados dos ficheiros *DadosViaturasRSB_2016 a 2020.xlsx* e *DadosViaturasRSB_Locais_2016 a 2020.xlsx* respetivamente. A tabela 3.4 tem ambos dados espaciais como temporais, embora se pareça focar no aspeto temporal, e o contrário pode ser verificado na tabela 3.5.

A coluna chamada de "Ocorrência" é o código identificador do sistema interno de ocorrências do Regimento de Sapadores Bombeiros. Aparece por este identificador

Tabela 3.5: A tabela mostra a lista dos atributos do segundo ficheiro Excel, textitDadosViaturasRSB_Locais_2016 a 2020.xlsx, e uma breve descrição do que representa.

Nome da Coluna	Descrição
N.º Ocorrência	Identificador interno da ocorrência
Data / Hora	Data e hora da ocorrência
Longitude	Longitude do local da ocorrência
Latitude	Latitude do local da ocorrência
Local	Morada do local da ocorrência

Tabela 3.6: A tabela mostra que o ficheiro Excel das viaturas contém dados numéricos e simbólicos, dos quais mais de metade são referentes a dados temporais. A tabela mostra um exemplo real retirado aleatoriamente do conjunto de dados.

Nome da Coluna	Tipo de valores	Exemplo
Ocorrência	Numérico	1927601012016
Viatura	Simbólico	VUCI 31
Natureza	Simbólico	7401 - Serviços - Abertura de Porta - Com Socorro
Freguesia	Simbólico	São Domingos de Benfica
Data	Simbólico	1/1/2016
Chamada	Simbólico	21:38
Despacho	Simbólico	21:38
Saída	Simbólico	21:41
Chegada	Simbólico	21:48
Retirada	Simbólico	1/1/2016 22:00

Tabela 3.7: A tabela mostra que o ficheiro Excel dos locais contém dados numéricos e simbólicos, dos quais a maioria se refere a dados espaciais. A tabela mostra um exemplo real retirado aleatoriamente do conjunto de dados.

Nome da Coluna	Tipo de valores	Exemplo
N.º Ocorrência	Numérico	1838401012020
Data / Hora	Simbólico	1/1/2020 0:14
Longitude	Numérico	-9.15648072
Latitude	Numérico	38.78162189
Local	Simbólico	Estrada de São Bartolomeu

que ambos os ficheiros podem ser unidos e correlacionados, para uma análise conjunta de todas as colunas dos dois ficheiros.

O ficheiro de Viaturas tem 115 849 linhas, um total combinado entre os ficheiros de 28 colunas de interesse.

3.3.2 Limpeza e Transformação dos Dados

Este passo foi executado usando o Tableau Prep. O Tableau Prep é uma ferramenta que ajuda a limpar e preparar dados não processados para posterior análise e criação de visualizações. No nosso caso, usámos outra ferramenta da mesma empresa, o Tableau, cuja visualizações iremos ver noutra secção do relatório. Costuma também fornecer uma

ideia dos dados, sendo auxiliar na exploração inicial dos mesmos.

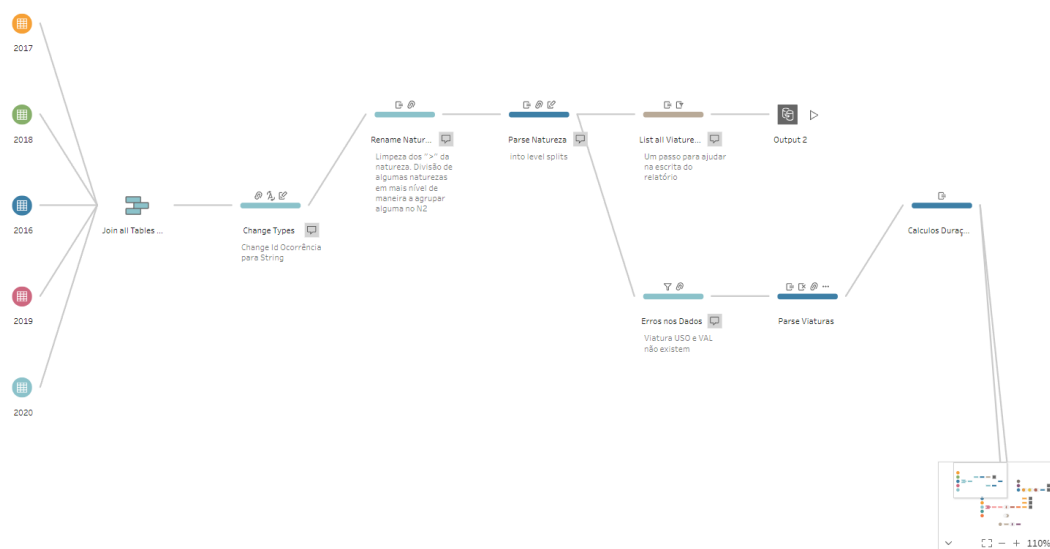


Figura 3.2: *Flow* de limpeza usado no Tableau Prep para o ficheiro sobre as viaturas.

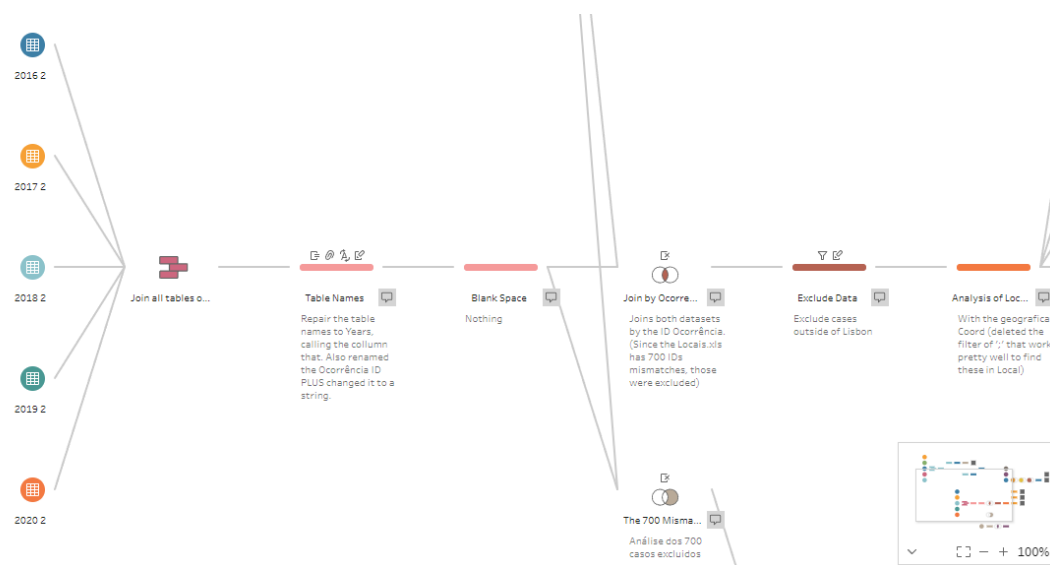


Figura 3.3: *Flow* de limpeza usado no Tableau Prep para o ficheiro sobre os locais e aonde junta com o ficheiro das viaturas no Join.

O primeiro passo para ambos os ficheiros foi juntar os dados numa única coleção. Esta etapa foi atingida através da junção das folhas por anos e adicionando uma nova coluna para identificar a folha original da ocorrência ou chamada *Ano*. Como podemos observar na figura 3.2 e na figura 3.3, do lado esquerdo estão as folhas com os anos a serem juntas por um passo de agregação. Estas figuras representam os vários passos e respetivas mudanças realizadas aos ficheiros de Viaturas e Locais que são juntos na figura 3.3. De seguida, ao fazer mudanças a todas as folhas de cada ficheiro simultaneamente,

corrigimos as tipologias que o Tableau atribui aos dados. Em particular existe dificuldade em distinguir dados que possam ser interpretados como números ou strings e ainda nos dados de tipo espacial. Verificar que os identificadores (que são numéricos) são interpretados de maneira igual entre os dois ficheiros é importante para os conseguir ligar, porque o Tableau não aceita relações com campos de tipos diferentes. Os identificadores foram marcados como strings porque são como nomes para as ocorrências, não queremos fazer cálculos com eles, e mudámos também o nome dessa coluna para *Id Ocorrência*.

Dentro do atributo *Natureza* havia duplicados da mesma natureza porque a simbologia utilizada nos divisores mudou. Por exemplo: 1100 -> Incêndio -> Povoamento Florestal e 1100 - Incêndio - Povoamento Florestal. Ambos os valores são reais e foram retirados dos dados, portanto escolheu-se manter apenas o traço.

Criação de várias colunas com o intuito de dividir a *Natureza* pelos seus níveis hierárquicos. Um para o Código, outro para o primeiro nível da *Natureza*, para o segundo nível da *Natureza* e o raro terceiro nível. Se não houver terceiro nível, o valor nesta coluna será null (37% das ocorrências). No final existem 111 naturezas diferentes, entre as combinações dos 3 níveis.

Foi necessário pesquisar e eventualmente pedir dados adicionais sobre as viaturas utilizadas. Esta fase é de *Business Understanding* no CRISP-DM. Foram realizadas algumas reuniões com o Regimento Sapadores de Bombeiros de Lisboa. Estes referiram que certas viaturas especializadas são colocadas estrategicamente pela cidade, pois existem em pouca quantidade. Em específico, as VEXX (onde o XX é a altura da escada) para incêndios em prédios maiores e as duas VSAT para acidentes rodoviários com encarceramentos mais graves. Estas viaturas costumam estar bem posicionadas. O resultado da investigação própria, com esclarecimentos do RSB, foi a tabela 3.8. A lista da tabela apenas inclui os veículos que aparecem na janela temporal de dados fornecida, o RSB poderá ter outras viaturas não utilizadas durante esse tempo. Uma nota importante é que os dados incluem duas viaturas que não existem, *USO* e *VAL*, que foram declarados erros pelos próprios RGB e portanto excluídos dos dados.

De novo foi dividido o nome das viaturas por tipologia do carro, subtipo e ambos, denominado de "Viatura Classe". A coluna original passou a ser identificada por "Id Viatura", pois contém tanto a classe como o número da viatura (e não existem duas viaturas com o mesmo modelo/número em Lisboa).

Nesta parte foram criados mais campos relativos a tempos de resposta: Tempo de Reação (Saída-Chamada); Tempo em Rota (Chegada-Saída); Duração do Tratamento da Ocorrência (Retirada-Chegada); e Duração Total (Chamada - Retirada). Nesta fase notou-se dois problemas: existem casos no *dataset* cujas horas não estão ordenadas na ordem cronológica esperada, criando negativos nestes cálculos; E o facto de as colunas da Data e Retirada serem as únicas com a data do dia e mês, ou seja, que as colunas Chamada, Despacho, Saída e Chegada apenas têm a hora, se houver mais que 48 horas é impossível saber-se a qual dia se refere a Saída, Despacho e Chegada sem assumir, talvez erroneamente. No primeiro caso foram excluídos esses dados, no total de 13 linhas. No último caso foi

Tabela 3.8: Resultado da cooperação com o RSB no entendimento das siglas e funções de cada viatura.

Sigla da Viatura	Significado da Sigla
AA	Auto Administrativo
AAI/GIB 1 ^a	Atrelado de Apoio à Intervenção - Gerador de Iluminação e Bombagem (1. ^a Companhia)
ABSC	Ambulância de Socorro
ABTM	Ambulância de Transporte Múltiplo
AC	Auto Comando
ACC	Auto Comando de Companhia
BSRS	Bote de Socorro e Resgate Semi-rígido
ERST	Embarcação de Reconhecimento Socorro e Transporte
VALE	Veículo de Apoio Logístico Especial (transporte de água 12 000 e 13 000L)
VAME	Veículo de Apoio a Mergulhadores
VC	Veículo de Comando
VCOC	Veículo de Comando e Comunicações
VCOT	Veículo de Comando Tático
VE25	Veículo com Escada Giratória de 25 metros
VE30	Veículo com Escada Giratória de 30 metros
VE37	Veículo com Escada Giratória de 37 metros
VECI	Veículo Especial de Combate a Incêndios
VFCI	Veículo Florestal de Combate a Incêndios
VLCI	Veículo Ligeiro de Combate a Incêndios
VLTP	Veículo Ligeiro Transporte Pessoal
VOPE	Veículo para Operações Específicas
VOPE/C	Veículo para Operações Específicas - Comando
VOPE/CAT	Veículo para Operações Específicas - Camioneta
VOPE/CAT CG	Veículo para Operações Específicas - Camioneta de Cargas Gerais
VOPE/CAT ESC	Veículo para Operações Específicas - Camioneta de Escoramentos
VOPE/MA	Veículo para Operações Específicas - Material de Apoio
VOPE/MANITOU	Veículo para Operações Específicas - Manitou
VOPE/MEG	Veículo para Operações Específicas - Mergulho
VOPE/RE	Veículo para Operações Específicas - Retroescavadora
VOPE/RL	Veículo para Operações Específicas - Reboque Ligeiro
VOPE/RP	Veículo para Operações Específicas - Reboque Pesado
VOPE/UCR	Veículo para Operações Específicas - Unidade Cinotécnica de Resgate
VP	Veículo com Plataforma Giratória
VPMA	Veículo de Proteção Multirriscos e Ambiente
VPMT	Veículo de Proteção Multirriscos Tático
VSAE	Veículo de Socorro e Assistência Especial
VSAT	Veículo de Socorro e Assistência Tático
VTO	Veículos Operacionais no Terreno
VTPG	Veículo de Transporte de Pessoal Geral
VTTF	Veículo Tanque Tático Florestal
VTTU	Veículo Tanque Tático Urbano
VUCI	Veículo Urbano de Combate a Incêndios
USO	Erro
VAL	Erro

decidido assumir que estes se passaram no dia mais cedo possível.

No outro *dataset* começou-se por juntar as tabelas como descrito inicialmente nesta secção, juntando os dados de todas as tabelas (que estavam organizadas por ano) e corrigindo tipos de dados.

No *dataset* existiam algumas ocorrências fora de Área Metropolitana de Lisboa (onde se encontra a grande parte das ocorrências) que foram excluídas por não existirem dados suficientes para fazer análises. É o passo vermelho após o Join na figura 3.3. Entre estas:

- uma ocorrência no Brasil em 2017, de prevenção de queimadas (possível erro na documentação), apresentava apenas coordenadas geográficas. ID: 899626072020;
- três ocorrências em Castelo Branco, em 2020 de Julho a Setembro, de Serviços de Prevenções com o intuito de Pré-Posicionamento Meios (apoio ao Regimento de Castelo Branco). IDs: 1130713092020, 1131614892020, 141227012017;
- uma ocorrência em Faro, em 2017 que envolvia a queda de uma árvore. ID: 319504032017;
- uma ocorrência em Setúbal de Operação Nacional de Assistência em 2017. ID: 232914022017;
- uma ocorrência em Santarém de Operação Nacional de Socorro em 2016. ID: 1461321102016.

E os seguintes no distrito de Lisboa, embora fora da área metropolitana:

- quatro ocorrências de Desinfecção COVID19 em 2020. IDs: 1186825092020, 1358703112020, 1053329072016, 902727072020;
- três ocorrências de Rebocagem ou Desempanagem em 2016 e 2020. IDs: 1586629122020, 1363304112020, 1119511092020;
- uma ocorrência de Doença súbita em 2019. ID: 1056429072016;
- uma ocorrência de Deslocações serviço geral em 2020. ID: 1109508092019.

No total foram excluídas 16 ocorrências nesta etapa.

A junção dos dois ficheiros, com a *clause* comum sendo o identificador numérico da ocorrência, resultaram em 115.848 linhas no *dataset*. Das 93.496 ocorrência no ficheiro dos locais, 700 foram excluídas neste *join*, visto que o ficheiro das viaturas não tinha *match*.

Após investigação com um terceiro *dataset* cedido anteriormente à antiga aluna Mafalda Camilo no seguimento da sua dissertação [20], foi possível analisar a razão para estas ocorrências não terem viaturas. A conclusão é que estes eventos são de alguma forma duplicados de outros, delegados a outra entidade (como o INEM) ou que o reclamante/entidade prescindiu mais tarde dos serviços. Tendo em consideração esta informação, e visto que a

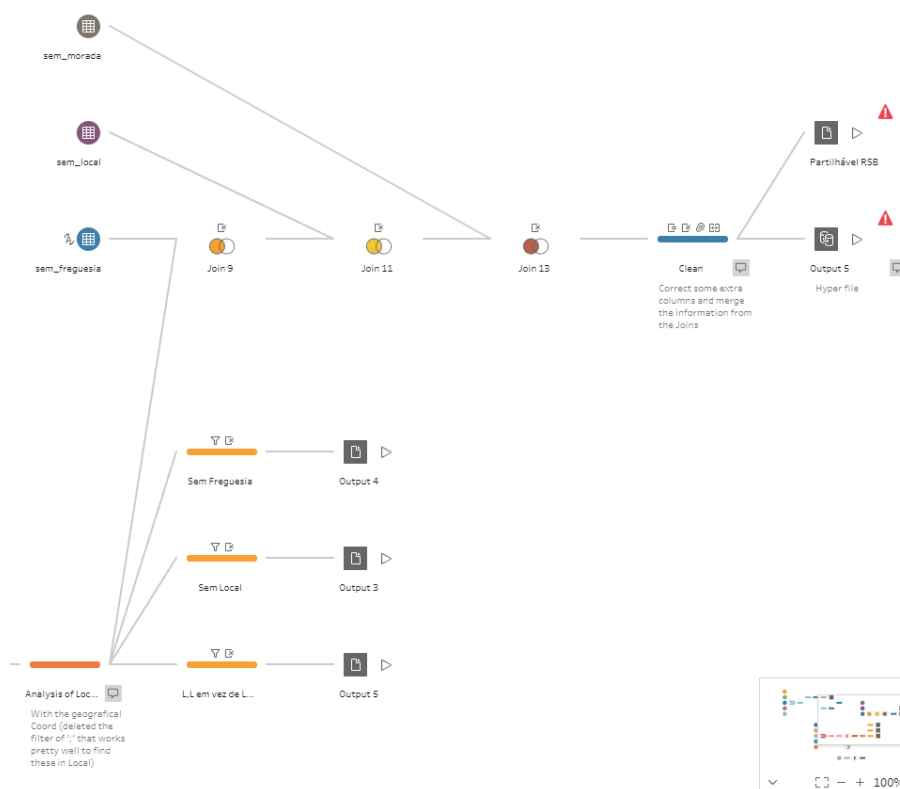


Figura 3.4: *Flow* de limpeza usado no Tableau Prep para juntar as correções aos dados feitas manualmente.

falta de correspondência no ficheiro quer dizer que não há informação, como por exemplo a natureza da ocorrência, foram excluídas por não serem relevantes.

