



Análise e Visualização de Incidências de Doenças Transmissíveis

Nuno Ricardo Gomes Pires

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática e de Computadores

Orientadores: Prof. Daniel Jorge Viegas Gonçalves
Prof. Mário Jorge Gaspar da Silva

Julho 2017

Agradecimentos

Em primeiro lugar gostaria de agradecer aos meus familiares por todo o apoio que me deram ao longo do meu percurso académico e sem o qual este projeto não seria possível.

Também gostaria de agradecer à minha orientadora Dr^a Cátia Sousa Pinto por me ter dado a oportunidade de ingressar num estágio no âmbito da tese na divisão de epidemiologia da Direção-Geral da Saúde (DGS) e por toda a orientação durante a realização do mesmo. Não esquecendo também toda a equipa da divisão de epidemiologia com quem partilhei muitos momentos na DGS, nomeadamente, José Loff, Paula Vicêncio, Célia Gaspar, Maria João e Lurdes Morgado e ainda outros dois estagiários com quem muitas experiências vivenciei, Daniela Pimentel e Francisco Duarte.

Queria agradecer também aos orientadores Prof. Daniel Gonçalves e Prof. Mário Gaspar por toda a orientação e apoio prestado, conselhos, disponibilidade e partilha de conhecimentos que sem os quais esta tese não seria possível.

Por último, mas não menos importante, gostaria de agradecer a todos os meus amigos e colegas por todos os momentos partilhados, os bons e os menos bons, e que sem eles não seria a mesma coisa. Gostaria de acrescentar que todos os momentos de pressão que passamos ao longo do curso valem a pena, pois com eles evoluímos e adquirimos novas capacidades que serão certamente uma mais valia num futuro próximo.

A todos e a cada um de vocês, o meu obrigado e um bem-haja.

Abstract

Directorate-General for Health (DGS) has an information system that allows reporting cases of mandatory notification surveillance in Portugal, the National System of Epidemiological Surveillance (SI-NAVE). This system allows monitoring of epidemiological data. However, it currently does not allow any kind of visualization for that data. This document presents a complementary system, the eVD Lab (E-communicable Diseases Surveillance) which allows the visualization of information on several variables of analysis such as disease, location, age and gender, using the potentialities of several visual elements such as *heatmap*, line chart and *choropleth*. With this system it is now possible to have an immediate overview of the incidence of communicable diseases. An evaluation of the system demonstrated good usability and good performance in the proposed tasks, and its usefulness for analysis in the DGS is also proven.

Keywords

Epidemiological surveillance; Visualization information; Real-time surveillance

Resumo

A DGS tem um sistema de informação que permite notificar a ocorrência de casos de doenças de declaração obrigatória em Portugal, o Sistema Nacional de Vigilância Epidemiológica (SINAVE). Este sistema permite monitorizar informação epidemiológica sobre doenças transmissíveis sujeitas a notificação obrigatória. No entanto, atualmente não permite visualizar os dados. O trabalho descrito neste documento apresenta um sistema complementar, o eVD Lab (vigilância eletrónica de doenças transmissíveis), que permite visualizar informações sobre diversas variáveis de análise como doença, localização geográfica, grupo etário e género, utilizando as potencialidades de diversos elementos visuais, tais como *heatmap*, gráfico de linhas e *choropleth*. Com este sistema é agora possível ter uma percepção imediata do panorama geral sobre a incidência de doenças transmissíveis. Uma avaliação ao sistema demonstrou uma boa usabilidade e um bom desempenho nas tarefas propostas, sendo ainda comprovada a sua utilidade para análise na DGS.

Palavras Chave

Vigilância epidemiológica; Visualização de informação; Vigilância em tempo real;

Conteúdo

1	Introdução	1
1.1	Motivação	3
1.2	Objetivos	4
1.3	Estrutura do documento	5
2	Trabalho relacionado	7
2.1	Técnicas de visualização	9
2.1.1	3D	9
2.1.2	Espaço	11
2.1.3	Tempo	13
2.1.4	Gráficos	14
2.2	Sistemas de visualização existentes	19
2.2.1	Análise de 3 protótipos	19
2.2.2	VoroGraph	20
2.2.3	Epinome	21
2.2.4	GIDEON	22
2.2.5	EpiCanvas	23
2.2.6	EPIPOI	24
2.2.7	EpiCaster	26
2.2.8	PanViz	27
2.3	Discussão	28
3	Solução	31
3.1	SINAVE	33
3.2	Levantamento de requisitos	35
3.3	Arquitetura de software da solução	36
4	Análise de dados (<i>backend</i>)	39
4.1	Acesso e análise aos dados	41
4.2	Organização da base de dados	42

4.3	ID das notificações	42
4.4	Dados referente à data de notificação	43
4.5	Dados duplicados e especificações de doenças	44
4.6	Dados referente ao grupo etário	45
4.7	Dados referente à localização	46
4.8	Dados referente ao género	47
4.9	Funcionamento do <i>backend</i>	47
5	Visualização (<i>frontend</i>)	49
5.1	Relação com o E-Mortality Surveillance (eVM) - E-Mortality Surveillance	51
5.2	E-communicable Diseases Surveillance (eVD Lab) - <i>frontend</i>	51
5.2.1	A interface de navegação	52
5.2.2	Página principal - SINAVE hoje	53
5.2.3	Restantes menus e visualizações	54
5.2.4	Visualização por grupo etário e género	56
5.2.5	Visualização por doença e dia de notificação	58
5.2.6	Visualização por distribuição geográfica	60
5.3	Demonstração do potencial	61
6	Avaliação	65
6.1	Características da avaliação	67
6.2	Testes com utilizadores	67
6.2.1	Os utilizadores	68
6.2.2	Dados recolhidos	69
6.2.3	Questionário de satisfação	71
6.3	Cenários de teste com utilizadores da DGS	73
6.4	Discussão	74
7	Conclusão	75
7.1	Contribuição	77
7.2	Trabalho futuro	77
A	Lista de doenças transmissíveis de notificação obrigatória	81

Lista de Figuras

2.1	Utilização da framework Digital Earth, Weng et al. [2012] .	9
2.2	Histograma, Nguyen et al. [2013] .	10
2.3	Por tipo de doença, Ran et al. [2013] .	12
2.4	Voronoi, Ran et al. [2013] .	12
2.5	SIET, Avruskin et al. [2004] .	13
2.6	(a) Choropleth map, (b) cartograma, (c) mapa de símbolos graduados, (d) mapa de densidade de pontos, Plaza-Rodríguez et al. [2016] .	15
2.7	Visualização interativa dados sobre Vírus da Imunodeficiência Humana (VIH), Blevins et al. [2016] .	16
2.8	Gráfico multi-painel, Chui et al. [2011] .	17
2.9	Centroidal Voronoi Tessellation (CVT) morphing, Dunne et al. [2015] .	20
2.10	CVT meta-layout, Dunne et al. [2015] .	20
2.11	Vista do sistema Epinome, Livnat et al. [2012] .	21
2.12	Informações geográficas, Alonso and McCormick [2012] .	24
2.13	Matriz de intervalos temporais, Alonso and McCormick [2012] .	25
2.14	Grelha de calor, Alonso and McCormick [2012] .	25
2.15	Epicafter, Deodhar et al. [2015] .	26
2.16	PanViz, Maciejewski et al. [2011] .	27
3.1	Notificação do SINAVE.	33
3.2	As vias possíveis para fazer uma notificação no SINAVElab.	34
3.3	Dashboard.	37
3.4	Arquitetura da solução.	38
5.1	eVM.	51
5.2	Dashboard com menu lateral retrátil e painel central informativo.	52
5.3	SINAVE hoje.	54

5.4	Menu.	55
5.5	Esboço da disposição gráfica.	55
5.6	Versão final da disposição gráfica do eVD Lab.	56
5.7	Esboço de gráfico para visualizar grupo etário e sexo.	57
5.8	Versão final do gráfico para visualizar grupo etário e sexo.	57
5.9	Exemplo de gráfico de linhas do eVD Lab.	58
5.10	Gráfico de linhas referente ao Top 8 atual das doenças com mais notificações no SINAVE.	59
5.11	<i>Heatmap</i> referente ao Top 8 atual das doenças com mais notificações no SINAVE.	60
5.12	Mapa com dados referentes à incidência de notificação de doenças de declaração obrigatória por Administrações Regionais de Saúde (ARS).	61
5.13	Painel de vigilância diária, por doença, estando selecionada a doença de Sífilis excluindo sífilis congénita.	62
5.14	Painel de vigilância diária, por grupo etário e sexo, estando selecionada a doença de Sífilis excluindo sífilis congénita.	63
5.15	Painel de vigilância por divisão territorial, por Nomenclatura das Unidades Territoriais para Fins Estatísticos (NUTSIII), no mês de Abril, averiguando na região com maior incidência de notificações por 100.000 habitantes, quantas notificações de sífilis excluindo sífilis congénita ocorreram.	64
6.1	Grau de instrução completo dos 20 utilizadores.	68
6.2	Boxplot dos tempos de realização das tarefas 1, 2, 3 e 4.	70
6.3	Boxplot dos tempos de realização da tarefa 5.	70

Lista de Tabelas

2.1	Tarefas e técnicas de visualização	29
6.1	Tempos, em segundos, e número de erros cometidos por utilizador em cada uma das tarefas realizadas na sessão de testes.	69
6.2	Pontuação System Usability Scale (SUS) das 10 perguntas.	72

Acrónimos

ARS Administrações Regionais de Saúde

CDC Center for Disease Control and Prevention dos Estados Unidos

CNPD Comissão Nacional de Proteção de Dados

CSS Cascading Style Sheets

CVT Centroidal Voronoi Tessellation

D3 Data-Driven Documents

DGS Direção-Geral da Saúde

ECDC European Center for Disease Control and Prevention

EHR Electronic Health Record

eVD Lab E-communicable Diseases Surveillance

eVM E-Mortality Surveillance

HTML HyperText Markup Language

INE Instituto Nacional de Estatística

NUTSIII Nomenclatura das Unidades Territoriais para Fins Estatísticos

PHC Primary Healthcare Centers

SGBD Sistema de Gestão de Base de Dados

SIET Sistemas de Informação Espaço Temporal

SIG Sistemas de Informação Geográfica

SINAVE Sistema Nacional de Vigilância Epidemiológica

SPMS Serviços Partilhados do Ministério da Saúde

SUS System Usability Scale

SVG Scalable Vector Graphic

VIH Vírus da Imunodeficiência Humana

1

Introdução

Conteúdo

1.1	Motivação	3
1.2	Objetivos	4
1.3	Estrutura do documento	5

1.1 Motivação

Num mundo globalizado onde há cada vez mais viagens e a transmissão de doenças infecciosas é uma grande preocupação, há a necessidade de controlar e vigiar a propagação de doenças. Nesse sentido, tem havido um aumento no desenvolvimento de aplicações informáticas para monitorizar os dados respeitantes a várias doenças transmissíveis. Mundialmente há várias entidades que têm como objetivo a curto prazo melhorar os sistemas de controlo de doenças, como por exemplo os Center for Disease Control and Prevention dos Estados Unidos (CDC) e o European Center for Disease Control and Prevention (ECDC) e que pretendem estar na vanguarda do controlo e prevenção de doenças. Para isso, querem modernizar os sistemas de recolha e visualização de dados para que as informações sobre o estado da saúde pública sejam de melhor qualidade e que permitam respostas rápidas e incisivas na contenção de doenças epidemiológicas.

A CDC e a ECDC possuem sistemas que mostram alguns dados de doenças transmissíveis e que, apesar de não serem ainda aplicações robustas e complexas, são um primeiro passo para atingir o objetivo de terem um sistema capaz de dar uma maior cobertura no controlo e vigilância de doenças transmissíveis. À semelhança da CDC e da ECDC, a DGS em Portugal também pretende melhorar o controlo de doenças epidemiológicas. Para isso implementou um sistema, o SINAVE. Nestes sistemas encontram-se dados referente aos utentes com sintomas de doenças de declaração obrigatória, nomeadamente data de sintomas, doença, data de nascimento e género do paciente. Esses dados permitem fazer um estudo sobre o panorama geral da incidência de doenças de declaração obrigatória.

O SINAVE é um sistema de vigilância em saúde pública, que identifica situações de risco, recolhe, atualiza, analisa e divulga os dados relativos a doenças transmissíveis. Desde 1 de Janeiro de 2015 que é obrigatória a sua utilização para notificar doenças transmissíveis. Isso assegura a base de dados que suporta o sistema, sendo possível recolher dados de diversas regiões do país. O SINAVE é composto por duas partes o SINAVE Clínico (SINAVEmed) e o SINAVE Laboratorial (SINAVElab). No SINAVE Clínico constam dados relativos a consultas e observações hospitalares, sendo inseridas no sistema pelos médicos quando diagnosticam uma doença de notificação obrigatória. No SINAVElab os dados dizem respeito aos resultados das análises laboratoriais que são feitas para confirmar a existência de doenças transmissíveis de declaração obrigatória. O trabalho desenvolvido tem como foco os dados provenientes do SINAVElab, sendo que simultaneamente, os dados provenientes do SINAVEmed foram foco de uma outra tese de mestrado, por parte de Daniela Pimentel de Oliveira, aluna do Mestrado Integrado em Engenharia Biomédica - Ramo de Engenharia Clínica da Universidade do Minho, tendo como objetivo o desenvolvimento de um sistema semelhante apenas para essa informação. No SINAVElab constam dados como a data de análises clínicas, doença, data de nascimento e género do utente, assim como local de residência. Estes dados são analisados de forma a que se obtenham informações sobre a incidência de doenças de declaração obrigatória a nível nacional. No entanto, para se ter uma melhor

percepção e interpretação dos dados, é necessário criar meios que permitam a sua visualização.

Com a criação de sistemas de informação que permitem a visualização através de gráficos, mapas, tabelas e outros elementos visuais, a tarefa de detetar padrões, picos e epidemias fica facilitada, pois consegue-se ver de forma simples e rápida alguma anormalidade, como por exemplo aumento do número de casos de uma determinada doença ou um pico sazonal através de um gráfico de linhas ou ainda quais as regiões mais afetadas através de um *choropleth*. Para que esses sistemas possam ser utilizados e contenham informação relevante têm de ser suportados por dados fidedignos. Os dados em que esses sistemas se baseiam são dados recolhidos em centros hospitalares e clínicos.

1.2 Objetivos

De forma a que o trabalho a desenvolver tenha qualidade e esteja bem definido, foram delineados alguns objetivos. O principal objetivo é:

Desenvolvimento de um sistema de análise e visualização de incidência de doenças transmissíveis.

Para isso, o sistema deve permitir:

- Visualizar informações sobre a incidência de doenças transmissíveis por regiões, grupos etário e género.
- Visualizar a incidência de notificações, ajudando na identificação de valores extremos para controlo e prevenção de surtos.
- Visualizar tendências e evoluções ao longo do tempo.
- A exportação de dados para análise na DGS e para análises externas.
- A visualização pública da informação sobre doenças transmissíveis em Portugal.

O sistema construído permite analisar e visualizar a incidência de doenças transmissíveis através de variáveis como grupo etário, género e localidade dos utentes. As visualizações são conseguidas recorrendo a várias técnicas aliando o melhor de cada uma delas para se atingir os objetivos propostos. *Heatmap*, gráfico de linhas, *choropleth* são alguns exemplos das técnicas utilizadas. A avaliação do sistema consistiu em duas componentes, testes com utilizadores e cenários de teste. Ambas as avaliações demonstraram que o sistema é intuitivo e de boa usabilidade, obtendo uma pontuação média SUS de 77.035, ficando pouco abaixo dos 80.3 pontos dos sistemas considerados de excelente usabilidade. O sistema contribui para a modernização dos meios de controlo de doenças transmissíveis, sendo esse um dos objetivos a curto prazo da CDC e da ECDC.

1.3 Estrutura do documento

No capítulo 1 é possível encontrar a motivação da elaboração deste documento, fazendo-se acompanhar por uma revisão literária e sua avaliação no capítulo 2. No capítulo 3 é descrita a arquitetura do sistema construído, sendo possível encontrar mais detalhes sobre as duas componentes principais, *backend* e *frontend*, nos capítulos 4 e 5, respetivamente. No capítulo 6 encontra-se a avaliação do sistema, sendo discutido o resultado obtido nos testes com utilizadores e cenário de testes. No último capítulo são apresentadas as conclusões.

2

Trabalho relacionado

Conteúdo

2.1	Técnicas de visualização	9
2.2	Sistemas de visualização existentes	19
2.3	Discussão	28

Ao longo do tempo tem-se verificado um aumento do número de sistemas que recorrem aos mais variados dados sobre doenças infecciosas de forma a analisar e visualizar diversas informações tais como quais as regiões mais afetadas e grupos etários mais propensos a contrair doenças transmissíveis. Alguns desses sistemas são analisados de seguida, referindo pontos fortes e outros não tão vantajosos. São também abordadas algumas técnicas de visualização em vigilância epidemiológica.

2.1 Técnicas de visualização

Nesta secção é feita uma visão geral do panorama atual de ferramentas de visualização desenvolvidas para epidemiologia das doenças infecciosas. As ferramentas atualmente utilizadas foram analisadas com foco na necessidade de informação e preferências do utilizador, arquiteturas e características dos sistemas existentes, assim como considerações de usabilidade, [Carroll et al. \[2014\]](#), concentrando-se mais em visualizações de Sistemas de Informação Geográfica (SIG).

2.1.1 3D

Atualmente a maioria das análises epidemiológicas foca em variáveis como o número de casos, tempo de ocorrência, tipo ou intensidade da epidemia, entre outros e a visualização dos dados é feita recorrendo a gráficos a duas dimensões. Em Digital Earth (Figura 2.1), uma plataforma para visualização de dados epidemiológicos, [Weng et al. \[2012\]](#), estes encontram-se representados através de gráficos a três dimensões (3D). Os gráficos em 3D, apesar de visualmente parecerem mais apelativos, são de difícil análise, uma vez que é extremamente difícil comparar gráficos que se encontram sobrepostos e com diferentes níveis de perspetiva por parte do utilizador.