Por último foi realizada alguma investigação sobre algumas ocorrências com dados incompletos, manualmente. Existiam três tipos destes erros corrigíveis: sem freguesia, sem morada e sem local. Por exemplo ocorrências sem locais, mas com coordenadas geográficas, foram pesquisados num motor de pesquisa os locais a que se referiam as coordenadas e assim completar os dados. Estas mudanças estão descritas na figura 3.4.

3.3.3 Análise Exploratória dos Dados

Como já foi referido, os dados fornecidos são um registo de todas as ocorrências que o RSB atendeu durante 5 anos, desde 2016 a 2020. Começando na distribuição pelo aspeto temporal, na figura 3.5 podemos ver a distribuição da quantidade de ocorrência por ano. O ano 2018 é o que tem mais ocorrências com um total de 25889 eventos registados, enquanto o 2020 é o que tem menos com 19554 ocorrências. Existe uma variância entre todos os anos.

Avançando para uma análise mais profunda sobre como é a distribuição das ocorrências ao longo do ano, foi realizada uma análise da média mensal, como se pode observar na

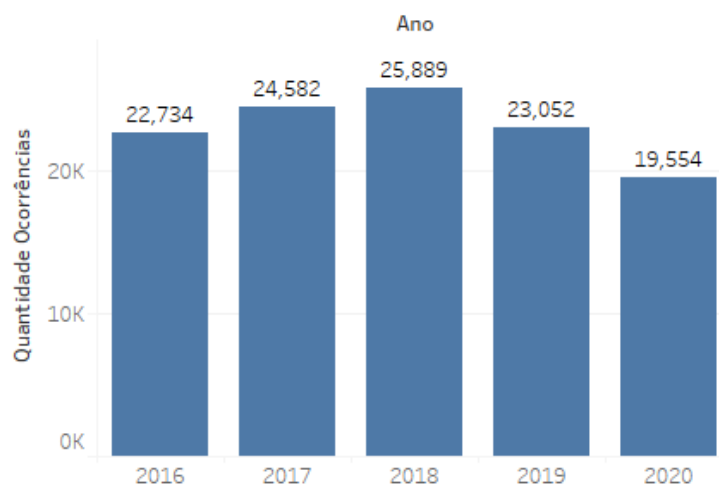


Figura 3.5: Distribuição da quantidade de ocorrências por ano. Pode-se observar que há alguma variância entre os anos.

figura 3.6. O mês com a média mais alta é Outubro, seguido de Setembro. O mês com menos ocorrências, em média, é Junho seguido de Abril e Fevereiro.

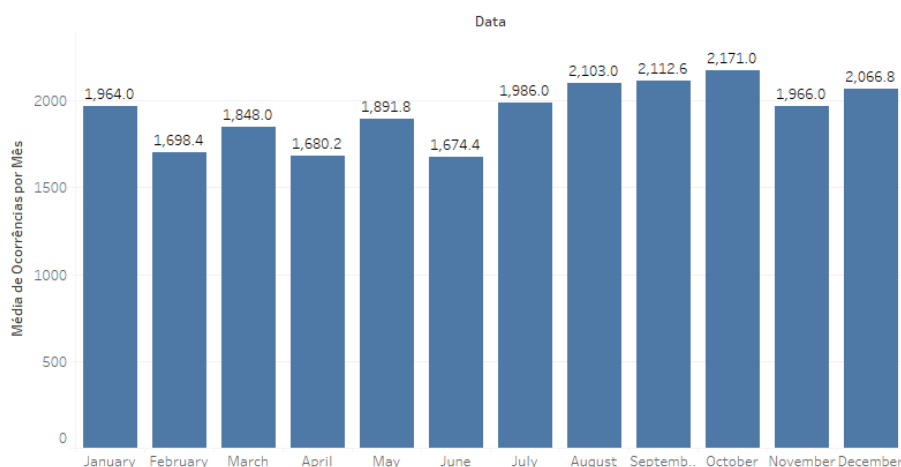


Figura 3.6: Distribuição da média, dos cinco anos, de ocorrências por mês. Pode-se observar que há alguma variância entre os meses.

Analisando a granularidade temporal menor seguinte, a semana, podemos observar os resultados das médias das ocorrências pelo dia da semana na figura 3.7. Os dias com menos média de ocorrências é o domingo e sábado, nessa ordem. Os dias com mais ocorrências são sexta e quinta, embora todos os dias úteis da semana tenham em média mais ocorrências que o fim de semana.

Podemos também ver os dados por média de ocorrência por hora, como se pode observar na figura 3.8. Neste gráfico podemos observar que o máximo da média de ocorrências por hora é as 10:00 da manhã, e o mínimo da média de ocorrências é as 5:00 da manhã, seguido por um décimo da média das 4:00. Podemos também notar uma grande

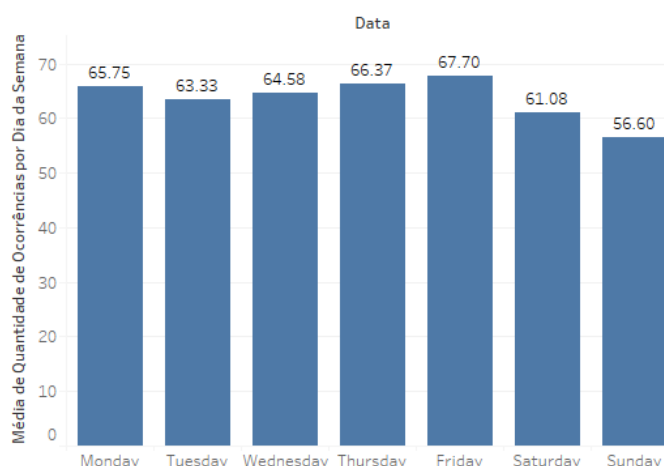


Figura 3.7: Distribuição da média, dos cinco anos, de ocorrências por dia da semana. Pode-se observar que há alguma variância entre a média de ocorrências dependendo do dia.

diminuição de ocorrências durante a noite, tendo a tendência decrescente a partir das 15:00.

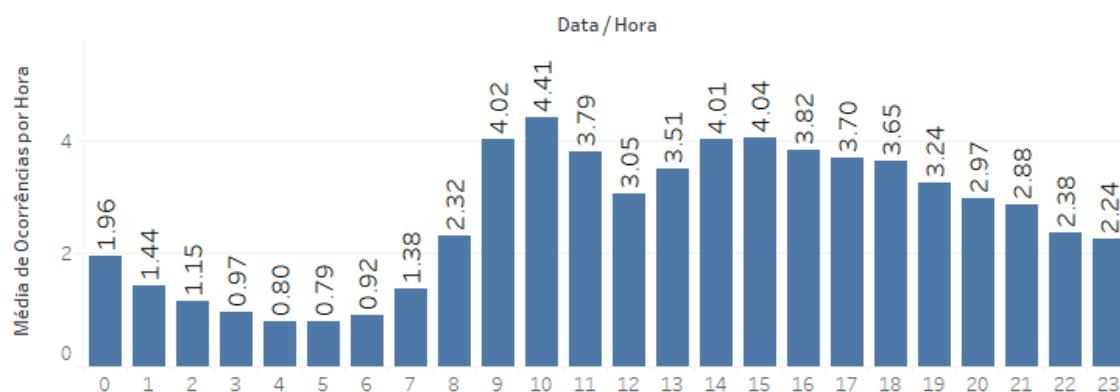


Figura 3.8: Distribuição da média, dos cinco anos, de ocorrências por hora. Pode-se observar que há alguma variância entre a média de ocorrências dependendo da hora do dia.

Através dos campos fornecidos pelos RSB foi possível calcular outros quatro campos temporais: Tempo de Reação, Tempo em Rota, Duração do Tratamento da Ocorrência; e Duração Total. O campo mais vital é o tempo de duração total, que é o tempo que uma ocorrência demorou desde a chamada à conclusão da mesma. Olhando para o gráfico na figura 3.9, podemos concluir que a maioria das ocorrências demoram menos de 200 minutos. Também é possível notar que o pico de duração com 2 125 ocorrências, ou seja, a maior quantidade de ocorrências no minuto, na duração de 30 minutos é o maior. A exploração para este gráfico foi descoberto que a duração total máxima é 125 695 minutos.

Como se pode observar na figura 3.10 a duração das ocorrências foi oscilando um pouco, entre mais/menos 2 a 4 minutos por ano, à exceção do ano 2019 que parece ser um

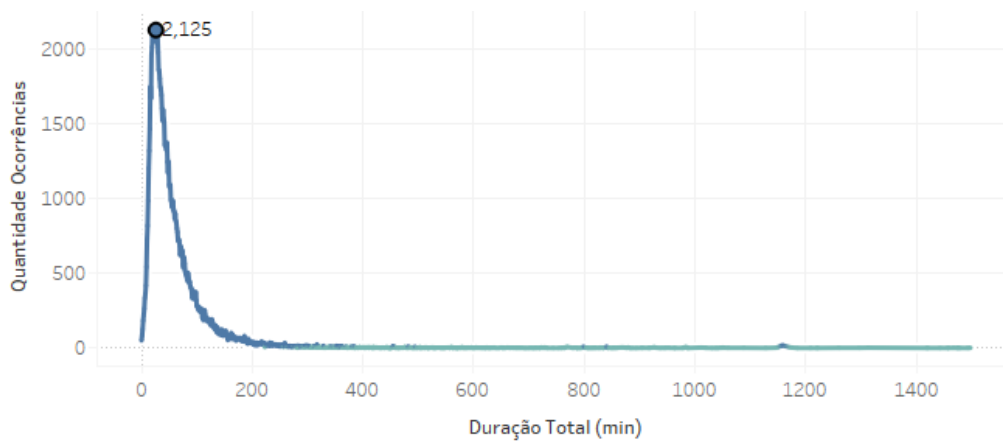


Figura 3.9: Frequência global da duração total das ocorrências (em minutos) com a quantidade de ocorrências. O azul-petróleo realça onde há menos de 10 ocorrências.

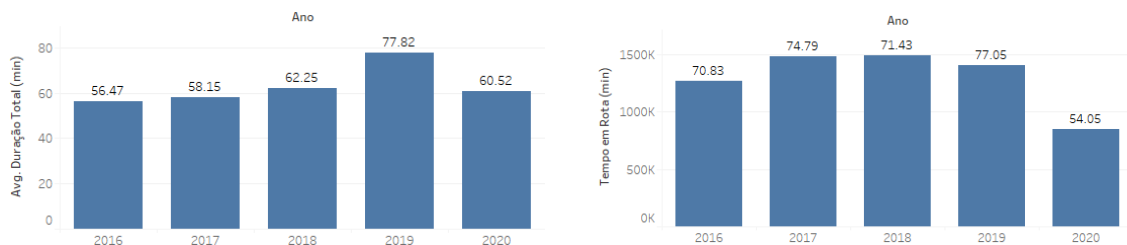


Figura 3.10: Duração total média por ano. Figura 3.11: Tempo médio em Rota por ano.

ano atípico na média geral e sobe 15 minutos na média em relação ao ano anterior. No ano 2019 há uma diferença de 18 minutos em relação aos restantes anos. Na figura 3.11 está o tempo médio demorado a caminho das ocorrências. A média deste está por volta 69.63 minutos, à exceção de 2020 que tem uma diferença de 19 minutos em relação à media dos restantes anos. Isto deu-se, possivelmente, devido à pandemia de COVID19 e o facto de se ter recorrido a trabalho remoto, que reduziu o tráfego automóvel, efetivamente tornando viagens mais rápidas.

Analisando a figura 3.12, podemos notar quais são as naturezas com mais ocorrências e como mudam ao longo do ano. Existem nove categorias de primeiro nível relativas à natureza da ocorrências, listadas na legenda no lado direito da figura. A maior fatia é a dos serviços, consistentemente sendo cerca de 40% das ocorrências em todos os anos, ao longo do ano. A seguir podemos ver os Acidentes, seguido dos Incêndios e problemas com Infraestruturas e Vias de Comunicação. A natureza dos Pré-Hospitalar parece quase desaparecer em 2020, ao contrário dos Eventos de Proteção Civil que começam a meio de 2018. Em 2018, as Atividades começam a ser comparáveis e quase iguais aos Incêndios e problemas de Infraestruturas.

Na figura seguinte 3.13, podemos ver a quantidade de eventos, por trimestres, ao longo dos anos do *dataset*. Podemos notar que o máximo de ocorrências é no terceiro trimestre

de 2019 com 7 109 ocorrências, enquanto o mínimo é no segundo trimestre de 2020 com 4 101 ocorrências. São mínimos drásticos, com uma diferença de menos 2 000 ocorrências relativos ao trimestre anterior. O trimestre de 2020 é comparável ao primeiro trimestre de 2019 com 4 448 ocorrências. Voltando à figura anterior, não parece existir nenhuma tendência que acompanhe as mudanças na figura 3.13. Pelo menos no que é relativo às naturezas de nível 1, não parece haver relação com um aumento ou decréscimo nestas categorias.

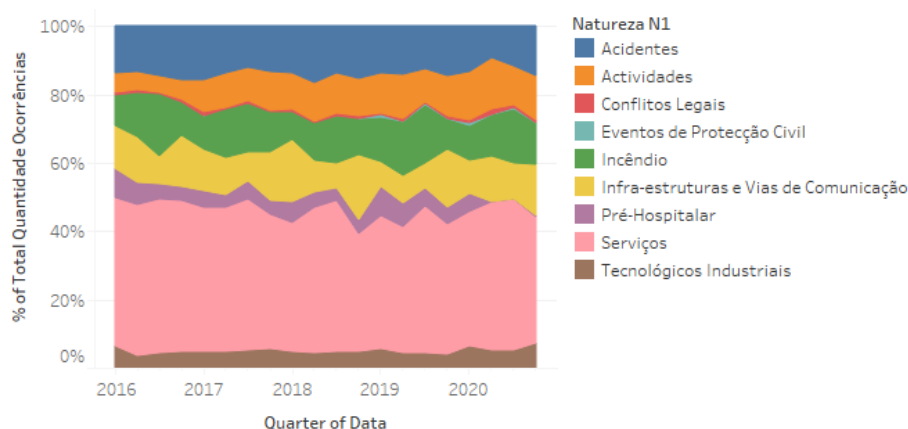


Figura 3.12: Percentagem das ocorrências que ocorreram, por trimestres, que pertencem à natureza de primeiro nível.

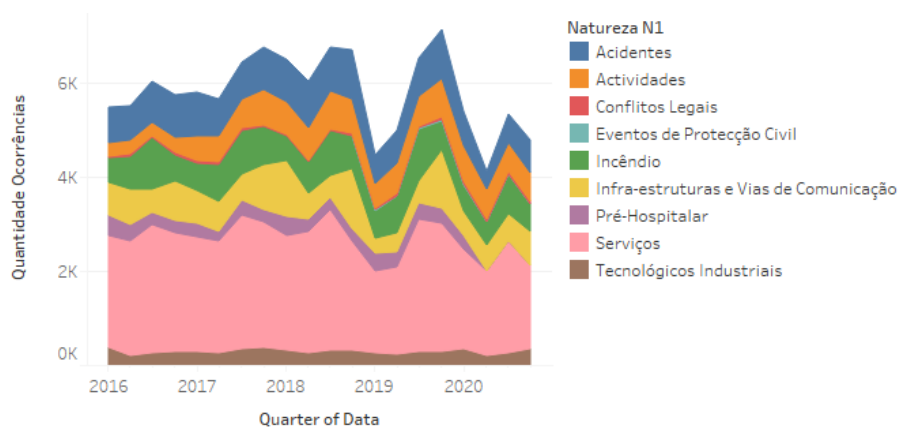


Figura 3.13: Quantidade de ocorrências que ocorreram, por trimestres, que pertencem à natureza de primeiro nível.

Na figura 3.14 podemos ver uma visualização do tipo de caixas de bigodes. Este tipo de gráfico mostra a distribuição dos dados em quartis e destaca a média e os valores atípicos. A mediana de todas as naturezas varia entre os 27 e os 86. Podemos já notar que existe assimetria entre os quartis. O quartil um e dois são menores que os quartis opostos, quatro e três, são menores, o que indica que as naturezas têm uma menor variância de tempo para o lado inferior da mediana, ou seja de 50% dos dados. O facto dos Eventos de Protecção

Civil ter a maior amplitude, com as maiores barreiras quer dizer que é em comparação a natureza com maior diferença de tempos nos dados, tal como também os maiores valores. A versão desta visualização que mostra *outliers* (não mostrado) demonstra que existem vários *outliers* a estes intervalos, alguns chegando a ter 130 000 minutos de duração total.

As ocorrências de Incêndios e Tecnológicos Industriais são semelhantes, ambos com uma média de pontos semelhante, e ambos as categorias têm 17 e entre 43 minutos. E esses tempos são dos mais baixos entre todas as naturezas e com menor altura, que indica que todas as ocorrências têm uma menor variação dos tempos.

Por fim este gráfico revela que as ocorrências que demoram mais tempo são os Eventos de Proteção Civil, com 50% das ocorrências estando entre os 53 e os 147 minutos.

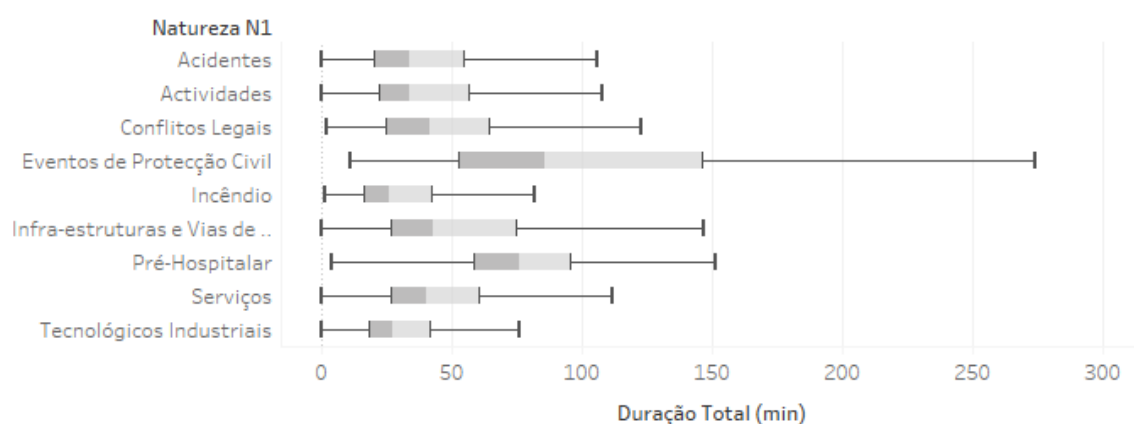


Figura 3.14: Caixa de bigodes da duração total de cada ocorrência, em minutos, com a quantidade de ocorrências.

Avaliando a participação de viaturas em todas as ocorrências, as presentes na figura 3.15 são as top cinco do total de trinta tipos de viaturas. Relembrando que a VUCI é um Veículo Urbano de Combate a Incêndios; seguido de VOPEs, que são Veículos de Operações Específicas e como o nome indica tem as mais variadas funções; e da VLCI ou também conhecido por Veículo Ligeiro de Combate a Incêndios. Por último temos as ABSC e ABTM que são ambas ambulâncias de Socorro e de Transporte Múltiplo respetivamente. As restantes viaturas representam 2% da participação, e estão presentes na figura apenas para contexto.

O tipo de viatura mais utilizado é portanto as VUCI, em todos os anos dos dados, havendo um aumento de cerca de 2% da média nos anos 2018 e 2019. Podemos observar que a participação não muda ao longo dos anos de maneira significativa, mantendo a importância destas viaturas nesta ordem.

Os dados contêm também dados espaciais. Para observar a distribuição das ocorrências mapas coropletos são um bom começo. Nestes são embutidos cores a parte do espaço que representem, numa certa graduação escolhida, os valores a mostrar. Podemos ver na figura 3.16 uma mapa coropletos das freguesias de Lisboa. Na sua análise, podemos ver

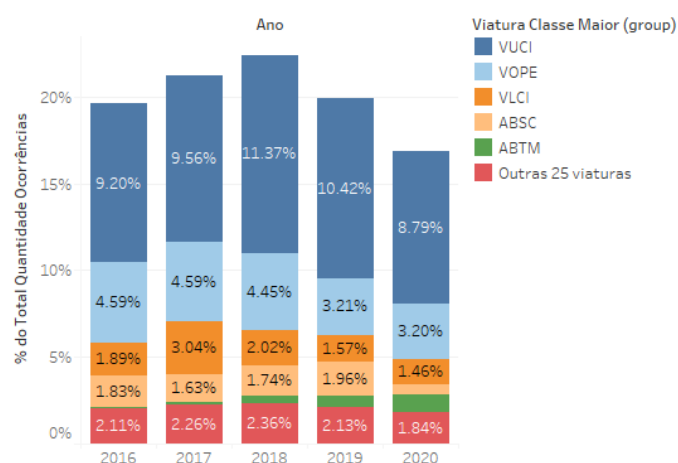


Figura 3.15: Gráfico da participação por viaturas em relação ao total de todos os anos das ocorrências do RSB.

que as freguesias com mais ocorrências são quatro: Benfica, Alvalade, Arroios e Santa Maria Maior. Não há uma padrão explicativo nestes dados, portanto mais análise seria necessária para saber porque que são consideradas as freguesias com mais ocorrências. As freguesias com menos ocorrências são também quatro: Carnide, Santa Clara, Ajuda, Beato.

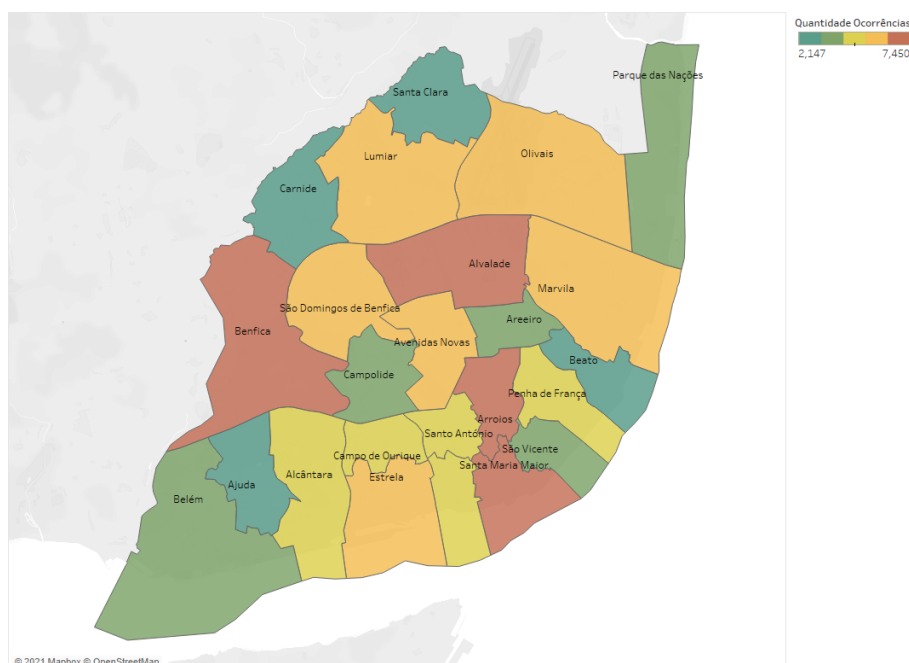


Figura 3.16: Mapa coropleto de Lisboa com a quantidade de ocorrências nos cinco anos do dataset por freguesia.

De seguida apresentamos um mapa de densidade na figura 3.17. Nesta figura, quanto as cores vão de azul, verde a branco, representando respetivamente por essa ordem

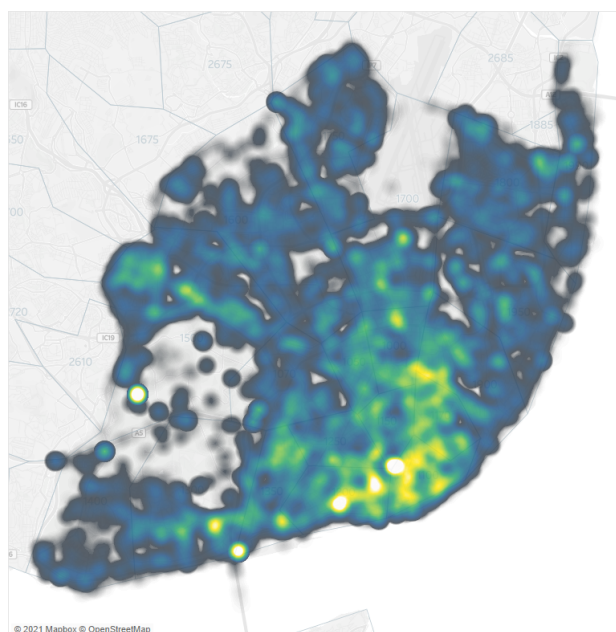


Figura 3.17: Mapa de Densidade com as ocorrências nos cinco anos dos dados. Áreas mais claras têm uma maior densidade de ocorrências nesse local.

crescente um maior número de ocorrências (que a cor anterior). Aqui podemos observar alguns *hotspots*, nomeadamente em Benfica perto do Parque Florestal de Monsanto, que é maior que os restantes. Outros mais pequenos existentes em Alcântara, perto da Doca de Santo Amaro, na Estrela Avenida Dom Carlos I; nos arredores da Travessa Queimada, na Misericórdia e o último em Rua de Barros Queirós e Rua das Portas de Santo Antão, em Santa Maria Maior, no limiar de Arroios. Este mapa está em concordância com o mapa coropleto.

Por último, foi criado um mapa coropleto com a duração total média das ocorrências, também por freguesia, como se pode observar na figura 3.18. Aqui há uma média entre 50 minutos a 95 minutos. Neste intervalo, Olivais e Parque das Nações são as freguesias com ocorrências com maior média de tempo para resolução, com 95 minutos em média. As ocorrências mais rápidas, em média, ocorrem em Alvalade e Campo de Ourique, mais perto de 50 minutos. Em contraste com o mapa coropleto anterior, apesar de haver mais ocorrências noutras freguesias, poderão não ser as mais graves ou difíceis de resolver.

3.4 Dashboards Desenvolvidos

O principal objetivo desta dissertação é construir ferramentas para se conseguir chegar a conclusões sobre os dados fornecidos pelo RSB de Lisboa. Estas conclusões podem ou não vir a ser úteis para o RSB.

É nesta secção que apresentamos então as ferramentas, conjuntos de visualizações com suporte interativo que são os *dashboards*, ou seja, o conjunto de visualizações apoiam-se mutuamente para informar o visualizador para conseguir processar e construir a imagem

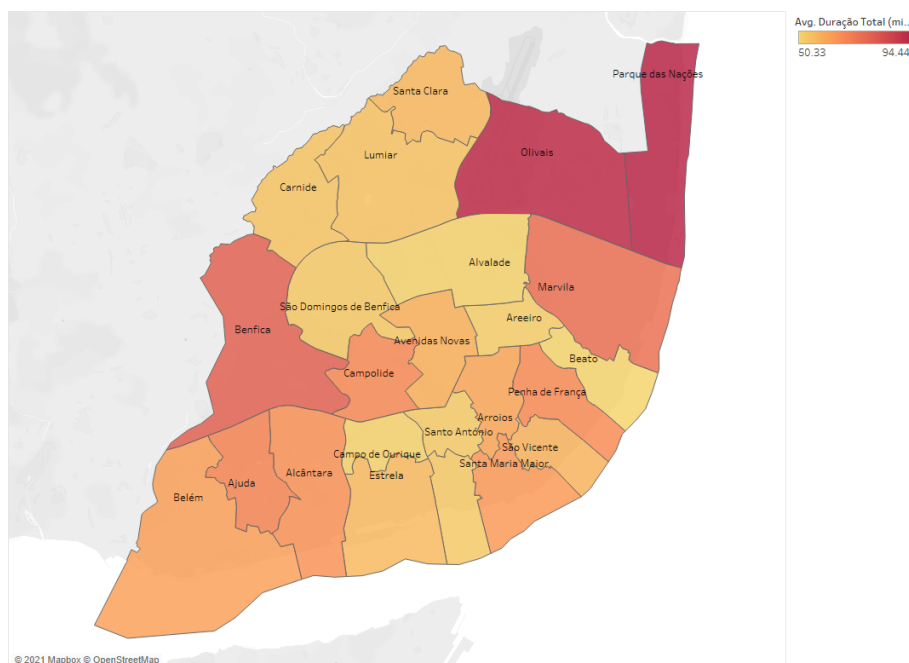


Figura 3.18: Mapa coropleto de Lisboa com a duração média das ocorrências nos cinco anos do *dataset* por freguesia.

dos dados. Isto permite a sua análise e consequentes conclusões sobre as observações feitas com estas ferramentas. Conjuntos de dados têm sempre padrões, mas para efeitos de pertinência procuramos padrões que sejam previsíveis, seja por comportamento humano (falsas construções humanas como por dias da semana ou horas) tanto como por mudanças no ambiente (chuva, estações do ano). Encontrar estes padrões é particularmente útil para futuras previsões, contudo a falta de padrões também é útil, no sentido que se conclui que a variável não está relacionado com o efeito que estamos a estudar.

Por regra geral, estes *dashboards* focam-se em responder a uma questão específica, que depois podem ramificar-se em mais perguntas. Esta regra ajuda a que o *dashboard* fique coeso para dar contexto e apoio apenas àquela pergunta, para que a possível resposta se torne mais evidente. E é por isso que nesta secção descreverei os *dashboards* criados, o que cada elemento faz, porquê aquele tipo de visualização (se for o caso) e listarei a que pergunta(s) da secção 3.1 é que criou a sua necessidade.

3.4.1 D1 - Análise geral da distribuição temporal e espacial das ocorrências RSB

Começando pelo primeiro *dashboard* elaborado, cujo objetivo era dar uma visão global dos dados. Este responde às questões gerais Q1, Q2 e Q4, sendo as questões modeladoras "Há padrões temporais?" e "Há padrões espaciais?". Na figura 3.19, onde está apresentado uma captura da ferramenta, notar-se-á os padrões e valores atípicos.

Observando a figura 3.19 podemos ver que a secção (A) são todos os filtros e parâmetros,

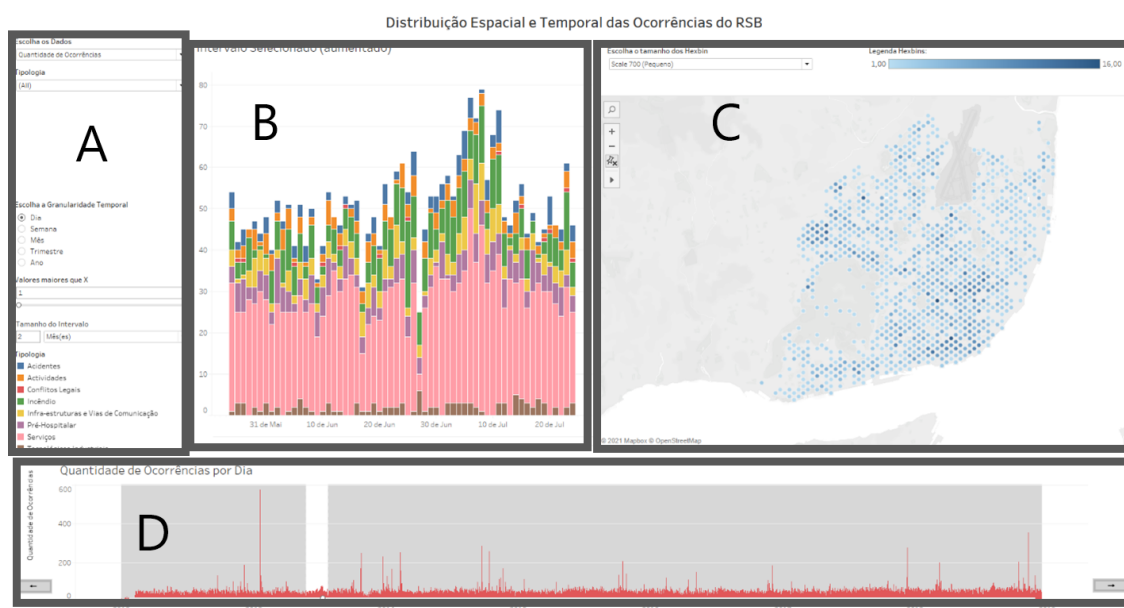


Figura 3.19: Captura de ecrã do *dashboard* D1, intitulado Análise geral da distribuição temporal e espacial das ocorrências RSB.

tal como a legenda do gráfico de barras que se encontra na zona (B), que muda conforme a granularidade do tamanho da janela da *timeline*, que é o espaço branco *highlighted* na zona (D). Se se escolher um tamanho de janela maior que 4 meses o gráfico (B) muda para um gráfico de linhas, com a área preenchida. Isto pode se observar na figura 3.20. Na zona (C) há um mapa coropleto de *hexabins* com a quantidade de ocorrências e a respetiva legenda, e na (D) um gráfico de linhas com a quantidade de ocorrências por dia para todo o conjunto de dados.

Como o objetivo era para se obter uma visão geral dos dados, visão essa que é tanto espacial como temporal, a granularidade da informação apresentada é bastante grossa inicialmente. No entanto, é possível controlar o tamanho da janela na *timeline* temporal, para além de se conseguir ajustar a sua posição, sendo que mudanças aí atualizam as restantes visualização do *dashboard*. Já mencionado anteriormente, o tamanho da janela influencia o aspeto do gráfico B. O tamanho da janela que muda o gráfico foi escolhido, pois a partir dos 5 meses o gráfico de barras não era visível. Também é possível mudar o tamanho dos *hexabins* do mapa, controlando assim a granularidade dos dados apresentados, portanto para uma visão mais fina queremos selecionar um tamanho da escala maior. As opções são: *Scale 700 (Pequeno)*, *Scale 400 (Médio)* e *Scale 100 (Grande)*, como demonstrado na figura 3.21.

Neste *dashboard* é possível mudar os dados que são mostrados, que é o primeiro *drop-down menu* da (A), sendo as opções Quantidade de Ocorrências, Soma de Viaturas, Média de Viaturas, Soma de Homens e Média de Homens. Por "Homens" entende-se que é o número de operacionais.

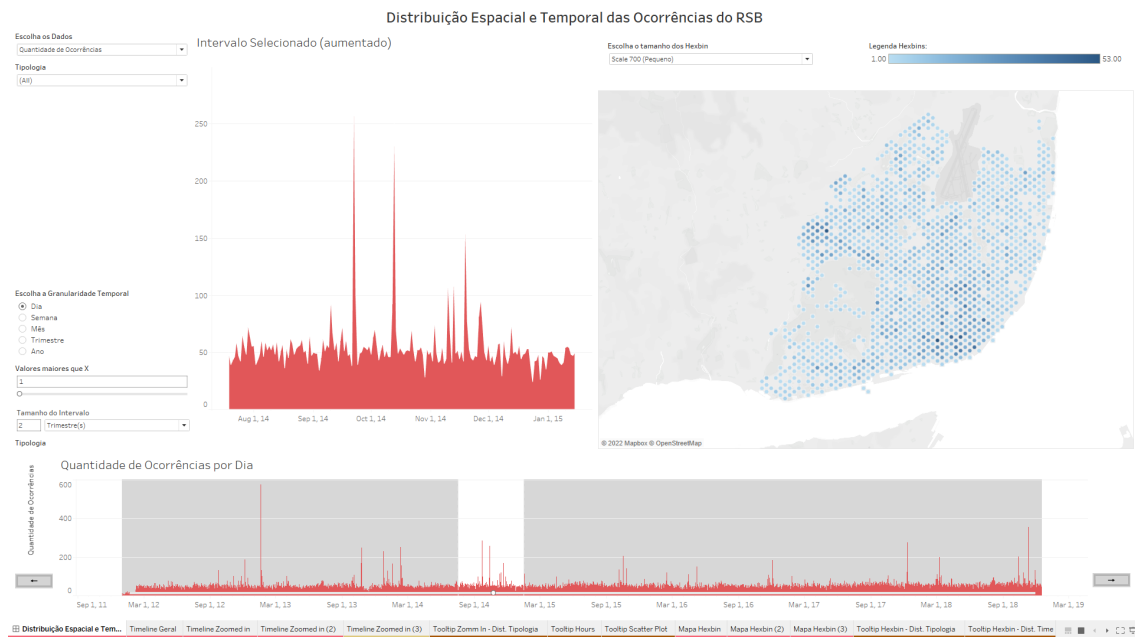


Figura 3.20: Captura de ecrã do *dashboard* D1, com granularidade da janela maior que 4 meses.