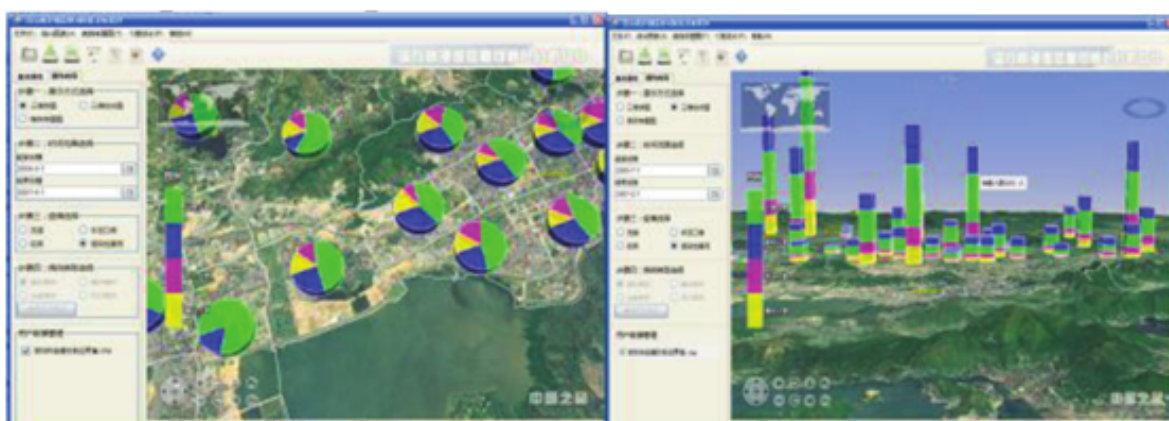


Figura 2.1: Utilização da framework Digital Earth, [Weng et al. \[2012\]](#).

Outra questão é o modo como os dados epidemiológicos estão referenciados em relação a tempo, localização e número de ocorrências. O facto dos dados terem essas informações prende-se com a

necessidade de numa epidemia ter que se conseguir responder às questões “o quê?”, “quando?” e “onde?”, [Nguyen et al. \[2013\]](#), [Avruskin et al. \[2004\]](#). Para visualizar esses dados, [Nguyen et al. \[2013\]](#) propuseram um modelo de visualização de um cubo multi-variável para representar dados epidemiológicos incluindo tempo, localização e número de ocorrências, sendo representados através de zona epidémica e número de infetados durante uma unidade de tempo.

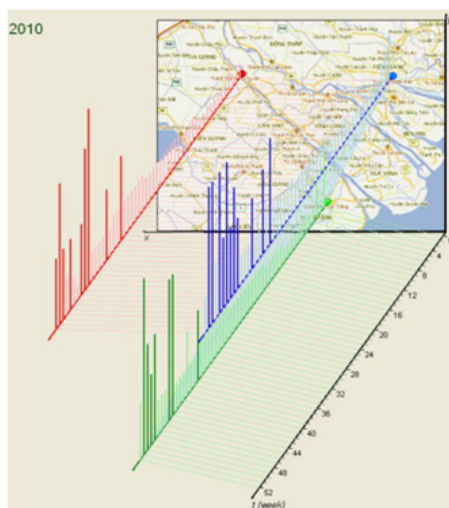


Figura 2.2: Histograma, [Nguyen et al. \[2013\]](#).

Aplicaram este modelo a um caso de estudo de epidemia de dengue, atentando ao número de pessoas infetadas numa semana e período de epidemia. Consequentemente, o dataset consiste em variáveis de localização, tempo e número de pessoas infetadas.

O cubo multi-variável é uma metodologia para representar dados espaço-temporais e é constituído por um sistema de coordenadas cartesiano 2D com um sistema de coordenadas espaço-temporal 3D, onde o eixo do tempo é partilhado para indicar dados espaço-temporais no cubo. O número de infetados é considerado como um atributo que muda ao longo do tempo nas localizações e é representado como um histograma no sistema de coordenadas 2D, que está agregado ao mapa no cubo espaço-temporal em zonas epidémicas de forma a que o eixo do tempo seja paralelo e esteja sincronizado. As barras do histogramas têm várias cores (Figura 2.2) para distinguir o número de pessoas infetadas em várias localizações. Para além disso, fornece ao utilizador a possibilidade de variar a escala do tempo ao longo do eixo através de uma ferramenta de deslizamento.

No caso de estudo utilizaram o modelo do cubo multi-variável para analisar vários aspetos. Através do cubo, atentando ao ano no sistema de coordenadas 2D e ao histograma, é possível ver quantas pessoas foram infetadas, sendo a unidade de tempo a semana. Outro aspeto analisado foi a possível relação entre surtos de dengue em duas regiões. Para isso, atenta-se ao mapa e ao número de pessoas infetadas em ambas as regiões ao longo do tempo, verificando-se que as regiões são próximas do

mesmo rio e que se a epidemia for transmissível pela água, dado que a diferença temporal não é muita, é possível que a epidemia se tenha propagado entre as regiões.

Este modelo tem a vantagem de se poderem relacionar dados espaço-temporalmente, permitindo também uma análise entre regiões. Contudo, as análises feitas são pouco profundas, só sendo utilizados os números de casos, não sendo possível saber por exemplo os grupos etários. Outra questão a apontar é o facto de que ao relacionar regiões, ao seleccionar muitas, os dados sobrepõem-se, tapando a visão do mapa e quando as regiões são muito próximas, a própria visualização do histograma torna-se difícil pois podem sobrepor-se.

2.1.2 Espaço

Um dos maiores fatores em falta nos projetos atuais de investigação de análise em dados epidemiológicos é a visualização do local da ocorrência ou “hot spot”, por isso, [Weng et al. \[2012\]](#) criaram a “China Star Epidemic”, uma ferramenta de análise e visualização de dados epidemiológicos.

Esta ferramenta integra a maior parte dos atributos dos dados epidemiológicos num só sistema. No desenvolvimento desta ferramenta implementaram o conceito de “Thinking Spatially” que consiste numa coleção de capacidades cognitivas. Estas consistem em formas declarativas e perceptuais do conhecimento e algumas operações cognitivas que podem ser utilizadas para transformar, combinar ou outras operações com base nesse conhecimento. Consiste basicamente em três elementos: conceitos de espaço, ferramentas de representação e processos de raciocínio. Estes conceitos permitem ter uma maior noção de certas propriedades como dimensão, continuidade, proximidade e separação que são um veículo para estruturar problemas e encontrar soluções e comunicá-las.

[Weng et al. \[2012\]](#) recorreram à framework Digital Earth que providencia recursos que permitem localizar informações espaço-temporalmente, tendo a capacidade para organizar e calendarizar em massa a sincronização de dados, e ainda visualizar geograficamente os dados em mapas reais numa perspetiva 3D, conferindo assim maior realismo e precisão geográfica dos dados ao utilizador. Com este modelo de visualização também é possível fazer contraste do número de casos em *hot spots*.

Apesar da informação geográfica não ter sido fortemente explorada no desenvolvimento dos primeiros sistemas de análise e visualização de dados epidemiológicos, é notória a sua evolução e utilização em várias ferramentas de análise atuais. [Ran et al. \[2013\]](#) abordam um caso de estudo de prevenção de doenças infecciosas na China recorrendo a um sistema de informação geográfica. De forma a ver detalhadamente as regiões, estas foram segmentadas em distritos no mapa do país, sendo cada distrito representado pela sua forma geográfica. Foram analisados dados de várias doenças infecciosas entre 2011 e 2013 para a região de Ningbo. Os dados foram analisados de diversas maneiras, agregando primeiro a população em diversas variáveis como idade, profissão, género, local de ocorrência e tipo de doença (Figura 2.3). Estes dados foram representados através de gráficos circulares. Este tipo de

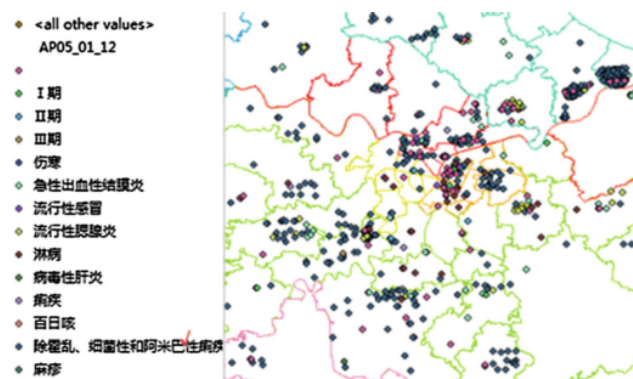


Figura 2.3: Por tipo de doença, [Ran et al. \[2013\]](#).

gráficos funciona bem quando há poucas partes, como é o caso do gênero, masculino e feminino.

No entanto, no caso do tipo de doença, onde há vários, o gráfico circular não é a técnica mais adequada porque é difícil ver cada uma das porções. Para mostrar a capacidade informativa dos sistemas de informação geográfica, representaram os mesmos dados mas desta vez através de *dot points* num mapa o local os casos. No mapa foram também marcados as restantes variáveis de maneira a que o utilizador conseguisse observar diversas relações como por exemplo a região onde mais jovens eram afetados pelas doenças ou os tipos de doenças associados às diversas regiões. Este tipo de informação é muito mais fácil obter quando representado geograficamente ao invés de com gráficos circulares.

Para complementar o estudo da cartografia, utilizaram diagramas de Voronoi (Figura 2.4), que resumidamente é a divisão de um plano em várias regiões tendo por base a distância a um ponto específico do plano, para localizarem os hospitais e centros de saúde de cada região, permitindo saber se existia uma distribuição razoável dos estabelecimentos de saúde ou se seria necessário um reajustamento para dar o apoio necessário à população. Este estudo permitiu mostrar a importância de representar geograficamente os dados epidemiológicos, uma vez que é mais fácil e rápida a obtenção de dados que são de extrema importância na prevenção e monitorização de doenças infecciosas.



Figura 2.4: Voronoi, [Ran et al. \[2013\]](#).

2.1.3 Tempo

Outro fator de análise é a utilização de Sistemas de Informação Espaço Temporal (SIET). A importância dos SIG para investigação médica e epidemiológica é há muito reconhecida e é frequentemente utilizada para reconstruções retrospectivas à exposição. Contudo, a aplicação dos SIG ao risco e avaliação da exposição de doenças infecciosas tem sido focada em zonas industriais contaminadas com altos níveis cancerígenos em vez de se focar no indivíduo [Avruskin et al., 2004]. Por isso, é necessário estender os SIG de forma a que para além de responder às questões "o quê" e "onde" passe a responder às questões "o quê?", "onde?" e "quando?". Para isso, implementa-se o SIET que possui dados referentes a objetos (entidades modelados, por exemplo, uma pessoa), coordenadas espaço-temporais (localização que pode ser latitude, longitude, altitude, data, modelo de movimento, que regista o movimento do objeto, ou ainda polígonos de espaço-tempo como centróides, fronteiras ou datas), e atributos que são observações dos objetos, como por exemplo o rendimento da pessoa, etnia.

Avruskin et al. [2004] analisaram pessoas expostas ao arsénio, analisando a zona residencial, fornecimento de água e os hábitos de bebida de água. Para visualizar estes dados, recorreram a diversas técnicas. A primeira, utilizaram para representar o tempo, animações onde o utilizador percebe facilmente a evolução ao longo dos anos. Tal como os SIG, os atributos são representados por uma cor específica, forma e tamanho dos elementos gráficos (por exemplo símbolos representativos). Mas ao invés dos SIG, os SIET tornam a visualização mais fácil para o utilizador porque têm animações que mudam as formas dos polígonos e os valores dos atributos ao longo do tempo com mapas animados, histogramas e tabelas simultaneamente (Figura 2.5). Informação valiosa que antes podia ser perdida numa visualização estática dum GIS, é agora capturada e pode tornar-se foco de análise num SIET. Desta forma é possível analisar melhor o impacto que as viagens têm na propagação de doenças infecciosas, algo que só com o SIG era dificultado pois não se considerava a variável tempo que também é bastante importante no estudo epidemiológico.

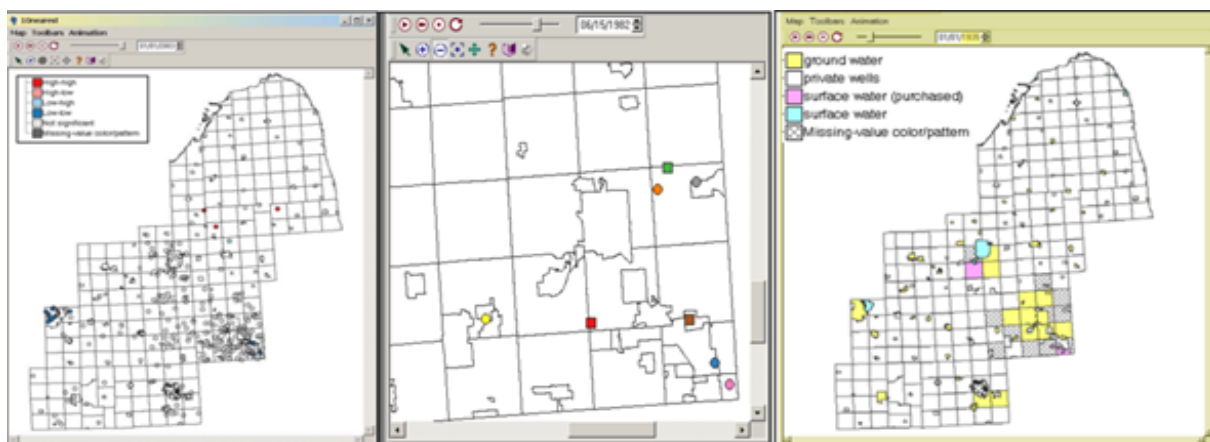


Figura 2.5: SIET, Avruskin et al. [2004].

2.1.4 Gráficos

Há várias formas de apresentar os dados graficamente. Por exemplo, os SIET apresentam quatro tipos de visualização [Avruskin et al., 2004] :

- Mapas
- Gráficos (histogramas, *scatterplots*, *box plots*)
- Tabelas
- *Time plots*

A vista do mapa mostra os dados geograficamente e tem a vantagem do utilizador interagir com este através de *zoom*, *panning*, seleção e consultas. Com os histogramas, *scatterplots* e *box plots* também animados ao longo do tempo, é possível selecionar um indivíduo ou grupo de indivíduos (por exemplo casos vs. controlo), selecionar um ponto no tempo, e dar a possibilidade ao utilizador de explorar, comparar e ver a distribuição dos valores selecionados ao longo do tempo.

As tabelas, também animadas, permitem, dado um valor da variável, que o utilizador visualize como é que um valor evoluiu ao longo do tempo, como por exemplo a concentração de arsénio na água de fornecimento municipal. Por último, usa *time plots* que têm o tempo no eixo dos xx e a exposição estimada ao arsénio no eixo do yy. Objetos de interesse como casos e controlos são mapeados no *time plot* para explorar a dependência do tempo na exposição ao arsénio.

Ao contrário das outras vistas, o *time plot* não é animado porque já mostra todo o intervalo de tempo ao longo do eixo dos xx. A vantagem dos SIET é a capacidade de se ver simultaneamente vários dados e de diversas formas, através da visualização simultânea de mapa, histograma, *scatterplot*, *box plot*, mostrando ao utilizador várias informações dos mesmos dados, tentando aproveitar ao máximo os dados que o sistema tem para retirar toda a informação útil e possível.

Ao longo dos anos, os dados epidemiológicos espaciais têm-se tornado fundamentais na geração de casos reportados, nomeadamente a representação das prevalências de certo elemento. Essa visualização tem sido feita através de *choropleth maps*, sendo recorrente a sua utilização pelas agências reguladoras europeias. Contudo, dadas as suas limitações, visto que não é a visualização adequada para dados brutos e contagens pois pode induzir em erro ao utilizar por exemplo escalas de cores para caracterizar contagens e em regiões contíguas, uma ter um caso e a outra ter dois, pode fazer com que na escala exista uma grande diferença, ou por exemplo ao identificar a região com um caso mas a região ser demasiado vasta e não conter informações precisas, é preciso procurar alternativas. As alternativas passam pela utilização de cartogramas, mapas com símbolos graduados ou mapas de densidade de pontos.

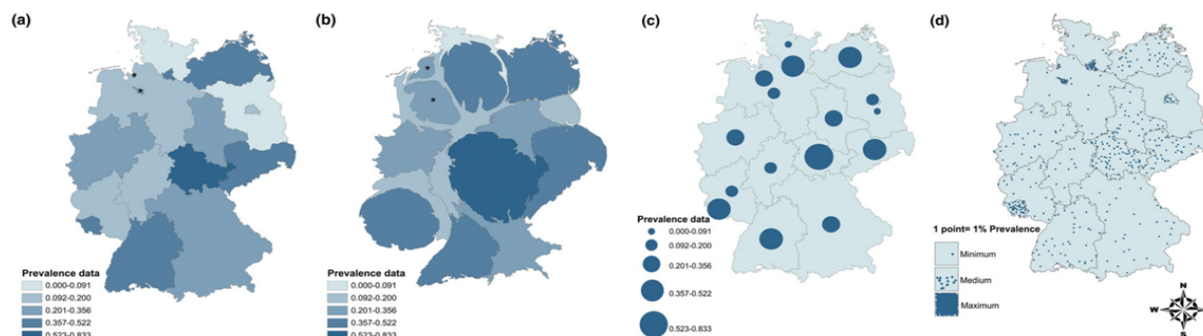


Figura 2.6: (a) Choropleth map, (b) cartograma, (c) mapa de símbolos graduados, (d) mapa de densidade de pontos, Plaza-Rodríguez et al. [2016].