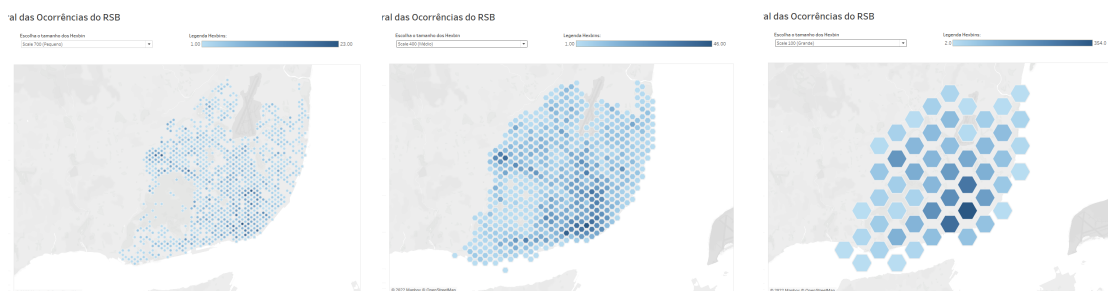


Figura 3.21: Granularidades do mapa de hexágonos: Scale 700 (Pequeno), Scale 400 (Médio) e Scale 100 (Grande) respetivamente.

Para se usar este *dashboard*, em primeiro lugar é necessário identificar o tipo de dados que queremos observar, selecionar um tamanho da janela temporal que nos seja relevante e explorar os valores na *timeline*, movendo a janela usando o *slider*, que se encontra por baixo da *timeline*, ou as setas de cada lateral da *timeline*. Ao mexer a janela, este atualiza o gráfico de barras e o mapa de hexabins. Aqui tentamos observar se existem quebras de padrão nas distribuições na natureza das ocorrências, ou se existe um *hotspots* no mapa de hexabins. Simultaneamente, caso a exceção seja visível da *timeline*, por exemplo identificamos logo na captura do ecrã que existem picos extremos de ocorrências, podemos também verificar se existem explicações na distribuição das naturezas ou se foram eventos localizados a um local em particular no mapa.

Este *dashboard* permitiu entender que existem padrões espaciais, algumas anomalias

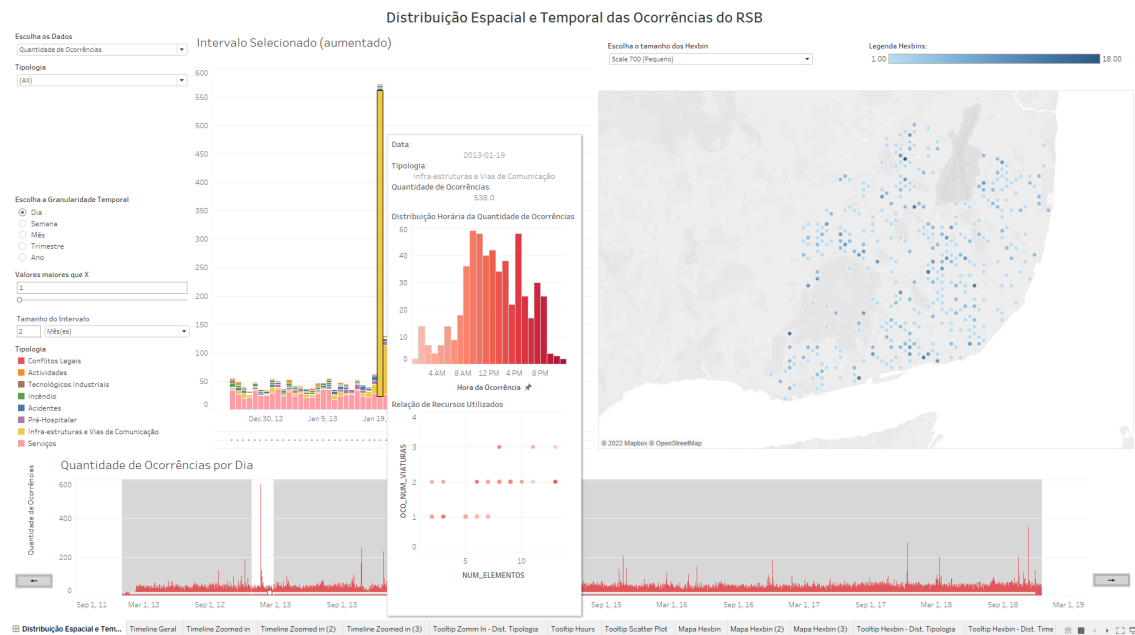


Figura 3.22: Análise ao dia atípico de 19 de janeiro de 2013 utilizando o D01.

temporais (tal como aqueles picos interessantes na *timeline*) e ainda explicar porque é que as mesmas existem. Por exemplo, no dia 19 de janeiro de 2013, no qual 538 das 573 ocorrências desse dia são do tipo "Infraestruturas e Vias de Comunicação". Uma breve pesquisa confirmou que foi um dia de tempestade que causou problemas na cidade. Esta análise foi feita com ajuda da *tooltip* que está inserida no gráfico de zoom. O exemplo referido pode ser observado na figura 3.22.

Por último, queria referir as técnicas de interações implementadas. Já mencionado anteriormente está o uso de *tooltips*. Através de *highlight*, seleção e filtragem, o gráfico de barras atualiza o mapa *hexabins*. Esta *filtragem* é feita tanto por *hover*, como por clique (seleção) em segmentos do gráfico de barras, que permite investigar padrões daquela tipologia. Se quisermos o dia, ou semana, existe uma caixa de quadrados por baixo deste gráfico de barras para se poder selecionar a granularidade escolhida por toda a sua integridade e diversidade de tipologias das ocorrências. Este comportamento pode ser observado na figura 3.22 na falta de hexágonos, porque existem muito menos dados selecionado em comparação ao normal que é mostrar o conjunto de dados selecionado na janela em baixo. Também existe uma *tooltip* em cada azulejo *hex* do mapa, cujo a estrutura pode ser observada na figura 3.23.

3.4.2 D2 - Análise de Padrões Cíclicos

Este *dashboard*, ao contrário do anterior, tem como objetivo explorar uma granularidade mais pequena de padrões temporais, nomeadamente ao nível semanal e diário. As perguntas que definiram o *dashboard* foram as Q2 "Há padrões temporais?" e Q4 "Existem

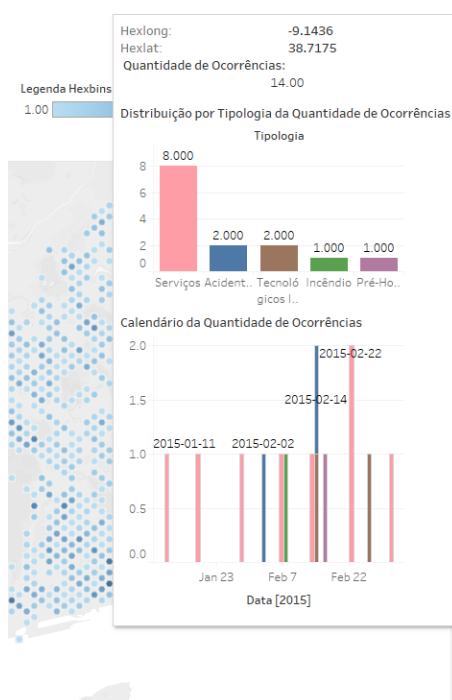


Figura 3.23: *Tooltip* do mapa de *hexabins* do *dashboard* D01.

padrões espaciais?".

Observando a figura 3.24, na zona (A) temos os filtros e parâmetros que controlam o *dashboard*, na (B) uma caixa de bigodes com os cálculos da média de quantidade de ocorrências nos diferentes dias da semana, (C) um gráfico de barras com a quantidade e consequente distribuição das ocorrências pela sua natureza, no (D) outra caixa de bigodes com os cálculos da média de quantidade de ocorrências ao longo das horas do dia e no (E) um mapa de hexabins com a média de quantidade de ocorrências e a respectiva legenda com uma escala de cores divergente.

Os filtros permitem filtrar alturas do ano, ao retirar meses ou usando os filtros por estações do ano. Permitem também filtrar os anos, sendo que estes são úteis para verificar se os padrões são anuais ou globais. E, por último, permite filtrar por tipologia da ocorrência. Imaginando que se quer comparar apenas os padrões de acidentes rodoviários e não os de serviços de abertura de porta, é usado este filtro para focar nas tipologias de interesse. Ao clicar no gráfico de caixa de bigodes dos dias da semana (B) também é possível filtrar pelo dia da semana selecionado. No entanto, este não filtrará o mapa.

Em termos de *tooltips*, as caixas de bigodes apresentam uma página com os vários valores dos limites das mesmas. A *tooltip* do gráfico de barras tem vários gráficos, como se pode ver na figura 3.25. Em primeiro lugar, existe uma lista dos valores absolutos mostrados no gráfico de barras, seguido de um gráfico da distribuição das ocorrências deste tipo pelo dia (em que cada linha e cor é um dia da semana diferente, sendo que se pode notar na figura que não há grande divergência). Posteriormente, existe um gráfico com a distribuição pelas subnaturezas da natureza especificada. É engraçado notar que

3.4. DASHBOARDS DESENVOLVIDOS

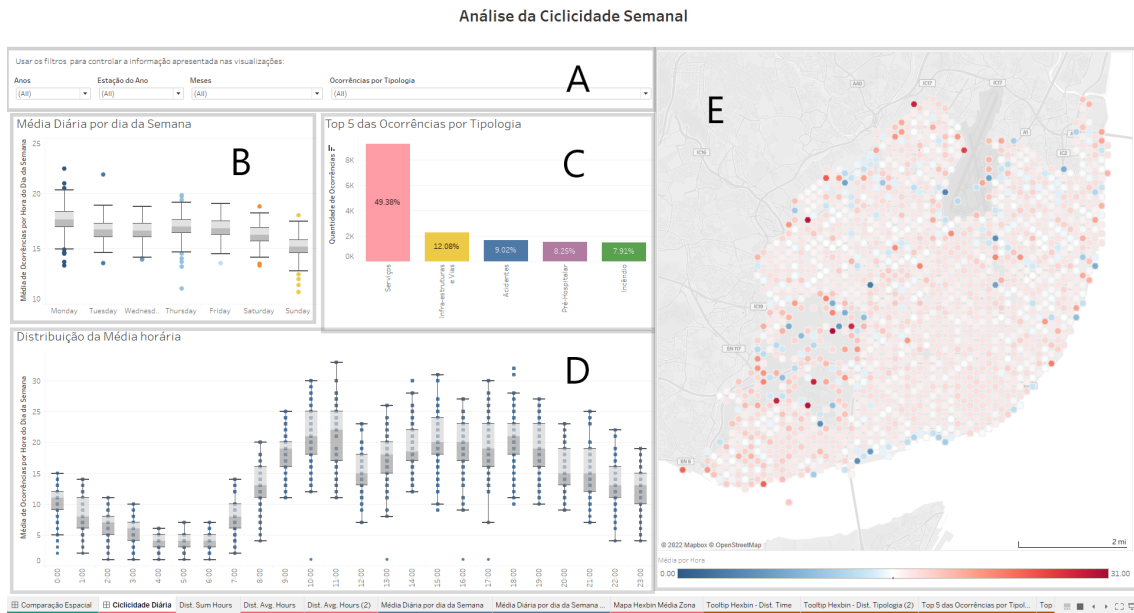


Figura 3.24: Captura de ecrã do *dashboard* D2, intitulado Análise da Ciclicidade Semanal.

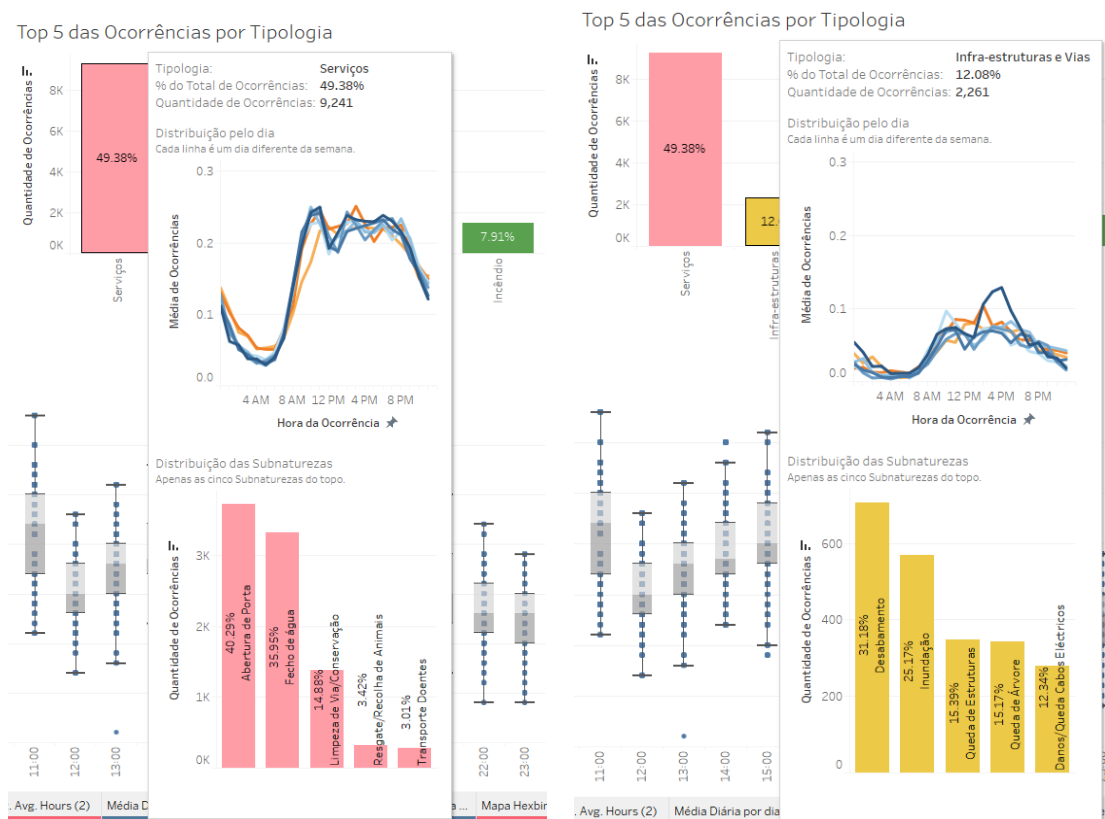


Figura 3.25: Informação apresentada na *tooltip* do gráfico de barras do D02.

apenas nestas *tooltips* se notam grandes diferenças nos padrões diários de cada tipologia. E por último, o mapa de hexabins tem uma *tooltip* para cada azulejo *hex*, onde se apresenta a informação geográfica primeiro (informação da localização daquele *hex*), seguido de um gráfico da distribuição das ocorrências pelo tempo (por dia), para todo o intervalo de tempo selecionado, e ainda um outro gráfico de barras com a distribuição, para aquele azulejo, das naturezas das ocorrências.

Este *dashboard* permitiu ver que, curiosamente, os piores dias (com mais ocorrências em média) de ocorrências são as segundas feiras, visto que apresentam a maior média de entre os dias, e o melhor dia (com menos ocorrências em média) de ocorrências são os domingos. O pico de ocorrências costuma ocorrer pelas 11:00 da manhã, demonstrando que a quantidade de ocorrências acompanha o típico horário diurno, isto é, com mais pessoas acordadas, mais ocorrências são reportadas. Ao explorar os dados também foi interessante comprovar que existem mais incêndios no verão, o que se confirma com este *dashboard*. Este facto pode ser observado nas imagens da figura 3.26, dado que no gráfico de barras os incêndios (que têm a cor verde) estão no topo das incidências, em contraste com o caso do Inverno, que nem aparece no topo 5 de ocorrências por tipologia. Devo mencionar que nesta análise foi omitido a tipologia de ocorrências "Serviços", isto porque devido à sua grande percentagem e quantidade de casos. Foi frequente fazermos exercícios de análise sem essa tipologia incluída. Além disto, os *dashboard* da figura 3.26 também permitem ver que existem outras diferenças nos padrões destas duas estações, por exemplo, a diminuição de casos em média à terça-feira.

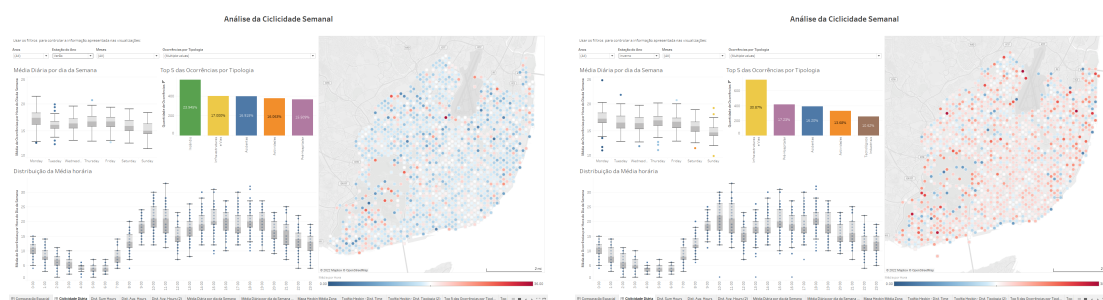


Figura 3.26: Exemplo de uma análise conduzida com o *Dashboard* 02, a comparar as diferenças entre os verões (esquerda) e os invernos (direita).

3.4.3 D3 - Tendências de Freguesias

A intenção deste *dashboard* era permitir analisar o comportamento por freguesia, ou seja, ver o seu perfil e padrões de ocorrências, podendo assim comparar. É portanto uma das maiores questões "Existem comportamentos diferentes entre as freguesias?"(Q3), sendo importante complementar essa questão com as ferramentas que usamos para encontrar padrões tanto espaciais como temporais. Neste caso o foco foi a distribuição das naturezas e o padrões temporais (Q2).

Observando a figura 3.27, na zona (A) temos os filtros de anos, controlo sobre as naturezas incluídas nas várias visualizações, como por exemplo se poder excluir naturezas não urgentes como aberturas de portas sem socorro, e a legenda sequencial do mapa coropleto que se encontra na zona (B). O gráfico (C) é uma caixa de bigodes com as distribuições relativas à duração por natureza em minutos, o (D) um gráfico de barras com a média de ocorrências pelos meses, o (E) é um gráfico de barras com as médias da quantidade de ocorrências pelos dias da semana e por fim o (F) é um gráfico com a distribuição da quantidade de ocorrências por natureza, ao longo do tempo.

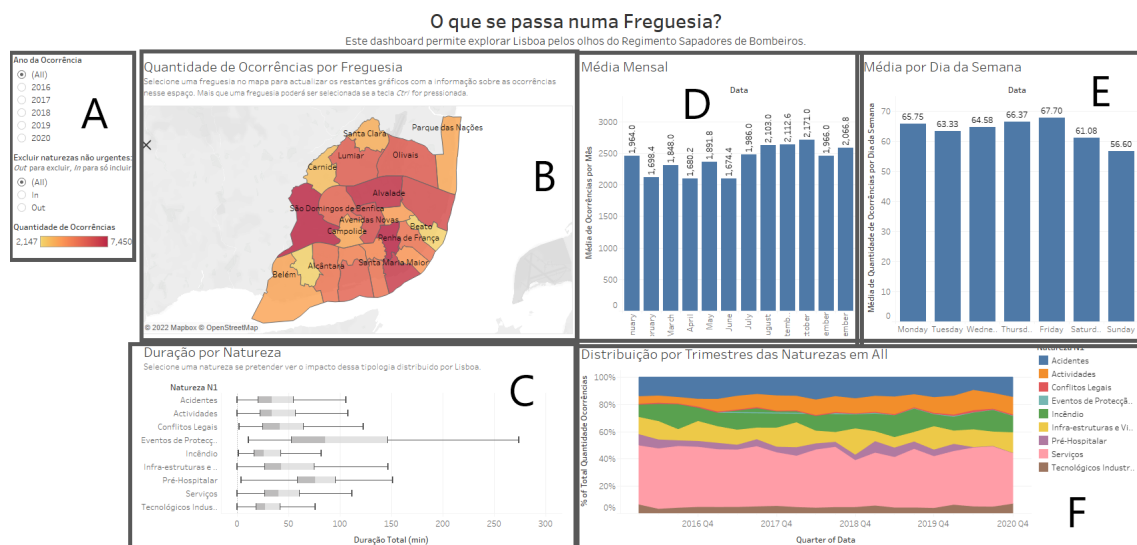


Figura 3.27: Captura de ecrã do *dashboard* D3, intitulado Tendências de Freguesias.

Este *dashboard* é suposto ser utilizado com interação que tira proveito do aspeto visual do mapa. Ou seja, é possível selecionar a(s) freguesia(s) que nos interessam e essa seleção cria a atualização e filtragem dos dados que ocorram na(s) freguesia(s) selecionada(s). Estes dados podem depois ser comparados, tanto através das animações entre as atualizações, como através de análise se notar as diferenças. Esta função pode ser observada na figura 3.28, onde a freguesia selecionada ficou *highlighted* no mapa e as restantes visualizações atualizam-se para mostrarem só os dados relativos à freguesia selecionada.

Algumas conclusões que podem ser tiradas com este *dashboard* é que as freguesias mais problemáticas são as de Benfica, Alvalade e Arroios. Os dias com mais ocorrências costumam ser à sexta, seguida de quinta e segundas. A altura do ano com mais ocorrências começa em Julho e termina em Outubro. Ao selecionar as freguesias mencionadas em cima como as top 3 com mais ocorrências, verifica-se que o máximo destas se encontra nesse intervalo, em Setembro, o que indica que elas contribuem significativamente para parte desse aumento nessa altura. E por fim, olhando para as caixas de bigodes, que os eventos de proteção civil são as ocorrências com mais duração por uma margem grande.

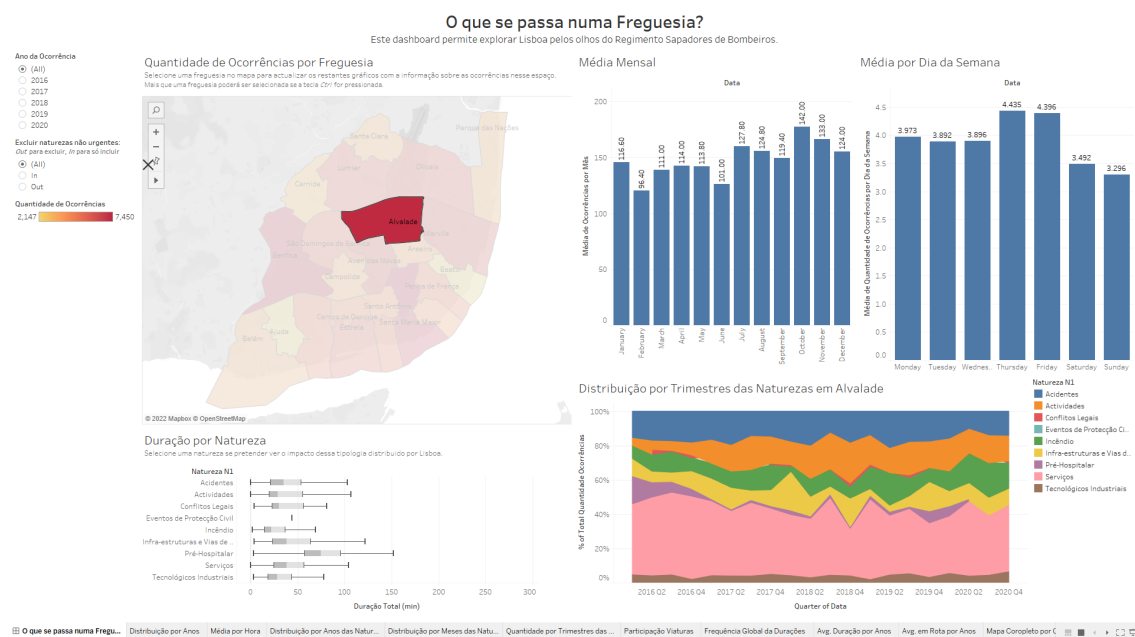


Figura 3.28: Exemplo da seleção de uma freguesia no mapa do *dashboard* D3.

3.4.4 D4 - Tendências ao longo dos Anos

Este *dashboard* foi desenhado para responder à questão geral 1 "Quais as tendências ao longo dos anos?". Tem várias opções interativas entre as visualizações que o formam, possibilitando filtragens e ações de zoom em zonas de interesses.

A figura 3.29 mostra o *dashboard* dividido pelas suas visualizações. A visualização (A) é um gráfico de barras que mostra a quantidade de ocorrências por ano, e permite filtrar as restantes visualizações ao selecionar um ano. Este serve para mostrar a distribuição e diferenças dos anos na quantidade de ocorrências registadas. A (B) é um gráfico de linhas, cada linha é um ano, da média diária de ocorrências por mês. A visualização (C) é um gráfico de linhas que mostra a Média Móvel ao longo de todo os anos registados no conjunto de dados, com parâmetro para selecionar a janela do cálculo da janela com as opções de 7, 14 e 21 dias. O (D) é um gráfico de barras que mostra as distribuições por percentagem total por naturezas, por ano. E em último, o (E) é um gráfico de barras com as distribuições, semelhante ao (D) mas das subcategorias das naturezas.

A figura C responde à questão se há padrões nos meses, ao longo de um ano. A visualização D é apenas relevante com uso de filtros na natureza.

Logo à partida nota-se em padrões, ou desvios dos padrões, *outliers* na visualização com as médias móveis, a (E). Algumas semanas com picos intensos de ocorrências, que selecionando esses dias revelam que normalmente são devido a tempestades, ou semanas com o contrário, em que há baixa ocorrências de casos. Em termos de padrões no ano, não há grande consistência entre as linhas, tanto em movimento como em direção. Embora pareça que os anos costumam começar em alta e baixar (exceto 2019) normalizar no verão

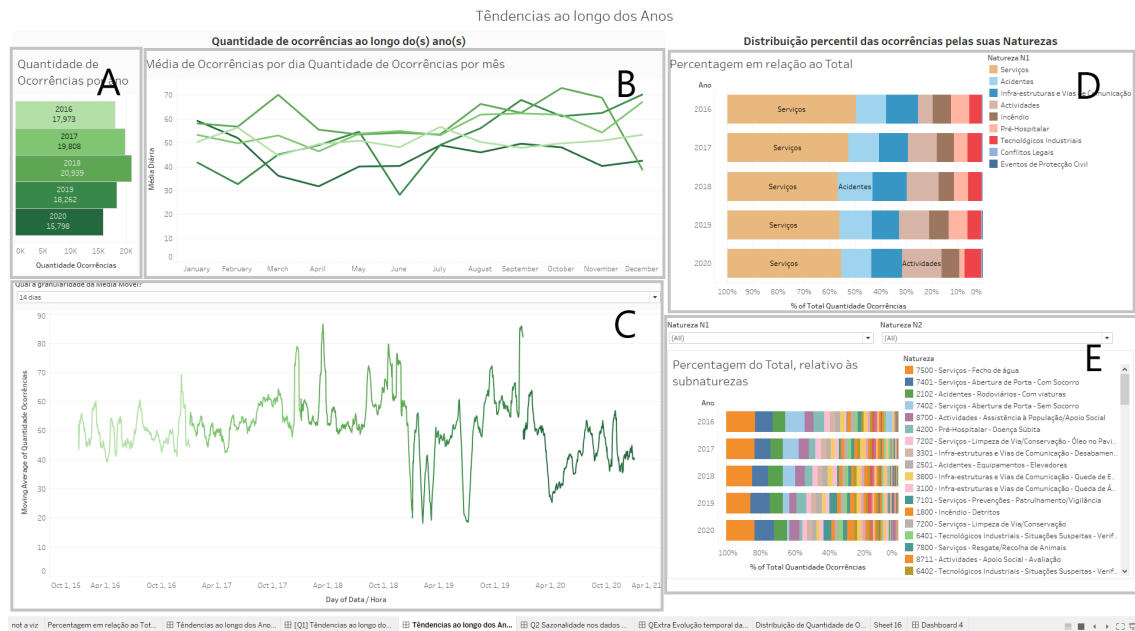


Figura 3.29: Captura de ecrã do *dashboard* 04, intitulado Tendências ao longo dos Anos.

e aumentar no outono de novo.

3.4.5 D5 - Análise de Acidentes Graves

Para se responder à questão das propostas dos RSB número 2, "Quais são as zonas com mais Acidentes Rodoviários com Encarcerados?", foi criado este *dashboard* simples para se ter uma noção do panorama dos dados, que revelou bastante. Esta é uma versão da questão geral 4 (Q04) relativa aos casos da natureza especificada.

Como se pode ver na figura 3.30, o *dashboard* é composto por duas visualizações. Um (A) gráfico de barras listando as ruas e ordenado pela quantidade de ocorrências do tipo de natureza "Acidentes Rodoviários com Encarcerados" e (B) um mapa de pontos, com as cores codificadas à quantidade de ocorrências do mesmo tipo, cuja legenda informa que é uma escala sequencial começando nos amarelos, que são as ruas com menos casos, e terminando nos vermelhos, que são as ruas com mais ocorrências. A rua no conjunto de dados com mais acidentes rodoviários com encarcerados é o Eixo Norte Sul com 95 ocorrências.

Em termos interativos, dar *hover* no gráfico de barras dá *highlight* aos pontos da rua seleccionada no mapa e seleccionar um ponto no mapa filtra o gráfico de barras. Isto foi feito para ajudar a localizar os lugares com os nomes de ruas e vice-versa. Esta mecânica de *brushing* e *linking* é demonstrada na figura 3.31, onde a seleção de ruas no gráfico de barras está a ser *highlighted* no mapa.

Este *dashboard* permite claramente notar que há um *hotspot* significativo, sendo uma conclusão coerente nas duas visualizações, na zona do Eixo Norte-Sul e a Avenida General

CAPÍTULO 3. TRABALHO REALIZADO

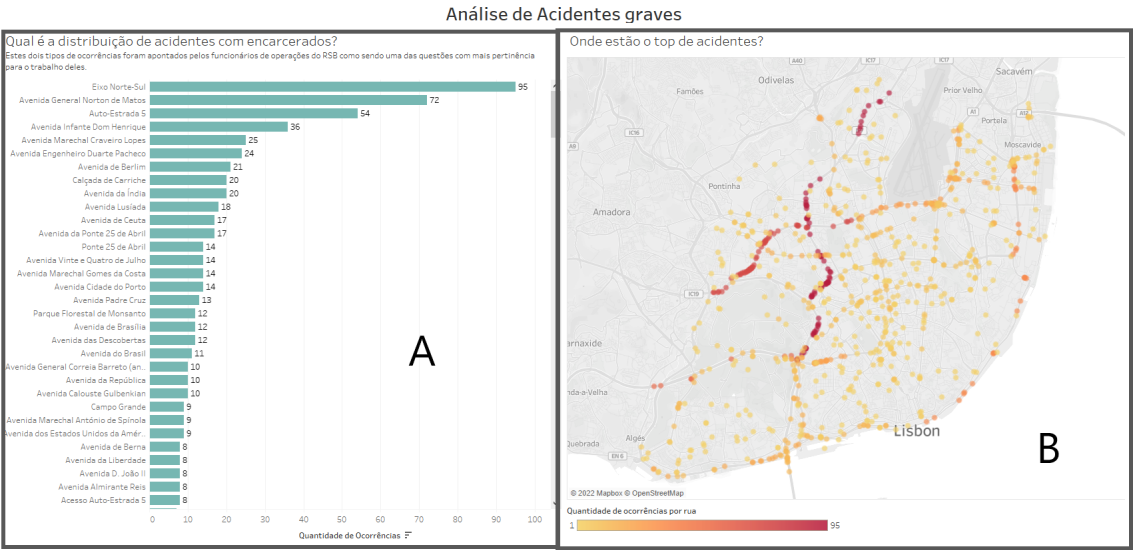


Figura 3.30: Captura de ecrã do *dashboard* 05, intitulado *Análise de Acidentes Graves*. Está dividido em duas visualizações, A e B.

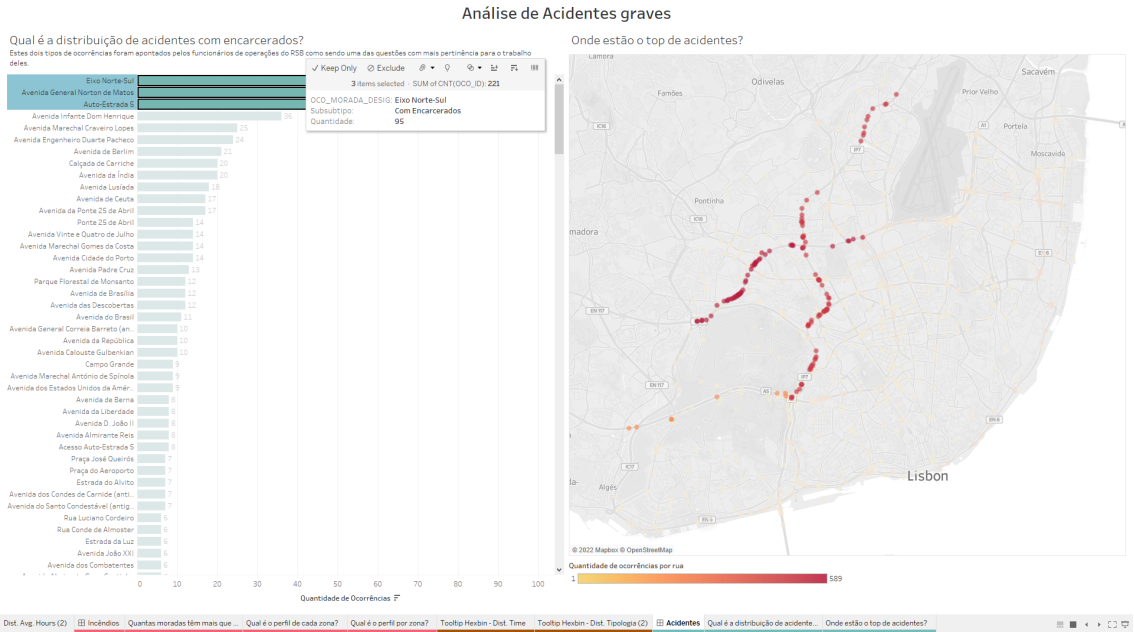


Figura 3.31: Exemplificação da função de *highlight* no *Dashboard* 05.

Norton de Matos, seguido da Auto-Estrada 5. O que significa que, na perspectiva de gestão de recursos dos RSB, esta é a zona de foco dos problemas, sendo um total combinado de 20.333% das ocorrências graves.

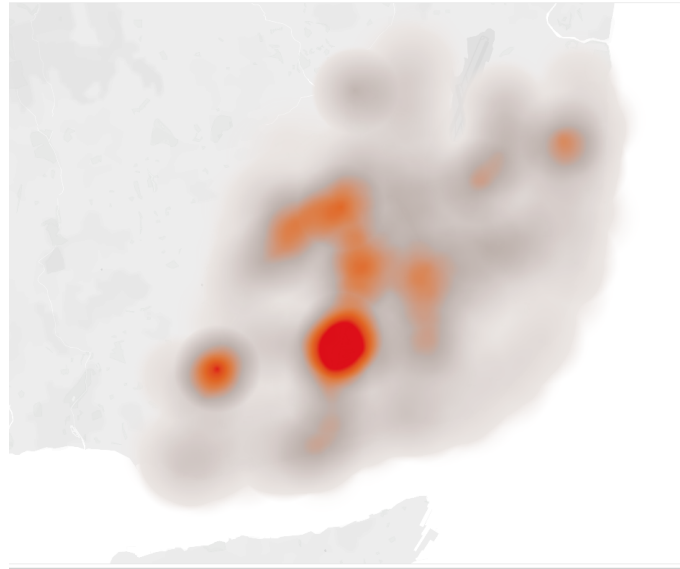


Figura 3.32: Gráfico de densidade com as ocorrências do tipo Acidente com Encarcerados.