Tendo por base a prevalência para a *Campylobacter*, Plaza-Rodríguez et al. [2016] compararam as diversas técnicas de visualização (Figura 2.6). Nessa comparação é possível ver que a extração de informação no *choropleth map* não é difícil, verificando que o uso de tons de azul diferenciam, sem esforço, os diferentes valores, correspondendo o azul mais escuro a maior prevalência. Contudo, tem principalmente duas limitações: número de unidades de enumeração e problema da unidade regional modificável. O número de unidades de enumeração pode variar muito. Por exemplo, no caso de estudo, na Alemanha a população residente em cada região varia muito e ao utilizar os dados em bruto, unidades de enumeração maiores dominariam a percepção do mapa, exagerando as amostras positivas.

No caso das regiões modificáveis, a interpretação do mapa depende nas fronteiras das unidades de enumeração. Por exemplo, dados agregados por países podem não representar diferenças regionais relevantes quando comparados com dados agregados por regiões. Atentando aos cartogramas, o tamanho dos estados federais foram modificados de acordo com o valor da prevalência pelo que a forma e topologia da geografia inicial foram distorcidas. Isto foi feito para facilitar a interpretação do mapa, pois áreas com maiores prevalências encontram-se agora facilmente identificáveis através de uma distorção que torna a área maior. Contudo tem a desvantagem de que as áreas com menor prevalência sejam praticamente invisíveis, o que não é desejável.

Outro problema é que a utilização da distorção pode, por vezes, tornar regiões desproporcionalmente largas e torná-las irreconhecíveis, ficando a leitura do mapa mais difícil para o utilizador. Para além disso, quando duas unidades de enumeração têm o mesmo tamanho mas formas diferentes, o leitor do mapa supõe que elas têm tamanhos diferentes. Com o mapa de símbolos graduados, foram criados círculos graduados para diferentes tamanhos dependendo do valor da prevalência, onde círculos maiores correspondem a maiores prevalências. Este tipo de mapas supera o problema do número de unidades de enumeração pois o tamanho dos símbolos depende unicamente da prevalência e não no tamanho da unidade de enumeração. Pequenas unidades de enumeração com valores altos de prevalência podem também ter um grande símbolo associado. Estes mapas têm grande flexibili-

dade, podendo ser utilizados para dados em bruto e dados estandardizados, permitindo mostrar várias variáveis usando símbolos compostos. No entanto tem a desvantagem da chamada ilusão de Ebbinghaus, que provoca a ilusão de que dois círculos são idênticos apesar de serem diferentes devido ao tamanho do contorno que os rodeia.

Outra questão que preocupa neste tipo de mapas é a escolha do tamanho dos símbolos, sendo preciso muito cuidado para evitar a sobreposição de símbolos que dificulta a interpretação correta de dados e no caso de regiões pequenas, com a utilização de símbolos, correm o risco de serem ocultadas, dificultando a interpretação dos resultados. Por outro lado, tem a vantagem de que ao classificar os dados em algumas classes, as diferenças serão facilmente interpretadas pelo utilizador, permitindo que a estimativa dos valores seja feita de forma adequada.

Por último, os mapas de densidade de pontos mostram os valores da prevalência em percentagem. Cada ponto corresponde a 1% de prevalência, e com isso, é fácil de ver no mapa as regiões com pouca, média e alta densidade de pontos, correspondendo a valores baixo, médios e altos de prevalência, respetivamente. Apesar disso, é difícil extrair informação numérica. Outro fator importante é a que a impressão visual continua a depender do tamanho das unidades de enumeração. Por exemplo, pequenas regiões são mais evidenciadas que outras com o mesmo número de pontos mas área maior.

Concluindo esta análise, percebe-se que apesar do *choropleth map* parecer a técnica de visualização mais adequada para dados epidemiológicos, por vezes pode não o ser. Para estudos mais aprofundados é preferível utilizar outro tipo de mapa como cartograma, mapa de símbolos graduados ou mapa de densidade de pontos, consoante a análise pretendida. Muitas análises epidemiológicas têm recorrido ao *choropleth map*, contudo, é difícil extrair alguma informação que é importante e que poderia ser facilmente obtida com a utilização de um dos outros tipos de mapas anteriormente enumerados.

Analisando Blevins et al. [2016] verifica-se haver três tipos de visualização para dados respeitantes a VIH: *longitudinal plots*, *bubble plots* e *heatmaps*. Os gráficos apresentados através de *longitudinal plots* como o *scatterplot* permitem ver a contagem de células CD4 (células responsáveis pela resposta imunitária do nosso corpo) após os dias de tratamento antiviral e a sua evolução (Figura 2.7).

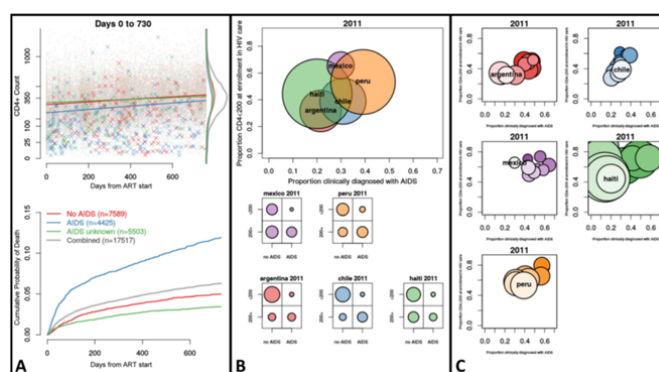


Figura 2.7: Visualização interativa dados sobre VIH, Blevins et al. [2016].

Para além disso, é possível ver através de um gráfico de linhas, a probabilidade de morrer após o dia do começo do tratamento antiviral. Este tipo de gráficos permite ver tendências e comparar valores entre os vários atributos. Por exemplo, é possível ver que os pacientes com maior contagem de CD4 no início do tratamento antiviral rapidamente se separam dos que têm menor contagem de CD4. Também é fácil perceber que os pacientes com VIH que começam o tratamento antiviral têm maior probabilidade de morrer do que os que não iniciam.

Através de *bubble plots* é possível visualizar mudanças nos indicadores do VIH ao longo do tempo. Têm a vantagem de mostrar três dimensões de dados, um em cada um dos dois eixos, um terceiro, o tamanho da bolha, e uma quarta dimensão é adicionada ao mostrar a mudança ao longo do tempo, permitindo por isso mostrar várias informações num só gráfico. Com este tipo de gráficos consegue-se encontrar tendências, distribuições e correlações entre os atributos e ainda encontrar *clusters* locais. Contudo, quando a quantidade de dados é demasiado grande, torna-se impercetível. Neste caso, é possível ver e relacionar, através do *bubble plot*, a proporção diagnosticada com VIH e a proporção de CD4 em diversas regiões, sendo o tamanho da bolha proporcional ao número de novos casos.

O outro tipo de visualização apresentada é o *heat map*. O *heat map* é utilizado para mostrar a proporção de CD4 no diagnóstico do VIH mundialmente com ênfase na América do Sul e Central. Através dos *heat maps* é possível ver a evolução da proporção ao longo dos anos, sendo esta assinalada através de cores, ficando mais escura quanto mais intensa for a proporção no país. Analisando esta ferramenta e o que possibilita ver, assim como as técnicas utilizadas, conclui-se que não é visível da melhor forma a transição entre os dados dos diferentes gráficos, apesar de algumas animações não fazem uma transição suave entre os diferentes estados. É muito mais perceptível para o utilizador se lhe for mostrado uma transição harmoniosa entre o dados, por exemplo, a variação ao longo dos anos ou os dados de diferentes anos, se houver animações da variação/trajeto da bolha, do que uma mudança rápida entre imagens estáticas, levando o utilizador a questionar o que se passou entre os anos.

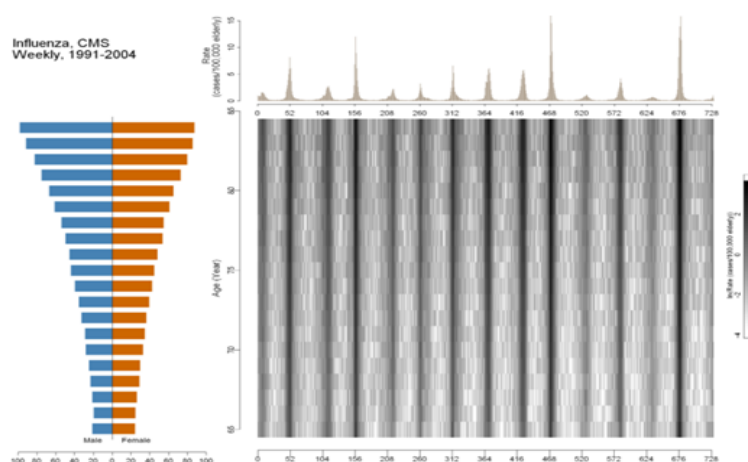


Figura 2.8: Gráfico multi-painel, Chui et al. [2011].

Um aspeto fundamental na apresentação de gráficos e tabelas é a disposição visual de vários elementos e a interação entre eles. [Chui et al. \[2011\]](#) utilizam um painel com vários gráficos na visualização de dados epidemiológicos, aliando idade, tempo a doenças (Figura 2.8). Para mostrar a distribuição de idades e género da população, recorrem à pirâmide de população que é construída através da justaposição de dois histogramas verticais, um para o sexo masculino e outro para o sexo feminino, tendo como eixo vertical comum a idade. Este tipo de gráficos permite ver e comparar a distribuição de género por idades. Através da forma permite saber se há uma distribuição uniforme, se há simetria, ou não, entre géneros e ainda perceber se há irregularidades ou picos de doenças. Contudo, este tipo de gráficos estáticos não permite ver mudanças temporais, um facto importante em doenças infecciosas com sazonalidades bem definidas.

Para visualizar tendências temporais e padrões sazonais é preciso escolher cuidadosamente a unidade temporal a apresentar nos gráficos. Havendo vários níveis de agregação temporais, e tendo todos benefícios e limitações, a medida standard é a semana. Este gráfico tem como eixo horizontal os intervalos de tempo e eixo vertical o número de casos ou rácios de doença. Com este tipo de gráficos é possível constatar picos de doenças ao longo do tempo e detetar sazonalidades. Construindo este gráfico para grupos etários específicos é possível ver quais os grupos mais afetados por algumas doenças, retirando-se informações de onde se deve atuar para prevenir a doença. No entanto, ao utilizar um intervalo de tempo reduzido, a informação retirada é pouco precisa, sendo necessário um maior número de anos para se verificar realmente tendências. Este gráfico e a pirâmide são úteis por si só para efeitos de vigilância, mas quando são construídos separadamente não são úteis para observar interações temporais-demográficas.

Por último, os gráficos de imagens são capazes de mostrar informações de pelo menos três variáveis. Aplicando a cada eixo uma variável e representando, por exemplo, o número de casos através de cores diferentes ou variações de tons da mesma cor, consegue-se ter três tipos de dados no mesmo gráfico. Com este gráfico consegue-se verificar relações entre os dados em análise, por exemplo, ver que uma certa doença tem tendência para ocorrer em determinada época num grupo etário específico. Mas, apesar desta vantagem, é um tipo de gráfico que para se fazer uma leitura correta precisa de algum tempo de treino e prática. Permite também a identificação de *clusters*. O objetivo da utilização destes gráficos é mostrar ao utilizador várias informações recorrendo a elos comuns entre os gráficos, partilhando eixos, como por exemplo idade e tempo. Com o aumento da complexidade de dados epidemiológicos, há cada vez mais uma maior necessidade de representar esses dados visualmente. E, melhor do que adicionar novas visualizações, os autores deste artigo recorrem ao enriquecimento do número de variáveis ou dimensões, combinando gráficos simples para detetar tendências e relações em certas doenças epidemiológicas.

2.2 Sistemas de visualização existentes

Nesta secção são analisados alguns sistemas de visualização de dados epidemiológicos existentes, atentando à forma como dispõem os dados ao utilizador.

2.2.1 Análise de 3 protótipos

Karlsson et al. [2013] fizeram um estudo para verificar o impacto do nível de abstração quando é utilizada visualização nas rotinas de controlo de doenças infecciosas. Para isso, desenvolveram três protótipos de visualização interativa com níveis crescentes de abstração para se comunicar subconjuntos de informações de vigilância de surtos de gripe. Nesses protótipos tiveram em conta os dois principais componentes das visualizações modernas, gráficos e interação. Os gráficos têm a vantagem de tornar flexível a percepção humana na identificação de padrões e a interatividade providencia uma camada de exploração que é necessária para apoiar o diálogo entre utilizadores e os dados. Para desenvolver os protótipos foram utilizados Data-Driven Documents (D3) e bibliotecas JavaScript de produção de visualização a partir de dados. O D3 manipula a HyperText Markup Language (HTML), Scalable Vector Graphic (SVG) e Cascading Style Sheets (CSS) e confere interatividade e dinamismo nas visualizações que se encontram em páginas web acessíveis a partir de um browser.

O primeiro protótipo, P1 - Disease map prototype, mostra uma visualização baseada em informação espaço-temporal, utilizando dados recolhidos através de um Electronic Health Record (EHR). O princípio deste protótipo é mostrar o progresso de carga da doença em Primary Healthcare Centers (PHC) específicos, através de uma lista, para além de ser possível visualizar os casos epidemiológicos por cada PHC. É possível ver os casos por dia e por grupo etário. No segundo protótipo, P2 - descriptive abstract prototype, foi desenvolvida uma visualização descritiva de dados de enfermaria recolhidos através de chamadas telefónicas, separadas por sintomas. É possível visualizar o total de chamadas por sintomas principais e a distribuição etária. Por último, o protótipo P3 - Analytic abstract prototype, mostra a visualização de deteção de surtos baseados em dados de enfermaria recolhidos através de chamadas telefónicas. É possível visualizar as chamadas por dia, por faixa etária e, através dos dados recolhidos, prever o número de chamadas dos próximos 5 dias.

Estes protótipos foram apresentados a diversos especialistas epidemiológicos e os resultados mostram que consideraram o P1 uma boa ferramenta para o controlo de surtos epidémicos pois mostrava a lista dos PHC sobrecarregados, permitindo uma melhor distribuição dos doentes e ainda a ocorrência dos casos por região, permitindo perceber melhor a evolução do surto. O P2 foi considerado bastante útil pelo facto de apresentar os dados por faixas etárias, permitindo atuar em faixas etárias de riscos como por exemplo crianças. No caso do P3, houve maior desconfiança devido à capacidade de prever, não sabiam se era confiável e as visualizações mais complexas poderiam ser úteis mas iriam requerer

algum tempo de treino por parte do utilizador. Concluindo, à medida que as visualizações iam sendo mais abstratas, a desconfiança dos epidemiologistas ia aumentando, pelo que um maior nível de complexidade nas visualizações iria requer vários processos de aprendizagem e não iria entrar facilmente na rotina prática.

2.2.2 VoroGraph

VoroGraph é uma ferramenta para análise epidemiológica que permite analisar a incidência e propagação da doença em relação à densidade populacional e outras condições demográficas em escalas geográfica abrangendo desde voos internacionais a deslocações locais [Dunne et al. \[2015\]](#). De modo a representar essas informações, foram ponderadas várias técnicas de visualização, nomeadamente mapas visuais, CVT e meta-layout baseado em CVT. O mapa visual utiliza codificações nas fronteiras entre regiões vizinhas para mostrar relações locais. O CVT é um caso específico de diagramas Voronoi. O diagrama Voronoi é o particionamento de um plano em várias regiões tendo por base a distância a um ponto específico do plano. No caso do CVT, o ponto de cada região é a média dos pontos do plano original. O CVT permite transformar o mapa numa forma de preenchimento do espaço, preservando posições relativas, de modo a realçar propriedades da região assim como as relações locais. Por último, o meta-layout é baseado em CVT exibindo o agregado de longa distância além das relações locais.

Em termos de design utiliza um *Border-Encoded Map* que, através da identificação clara de fronteiras no mapa, permite evidenciar as deslocações locais entre regiões (denominadas por "basin") contíguas, utilizando tamanhos e cores para representar o rácio de população infetada. O tamanho da fronteira codifica o número total de deslocações entre basins contíguos enquanto que a cor, numa escala de branco a vermelho mostra quantas pessoas foram infetadas. Utiliza também CVT Morphing (Figura 2.9), ou seja, transforma o mapa em CVT para lidar com as limitações das codificações de fronteira. CVT's otimizados têm várias propriedades que tornam as informações eficazes nas ferramentas de visualização. Utilizam ainda Labeling, Border Encodings, CVT Meta-Layout (Figura 2.10), transições animadas e Timeline.

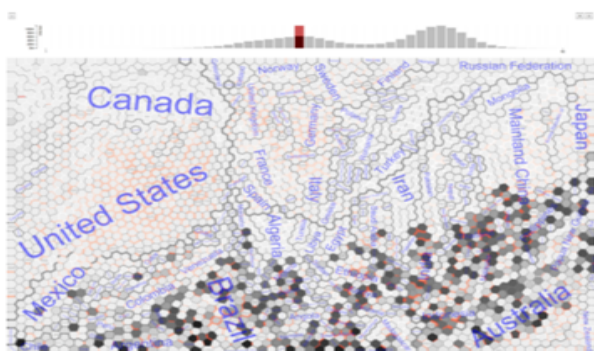


Figura 2.9: CVT morphing, [Dunne et al. \[2015\]](#).

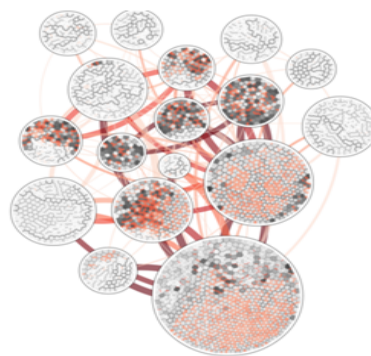


Figura 2.10: CVT meta-layout, [Dunne et al. \[2015\]](#).