Um dos problemas desta vista é que usa as ruas como escala, independentemente do seu tamanho. É efetivamente possível argumentar que tal erro, visto que a análise é sobre a eficácia da assistência dos RSB, é inócuo pois estarem na mesma rua diminui significativamente o tempo para chegar. No entanto, seria ingénuo ignorar o facto que distância é tempo. Para mitigar este problema, a hipótese de que o grande impacto de acidentes graves estão concentrados na zona apontada na conclusão desta visualização, pode ser analisada noutros *dashboards*. Versões anteriores deste *dashboard* tinham um gráfico de densidade, que não é o ideal para exploração dos dados com interação e consequentes ações nos mesmos (seja filtragem ou *highlighting*). O problema residia no facto dos dados já não estarem distribuídos uniformemente, mas sim por rua, que os concentrava em pontos específicos sem intermédios. Como consequência, as filtrações fazem com que o efeito de densidade de sobreposição dos pontos deixe de mostrar regiões problemáticas com fiabilidade. Para este tipo de visualização era necessário mais pontos e mais quantidade de dados de Acidentes Graves. Contudo, na visão geral e fixa de mostrar os dados, todos sem filtrações por anos ou zonas, como podemos ver na figura 3.32, é corroborada a hipótese que a zona problemática é a apontada. Na figura 3.32, as zonas cinzentas são as zonas menos densas de quantidade de ocorrências, enquanto as zonas mais laranjas são as com mais quantidade, ou densidade, de ocorrências.

3.5 Conclusões

Em suma foi realizado trabalho que produziu alguns resultados. Nomeadamente, através do relatório de qualidade de dados foram apontados erros encontrados nos dados e ainda apresentado um conjunto de dados limpos. Além disso, os ficheiros fornecidos foram também unidos. Este ficheiro consolidado, isto é o conjunto de dados, foi mais tarde partilhado com o RSB para que estes pudessem estar informados dos erros presentes, assim como ter acesso à versão corrigida, sendo este foi um dos contributos.

A outra componente que produziu resultados foi a dos *dashboards* construídos. Podemos notar que existe um sucesso ao obter ferramentas que solucionem as questões estabelecidas na secção 3.1. A tabela 3.5 foi criada para se poder visualizar melhor quais os objetivos atingidos, e em que elemento de trabalho está resolvido.

No entanto, houve também perguntas que ficaram por responder. Algumas, como estabelecido anteriormente, demonstravam uma falta de dados. As restantes houve falta de tempo e estabeleceu-se prioridades, nomeadamente para as questões que se encontram respondidas como se pode ver na tabela 3.5. As questões Q5, Q6, Q7, Q8 e a proposta P5 do RSB ficaram por se analisar. Houve em todas um trabalho exploratório inicial que revela que a sua análise é plausível e interessante, portanto, são adequadas para trabalho futuro.

Tabela 3.9: Tabela da eficácia do trabalho realizado.

Questões	D01	D02	D03	D04	D05
Q1.G	✓		✓	✓	
Q1.1	✓		✓	✓	
Q1.2					
Q1.3	✓		✓	✓	
Q1.4				✓	
Q2.G	✓	✓		✓	
Q2.1	✓			✓	
Q2.1.a	✓			✓	
Q2.4	✓	✓		✓	
Q2.5		✓		✓	
Q2.5.a				✓	
Q2.5.b		✓		✓	
Q3.G			✓	✓	
Q3.1			✓	✓	
Q3.2			✓*		
Q4.G	✓	✓			✓
Q4.1	✓				
Q4.2	✓				
Q4.2.a					
Q5.G					
Q6.G					
Q7.G					
Q8.G					
P1.1	✓		✓		
P1.2	✓		✓		
P2					✓

AVALIAÇÃO DO TRABALHO REALIZADO

Este capítulo é dedicado à explicação da avaliação experimental ao trabalho feito nesta dissertação. Começa pela explicação da metodologia da avaliação na secção 4.1, seguindo-se a secção 4.2 com informação demográfica sobre os participantes e, por último, na secção 4.3 são apresentados os resultados. O capítulo é fechado com a secção 4.4, na qual se podem ler as conclusões feitas a partir dos resultados e o seu impacto no trabalho realizado.

4.1 Metodologia de Avaliação

Os *dashboards* foram desenvolvidos até se considerar que atingiram um nível de satisfação no grupo de estudantes, tal como para os orientadores, do grupo da área de visualização de dados. Quando tal acontece, estes *dashboards* são considerados finalizados para passaram à fase de avaliação. O grupo de alunos com temas de tese na área de visualização de dados fizeram parte do processo iterativo dos *dashboards*, fornecendo *feedback* e recomendações que pensavam ser uma potencial melhoria na ferramenta. Considera-se que esta fase é uma avaliação continua do trabalho, permitindo reiniciar processos de modelação e rever as possíveis falhas, reforçando o trabalho desenvolvido e tornando-o mais robusto. Nesta fase também existiu uma tentativa de contacto e partilha dos conteúdos criados (*dashboards*, documentos e análises preliminares). Contudo, houve uma falha de comunicação com o RSB, que obrigou a repensar a forma como a avaliação poderia ser realizada sem utilizadores que dominem a temática dos dados.

A fase de avaliação tem como intenção averiguar se o trabalho está apto para cumprir a função para que foi desenvolvido. Queremos então avaliar a eficiência e eficácia dos *dashboards*, ou seja, não queremos apenas uma resposta binária para saber se cumprem o objetivo, mas avaliar a qualidade com que o fazem. Para isso foi desenvolvido um ambiente de testes, nomeadamente um questionário. O teste foi feito com acesso ao programa do Tableau, em reuniões virtuais com os utilizadores, para se observar as métricas necessárias. Este teste foi aplicado em duas fases.

Uma fase preliminar para recolher *feedback* geral com um utilizador, de maneira a antecipar alguns equívocos, recolher sugestões e fazer algumas melhorias ao ambiente de

teste. Os questionários e *dashboards* foram polidos para minimizar atritos à experiência dos utilizadores (devido à sua falta de familiaridade com o trabalho ou temática), para prever as dúvidas que poderiam surgir, e nesse caso preparar um esclarecimento prévio, ou, caso a dúvida fosse consistente, que essa informação fosse fornecida como contexto. A grande finalidade deste passo foi garantir que o teste experimental avaliava a qualidade do trabalho realizado e não o entendimento do teste. Esta experiência foi feita com um utilizador que não dominava a temática.

A fase seguinte foi a de testes experimentais. Como já foi referido anteriormente, foram observadas métricas que considerámos serem representativas do entendimento, ou falta deste, do utilizador tanto da questão como do *dashboard*. Consideramos então útil registar as seguintes métricas: a resposta, claro, visto que é importante saber se o utilizador conseguiu utilizar a ferramenta da maneira pretendida (sabendo que a solução existe), e se o *dashboard* é intuitivo o suficiente para se compreender como se utiliza; o tempo a atingirem cada resposta, prevê-se que quanto maior o tempo passado por pergunta, maior será o indício de quão pouco ou muito intuitivo é o *dashboard* e as suas *features*. No entanto, a métrica do tempo também poderá relacionar-se com o entendimento da questão, ou seja, se a questão é clara ou se o participante demora a perceber o que é pretendido; E, por último, o número de ações realizadas até ser fornecida a solução, novamente porque quanto maior o número de ações, maior a indicação se o utilizador esteve perdido, seja as questões ou o *dashboard* a causa. Com estes dados foi então possível avaliar a qualidade das ferramentas para cumprirem o objetivo para que foram desenvolvidas, tendo sido feitas análises e conclusões com esses mesmos dados na secção 4.3. O questionário foi apresentado aos utilizadores, em conjunto com os *dashboards* para que pudessem ser consultados e os usados para encontrar as respostas.

O questionário foi realizado com o apoio do Google Forms. Esta escolha foi feita para simplificar o processo de recolha de dados, sendo que as soluções do questionário foram apresentadas em formato de escolha múltipla para dar contexto à solução pretendida. Além disso, caso se optasse por perguntas abertas, seria provável que as respostas obtidas fossem variadas. Desta forma, com escolha múltipla conseguimos garantir a uniformidade das respostas e da sua cotação. A ordem das questões foi a mesma em todos os testes. O ecrã dos computadores foi gravado durante os testes para observar as métricas posteriormente. Antes de cada teste ser iniciado, foi apresentado o contexto, ou seja, explicações sobre o objetivo da dissertação, tal como informações sobre os dados e os mecanismos dos RSB que se revelam nos mesmos e são pertinentes para o entendimento dos trabalhos. À medida que o questionário foi avançando e novos *dashboards* foram introduzidos, foi feita uma apresentação breve para informar quais as características e capacidades do *dashboard*, assim como informações importantes sobre o significado dos dados, tendo sido fornecidos uns minutos para o utilizador se habituar ao mesmo. No final de cada teste foi pedido a opinião e sugestões aos utilizadores.

Como não se pretendia que os questionários ultrapassassem uma hora por pessoa, criou-se o limite de 20 questões analíticas. Os cálculos de tempo englobam 2 minutos por

questão, mais 20 minutos de apresentação dos dados e dos *dashboards*, e ainda a troca de *dashboards* conforme se ia progredindo no questionário. O questionário foi dividido em cinco seções, uma para cada *dashboard*. O questionário com as escolhas múltiplas e a grelha de respostas está disponibilizado no Apêndice B. A seguinte lista foram as questões escolhidas, por secções:

- D01 - Análise geral da distribuição temporal e espacial das ocorrências RSB
 - D01Q1.1 Qual é o segundo maior dia de ocorrências?
 - D01Q1.2 Qual é o primeiro maior dia de ocorrências?
 - D01Q1.3 Nesse dia com maior número de ocorrências, qual é a tipologia com maior número de ocorrências nesse dia?
 - D01Q1.4 Quantas ocorrências houve dessa tipologia?
 - D01Q1.5 E qual foi o total de ocorrências desse dia?
 - D01Q1.6 Na escala grande do mapa, qual são as coordenadas do *hextile* com mais quantidade de ocorrências?
- D02 - Análise de Padrões Cíclicos
 - D02Q2.1 Para todos os dados, qual é o dia da semana com maior mediana?
 - D02Q2.2 Para todos os dados, qual é a hora do dia com maior mediana?
 - D02Q2.3 Para as ocorrências com a tipologia de Acidentes, qual é o dia da semana com maior mediana?
 - D02Q2.4 Mantendo o filtro dos Acidentes: no ano 2015, qual foi o dia com maior mediana?
- D03 - Tendências de Freguesias
 - D03Q3.1 Em que freguesia ocorrem mais Eventos de Protecção Civil?
 - D03Q3.2 Em que mês existem mais Incêndios em média em Marvila?
 - D03Q3.3 Em que dia da semana ocorrem mais Incêndios em Marvila e Santa Clara?
 - D03Q3.4 Para as ocorrências com tipologia não urgente, quais dos seguintes anos têm a quinta feira como o dia com maior média de ocorrências?
- D04 - Tendências ao longo dos Anos
 - D04Q4.1 Qual a natureza com mais ocorrências?
 - D04Q4.2 Qual a sub-natureza com mais ocorrências?
 - D04Q4.3 Qual a segunda sub-natureza com mais ocorrências?

- D04Q4.4 Olhando para o gráfico de média móvel, com a opção de granularidade selecionada nos 21 dias, qual é o dia com menos ocorrências (nos 21 dias anteriores)?
- D04Q4.5 Olhando para o gráfico de média móvel, com a opção de granularidade selecionada nos 7 dias, quantas vezes a linha vai abaixo das 20 ocorrências em média?
- D05 - Análise de Acidentes Graves
 - D05Q5.1 Através da imagem acima, que mostra o top 3 de ruas com mais acidentes, qual diria ser o círculo que em média está mais perto dos acidentes *highlighted*?

4.2 Participantes

Foram feitas no total 4 sessões de testes, com 4 pessoas diferentes. Nenhum dos participantes tinha visto os *dashboards* anteriormente. A informação fornecida como contextualização foi semelhante para todos os participantes. Foram recolhidos os dados demográficos dos participantes, dos quais os resultados se podem ver na figura 4.1. Apenas um tem experiência académica ou profissional com a temática de visualização de dados. Os restantes estão familiarizados apenas como engenheiros e consumidores. Nenhum dos participantes conhecia o contexto dos dados e, como tal, todos foram introduzidos à temática com apresentações idênticas. Como visto na figura, 75% dos participantes tem 26 anos, e o restante 25% tem 30 anos. A recolha destes dados tem como objetivo perceber se alguma destas variáveis tem influência nos restantes fatores da experiência.

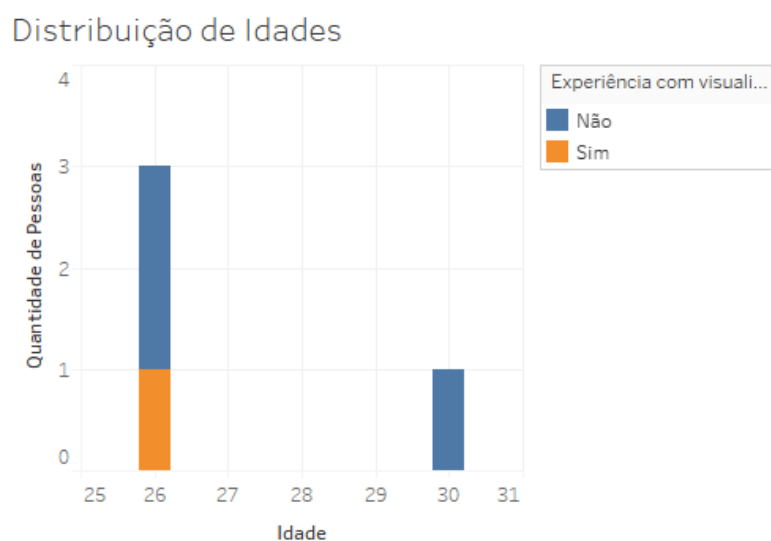


Figura 4.1: Distribuição das Idades e da sua experiência com Visualização de Dados.

4.3 Resultados

Como referido anteriormente na seção 4.1 foram recolhidas as seguintes métricas durante os testes experimentais: a resposta, o tempo até o utilizador chegar a essa conclusão, e o número de ações de interação para responder a cada questão. Neste último foi contado todo o tipo de ações que o utilizador tenha feito propositadamente, nomeadamente filtragens, ações de seleção, ou *hover* para ler uma *tooltip*.

Como podemos ver na figura 4.2, os participantes não tiveram dificuldades a responder corretamente às questões, sendo que 95% das questões têm 100% de taxa de sucesso. A questão D03Q3.4 foi a única com a taxa de sucesso a menos de 100% com uma taxa de 62.5%. Era a única questão no questionário com resposta de caixas de verificação, em vez da escolha múltipla, e 75% dos utilizadores não perceberam bem a questão o que levou a uma resposta incompleta, visto que selecionaram apenas uma das duas respostas válidas.

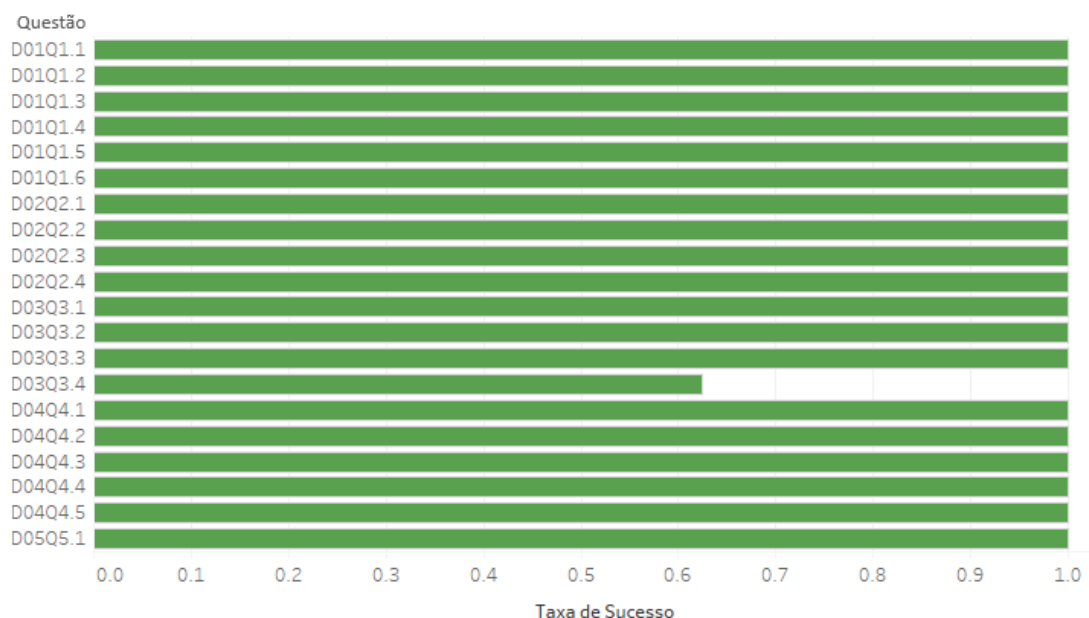


Figura 4.2: Gráfico com a taxa de sucesso de cada questão do questionário.

Como podemos observar na figura 4.3, em termos de tempo médio por resposta, conforme os utilizadores iam avançando nas questões de um *dashboard*, este parece aumentar, o que é o contrário do esperado com habituação da utilização da ferramenta. Contudo, avaliando melhor o questionário, revela-se que as questões vão crescendo em complexidade, o que justifica este aumento de ações e tempo. Em termos de ações, também se nota e confirma essa complexidade crescente. No entanto, sobressai que a primeira questão de todos os grupos tenha ações que não refletem a dificuldade da questão, mas sim a exploração inicial do utilizador do *dashboard*. Após a primeira questão, os participantes ao responderem as restantes questões, já têm alguma habituação ao *dashboard* e conseguem melhorar esse resultado.

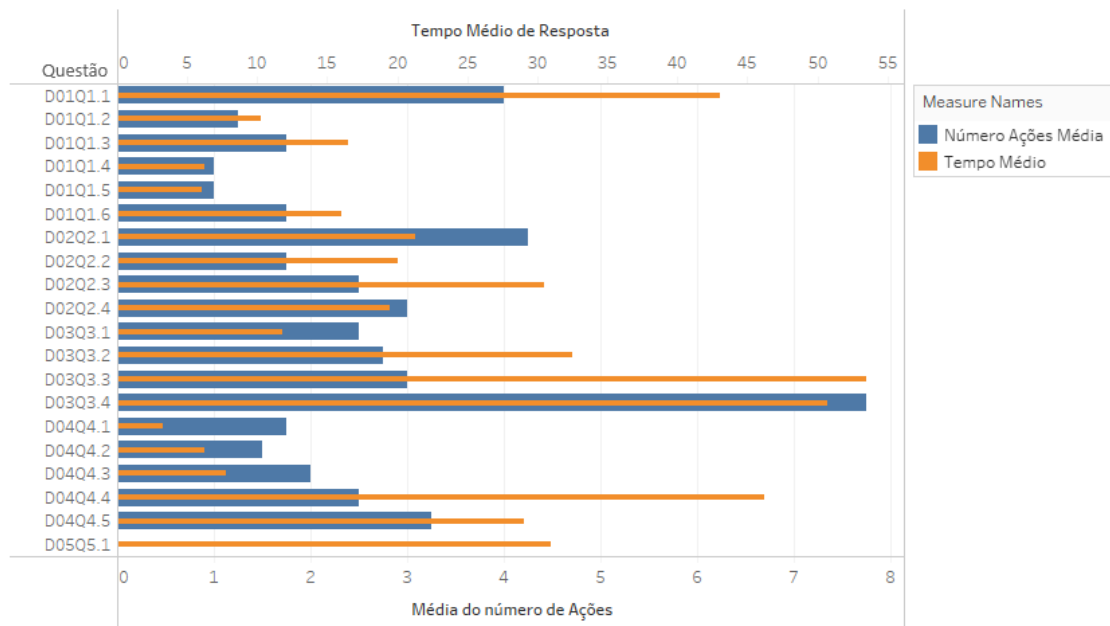


Figura 4.3: Distribuição do tempo médio e número de ações média por cada questão.

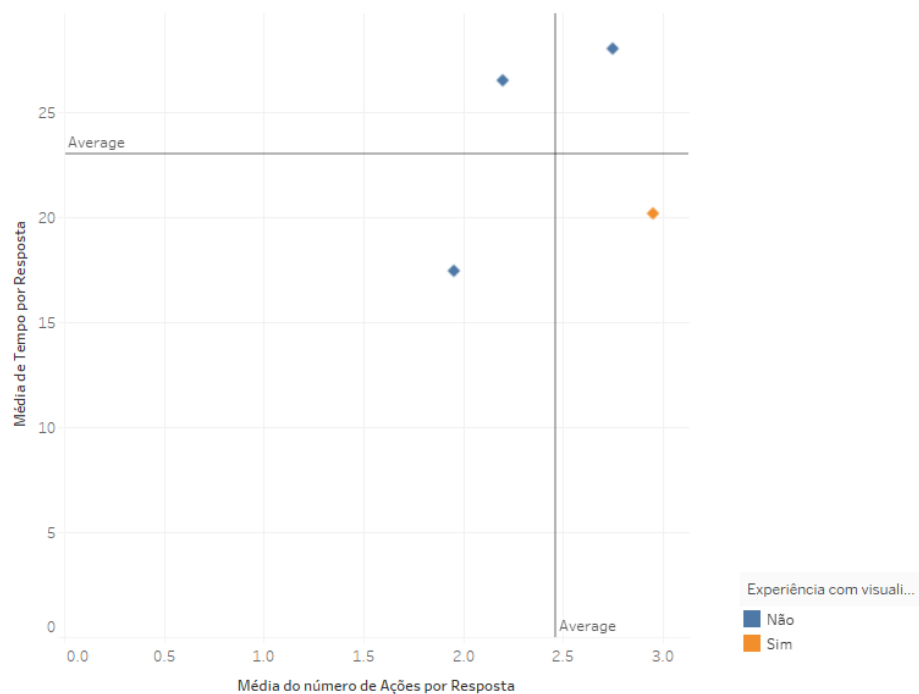


Figura 4.4: Distribuição das médias de tempo e número de ações por utilizador. Cada ponto é um utilizador.

Por último, na figura 4.4, podemos ver os resultados das métricas por utilizador. Não há

dados suficientes para se tirar conclusões concretas, mas ter experiência em visualização de dados não pareceu ter impacto na realização do teste. O que poderá querer dizer que as ferramentas criadas são intuitivas para todos os níveis de conhecimento na área.

4.4 Conclusões

Em suma, o questionário não revelou problemas no trabalho realizado. A única questão que não teve cotação máxima foi a D03Q3.4. Creio que esta pontuação é resultante de um lapso na fase preliminar de preparação e validação do questionário. De qualquer forma, este resultado é tanto revigorante, como preocupante, visto que os resultados obtidos poderão ter sido causados pelo baixo número de participantes, ou até por ter sido demasiado superficial para revelar quaisquer defeitos.

No entanto, o questionário foi uma experiência enriquecedora porque me permitiu aprender as maneiras como diferentes utilizadores, não familiares com o projeto, interagiram com os *dashboards*, por vezes de maneiras inesperadas. Em particular, no *dashboard* D01, todos os utilizadores responderam de maneira não prevista às três primeiras questões que envolviam a *timeline*. Como demonstrado na apresentação, alguns usaram a ferramenta de *highlighting* por valores maiores que o valor escolhido, mas todos utilizaram a *tooltip* da *timeline* para ver os valores totais por dia, em vez de preferirem o *slider*. Quando precisaram de mais informação, podiam usar o gráfico que fazia o zoom da janela selecionada na *timeline* (através do *slider*). Na questão que era necessário selecionar uma parte específica da *timeline* para se ver valores no gráfico *zoomed in*, vários tentaram clicar na linha do dia para fazerem esse zoom, em vez de usarem o *slider*, o que revela que pelo menos inicialmente a *timeline* não é intuitiva.

Foi gratificante ver os utilizadores a operarem os *dashboards* com eficiência e eficácia, ligando no seu mapa mental as visualizações e as interações possíveis para chegarem ao resultado pretendido da pergunta. Pode-se concluir que as ferramentas criadas são prometedoras a alcançarem os objetivos para que foram desenhadas.

CONCLUSÕES E TRABALHO FUTURO

O principal objetivo desta dissertação foi cumprido com sucesso. O estudo de técnicas de construção de ferramentas para estudo de dados espaciotemporais foi concluído e avaliado, demonstrando resultados prometedores. Os objetivos secundários foram mais complicados de avaliar, não obstante poder-se-á concluir que também foram alcançados com sucesso, pois foi possível obter algumas conclusões das análises dos dados dos RSB e ainda quais ferramentas respondiam a todas as questões analíticas geradas. Ainda que nem todas as questões tenham sido respondidas, foram definidas prioridades e essas perguntas foram respondidas com sucesso. As ferramentas que tentam responder a uma ou mais questões foram criadas usando a metodologia CRISP-DM e seguindo as boas práticas de visualização de dados.

No decorrer da dissertação também surgiram outros objetivos e oportunidades. Os dados fornecidos pelo RSB seguiram para um relatório de dados, tal como o conjunto de dados com as melhorias feitas, que foram partilhados com a entidade para poderem anotar os erros encontrados.

Pessoalmente também ajudei a desenvolver os inícios de uma linguagem para idealizar *dashboards* chamada de *Interaction Visualization Modeling Language*, ou IVML. Esta encontra-se descrita na dissertação da Maria Ferreira [21], uma colega do grupo de investigação de visualização de dados, cujo o documento produzido é o mais completo até à data.

A IVML foi um projeto elaborado pelos alunos de visualização de dados no ano 2019/20 para responder à necessidade de melhorar o planeamento de *dashboards* antes de começar a sua modelação. É possível estender as funcionalidades para ser uma ferramenta de documentação dos trabalhos, com a vantagem é que descreve partes normalmente invisíveis, como é o caso da parte interativa deste tipo de projetos. Em termos de modelação esta linguagem não esteve aprimorada a tempo de ajudar no processo de modelação. Fui um dos cinco alunos que ajudou a participar no início deste projeto, tendo sugerido a ferramenta do *Diagrams.net* [22], que é uma aplicação web *open-source* para desenhar diagramas. Em particular tem a funcionalidade de se poder criar e guardar bibliotecas próprias no formato XML. A gestão dos recursos que se juntaram na biblioteca, e partilha dos mesmos, foi feita através da Google Drive. No apêndice C pode-se consultar os

diagramas realizados para experimentar e ajudar a apurar a linguagem.

Como já referido, o processo de investigação e documentação desta ferramenta ainda está em desenvolvimento. Almeja-se que eventualmente esta linguagem venha a satisfazer todos os tipos de visualizações interativas, para todo o tipo de casos de uso, e que se torne a norma para modelação, desenho e documentação nesta área.

Por último, conclui várias formações sobre como utilizar o Tableau, onde ganhei toda a experiência que usei nesta dissertação. Usei esse conhecimento para ajudar outros colegas de dissertação, tal como fui ajudada por eles, sempre aprendendo.

Há várias partes mencionadas nesta dissertação que podem ser considerados para trabalhos futuros. A partir do levantamento de questões da secção da abordagem 3.1, onde se pode consultar uma listagem de problemas que o RSB partilhou, deu-se origem ao apêndice A com os problemas definidos em formato de propostas e com as justificações sobre como os dados correntes não são suficientes para analisar essas mesmas propostas. No entanto, caso haja oportunidade de obter os dados em falta, considero que são boas propostas para trabalhos futuros. Também foi mencionado que existiram questões para as quais os não foram desenvolvidas ferramentas, por falta de tempo, que se consideram bons temas sem condicionantes.

Por último, relacionado diretamente com o trabalho realizado nesta dissertação, considero que há sempre espaço para melhorias. Nomeadamente, poder-se-ia abrir o escopo do trabalho para outros casos de estudo e avaliar se as ferramentas continuariam a funcionar eficazmente para outros dados espaciotemporais noutras temáticas.

BIBLIOGRAFIA

- [1] J. M. Lourenço. *The NOVAtthesis L^AT_EX Template User's Manual*. NOVA University Lisbon. 2021. URL: <https://github.com/joaomlourenco/novathesis/raw/master/template.pdf> (ver p. ii).
- [2] W. Aigner et al. *Visualization of time-oriented data*. Springer Science & Business Media, 2011 (ver pp. 1, 2, 6, 7, 10–12, 25–33, 37).
- [3] Z.-W. Liu e M.-Y. Tsay. «Historical Overview of Data Visualization and its Attempts and Reflections in LIS». Em: (2022), pp. 930–942 (ver p. 1).
- [4] T. Munzner. *Visualization analysis and design*. CRC press, 2014 (ver p. 2).
- [5] S. Robert. «Information Visualization-Design for Interaction». Em: *UK Pearson Educ Ltd* (2007) (ver p. 6).
- [6] C. Ware. *Information Visualization: Perception for Design*. Interactive Technologies. Elsevier Science, 2012. ISBN: 9780123814654. URL: <https://books.google.pt/books?id=UpYCSS6snnAC> (ver p. 6).
- [7] M. O. Ward, G. Grinstein e D. Keim. *Interactive data visualization: foundations, techniques, and applications*. CRC Press, 2010 (ver p. 7).
- [8] S. Dos Santos e K. Brodlie. «Gaining understanding of multivariate and multidimensional data through visualization». Em: *Computers & Graphics* 28.3 (2004), pp. 311–325 (ver p. 7).
- [9] S. Rodrigues e A. Figueiras. «There and then: interacting with spatio-temporal visualization». Em: (2020), pp. 146–152. DOI: [10.1109/IV51561.2020.00033](https://doi.org/10.1109/IV51561.2020.00033) (ver p. 8).
- [10] C. Combi, E. Keravnou-Papailiou e Y. Shahar. *Temporal information systems in medicine*. Springer Science & Business Media, 2010 (ver pp. 12, 13).
- [11] M. Nöllenburg. «Geographic visualization». Em: *Human-centered visualization environments*. Springer, 2007, pp. 257–294 (ver p. 16).
- [12] N. Andrienko e G. Andrienko. *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media, 2006 (ver pp. 16, 17).

- [13] H. Howard et al. *Thematic Cartography and Geovisualization*. Prentice Hall, 2008 (ver p. 16).
- [14] J. Mackinlay. «Automating the design of graphical presentations of relational information». Em: *Acm Transactions On Graphics (Tog)* 5.2 (1986), pp. 110–141 (ver pp. 18–20).
- [15] CDC. *GIS Resources Types of Thematic Maps*. 2010. URL: <https://www.cdc.gov/dhds/maps/gisx/resources/thematic-maps.html> (acedido em 2020-07-27) (ver pp. 20–23).
- [16] C. Forlines e K. Wittenburg. «Wakame: Sense making of multi-dimensional spatial-temporal data». Em: (2010-01), pp. 33–40. DOI: [10.1145/1842993.1843000](https://doi.org/10.1145/1842993.1843000) (ver p. 26).
- [17] P. Gatalsky, N. Andrienko e G. Andrienko. «Interactive analysis of event data using space-time cube». Em: (2004), pp. 145–152. DOI: [10.1109/IV.2004.1320137](https://doi.org/10.1109/IV.2004.1320137) (ver p. 31).
- [18] *Introdução à CRISP-DM*. URL: <https://www.ibm.com/docs/en/spss-modeler/18.2.0?topic=guide-introduction-crisp-dm> (acedido em 2021-10-13) (ver pp. 35, 43).
- [19] *Tableau Products*. URL: <https://www.tableau.com/products> (acedido em 2022-04-06) (ver p. 41).
- [20] M. Camilo. «Análise de padrões dos pedidos de intervenção do Regimento de Sapadores Bombeiros». Tese de mestrado. Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, 2020 (ver p. 49).
- [21] M. Ferreira. «Visualização Interativa de Dados para Projetos de Monitorização Fitossanitária». Tese de mestrado. Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, 2021 (ver p. 79).
- [22] *Diagram software and Flowchart maker*. URL: <https://www.diagrams.net/> (acedido em 2020-11-24) (ver pp. 79, 92).

PROPOSTAS DE ABORDAGENS

Com o intuito de compilar todas as questões e as preocupações do Regimento Sapador de Bombeiros de Lisboa (RSB), foi criado este documento que delinea uma possível abordagem. A grande parte das questões focou-se na tipologia da ocorrência. No sentido que cada tipologia tem recursos e metodologias diferentes que levam a preocupações também diferentes, umas mais graves que outras. As dificuldades apresentadas envolvem as seguintes tipologias:

- 3100 - Infra-estruturas e Vias de Comunicação - Queda de Árvore
- 7401 - Serviços - Abertura de Porta - Com Socorro
- 7402 - Serviços - Abertura de Porta - Sem Socorro
- 3500 - Infra-estruturas e Vias de Comunicação - Inundação
- 3501 - Infra-estruturas e Vias de Comunicação - Inundação - Espaço Privado
- 3502 - Infra-estruturas e Vias de Comunicação - Inundação - Espaço Público
- 2102 - Acidentes - Rodoviários - Com viaturas
- 2103 - Acidentes - Rodoviários - C/ Encarcerados
- 1200 - Incêndio - Agrícola
- 1800 - Incêndio - Detritos
- 1408 - Incêndio - Edifício (Infra-estrutura/Instalação) - Comercial/Lojas/Feiras/Gare de Transporte
- 1409 - Incêndio - Edifício (Infra-estrutura/Instalação) - Cultura/Museu/Arte/Biblioteca
- 1430 - Incêndio - Edifício (Infra-estrutura/Instalação) - Edifício de grande altura (29 m)
- 1420 - Incêndio - Edifício (Infra-estrutura/Instalação) - Edifício Devoluto/Degradado

- 1404 - Incêndio - Edifício (Infra-estrutura/Instalação) - Escolar
- 1406 - Incêndio - Edifício (Infra-estrutura/Instalação) - Espectáculo/Lazer/Culto Religioso
- 1402 - Incêndio - Edifício (Infra-estrutura/Instalação) - Estacionamento
- 1401 - Incêndio - Edifício (Infra-estrutura/Instalação) - Habitação
- 1405 - Incêndio - Edifício (Infra-estrutura/Instalação) - Hospitalar/Lar
- 1407 - Incêndio - Edifício (Infra-estrutura/Instalação) - Hoteleira e similar
- 1411 - Incêndio - Edifício (Infra-estrutura/Instalação) - Indústria/Oficina/Armazém
- 1410 - Incêndio - Edifício (Infra-estrutura/Instalação) - Militar/Forças Segurança
- 1403 - Incêndio - Edifício (Infra-estrutura/Instalação) - Serviços
- 1500 - Incêndio - Equipamentos (sem afectação do ambiente)
- 1501 - Incêndio - Equipamentos (sem afectação do ambiente) - Contentores de lixo
- 1300 - Incêndio - Inculto
- 1100 - Incêndio - Povoamento Florestal
- 1702 - Incêndio - Transportes - Aéreo
- 1704 - Incêndio - Transportes - Aquático
- 1703 - Incêndio - Transportes - Ferroviário
- 1701 - Incêndio - Transportes - Rodoviário

Reduzindo às naturezas bases, são ocorrências de envolvem Incêndios, ou Acidentes Rodoviários, ou problemas nas Infraestruturas e Vias de Comunicação e por último, Serviços.

A.1 Acidentes Rodoviários e Incêndios Recorrentes

Começando por o problema de ocorrências recorrentes, especificamente de acidentes e incêndios. Apesar das naturezas serem muito diferentes, a abordagem é semelhante. Contudo, continuando no facto de serem dados distintos, a organização desses pontos poderá ser diferente. É conhecido que as estradas criam zonas de acidentes chamadas pontos negros. Estes são lanços de estrada, com máximo de 200 metros de extensão, no qual se registaram, pelo menos, cinco acidentes com vítimas no ano em análise, e cuja soma de indicadores de gravidade é superior a 20. Isto significa que deve haver na distribuição dos

dados uma concentração de pontos em certas estradas, que se chamam em visualização de dados de *hotspots*. Existem vários dados que poderemos analisar, nomeadamente as tipologias:

- Acidentes Rodoviários: 2102 e 2103

Nota sobre a natureza 2101 - Acidentes - Rodoviários – Atropelamento: foi-nos dito que a prioridade de atropelamentos não é dos bombeiros. Portanto a qualidade ou interesse da análise das 3101 é questionável, visto que os dados não são completos e não temos acesso à informação de qual o processo de decisão de quem vai socorrer.