Os principais pontos deste artigo e do VoroGraph são a utilização de CVT que ao serem otimizados produzem resultados eficazes na visualização de informação pois as regiões são uniformemente distribuídas e espacialmente bem preenchidas, as posições relativas são preservadas, as regiões são geralmente em formas de hexágonos regulares, têm relação de aspeto perto de 1 e têm posições e formas determinísticas. Isto torna a visualização mais clara e perceptível pois as formas são "rígidas" e é possível ver melhor as transições entre regiões, as passagens de uma para outra região. Para além das vantagens da utilização de CVT, o VoroGraph tem a vantagem de ter transições animadas que permite ao utilizador perceber, visualmente, a evolução da doença e a sua propagação.

Por outro lado tem algumas limitações no que à representação de deslocações locais diz respeito, focando-se mais em longas-distâncias, no entanto em epidemiologia por vezes requer mais atenção a estas relações de longa-distância para se perceber melhor a propagação e risco de contágio a um nível global.

2.2.3 Epinome

A deteção eficaz e a resposta a surtos de doenças infecciosas depende da habilidade para capturar e analisar informação e como os funcionários das entidades públicas conseguem responder perante essa informação. O Epinome, desenvolvido por [Livnat et al. \[2012\]](#), é um sistema integrado de investigação visual-analítica (Figura 2.11).

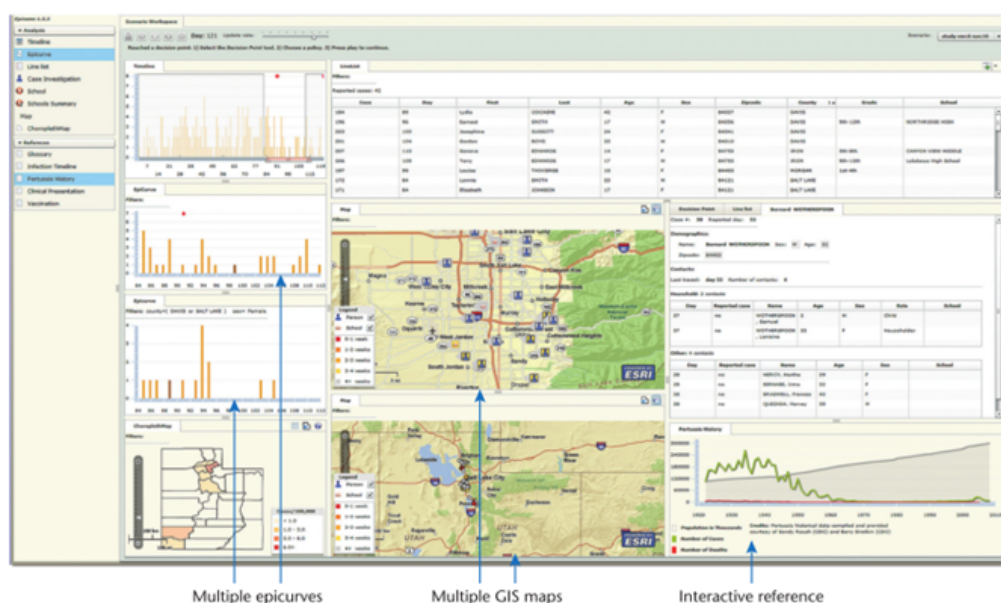


Figura 2.11: Vista do sistema Epinome, [Livnat et al. \[2012\]](#).

Este sistema tem um ambiente dinâmico que envolve perfeitamente e adapta as tarefas dos utilizadores e as suas necessidades. Possui quatro paradigmas de user-interaction em saúde pública:

- exibição visual evolutiva
- perfeita integração entre vistas distintas
- vistas múltiplas coordenadas livremente
- interação direta com os dados

Os principais pontos do Epinome são a facilidade com que se filtram dados em vistas múltiplas, permitindo arrastar (drag and drop) campos entre a várias vistas de modo a que o sistema se adapte aos filtros mesmo que provenha de uma tabela de dados e seja inserido num mapa de localização geográfica. Ao testarem o Epinome com diversos funcionários de saúde pública, os resultados mostram que o Epinome é uma valiosa ferramenta de apoio à decisão e altamente utilizável para análise visual.

2.2.4 GIDEON

No GIDEON é possível encontrar algumas capacidades que um sistema de informação para doenças infecciosas deve possuir de forma a que seja útil para o utilizador final [Edberg \[2005\]](#). Primeiro, é necessário que o sistema seja abrangente de forma a englobar as características clínicas e epidemiológicas de todas as doenças infecciosas, todos os patogénicos humanos e assim como todas as vacinas e medicamentos anti-infecciosos. O programa deve ser flexível e deve ser atualizado em "tempo real", para além de permitir a modificação de dados pela equipa de desenvolvimento/manutenção como pelo utilizador final. Adicionalmente, os dados entre as redes institucionais e alunos, devem ser de fácil transmissão e reprodução.

É também possível criar notas nos idiomas dos utilizadores e no seu alfabeto. O GIDEON possui 4 módulos: diagnóstico, epidemiológica, terapia e microbiologia. O módulo de diagnóstico é responsável por gerar um ranking de diferentes diagnósticos baseados nos sinais, sintomas, análises clínicas, período de incubação e país de origem da doença. Deste módulo resulta uma lista diferencial de diagnóstico que permite ao leitor deste relatório, que pode ser impresso ou enviado para um email, ter acesso um quadro comparativo das características clínicas das doenças indicadas e questões relacionadas com a omissão ou classificação de doenças específicas. O módulo de epidemiologia possui dados de muitas doenças infecciosas genéricas mas também específicas de alguns países. Todos os dados no GIDEON provêm do Ministério de Saúde, agências militares, listas especializadas que estão na Internet, artigos científicos e dados apresentados nas maiores conferências. O utilizador tem acesso a mais de 10.000 gráficos que permitem visualizar a incidência, predomínio e outros status específicos da doença. Possui adicionalmente submódulos que descrevem epidemiológica e clinicamente cada doença assim como a calendarização das vacinas de cada país.

De forma a que qualquer utilizador possa utilizar o GIDEON, este módulo possui uma lista de sinónimos em Espanhol, Alemão e Francês, entre outros idiomas. Contudo, o programa tem algu-

mas limitações pois apesar de os ficheiros de texto do GIDEON serem ricos em dados, seria bastante útil em futuras versões incluir tabelas, mapas e imagens que escasseiam na versão actual.

2.2.5 EpiCanvas

De forma a aproveitar a crescente quantidade de dados epidemiológicos disponíveis e os desafios de visualizar os mesmos e as correlações existentes, [Gesteland et al. \[2012\]](#) criaram o EpiCanvas. O EpiCanvas é um mapa meteorológico para doenças infecciosas permitindo deteção em tempo real, monitorização, exploração e descoberta de doenças epidemiológicas infecciosas a nível regional. Numa primeira fase entrevistaram e observaram atuais utilizadores epidemiologistas de sistemas de vigilância do estado do Utah. Através dessas entrevistas e observações chegaram à conclusão de que o sistema tinha de ser capaz de mostrar dados de várias fontes distintas, como por exemplo relatórios de síndromas e de doenças.

Para além disso tinha de ser capaz de mostrar as mudanças ao longo do tempo e espaço das variáveis mais importantes, assim como a visualizar as relações existentes entre os dados sem providenciar informação redundante quando o conhecimento já existe, por exemplo no caso da relação gripe com febre. Outro aspeto fundamental que chegaram à conclusão foi o da necessidade de existir uma ferramenta de visualização que fornecesse um espaço de trabalho onde diferentes ideias, conceitos, tipos de dados, localizações e outros meta-dados (denominadas por tags) possam ser organizados de maneira útil. Para isso, e de forma a mostrar as relações entre tags, o EpiCanvas organiza-as através de uma rede de conceitos que é uma cloud tag.

A cloud tag é uma técnica de visualização de dados do tipo texto através de keywords (tags), onde a importância/relevância das tags são evidenciadas através da cor ou tamanho. Nessa cloud tag é possível associar as tags a casos reportados, alertas ou outros itens dos dados. O tamanho de cada tag revela a importância da mesma, por exemplo quão maior for a tag, maior é o número de casos reportados, por exemplo de vírus respiratório. Através da cloud tag o utilizador pode selecionar duas tags, por exemplo, “vírus respiratório” e “criança”, e facilmente tem uma avaliação geral das tags selecionadas e a relação entre elas.

Outra vantagem da cloud tag é a capacidade de reduzir o esforço cognitivo necessário para transformar a visualização dos dados no modelo mental dos utilizadores. O utilizador é capaz de ajustar os parâmetros da correlação rapidamente para ocultar ou evidenciar as linhas da relação entre as várias tags e assim identificar casos similares entre tags. A utilização de tags como meta-dados facilita a separação do sistema das fontes de dados e, portanto, a integração de dados heterogéneos de fontes diferentes dentro da mesma framework. Sendo por isso o ponto forte do EpiCanvas.

No testes com utilizadores, 100% concordou que a interação com a cloud tag era intuitiva e a capacidade de consultar os dados através de seleção dinâmica e desselecionar tags seria bastante útil. Uma

grande maioria (90%) concordou que as linhas de correlação ajudavam a perceber as relações entre itens. Resumindo, o EpiCanvas é um software que permite visualizar diversos dados e as suas relações utilizando a cloud tag que tem como principal vantagem a capacidade de evidenciar as tags (sintomas, doenças, síndromas) através do tamanho e cor, permitindo ainda correlacionar com os diversos casos reportados através das linhas de correlação, evidenciando dessa forma as relações entre os diversos itens (tags).

2.2.6 EPIPOI

O EPIPOI é uma ferramenta user-friendly que permite a exploração e extração de parâmetros descrevendo tendências, sazonalidades e anomalias que caracterizam as doenças epidemiológicas [Alonso and McCormick, 2012]. Também é possível visualizar e explorar dados referentes a séries temporais que são fundamentais na análise epidemiológica, não descurando a inspeção de dados por regiões geográficas. Em termos de visualização de dados, uma das características do EPIPOI é a facilidade de traçar os parâmetros da série temporal extraídos, através da opção *scatterplot*. Com várias séries temporais, comparar esses parâmetros pode ser perspicaz, principalmente para dados com referências geográficas, com dados únicos de latitude e longitude. Com informações geográficas, os dados podem ser visualizados através de mapas, onde médias, amplitudes e tempos de picos epidemiológicos podem ser traçados para identificar tendências geográficas (Figura 2.12).



Figura 2.12: Informações geográficas, Alonso and McCormick [2012].

Se a opção do *scatterplot* for selecionada para mostrar anomalias, o tempo e a magnitude destas estará disponível para exploração no gráfico do intervalo temporal selecionado. No mapa são utilizados marcadores para representarem amplitudes epidemiológicas, sendo o seu tamanho proporcional. No entanto, há uma visualização alternativa, recorrendo a uma matriz dos intervalos temporais em vez de apresentar esses dados num mapa. Nesta matriz, os dados são visualizados com o tempo no eixo

horizontal e os intervalos temporais ordenados por latitude. Este tipo de visualização pode ter um impacto visual mais imediato do que a visualização de números pois as cores saltam à vista, dando mais percepção ao utilizador (Figura 2.13). Permite revelar padrões de sincronia no tempo dos picos das epi-

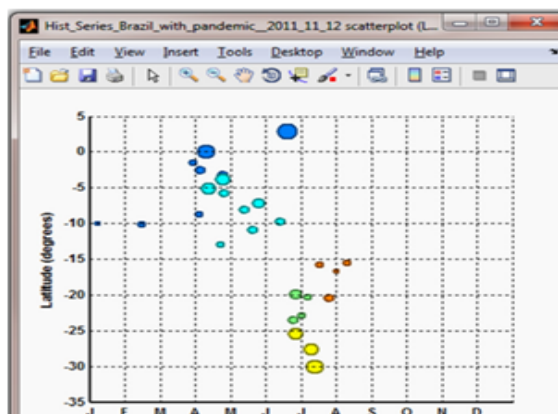


Figura 2.13: Matriz de intervalos temporais, [Alonso and McCormick \[2012\]](#).

demias através de diferentes séries temporais. Quando as séries temporais não representam dados recolhidos de diferentes locais, mas, de certa forma se relacionam com outras características, como por exemplo faixas etárias, a exploração das suas propriedades pode ser feita através de gráficos de dispersão e de grelhas de calor (Figura 2.14).

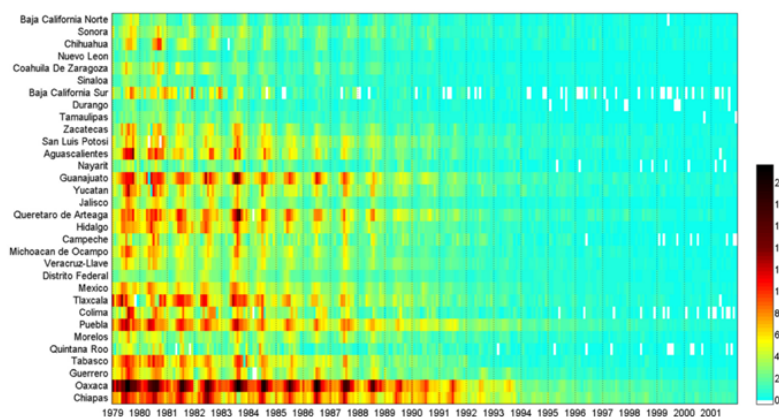


Figura 2.14: Grelha de calor, [Alonso and McCormick \[2012\]](#).

Outra característica do EPIPOI é a capacidade de exportação dos gráficos em vários formatos. Um ponto fraco do EPIPOI é a comparação entre gráficos, apesar de ser possível exportar gráficos, o utilizador tem de recorrer a esta opção para poder comparar gráficos, em vez de ter uma espécie de dashboard que permita comparar alguns gráficos, por exemplo, lado a lado.

2.2.7 EpiCaster

O EpiCaster é uma aplicação que permite a visualização de dados epidemiológicos, nomeadamente dados espaço-temporais e o número de pessoas infetadas [Deodhar et al., 2015]. É possível selecionar a doença a analisar através de uma lista. Isto tem a vantagem de estarem listadas todas as doenças presentes na aplicação mas tem a desvantagem de caso contenha muitas doenças a lista parecer infundável e o utilizador ter de percorrer a lista toda para encontrar, por exemplo, a última doença listada.

Assim que a doença é selecionada, é apresentado no mapa mundo, de forma destacada, com cores, onde há ocorrências da doença. O utilizador pode selecionar uma região em particular de forma a analisar mais precisamente essa zona. De forma a mostrar a evolução temporal, o EpiCaster providencia uma timeline cuja unidade é a semana que permite ao utilizador ter informação semanal sobre a doença previamente selecionada (Figura 2.15).

Através da timeline o utilizador consegue ver as alterações no número de infetados com determinada doença numa certa região, neste caso, o EpiCaster tem informações de até quatro semanas antes. O facto de ser interativo dá liberdade ao utilizador para escolher qual a semana que pretende obter mais informações. Também é possível ver o agregado de pessoas infetadas a nível geográfico, como estados ou países, recorrendo a um sistema de informação geográfica. Com este sistema é possível mostrar os dados num mapa, permitindo a seleção de regiões. É apresentado através de um *choropleth*, onde é possível ver, através de um esquema de diferentes cores, a severidade de uma epidemia nas diferentes regiões. Esta técnica de visualização permite ver de uma forma visualmente rápida quais as regiões com maior número de pessoas infetadas. No entanto, caso não haja muita diferença entre os valores dos vários países, o esquema de cores não irá ser a melhor opção, uma vez que irá ser apresentado um mapa praticamente monocromático.

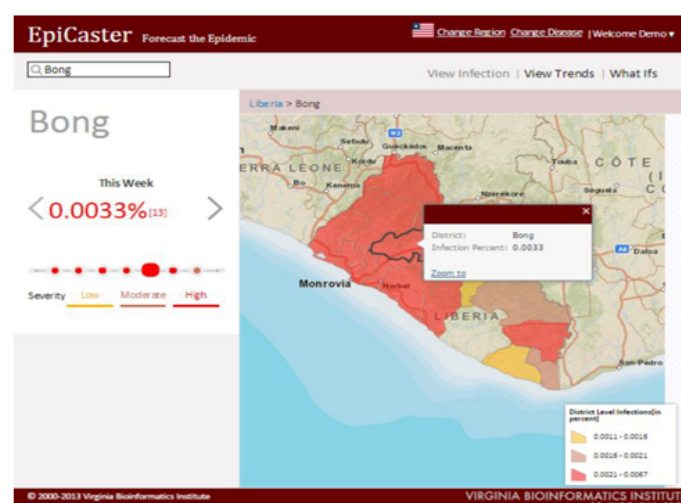


Figura 2.15: Epicaster, Deodhar et al. [2015].

Para finalizar, o EpiCaster permite ainda aos utilizadores ver tendências numa região ao longo de um dado período de tempo, assim como ver picos de pessoas infetadas num determinado intervalo de tempo. O utilizador pode seleccionar se pretende ver a tendência das semanas passadas ou da semana atual. Esta visualização é feita através de um gráfico chamado EpiCurves. O EpiCurves é um gráfico de barras que mostra o número de pessoas infetadas em cada semana. Um dos problemas do EpiCaster é ter poucas doenças pois ainda se encontra num estado inicial de desenvolvimento e ainda estar a ser integrado com base de dados de poucas doenças.

2.2.8 PanViz

De forma a modelar e visualizar dados de pandemias, [Maciejewski et al. \[2011\]](#) desenvolveram um sistema, o PanViz. O PanViz tem uma janela principal que é uma vista espaço-temporal e três gráficos laterais com vistas de estatística populacional, como o número de pessoas infetadas, número de hospitalizados e número de mortes causadas pela doença. Esta disposição de gráficos ligados, permite ao utilizador visualizar as mudanças ocorridas ao longo do tempo (Figura 2.16). O sistema permite um filtro interativo com base na demografia, mostrando o número de pessoas afetadas pela pandemia como percentagem do seu grupo etário. Com isto o utilizador consegue observar quais as regiões mais afetadas para uma dada faixa etária.

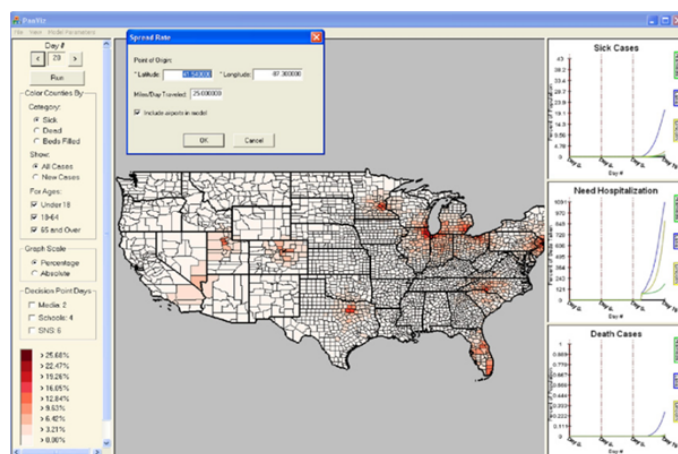


Figura 2.16: PanViz, [Maciejewski et al. \[2011\]](#).

No PanViz, as faixas etárias estão agrupadas em menos de 18, 18-65 e mais de 65 anos e as regiões em rurais, pequenas cidades e grandes metrópoles. Estes dados podem ser visualizados através de *choropleths*, tendo tons de cores como escala, maior o número de pessoas afetadas pela pandemia mais escuro é o tom da cor. Isto tem a vantagem de se ter uma rápida percepção visual, contudo, nas regiões fronteiriça, há uma mudança abrupta o que pode não corresponder à realidade dos dados, isto

e, numa região contígua, com dois tons de cores, a mudança de uma para a outra pode levar a um erro na interpretação, induzindo em erro de leitura o utilizador, pois vê que há uma mudança abrupta entre as duas cores.

No caso dos dados das vistas laterais, é possível ver o número de pessoas infetadas, hospitalizadas e mortes por dias. Estes dados são apresentados através de gráficos de linhas, sendo a variável apresentada no eixo vertical e os dados temporais no eixo horizontal. Com esta técnica de visualização é possível ver a evolução das três variáveis, dia a dia, percebendo se o número de pessoas infetadas, hospitalizadas ou mortas devido à pandemia está a aumentar ou diminuir e ao mesmo tempo permite identificar picos/surtos que possam ter ocorrido. O facto do PanViz ter este dashboard é uma vantagem pois permite num só painel ter diversas informações sobre uma pandemia numa dada região e ver no mapa quais as regiões mais afetadas. No entanto, é difícil comparar duas ou mais regiões, uma vez que só é possível ver uma região de cada vez, sendo difícil para o utilizador comparar o número de mortos entre regiões ou perceber entre duas regiões qual a que tem mais vítimas hospitalizadas da pandemia.

No mapa foram também marcados as restantes variáveis de maneira a que o utilizador conseguisse observar diversas relações como por exemplo a região onde mais jovens eram afetados pelas doenças ou os tipos de doenças associados às diversas regiões. Este tipo de informação é muito mais fácil obter quando representado geograficamente ao invés de gráficos circulares.

2.3 Discussão

Atualmente, em Portugal, o único sistema existente para notificação de doenças de declaração obrigatória é o SINAVE, que irá agora ser melhorado para que seja possível analisar e visualizar os dados de doenças infecciosas em tempo real. A nível europeu existe um sistema, o Surveillance Atlas of Infectious Diseases, que permite ver e analisar dados referentes a algumas doenças transmissíveis, contudo, apenas um pequeno número de doenças se encontra representado e a visualização não possui dados em tempo real.

Nos restantes sistemas analisados neste capítulo é possível ver as principais tarefas que um sistema apropriado para análise e visualização de dados epidemiológicos deve ser capaz de concretizar, assim como as respetivas técnicas presentes, encontrando-se representados na Tabela 2.1.

Analisando mais detalhadamente os sistemas que existem atualmente, constata-se que apresentam os dados de uma forma não muito diferente. No entanto há umas variantes, por exemplo os dados geográficos da incidência são representados, na sua maioria, através de mapas como é o caso do PanViz de [Maciejewski et al. \[2011\]](#), alguns, como por exemplo o VoroGraph de [Dunne et al. \[2015\]](#), com a vertente de CVT Voronoi, sendo uma boa opção visto que distribui bem as regiões. Também há uns sistemas que apresentam as regiões através de *bubble charts* onde cada região é uma bolha,

Tabela 2.1: Tarefas e técnicas de visualização

	Bubble Chart	Mapa	CVT Voronoi	Border encoded map	Timeline	Lista	Scatterplot	Gráfico barras
Analisar incidência	X	X	X				X	
Regiões fronteira				X				
Variação tempo	X				X		X	
Localizar incidência		X	X					
Ranking						X		
Principais sintomas	X					X		
Casos por dia						X		X
Grupo etário						X		X

tendo como exemplo o sistema desenvolvido por [Blevins et al. \[2016\]](#). Este tipo de apresentação não é a ideal, uma vez que é difícil avaliar a transmissibilidade pois não se percebe as regiões contíguas se não as conhecermos pelo nome representado nas bolhas. Um dos principais problemas encontrados nos diversos sistemas é a comparação entre regiões, havendo poucos que permitem a comparação entre regiões e anos, e é uma funcionalidade interessante para ser desenvolvida num sistema de visualização de doenças transmissíveis para averiguar se as medidas tomadas estão a surtir efeito. Por último, há a questão da apresentação de vários dados num só painel, onde apenas alguns sistemas apresentam uma espécie de dashboard como são os casos do PanViz de [Maciejewski et al. \[2011\]](#) e do EPIPOI de [Alonso and McCormick \[2012\]](#), sendo mais fácil para o utilizador ver os dados, mantendo-os num só sítio e de forma organizada.

Atentando na tabela 2.1 constata-se que várias técnicas conseguem mostrar, de forma diferente, os dados, contudo nenhuma consegue disponibilizar uma visualização que abranja todas. Por isso, é necessário conjugar vários paradigmas, aproveitando os pontos fortes de cada um deles de forma a visualizar o que se pretende.

3

Solução

Conteúdo

3.1	SINAVE	33
3.2	Levantamento de requisitos	35
3.3	Arquitetura de software da solução	36

3.1 SINAVE

Os dados que dão suporte ao sistema provêm do SINAVE. O SINAVE é um sistema de vigilância em saúde pública, que identifica situações de risco, recolhe, atualiza, analisa e divulga os dados relativos a doenças transmissíveis (www.dgs.pt/paginas-de-sistema/saude-de-a-a-z/sinave.aspx) e é suportado por uma aplicação informática disponível na *world wide web*, sendo apenas acedida por profissionais de saúde e laboratórios de análises clínicas registados, utilizando para o efeito credenciais pessoais. Atualmente, o SINAVE apenas recolhe a informação, sendo um portal para a notificação de casos de doença de declaração obrigatória por parte dos profissionais de saúde (SINAVEmed) e por parte dos laboratórios de análises clínicas (SINAVElab). Não tendo por isso uma forma de apresentar e visualizar os dados, apenas internamente, na DGS, é possível extrair os dados diretamente da base de dados, obtendo um ficheiro no formato excel com os resultados pretendidos. Isso leva a uma análise mais lenta dos dados e é difícil detetar à primeira vista surtos e informações específicas como qual o grupo etário com maior incidência de uma certa doença ou ainda o estado vacinal dos doentes. Os dados do SINAVE, med e lab, estão centralizados numa base de dados alojada nos Serviços Partilhados do Ministério da Saúde (SPMS) (Figura 3.1), sendo utilizado para efeitos de análise uma réplica dessa base de dados. Todo o acesso à base de dados é efetuado através desta instituição com a autorização da DGS. Esta réplica da base de dados está a ser atualizada apenas uma vez por dia, às quatro horas da manhã. Nela constam os dados referentes às notificações efetuadas pelos médicos, no caso do SINAVEmed, e às efetuadas pelos laboratórios clínicos, no caso do SINAVElab. Atendendo ao SINAVE lab, que é a componente em análise nesta dissertação, uma notificação é a informação que é comunicada à DGS através do SINAVE. Cada notificação contém dados referentes ao paciente que originou a notificação, como o seu nome, data de nascimento, local de residência e ainda informações clínicas como a doença a reportar, análises laboratoriais realizadas e respetivos resultados.



Notificações no SINAVEmed e SINAVElab.

Figura 3.1: Notificação do SINAVE.

Os utilizadores do SINAVElab, para notificarem a DGS, têm duas formas para o fazer. Podem fazê-lo diretamente no portal do SINAVE, onde após entrarem no sistema através das credenciais obtidas no registo prévio, lhes é apresentado um questionário para inserirem os dados referentes à notificação, ou seja, os dados pessoais e clínicos do paciente, como descritos anteriormente. Outra forma possível para notificar é através de interoperabilidade nos próprios sistemas que possuam nos laboratórios clínicos, sendo o esquema de notificação exatamente o mesmo do portal do SINAVE (Figura 3.2). Os dados provenientes quer diretamente do portal do SINAVE quer de interoperabilidade, são armazenados na mesma base de dados em servidores da SPMS. A base de dados foi desenvolvida com tecnologia Oracle e apenas pode ser acedida com as credenciais fornecidas pela SPMS e estando conectado à rede da SPMS ou da DGS.

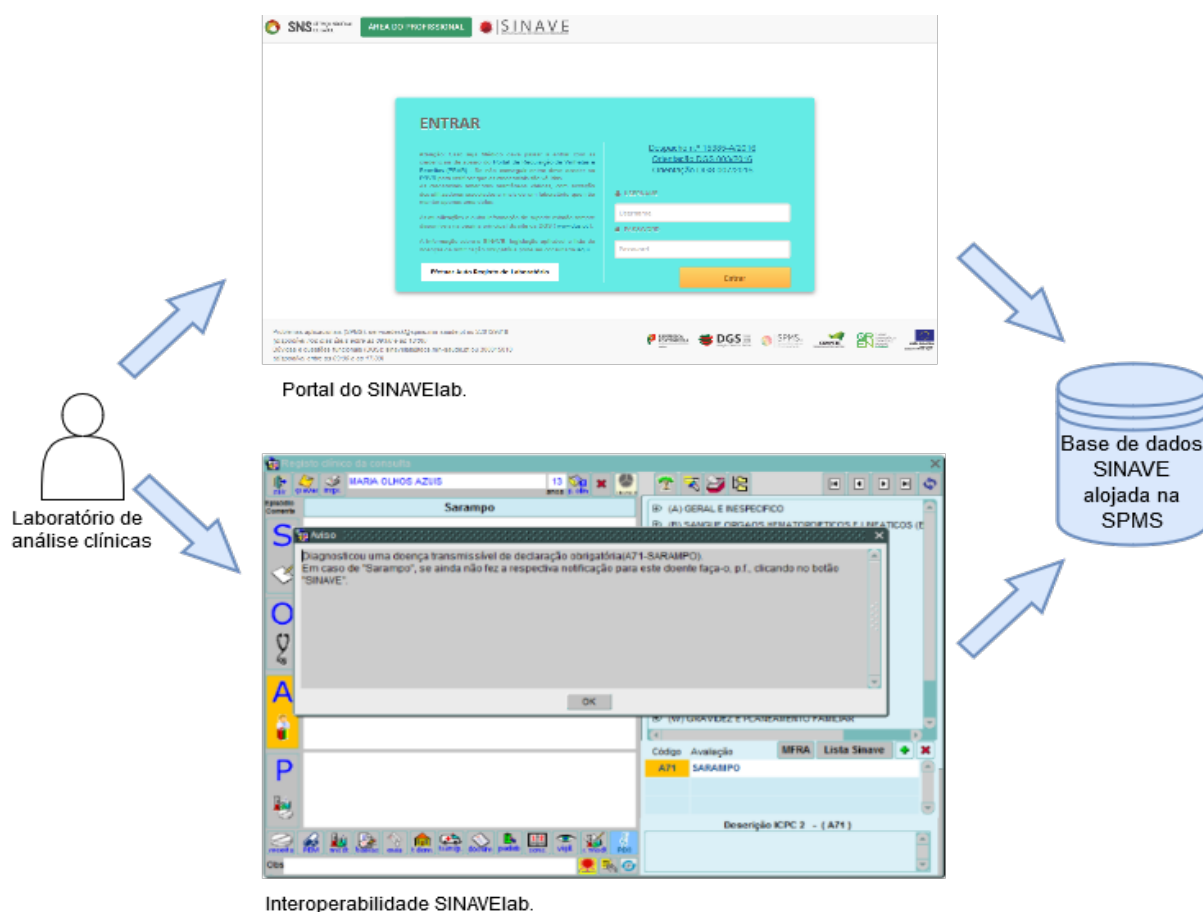


Figura 3.2: As vias possíveis para fazer uma notificação no SINAVElab.

3.2 Levantamento de requisitos

De forma a perceber melhor o domínio, estive integrado na Divisão de Epidemiologia e Vigilância da DGS, à responsabilidade da Dr^a Cátia Sousa Pinto. O sistema implementado tem como base os dados do SINAVE. Para que pudesse compreender o pretendido para o sistema, tive cinco reuniões com a Dr^a Cátia Sousa Pinto, chefe da Divisão de Epidemiologia e Vigilância da DGS, e com José Loff, consultor de estatística externo contratado pela DGS para acompanhar e auxiliar a implementação de sistemas e serviços na DGS, como é o exemplo do sistema desenvolvido. As duas primeiras foram reuniões mais informativas, espaçadas por duas semanas, que permitiram ter um maior conhecimento do domínio e identificar os objetivos do sistema, nomeadamente as especificações que deveria cumprir, sendo estas enumeradas abaixo:

- Deveria identificar as notificações por doença, tendo em atenção que em doenças como Hepatite C e VIH, apenas a primeira notificação para um dado paciente é considerada, a partir daí, todas as notificações para o mesmo paciente para a mesma doença são consideradas duplicados pois são doenças sem cura, pelo que havendo o registo da primeira notificação, o sistema já a contabilizou nessa altura.
- Em termos de distribuição geográfica, os dados deveriam ser apresentados por local de residência, quando conhecido, e por local de ocorrência, quando desconhecido. No caso do local de residência ser conhecido, deveria ser apresentada primeiramente a freguesia e só depois o concelho. A mesma ordem de apresentação deveria ser seguida no caso do local de residência ser desconhecido, fazendo referência à freguesia e concelho do local de ocorrência. Posteriormente, devido a questões de privacidade e da sensibilidade da informação apresentada, tomou-se a decisão de não considerar os dados das freguesias e só considerar os concelhos. A apresentação dos dados geográficos deveriam seguir os dados das ARS e NUTSIII. Notando as ARS são administrações regionais de saúde e coordenam a prestação de cuidados de saúde a todos os níveis na respetiva área geográfica. E NUTS é o acrónimo de “Nomenclatura das Unidades Territoriais para Fins Estatísticos”, sistema hierárquico de divisão do território em regiões, sendo a NUTSIII definida de acordo com critérios geográficos.
- Temporalmente os dados deveriam ser apresentados primeiro por data do resultado laboratorial e posteriormente por data da notificação.
- Deveria apresentar os dados também por género e grupo etário. Por género deveria apresentar os casos masculinos e femininos. Por grupo etário deveria apresentar os dados para as idades 0-1, 1-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+ anos.

- Deveria permitir que o utilizador do sistema conseguisse extrair os dados para efetuar a sua própria análise.

As restantes ocorreram com um mês de intervalo e foram reuniões de avaliação que serviram para avaliar o ponto de situação do desenvolvimento do sistema, esclarecer dúvidas em relação aos dados constantes na base de dados e discutir técnicas de visualização a utilizar.

3.3 Arquitetura de software da solução

O sistema irá estar disponível através de um *browser*, estando acessível através da página web da DGS, sendo que atualmente ainda não está online. O sistema é composto por três componentes: a base de dados, na qual estão armazenados os dados a utilizar, o *backend*, camada responsável por aceder aos dados e fazer a análise dos mesmos, e o *frontend*, camada responsável pela apresentação gráfica dos dados ao utilizador.

A base de dados está alojada na SPMS, sendo possível aceder através das credenciais cedidas para acesso da DGS que foram solicitadas nas três primeiras reuniões na DGS. Apenas se consegue aceder à base de dados se, para além de possuir as credenciais, estiver conectado à rede da SPMS ou da DGS através de VPN, uma rede privada virtual que conecta dois ou mais computadores, cifrando as ligações de forma a que só estes tenham acesso ao seu conteúdo. A base de dados foi criada com o Sistema de Gestão de Base de Dados (SGBD) da Oracle. Um SGBD é um programa que permite criar e manipular base de dados, mantendo a independência entre o modo como os dados são armazenados e a sua estrutura, e o modo como são utilizados num programa ou aplicação que os manipulam. Para aceder aos dados utiliza-se queries em SQL. O SQL é uma linguagem que permite fazer consultas (queries) em base de dados e modificar ou eliminar elementos de uma base de dados. O modo como a base de dados armazena os dados e está estruturalmente organizada está descrito mais detalhadamente no capítulo 4.

A camada de análise, o *backend*, é a camada do sistema que recebe os dados da base de dados e faz a análise de forma a que a camada seguinte, o *frontend*, apenas seja responsável por apresentar os dados ao utilizador. É nesta camada que os dados são obtidos diretamente da base de dados e lhes é aplicado um processo de análise para assegurar que após o mesmo, os dados se encontram num estado de coerência e consistência que permita criar visualizações a partir dos mesmos. É também nesta camada que é criado um servidor que vai servir os dados à camada de visualização, o *backend*. Esta camada de análise está implementada em R (<http://www.r-project.org>). O R é uma linguagem muito utilizada principalmente por estatísticos e analistas de dados, uma vez que permite fazer cálculos estatísticos e análise de dados facilmente e não requer conhecimentos aprofundados de programação.