- Incêndios: 1200, 1800, 1408, 1409, 1420, 1430, 1404, 1406, 1402, 1401, 1405, 1407, 1411, 1410, 1403, 1500, 1501, 1300, 1100, 1702, 1704, 1703, 1701.

O primeiro passo é encontrar esses *hotspots*, ou seja, se se verificar uma distribuição uniforme pelo mapa de Lisboa, ou se existe um padrão nos dados. Se existirem, o passo seguinte é verificar se estão presentes mudanças com o tempo, possivelmente na granularidade dos anos para averiguar se sempre existiram esse padrão e/ou se moveram ao longo do tempo. Posto isto, podemos analisar se existe algum padrão na parte temporal, se são ou existem mudanças sazonais. Seria também interessante tentar analisar o impacto da meteorologia nestes dados.

A.2 Acidentes Rodoviários com Encarcerados

Dentro da natureza específica dos Acidentes Rodoviários, foi expresso específico interesse nos Acidentes com Encarcerados, pois à semelhança da situação com os VE37, existem também um número limitado dos veículos equipados com as ferramentas de resgate hidráulico para os acidentes mais graves. “Ocorrências de acidente com encarcerados, a ordenança do RSB é a seguinte: VUCI, VSAT, VCOT e ABSC (se operacional e depende do pedido realizado pelo 112, ou INEM - se estão a ambulâncias deles no local);”

Neste tipo podemos primeiro observar qual é a distribuição espacial deste tipo de acidentes e também analisar qual o tempo médio de resposta para cada um desses *clusters*. Tendo acesso a dados meteorológicos, visto que a qualidade das estradas muda com a temperatura, chuvas e outros fatores atmosféricos, podemos tentar correlacionar a mudança desses fatores com aumentos em acidentes graves. Há uma natureza diretamente relacionada com este tópico, 2103. Segundo o que foi discutido, poderá incluir encarcerados menos graves, que não precisem do tal mecanismo hidráulico limitado, mas deverá revelar alguns padrões de onde acontecem este tipo de acidentes.

A.3 Serviço de Abertura de Porta

Os RSB hipotizam que houve um aumento das ocorrências desta natureza no 2020, possivelmente devido a pessoas idosas que se encontram mais isoladas. Não é possível

saber a informação pessoal de cada pessoa em cada edifício, o mais próximo que se pode fazer é inferir com base, por exemplo, usando os dados dos Censos mais recentes para saber qual a idade média numa freguesia e tentar tirar paralelos sobre os padrões dessa região. O que podemos fazer sem ter que pedir mais dados, portanto acelerando as possíveis respostas à questão, é verificar se realmente houve algum aumento desta natureza em comparação a outros anos. Se a variância se verificar realmente fora do comum, podemos então prosseguir com o pedido de dados e tentar investigar por esse lado se há alguma variável explicativa.

A.4 Incêndios e VE37 – Veículo com Escada Giratória de 37 metros

Este é um tema interessante cuja exploração seria útil. O que nos foi explicado é que existe um número limitado de camiões com as grandes escadas, capazes de assistir incêndios em edifícios altos. Logo, se souberem onde há edifícios maiores, a gestão de recursos, quando ocorre uma chamada, poderia ser mais eficiente. De momento não é possível, visto que há falta de dados relativos às alturas das edificações de Lisboa. Apesar de haver a natureza de código 1430, cujo título é *Incêndio (...) Edifício de grande altura (29 m)*, existem apenas 7 ocorrências no total do conjunto de dados, não sendo possível uma análise com tão poucos dados. No entanto, é um trabalho futuro interessante com potencial de utilidade diária. Dado que seja possível procurar ou criar estes dados, que poderá ser um tema por si só.

A.5 Queda de Árvores

A queda de árvores pode ocorrer por vários motivos, dos quais o maior fator são alguma forma de tempestade e ventos fortes. Os RSB acreditam que poderá ser relacionado com grandes secas seguidas de grandes chuvadas, que fragilizam a estrutura de árvore. Embora esta hipótese poderá não se revelar nos dados, é certamente uma dica que podemos seguir: qual a relação entre esta ocorrência e mudanças em fatores meteorológicos? Logo, nesta natureza podemos analisar qual o impacto da meteorologia (do qual já temos acesso a alguns dados provenientes do IPMA), visto ser um grande fator neste tipo de ocorrências. Também, como na seção de ocorrências recorrentes, é útil analisar se há algum padrão na distribuição da incidência de queda de árvores em Lisboa porque podem indicar área mais fragilizadas a este tipo de perigos. Uma análise rápida deste tema poderá revelar se é interessante investigar mais. A natureza relativa a este tópico tem o código: 3100.

A.6 Cheias Rápidas

Foi mencionado que era útil saber onde poderiam estar a ocorrer cheias rápidas, visto que causam danos igualmente rápidos e o seu processamento é fácil. Ao contrário de

outros temas, este não tem uma ligação direta a nenhuma das naturezas, sendo as mais próximas as ocorrências com o código 3500, 3501 e 3502, do tipo Inundação. Contudo, como mencionado na reunião, é praticamente impossível saber onde na cidade irá chover mais. Outras dificuldades são a falta de dados sobre as grelhas de águas pluviais, tal como a inclinação de ruas (para prever onde escorrerá a água) a uma granularidade por rua. É, no entanto, um trabalho possível, e com resultados importantes e interessantes para o RSB.

QUESTIONÁRIO COM SOLUÇÕES

- D01 - Análise geral da distribuição temporal e espacial das ocorrências RSB
 - D01Q1.1 Qual é o segundo maior dia de ocorrências?
 - A 2018-11-11 ✓
 - B 2013-01-19
 - C 2017-04-25
 - D01Q1.2 Qual é o primeiro maior dia de ocorrências?
 - A 2013-01-19 ✓
 - B 2018-11-11
 - C 2017-04-25
 - D01Q1.3 Nesse dia com maior número de ocorrências, qual é a tipologia com maior número de ocorrências nesse dia?
 - A Infra-estruturas e Vias de Comunicação ✓
 - B Serviços
 - C Acidentes
 - D01Q1.4 Quantas ocorrências houve dessa tipologia?
 - A 538 ✓
 - B 581
 - C 394
 - D01Q1.5 E qual foi o total de ocorrências desse dia?
 - A 573 ✓
 - B 541
 - C 620
 - D01Q1.6 Na escala grande do mapa, qual são as coordenadas do *hextile* com mais quantidade de ocorrências?
 - A 38.7200 -9.1350 ✓

B 38.7287 -9.1500

C 38.7373 -9.1350

- D02 - Análise de Padrões Cíclicos

- D02Q2.1 Para todos os dados, qual é o dia da semana com maior mediana?

- A Segunda-feira ✓

- B Quinta-feira

- C Sexta-feira

- D02Q2.2 Para todos os dados, qual é a hora do dia com maior mediana?

- A 11:00 ✓

- B 15:00

- C 18:00

- D02Q2.3 Para as ocorrências com a tipologia de Acidentes, qual é o dia da semana com maior mediana?

- A Segunda-feira ✓

- B Terça-feira

- C Quinta-feira

- D02Q2.4 Mantendo o filtro dos Acidentes: no ano 2015, qual foi o dia com maior mediana?

- A Segunda-feira

- B Terça-feira ✓

- C Quinta-feira

- D03 - Tendências de Freguesias

- D03Q3.1 Em que freguesia ocorrem mais Eventos de Protecção Civil?

- A Avenidas Novas ✓

- B Benfica

- C Alvalade

- D03Q3.2 Em que mês existem mais Incêndios em média em Marvila?

- A Setembro ✓

- B Julho

- C Agosto

- D03Q3.3 Em que dia da semana ocorrem mais Incêndios em Marvila e Santa Clara?

- A Segunda-feira

- B Quarta-feira

C Sábado ✓

- D03Q3.4 Para as ocorrências com tipologia não urgente, quais dos seguintes anos têm a quinta feira como o dia com maior média de ocorrências?

A 2016

B 2017 ✓

C 2018

D 2019 ✓

E 2020

- D04 - Tendências ao longo dos Anos

- D04Q4.1 Qual a natureza com mais ocorrências?

A Serviços ✓

B Acidentes

C Atividades

- D04Q4.2 Qual a sub-natureza com mais ocorrências?

A Serviços - Fecho de Água ✓

B Serviços - Abertura de Porta - Com Socorro

C Acidentes - Rodoviários - Com viatura

- D04Q4.3 Qual a segunda sub-natureza com mais ocorrências?

A Serviços - Fecho de Água

B Serviços - Abertura de Porta - Com Socorro ✓

C Acidentes - Rodoviários - Com viatura

- D04Q4.4 Olhando para o gráfico de média móvel, com a opção de granularidade selecionada nos 21 dias, qual é o dia com menos ocorrências (nos 21 dias anteriores)?

A 2 de janeiro de 2019

B 4 de julho de 2019 ✓

C 18 de setembro de 2020

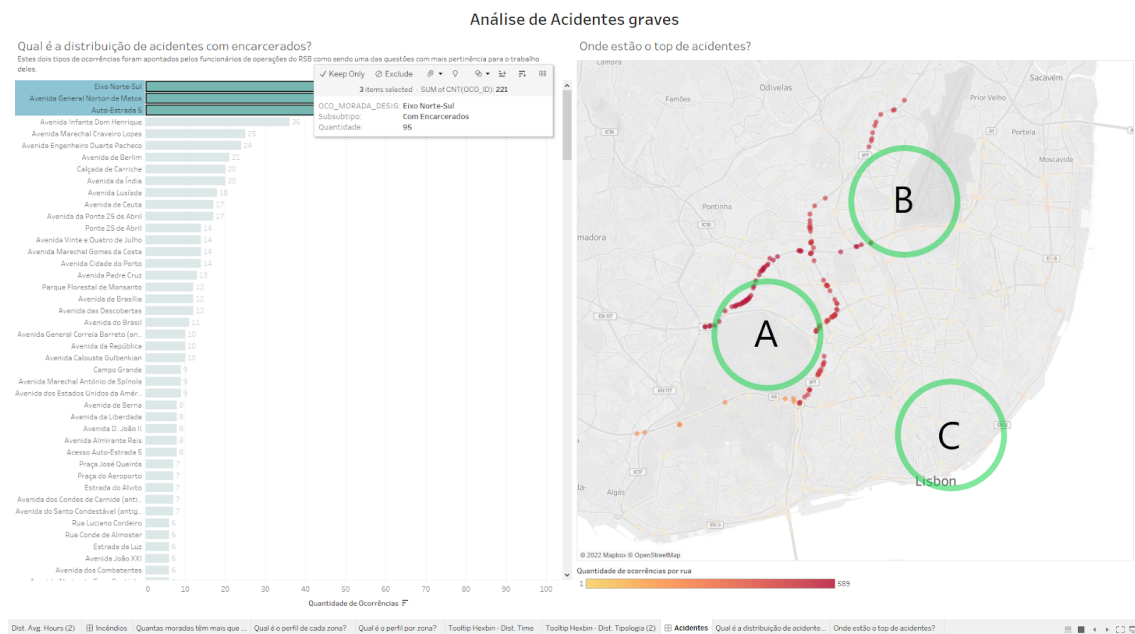
- D04Q4.5 Olhando para o gráfico de média móvel, com a opção de granularidade selecionada nos 7 dias, quantas vezes a linha vai abaixo das 20 ocorrências em média?

A 3

B 4 ✓

C 5

- D05 - Análise de Acidentes Graves



– D05Q5.1 Através da imagem acima, que mostra o top 3 de ruas com mais acidentes, qual diria ser o circulo que em média está mais perto dos acidentes *highlighted*?

- A ✓
- B
- C

INTERACTION VISUALIZATION MODELING LANGUAGE

No âmbito desta investigação criei os seguintes diagramas para representar o *Dashboard 01*, para poder testar e dar *feedback* na linguagem.

A utilização da aplicação diagrams.net [22] é feita através do carregamento da biblioteca XML da IVML. Isto é possível através do menu File -> Open Library from -> Google Drive or Device. A partir daí, arrastam-se os elementos para a área de modelação e utilizam-se como descrito na documentação (cada elemento tem uma *tooltip* com instruções).

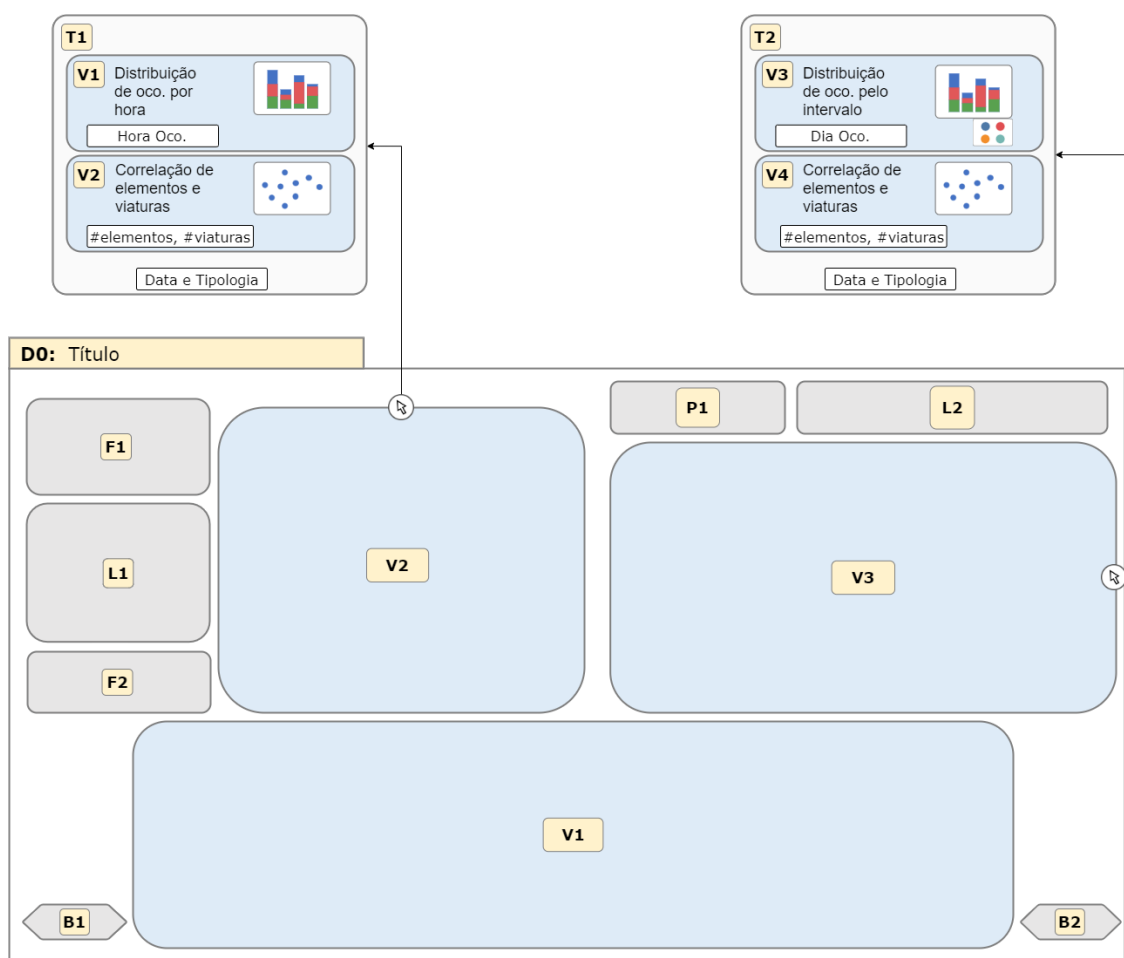


Figura C.1: Diagrama de *tooltips* do dashboard 01, com duas *tooltips*.

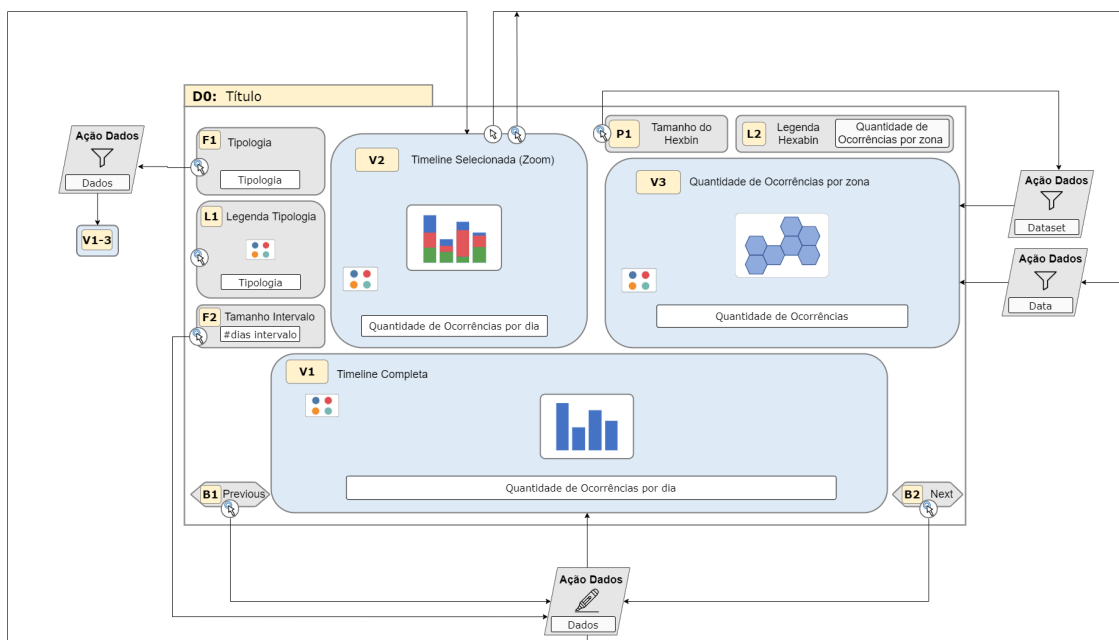


Figura C.2: Diagrama de interações do dashboard 01, com quatro interações.





Newspaper is a first hand account of the events of the day