O *backend* encontra-se mais detalhado no capítulo 4.

A camada de visualização está também implementada em R. Os gráficos foram implementados através da biblioteca Highcharts (<http://www.highcharts.com>) que permite criar gráficos de forma simples em R sem manipular diretamente JavaScript, sendo essa a linguagem que de certa forma está encapsulada nesta biblioteca para possibilitar que se crie gráficos sem grandes conhecimentos de programação. Nos primeiros dois meses debateu-se sobre qual a linguagem a utilizar, tendo chegado a considerar-se o uso do JavaScript. O JavaScript é uma linguagem de scripts principalmente no lado do cliente, embora também possa ser utilizada no lado do servidor, sendo muito utilizada na web para moldar o comportamento de elementos HTML e CSS. Esta opção de utilizar o JavaScript, aliada a HTML e CSS, seria, do ponto de vista de um engenheiro informático, a escolha mais natural, visto que são linguagens adequadas ao pretendido. Contudo, devido à ausência de pessoal qualificado em informática na DGS, não existindo quem conseguisse manter o sistema caso este fosse implementado através desta opção, fez com que se optasse por implementar em R, pois desta forma já poderia ter uma maior acompanhamento no processo de desenvolvimento do sistema por parte da DGS e a sua instalação e manutenção já poderiam ser asseguradas, nomeadamente por técnicos superiores da área de estatística na DGS. O *frontend* encontra-se mais detalhado no capítulo 5.

Em termos visuais, o sistema consiste num dashboard interativo, apresentando um menu à esquerda e a zona da visualização dos dados à direita (Figura 5.2).

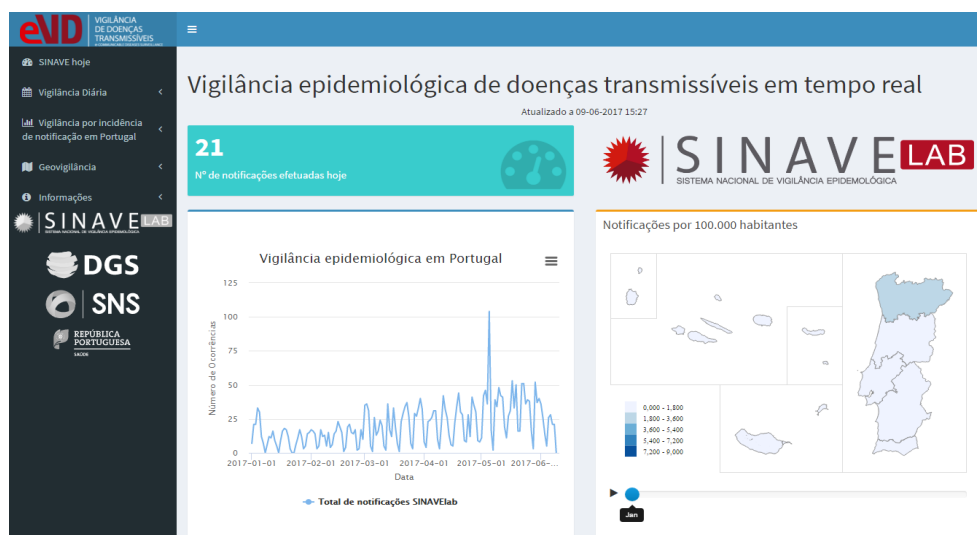


Figura 3.3: Dashboard.

Na camada de visualização, o *frontend*, os dados já se encontram processados, pelo que é nesta camada que são criadas as visualizações de forma a mostrar informações sobre os dados ao utilizador. De forma a que esta camada tenha os dados para visualizar, são feitos pedidos HTTP ao servidor criado na camada de análise para obter os dados. Os gráficos implementados nesta camada podem ser

accedidos através de um URL único, sendo possível visualizá-lo incorporado na página web do sistema ou isoladamente, uma vez que cada gráfico corresponde a um pedido HTTP diferente. Isto faz com que possa haver problemas de redundância, uma vez que não há controlo na escolha de servidores para cada pedido, há o risco dos vários elementos virem de servidores diferentes e ainda que os dados estejam dessincronizados. Uma solução possível a implementar aquando da instalação do sistema nos servidores da SPMS é a criação de uma cache entre a camada de análise e a base de dados que guarda os dados de forma síncrona para que todas as visualizações criadas sejam coerentes, e ao mesmo tempo irá reduzir a carga sobre a camada de análise. Outro problema que poderá ocorrer é a concorrência, uma vez que o R só responde a um pedido de cada vez. Uma possível solução será a utilização de uma cache na camada de visualização onde os resultados dos pedidos são guardados por um período não inferior a 10 minutos porque não há mudanças significativas durante esse tempo (Figura 3.4). De salientar que atualmente o sistema não possui qualquer esquema de caches pois ainda não se encontra instalado no servidor da SPMS, no entanto está já preparado para a instalação das caches. O sistema de caches irá executar sobre os dados obtidos da base de dados, na periodicidade desejada, a camada de análise, guardando o seu resultado na cache que se encontra no *backend*. De seguida, a cache no *frontend* executa a camada de visualização sobre os dados que se encontram na cache do *backend*, guardando os gráficos resultantes já prontos a visualizar.

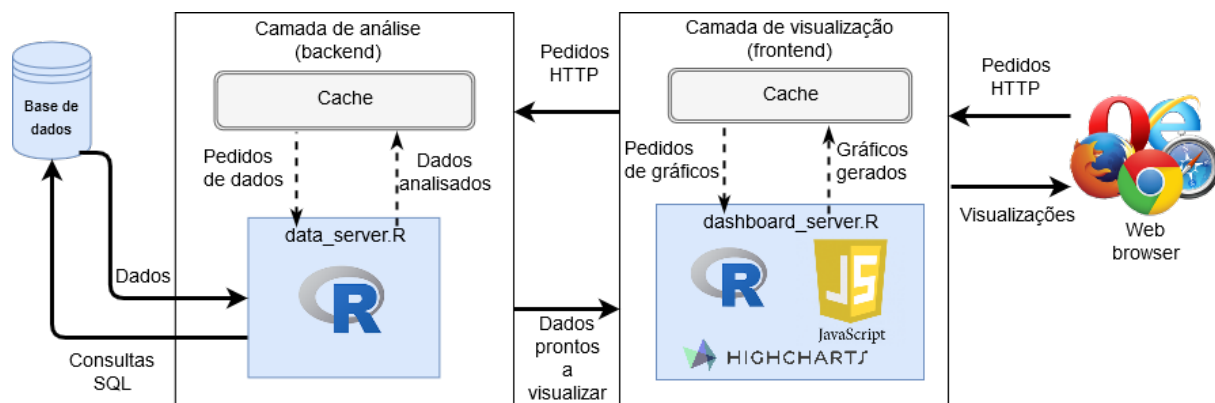


Figura 3.4: Arquitetura da solução.

4

Análise de dados (*backend*)

Conteúdo

4.1	Acesso e análise aos dados	41
4.2	Organização da base de dados	42
4.3	ID das notificações	42
4.4	Dados referente à data de notificação	43
4.5	Dados duplicados e especificações de doenças	44
4.6	Dados referente ao grupo etário	45
4.7	Dados referente à localização	46
4.8	Dados referente ao género	47
4.9	Funcionamento do <i>backend</i>	47

Para que um projeto seja bem implementado e tenha sucesso, é necessário que o mesmo esteja bem estruturado e que esteja dividido por etapas bem definidas. É fundamental caminhar-se com passos firmes e finalizar cada etapa antes de passar à próxima. Pelo que antes de começar a visualizar qualquer informação sobre os dados do SINAVElab, foi preciso ter acesso à base de dados fornecida pela SPMS para perceber realmente que informação se poderia obter.

4.1 Acesso e análise aos dados

Inicialmente foi permitido aceder a uma base de dados de qualidade, através da ferramenta SQL Developer da Oracle, onde constavam os dados que os laboratórios clínicos forneciam através das notificações no SINAVElab. O SQL Developer é um programa que possui uma interface gráfica para facilitar o acesso aos dados de uma base de dados. Contudo, esta base de dados continha também dados de testes, uma vez que ainda não se tratava do repositório final e mesmo o modelo da base de dados ainda não se encontrava fechado. Isto levou existissem alguma incertezas quanto à estrutura dos dados e como iriam de facto ficar armazenados.

Essa base de dados possuía informações sobre as notificações dos laboratórios clínicos em que constavam os dados da pessoa sujeita à análise clínica, como nome, data de nascimento, número de utente ou outro número de identificação pessoal como cartão do cidadão ou número de contribuinte, e local de residência, isto, desde que a análise não fosse feita sob anonimato, nesse caso, estes campos não possuíam qualquer tipo de dados. Em relação à informação clínica, tinha dados referentes à doença a que a notificação dizia respeito, data da notificação no sistema, laboratório (número de identificação fiscal, localidade), análises realizadas, resultado da análise (se positivo, negativo ou inconclusivo), produtos e agentes utilizados nas análises. Estes dados encontravam-se guardados em várias tabelas diferentes, sendo que para aceder à totalidade dos dados tinha que se cruzar as tabelas necessárias para obter o pretendido. Muitos dos campos não eram relevantes obter, pois não se pretendia apresentar algumas dessas informações, umas porque não eram relevantes para o público alvo do sistema (público em geral), como por exemplo o agente ou produto utilizado nas análises, outras por questão de privacidade e legais impostas pela Comissão Nacional de Proteção de Dados (CNPD), que o próprio sistema SINAVE respeita, como é caso do nome das pessoas que não podem ser revelados publicamente. A base de dados apesar de ser atualizada em tempo real, a réplica atualmente apenas é atualizada uma vez por dia, mais concretamente às quatro da manhã.

4.2 Organização da base de dados

Um pormenor encontrado, e que não era esperado, foi o facto de como os dados provenientes das diferentes vias de notificação por parte dos laboratórios de análises clínicas, notificação por interoperabilidade e diretamente no portal do SINAVE, estarem armazenados em tabelas diferentes, tendo inclusive campos distintos. Os laboratórios clínicos podem notificar no SINAVElab através de duas formas distintas, através do portal do SINAVE, acedendo com as suas credenciais ao sistema e através de interoperabilidade. O modo como estes dados estão armazenados, isto é, estão em tabelas diferentes e não possuem exatamente os mesmos campos, faz com que seja necessário proceder a uma limpeza nos dados para que no final deste processo os dados sejam coerentes e consistentes.

Esse processo de limpeza passa por tornar os dados idênticos, isto é, os dados que se obtêm quer da tabela que contém os dados diretamente do portal do SINAVE quer os da tabela que provêm por interoperabilidade, terem os mesmos campos de forma a que seja possível juntá-los, com o objetivo de fazer a análise pretendida a todos os dados uma única vez, pois não é relevante, na camada de visualização, a proveniência dos dados, se do portal do SINAVE se da interoperabilidade dos sistemas dos laboratórios clínicos.

No processo de limpeza, realizado em R, constatou-se que existia um campo, data de colheita, que apenas constava na tabela dos dados provenientes da interoperabilidade dos laboratórios clínicos, estando ausente na tabela dos dados do portal do SINAVE. Este campo deveria existir em ambas as tabelas, uma vez que é um dado importante para a DGS, sendo um dos critérios para apurar com mais precisão a data da notificação. Por isso, foi reportado à SPMS, entidade responsável pela criação e manutenção da base de dados, a ausência desse campo e no respetivo formulário de notificação no portal do SINAVE, solicitando que o mesmo fosse acrescentado, tendo sido prontamente corrigido.

4.3 ID das notificações

Outro pormenor encontrado aquando da realização do processo de limpeza, foi o facto do identificador único (ID) de cada notificação, ter um contador para cada uma das vias de notificação. Ou seja, poderia haver duas notificações com o mesmo ID. Isso fazia com que se sobrepussem notificações provenientes do portal do SINAVE com as notificações provenientes da interoperabilidade dos sistemas das clínicas laboratoriais. Esse facto foi comunicado à SPMS, e foi corrigido o modo como os ID's estavam a ser atribuídos, passando a solução por atribuir diferentes inícios aos contadores de cada uma das vias de notificação. Ao mesmo tempo que esta alteração foi implementada, foi comunicado pela SPMS que futuramente as tabelas provenientes do portal do SINAVE e da interoperabilidade dos sistemas das clínicas laboratoriais, iriam convergir numa só, ou seja, os dados provenientes de ambas as vias seriam armazenados numa só tabela. Contudo, não foi dado prazo para que essa alteração fosse

efetuada, pelo que se deu seguimento ao processo de limpeza dos dados para que estes se tornassem consistentes e coerentes.

Posteriormente, foi solicitado à SPMS uma réplica da base de dados final, já com as correções efetuadas e sem os dados de teste para que fosse possível aceder aos dados quanto mais reais possíveis. Após a criação dessa réplica foram criadas credenciais, procedendo-se da mesma forma que a base de dados de qualidade. Esta base de dados, à semelhança da de qualidade, continua a ser atualizada uma vez por dia às quatro da manhã, e, até à data, ainda não possui uma tabela única com os dados provenientes das duas vias de notificação por parte das clínicas laboratoriais.

4.4 Dados referente à data de notificação

Obtendo os dados da base de dados através de *queries* SQL, foram obtidas apenas as informações que eram necessárias para cumprir os requisitos, como os IDs de cada notificação, dados das pessoas sujeitas às análises clínicas (data de nascimento, nome, apenas numa fase inicial para análise, sendo posteriormente eliminado para garantir o anonimato, privacidade e respeitar a legislação da CNPD, e ainda o local de residência), doença notificada, data de colheita, análise realizada na notificação e respectivo resultado, data de realização da análise e ainda a data de entrada da notificação no SINAVE. A *query* que obtém estes dados está aplicada quer aos dados provenientes do SINAVE quer aos provenientes por interoperabilidade.

Um dos requisitos que o sistema deveria cumprir era a atribuição da data à notificação, em que a data final atribuída pelo sistema deveria ser a menor de três casos: data de colheita, data de realização e data de entrada da notificação no SINAVE. Uma vez que todos estes dados são obtidos diretamente da base de dados, esta análise está incorporada na própria *query*, aproveitando as capacidades do SQL em calcular o menor destes três campos, quando existentes. Nesta fase, foi encontrado um problema, apesar destes campos serem datas, estão a ser guardados com o formato de texto, *string*. Desta forma, não é possível calcular diretamente a menor data dos três campos através do SQL, uma vez que a função LEAST recebendo como argumento caracteres de texto tem um comportamento diferente do que na comparação de datas. Por isso, foi necessário converter os três campos, data de colheita, data do resultado e data da entrada da notificação no sistema, do formato texto para o formato data, *date*, recorrendo à função do SQL TO_DATE . No entanto, nem todos os casos notificados possuem os três campos das datas preenchidos, havendo casos onde estão vazios. Sendo que, antes de se proceder à conversão do formato texto para data, foi necessário atribuir uma data default (01-01-9999) aos campos que não a possuem. Esta atribuição é necessária para que na comparação entre as três datas sejam consideradas todas elas. Foi escolhido o dia 01-01-9999 para garantir que a data que não está preenchida nunca será a menor das três datas. A menor ficou denominada por “data final”.

4.5 Dados duplicados e especificações de doenças

Outro dos requisitos que o sistema devia cumprir era a identificação de casos duplicados, apenas considerando-os uma única vez. Há doenças que não são curáveis atualmente, como a Hepatite C e o VIH, pelo que caso exista uma notificação para uma determinada pessoa com essa doença no SINAVE, o sistema não deve considerar todas as análises que essa mesma pessoa venha a realizar e acuse positivo, pois assim, estará a dar entrada no sistema casos que não são mais do que uma repetição de um caso já reportado. No entanto, é importante salientar que os laboratórios devem reportar todos os casos ao SINAVE para que a DGS tenha toda a informação existente, apenas no novo sistema desenvolvido, para efeitos de análise, não são considerados duplicados, pois é uma informação que irá estar visível publicamente e caso fossem considerados todos os casos, uma vez que seria errado sob o ponto de vista epidemiológico.

Para que os casos duplicados fossem detetados pelo sistema, foi necessário obter o nome das pessoas que constam na notificação para se detetar que de facto se tratam de duplicado, sendo posteriormente eliminada a informação do nome da pessoa por questões de privacidade. Como pode ocorrer o caso de duas ou mais pessoas partilharem o nome, considerou-se uma aglutinação do campo do nome da pessoa com o campo da data de nascimento, uma vez que a probabilidade de duas ou mais pessoas partilharem o nome e a data de nascimento é mais reduzida. Após a junção do nome com a data de nascimento, apenas se considerou a notificação com a data de notificação mais antiga, assegurando assim uma maior precisão sobre a primeira data que aquela pessoa acusou positivo num teste laboratorial. No entanto, este processo apenas consegue detetar e filtrar os casos de doenças que não possuem cura, ou seja, apenas deve ser considerado a primeira data em que o resultado da análise seja positivo uma vez que não é curável. Todas as análises laboratoriais realizadas posteriormente a essa data irão ter resultado positivo, mas há doenças que têm cura só que durante um determinado período de tempo, caso a pessoa realize novas análises clínicas, e que seja positivo, não devem ser consideradas como novos casos por ainda se considerar respeitante ao caso anterior por estar temporalmente próximo. Esse período em que se deve considerar que diz respeito ao mesmo caso anteriormente reportado, considerando assim um duplicado, varia consoante as doenças, sendo que há doenças cujo período é de um ano e doenças em que o período é de cinco anos. Inicialmente foi guardada uma lista com as doenças que devem ser consideradas novos casos apenas após um ano da primeira data em que ocorra uma análise laboratorial que seja positiva, outra lista com as doenças que só se deve considerar novos casos passados 5 anos desde a primeira data de uma análise positiva, uma outra lista com as doenças que só devem ser considerada a primeira notificação e uma nova lista com as restantes doenças que devem ser considerados novos casos a cada nova notificação no SINAVE. Nessa função, é verificada a que lista pertence a doença de cada notificação e são devolvidas as notificações que cumprem o intervalo temporal. Dependendo da lista em que estiver, será atribuído

um intervalo, *gap*, em dias, considerando que um ano corresponde a 365 dias e 5 anos a 1825 dias, para que se possa utilizar as potencialidades do R em calcular a diferença, em dias, entre duas datas. Após a atribuição do *gap*, são selecionadas todas as notificações que tenham como diferença de datas um número superior ao *gap*, isto é, ao iterar sobre cada notificação, verifica-se a data de notificação das notificações já existentes no sistema, e caso o sistema já possua uma notificação para aquela pessoa, e, ao mesmo tempo, se a diferença entre esta data e a data da nova notificação for superior ao *gap*, significa que não é duplicado. A data desta nova notificação passa a ser considerada a data mais recente.

De salientar que apesar deste processo de detecção de duplicados cobrir muitos dos casos, há ainda a considerar os casos em que as análises laboratoriais podem ser realizadas de forma anónima. Nestes casos, é totalmente impossível descortinar qualquer tipo de duplicados, pois não há dados suficientes que permitem associar uma notificação a uma pessoa de forma a saber que duas notificações dizem respeito à mesma pessoa, sendo este viés considerado na fase de interpretação dos resultados.

4.6 Dados referente ao grupo etário

Para que o sistema consiga considerar os grupos etários pretendidos, foi necessário calcular a idade a partir da data de nascimento. Tal como as outras datas armazenadas na base de dados, a data de nascimento, não está com o formato *date*, mas sim como texto, *string*. Pelo que também aqui houve a necessidade de converter em formato *date*, desta vez não através do SQL mas de funções em R, pois não é necessário comparar datas e é simples fazer a conversão no R. No entanto, foi encontrado outro problema. A data de nascimento armazenada na base de dados só possui dois caracteres para o ano, estando guardada no formato DD-MM-AA, sendo DD, dias de 1 a 31, MM, mês de 1 a 12 e AA os dois últimos dígitos do ano. Isto faz com que haja incerteza na determinação do ano de nascimento, por exemplo, alguém nascido a "15-06-17", não se sabe se diz respeito a uma pessoa que nasceu no ano de 1917 ou em 2017. Considerou-se que todas as datas posteriores à data atual, os caracteres do ano dizem respeito a anos de 1900's e não 2000's. Isto faz com que se levante outra questão. No sistema não existirão pessoas com mais de 99 anos, nos casos em que a pessoa tenha mais do que essa idade, o sistema irá considerar menos 100 anos à idade que tem, por exemplo, uma pessoa com 104 anos, será considerada como tendo 4. Contudo, foi a melhor solução encontrada, pois, face ao número de pessoas centenárias em Portugal, teria pouco impacto na análise.

Isto faz com que se levante outra questão. No sistema, por agora, não haverá casos com pessoas cuja data de nascimento seja posterior a "31-12-2016". Contudo, dadas as limitações existentes, forma como está guardado a data de nascimento na base de dados e retorno da função R utilizada para converter texto em datas, foi a melhor solução encontrada, porque depois disso, a função do R devolve

sempre anos 2000's, por exemplo, alguém que tenha nascido a "16-05-18" será convertido para "16-05-2018", o que será impossível, pois a data de hoje é anterior a essa data. Após obter a data de nascimento em formato *Date*, foi calculada a idade de cada pessoa, obtendo para isso a diferença entre a data de nascimento e a data final calculada anteriormente, no primeiro passo de análise. No entanto, o sistema deveria apresentar a idade em grupos etários, por isso, utilizando a função *cut* do R, em que fornecendo como argumentos um vector de inteiros, as idades de cada notificação, atribui um intervalo de idades consoante o argumento recebido. Sendo os grupos etários a considerar 0-1, 1-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74 e 75+ anos, um exemplo do funcionamento desta função seria, para uma notificação cuja idade da pessoa seja 23, será atribuído o grupo etário 15-24 anos, desta forma é cumprido o requisito de apresentar a idade por grupos etários.

4.7 Dados referente à localização

Para cumprir o requisito do sistema mostrar informações sobre a localidade da ocorrência dos casos notificados, era necessário obter o local de residência da pessoa em causa na notificação. Essa informação é obtida, na *query* SQL, através do campo da freguesia do utente das tabelas quer dos dados provenientes diretamente do portal do SINAVE quer dos dados provenientes por interoperabilidade. Nesse campo consta um código. Esse código é relativo à codificação de localidades elaborado pelo Instituto Nacional de Estatística (INE), sendo que é um código único e através do qual se obtém a freguesia, concelho, distrito, NUTS I, NUTS II e NUTS III uma vez que o código é composto por nove dígitos, no formato, "ABCxxyyz", onde "A" diz respeito à NUTS I, "AB" à NUTS II, "ABC" à NUTS III, "xx" ao distrito, "xxyy" ao concelho e "xxyyz" à freguesia. Na base de dados não existe informação sobre a correspondência dos códigos, isto é, não há nenhuma tabela que relacione códigos com as localidades, por exemplo, o nome da freguesia. Por isso, foi necessário criar um ficheiro em formato csv onde figuram os dados do INE para que seja possível relacionar os códigos presentes na base de dados com o nome da localidade. Um ficheiro csv é um ficheiro sem formatação cujos valores estão separados por vírgulas. Primeiro é necessário proceder à leitura deste ficheiro no R. Após a leitura do ficheiro, atribui-se os valores lidos a um vector para que se possa cruzar os dados nele contidos com os dados obtidos da base de dados através das *queries* SQL, devidamente analisados com os processos descritos anteriormente, estando estes dados também guardados num vector. O cruzamento dos dois vectores tem como objetivo a obtenção de um vector final com os dados do nome das localidades e respetivos códigos definidos pelo INE.

Para além do sistema ter como requisito disponibilizar informações por distribuição geográfica territorial, também deveria apresentar informações sobre distribuição geográfica ao nível da saúde, nomeadamente ao nível de ARS. Contudo, na base de dados não existe qualquer informação sobre essa

distribuição geográfica, pelo que, no ficheiro em formato csv foi adicionado a relação entre as ARS com as freguesias e concelhos. A maioria dos códigos que estão na base de dados, quando este campo se encontra preenchido, pois não é um campo que seja de preenchimento obrigatório, diz respeito a códigos de freguesias, no entanto, devido a questões de privacidade das pessoas, nomeadamente em freguesias com pouca população, e legais impostas pela CNPD para o SINAVE (anonimização dos dados), foi decidido não guardar qualquer tipo de dados referentes à freguesia, apenas ao concelho. Protegendo assim, alguma informação que não deva ser revelada publicamente.

4.8 Dados referente ao género

Em relação ao requisito de mostrar informações por género, os dados não estão armazenados na base de dados da mesma forma, isto é, os dados provenientes do SINAVE não estão representados da mesma maneira que os provenientes por interoperabilidade. O género nos dados do portal do SINAVE estão representados por “F”, “M” e “N”, correspondendo aos géneros feminino, masculino e desconhecido, respetivamente, enquanto que nos dados da interoperabilidade estão representados, pela mesma ordem, por “18”, “17” e “N”. De forma a juntar os dados e haver coerência, adotou-se que todos os dados representados por “F” e “18” passassem a designar por “Feminino”, “M” e “17”, por “Masculino” e, “N” e “0”, que correspondem aos casos em que não há informação do género, designados por “Desconhecido”.

Após a análise dos dados, tornando os dados que provêm diretamente do portal do SINAVE e por interoperabilidade, coerentes e consistentes, agregam-se os dados num só vetor.

4.9 Funcionamento do *backend*

Atualmente, sempre que um utilizador aceda ao sistema, o *backend* procede à execução da análise, obtendo em primeiro lugar os dados da base de dados através de pedidos http com as *queries* SQL e aplicando-lhes a análise descrita anteriormente. Após a execução da camada de análise, os resultados são enviados por http ao *frontend* que irá proceder à criação das visualizações. Quando o sistema de caches estiver implementado, este processo será apenas executado na periodicidade desejada e não sempre que um utilizador aceda ao sistema, armazenando o resultado da análise na cache do *backend*, fornecendo diretamente o *frontend* dentro do período de não execução.

Com a obtenção dos dados e com a análise feita, é agora possível começar a construir visualizações de forma a mostrar informações ao público.

5

Visualização (*frontend*)

Conteúdo

5.1	Relação com o eVM - E-Mortality Surveillance	51
5.2	eVD Lab - <i>frontend</i>	51
5.3	Demonstração do potencial	61

5.1 Relação com o eVM - E-Mortality Surveillance

Este sistema desenvolvido para visualizar informações sobre os dados do SINAVElab pretende dar continuidade ao trabalho desenvolvido pela DGS no fornecimento de ferramentas ao público para obter informações sobre a saúde em Portugal, em tempo real. Um sistema desenvolvido antes deste, foi o SICO/eVM – Vigilância eletrônica de mortalidade em tempo real (E-Mortality Surveillance). Este sistema, através dos requisitos propostos e da tecnologia utilizada, conseguiu ser semelhante ao eVM (Figura 5.1). Esse fator é positivo pois assim mantém uma coerência nos sistemas desenvolvidos pela DGS. Este novo sistema desenvolvido à semelhança do SICO com o eVM, passou a designar-se por eVD Lab - Vigilância eletrônica de doenças transmissíveis em tempo real (E-communicable Diseases Surveillance), sendo que Lab diz referência aos dados constantes no SINAVElab.

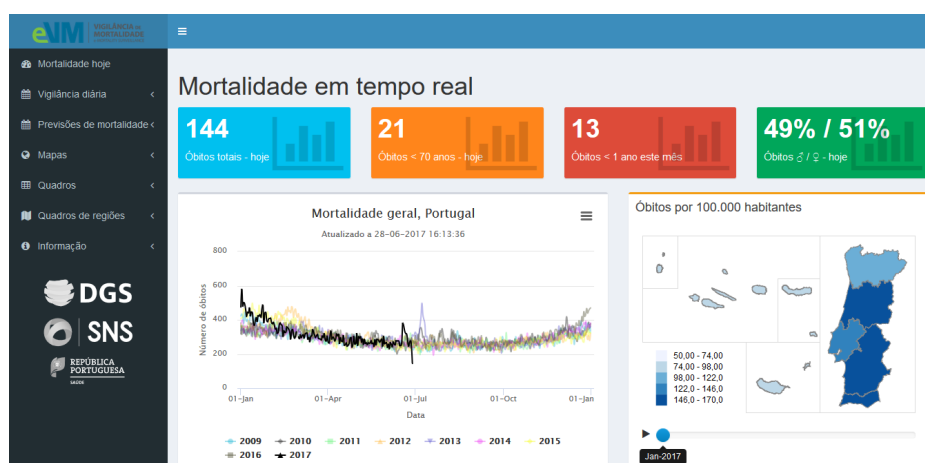


Figura 5.1: eVM.

5.2 eVD Lab - *frontend*

Esta camada, a de visualização, é também, pelos motivos referidos no capítulo 3, desenvolvida em R, utilizando a biblioteca Highcharts.js para criar os gráficos. O Highcharts é uma biblioteca de gráficos escrita em JavaScript puro, oferecendo uma maneira fácil de adicionar gráficos interativos em websites ou aplicações web (<http://www.highcharts.com>). Inicialmente foi ponderada outra opção para criar gráficos em R, o *ggvis*. O *ggvis* é um pacote de visualização de dados em R que permite declarativamente descrever gráficos de dados. A manipulação e a transformação de dados são feitas em R, e os gráficos são renderizados num *browser*, usando a Vega (<http://vega.github.io/vega/>). A Vega é uma gramática de visualização, uma linguagem declarativa para criar, guardar e compartilhar projetos de visualização interativa. Contudo, o *ggvis* tem uma interação limitada, apenas a permite ter através de uma barra deslizante de opções ou botões, e tem de ser utilizado juntamente com a Shiny

(<http://shiny.rstudio.com/>). A Shiny é uma framework web para o R que permite criar aplicações interativas web e alojá-las num servidor Shiny. O servidor Shiny tem uma versão *open-source*, mas esta versão é muito limitada, uma vez que não permite ter controlo sobre o servidor e permite poucos acessos. Embora na versão profissional essas limitações já não existam, é uma versão paga, e a DGS não manifestou interesse nessa opção, recorrendo por isso à função *jug* do R que permite criar um servidor local, sendo posteriormente instalado nos servidores da SPMS. Por esse motivo, foi utilizado o Highcharts, já que não está dependente do servidor Shiny. No entanto, nenhuma dos pacotes ou bibliotecas permite ter um controlo total sobre o modo como se criam os gráficos como em JavaScript onde, através de bibliotecas, como por exemplo o D3.js, existe uma liberdade maior para a criação de gráficos. O D3 é uma biblioteca de JavaScript para visualizar dados usando padrões da web. O D3 combina técnicas visualização e interação com uma abordagem orientada por dados para a manipulação de DOM, fornecendo liberdade de criação da interface visual dos dados. Sendo que DOM significa Modelo de Objeto de Documento (Document Object Model) e é uma convenção multiplataforma e independente de linguagem para representação e interação com objetos em documentos HTML. No Highcharts e bibliotecas semelhantes, não existe controlo total na criação dos gráficos, estando limitando às opções existentes na API disponibilizada (<http://api.highcharts.com/highcharts>).

5.2.1 A interface de navegação

O sistema assenta num dashboard com um menu lateral retrátil (figura 5.2 número 1) e um painel central (figura 5.2 número 2) onde são apresentados os dados. O facto de se basear num dashboard permite apresentar o panorama geral dos dados, resumindo a informação mais importante num único local, dando ao utilizador a possibilidade de escolher o que pretende visualizar e o seu nível de granularidade, através de interações com o sistema.

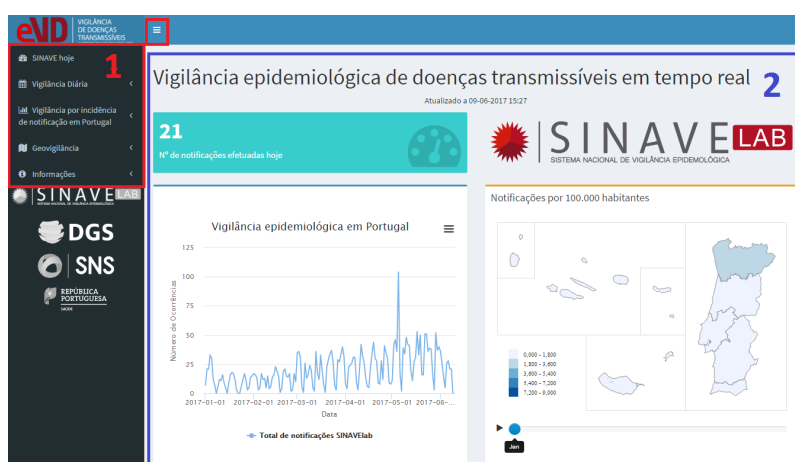


Figura 5.2: Dashboard com menu lateral retrátil e painel central informativo.

No menu lateral é possível navegar entre as seguintes páginas:

- SINAVE hoje
- Gráficos de vigilância diária
 - Vigilância diária por doença
 - Vigilância diária por grupo etário e sexo
- Gráficos de vigilância por incidência de notificações
 - Ranking da incidência
 - Doenças de baixa incidência
- Vigilância por área geográfica
 - Incidência por ARS
 - Incidência por NUTSIII
 - Incidência por distrito
 - Incidência por concelho
- Informações
 - SINAVE
 - Metadados
 - Legislação

5.2.2 Página principal - SINAVE hoje

A página principal, SINAVE hoje, tem como objetivo permitir que o utilizador consiga obter rapidamente o panorama geral da notificação de doenças de declaração obrigatória, sendo por isso apresentado o valor total de notificações no sistema no corrente dia, não estando discriminado por doença (Figura 5.3, número 1). Com isto, o utilizador sabe imediatamente quantos casos de doenças de declaração obrigatória foram notificados no próprio dia. É também apresentado um gráfico de linhas que mostra o número de notificações no sistema por dia de ocorrências, mais uma vez, não se encontra discriminado por doença, apenas são os valores totais (Figura 5.3, número 2). Desta forma, é possível averiguar a evolução do número de notificações por dia, constatando ainda padrões que possam ocorrer, como por exemplo se as notificações seguem algum padrão cíclico semanal ou mensal e permite ainda detetar picos de notificações. Para além disso é ainda apresentado um mapa colorido (*choropleth*) de Portugal dividido por áreas regionais de saúde (ARS), tal como na análise feita no capítulo 2,

é uma boa opção para localizar as incidências, permitindo visualizar as incidência de notificações por 100.000 habitantes por mês, sendo que a escala de cores aplicada tem como correspondência, cores mais escuras, maior a incidência (Figura 5.3, número 3).

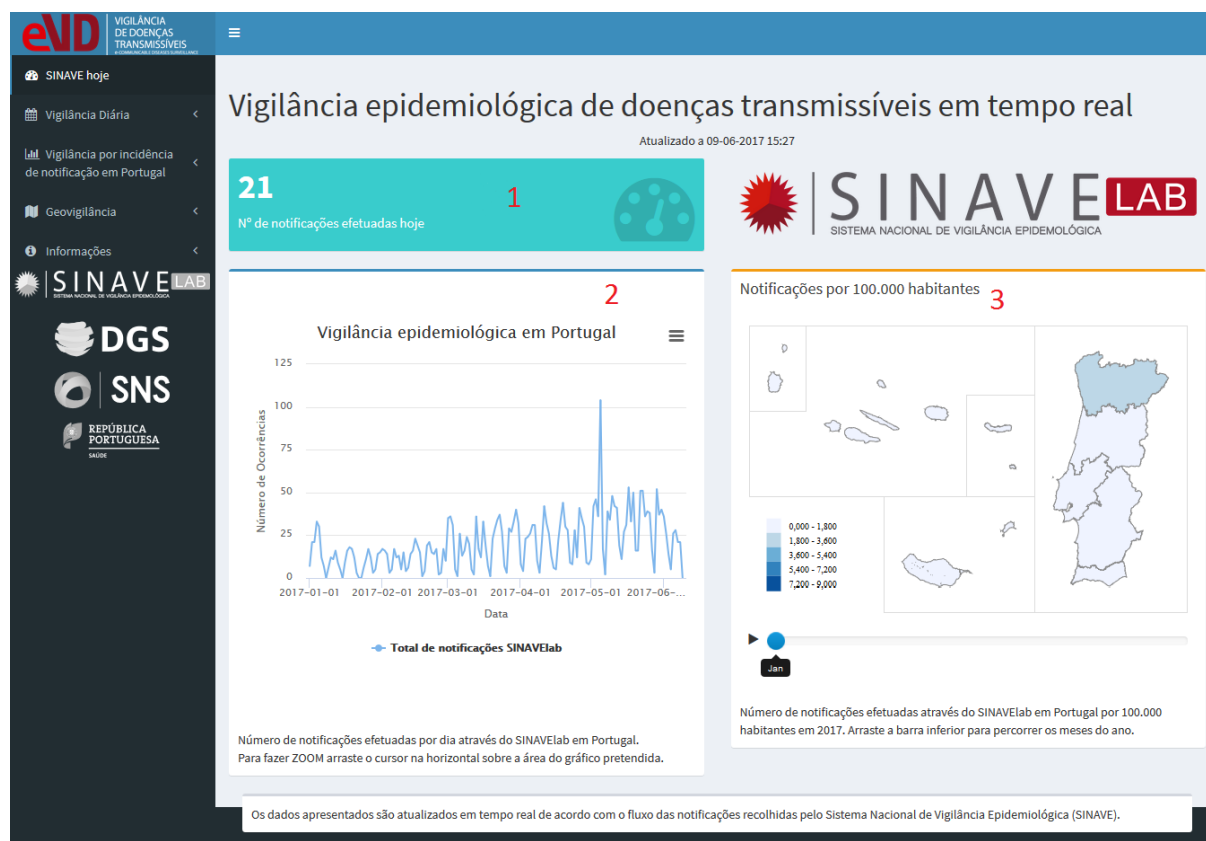


Figura 5.3: SINAVE hoje.

5.2.3 Restantes menus e visualizações

Os restantes menus (5.4) foram criados de forma a separar a informação que não fazia sentido estar agregada na mesma página ou separador. Os critérios utilizados basearam-se no pretendido pela DGS, demonstrados em 3 reuniões que tiveram lugar nas instalações da DGS e que contaram com a presença da Dr^a Cátia de Sousa Pinto, chefe da Divisão de Epidemiologia e Vigilância.

Os principais critérios utilizados são a granularidade de informação e a experiência da DGS na análise destes dados, isto é, permitir que o utilizador, à medida que navegue pelos menus, consiga obter mais informações sobre as notificações do SINAVElab e atentando na experiência que a DGS tem nas análises internas e nas análises que por vezes são requeridas para estudos externos. Por isso, foram criados menus onde apenas é possível visualizar informações mais gerais sobre as notificações, sendo possível navegar nos seus sub-menus para consultar informações mais precisas sobre os dados.

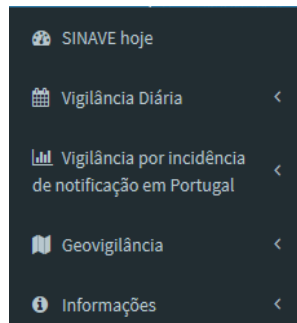


Figura 5.4: Menu.

Por exemplo, no menu "Vigilância Diária" é possível observar a separação "Por doença" e "Por grupo etário". Embora a informação pudesse estar contida num só painel, iria adensar a informação e um utilizador que apenas queira saber o número de notificações de determinada doença mas não queira saber a especificação por grupo etário e sexo, iria encontrar um painel repleto de informação que não pretende, estando assim segmentada para permitir uma maior filtragem do que se pretende visualizar.

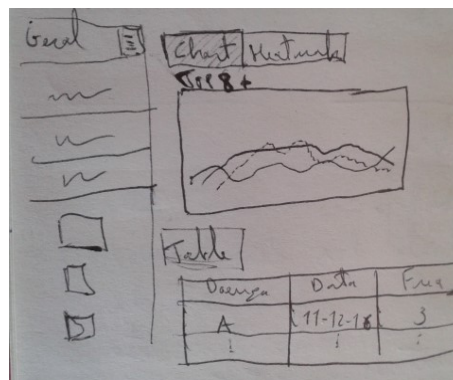


Figura 5.5: Esboço da disposição gráfica.

A disposição gráfica de cada painel de informações baseou-se, uma vez mais, na experiência da DGS na análise e no que habitualmente é pretendido obter. Era pretendido que fosse possível obter informações de forma imediata, no entanto, também teria que ser possível obter informações mais detalhadas e permitir que o utilizador, extraíndo os dados, possa fazer a sua própria análise. Desta forma, foi realizado um esboço da disposição gráfica da informação (figura 5.5), apresentando em forma de gráfico, de linhas, *heatmap* ou apenas numérico, consoante o caso, os dados, e por baixo, em forma de tabela, com filtros em cada coluna, os dados que suportam os gráficos acima, encontrando assim um equilíbrio entre o que se pretende mostrar de forma imediata e o detalhe necessário para uma análise mais precisa. Para que não seja mostrada demasiada informação num só painel, inicialmente as tabelas encontram-se ocultas, sendo possível expandi-las através de um *click* no sinal "+". Como um dos objetivos do sistema era proporcionar ao utilizador a possibilidade de extrair os dados para análise

própria, as tabelas têm a funcionalidade de exportação dos dados num ficheiro em formato CSV, em formato Excel ou em formato PDF. É possível ver a versão final na figura 5.6, sendo este apenas um exemplo mas a abordagem foi aplicada a todos os menus de informação à excepção do menu "Vigilância por incidência de notificação em Portugal", onde no sub-menu "Ranking" só se encontram os dados representado por gráficos de linhas e *heatmaps*, enquanto que no sub-menu de "Doenças de baixa incidência" os dados apenas estão representados sob a forma de uma tabela.



Figura 5.6: Versão final da disposição gráfica do eVD Lab.

5.2.4 Visualização por grupo etário e género

Nos restantes gráficos, o processo começou pela análise das opções utilizadas em sistemas semelhantes (análise referida no capítulo 2) e qual a melhor visualização tendo em conta os dados presentes e que informação se pretende obter. Por exemplo, considerando o objetivo de que neste sistema o utilizador seria capaz de visualizar os dados por grupo etário, atentando à tabela 2.1, a solução utilizada por sistemas semelhantes passa por utilizar listas e gráficos de barras. Isto seriam de facto boas soluções, contudo, uma vez que o sistema também deveria apresentar informações por género, nenhuma dessas soluções seria ideal, pois iria fazer com que houvesse repetição de gráficos, pois se para cada género o objetivo é mostrar também os dados por grupo etário, o mesmo gráfico relativo à informação iria estar representado duas vezes, apenas mudando a variável do género. A solução encontrada baseou-se no gráfico utilizado por Chui et al. [2011], onde mostra várias informações num gráfico multi-painel (figura 2.8). Desta forma, num só gráfico é possível obter informações referentes ao género e ao grupo etário, evitando que existam gráficos repetidos n vezes, consoante o número n de variáveis existentes.

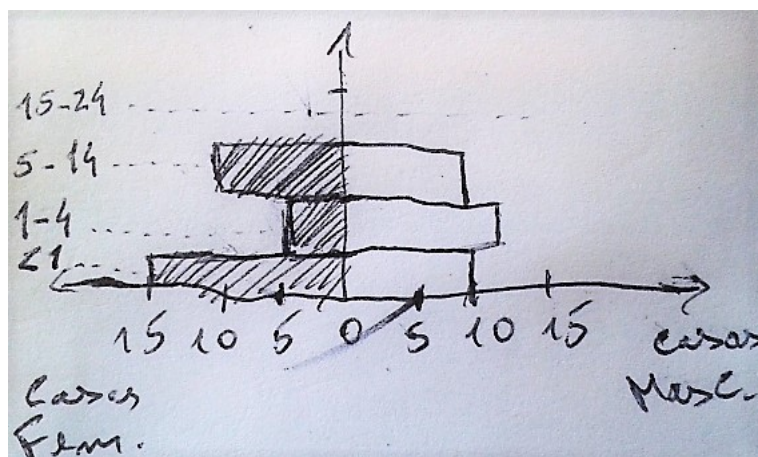


Figura 5.7: Esboço de gráfico para visualizar grupo etário e sexo.

Após a escolha do gráfico escolhido, foi criado um esboço em papel para clarificar o caso de como seria utilizado no contexto (figura 5.7). O eixo horizontal está dividido em dois semi-eixos, sendo ambos positivos e tendo como início o mesmo ponto, ponto 0, interseção com eixo vertical. Os semi-eixos utilizam a mesma unidade de escala, sendo que o semi-eixo do lado esquerdo representa o número de notificações dos casos femininos e o semi-eixo do lado direito o número de notificações dos casos masculinos. No eixo vertical encontram-se marcados os grupos etários, sendo cada unidade um dos grupos etários que se pretende categorizar, estando ordenados por ordem ascendente. Adicionando ao gráfico alguma interação como o aparecimento de uma *tooltip* ao passar o rato por cima de cada uma das barras dos gráfico, mostrando o número de casos notificados nesse grupo etário, quer do sexo feminino quer do sexo masculino, ao mesmo tempo que mostra o total dos casos para que o utilizador não tenha que fazer contas caso queira saber o valor total para aquele grupo etário. Desta forma é possível encontrar num só gráfico informações referentes ao grupo etário e ao sexo das notificações de uma doença. É possível ver a versão final do gráfico na figura 5.8.

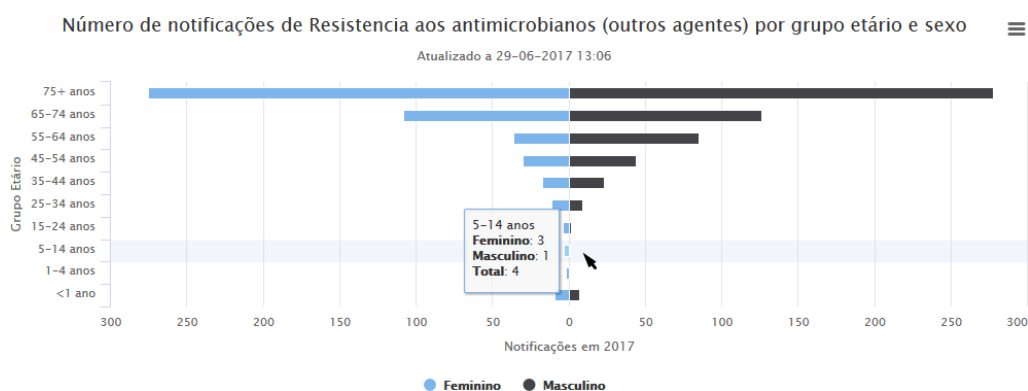


Figura 5.8: Versão final do gráfico para visualizar grupo etário e sexo.

5.2.5 Visualização por doença e dia de notificação

Outro exemplo de gráficos utilizado foi o gráfico de linhas. Este tipo de gráfico é adequado para mostrar tendências ao longo do tempo, como é o caso das notificações de doenças de declaração obrigatória por dia. No painel de cada uma das doenças, do sub-menu "Por doença" do menu "Vigilância diária", encontra-se um gráfico de linhas, mostrando a tendência das notificações por dia, sendo possível ver quando há mais notificações e a evolução da incidência da doença em estudo. Na figura 5.9 é possível ver um exemplo da aplicação do gráfico de linhas. Uma vez que estão representados todos os dias, o eixo é ajustado automaticamente pelo Highcharts de forma a marcar o eixo com datas equidistantes. Esta situação acontece porque ainda só há dados desde 1 de Janeiro de 2017, o número de dias está constantemente a aumentar até que o ano de 2017 termine. Quando o ano de 2018 começar, será acrescentada uma nova linha no gráfico correspondente ao ano, desta forma é possível comparar anos. Nesta situação, o eixo horizontal deixará de ser dinâmico, uma vez que o valor máximo dos dias já foi alcançado. Para permitir que o utilizador consiga ver com mais detalhe quantas notificações houve num determinado dia, tem ao seu dispor *zoom*, carregando e arrastando o cursor na área pretendida. Mais uma vez, um dos objetivos era a exportação dos dados, por isso, clicando no botão que se encontra no canto superior direito do gráfico é possível exportar o gráfico, devidamente identificado com a imagem do SINAVE, numa imagem em formato PNG, JPEG, SVG ou ainda num ficheiro em formato PDF.

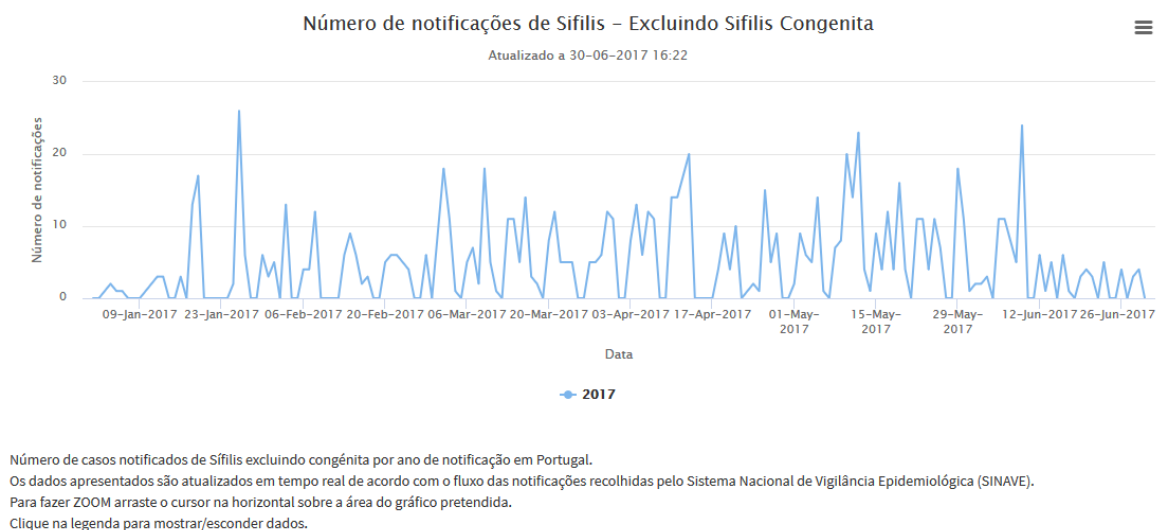


Figura 5.9: Exemplo de gráfico de linhas do eVD Lab.

O gráfico de linhas foi também aplicado para mostrar a evolução das oito doenças com mais notificações. Foi decidido mostrar apenas este número de doenças porque no último relatório público da DGS eram evidenciadas as oito doenças com mais casos. Nesta situação há dois casos diferentes para mostrar, primeiro, um gráfico com as oito doenças com mais notificações no SINAVE à data do

último relatório público da DGS mas com as notificações do ano atual, neste caso, 2017, sendo o gráfico designado por “Gráfico de linhas - TOP8” onde existem oito linhas, uma para cada doença. Contudo, de forma a cumprir os objetivos do sistema, de mostrar a informação em tempo real e o mais atual possível, faz sentido em vez de se mostrar as oito doenças tendo por base o último relatório público da DGS, mostrar as oito doenças com mais notificações desde o início do ano, no preciso momento. Este gráfico está em permanente atualização, sendo que, consoante os dados obtidos em cada atualização da base de dados, as doenças presentes no gráfico podem alterar. Este gráfico designa-se “Gráfico de linhas - TOP8 Atual”(figura 5.10).

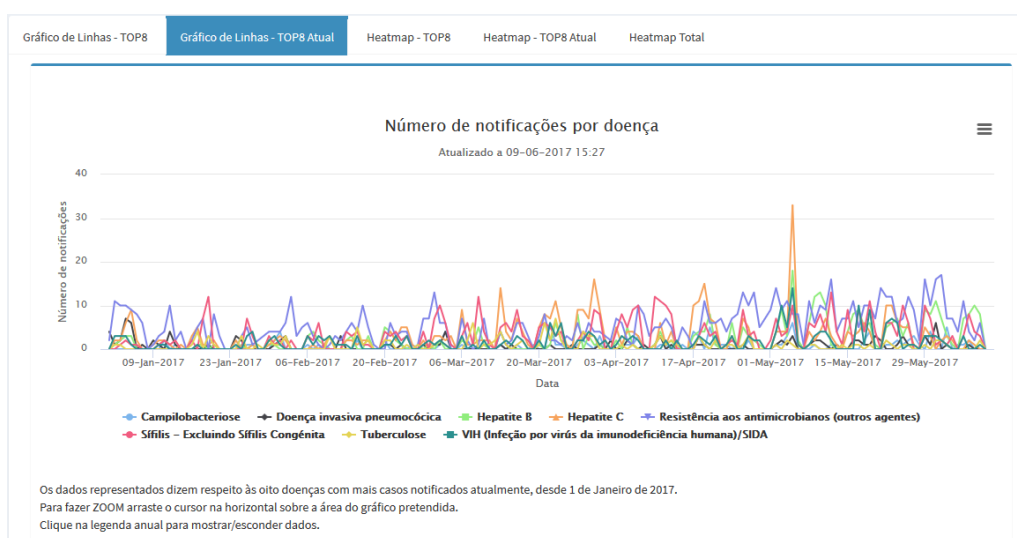


Figura 5.10: Gráfico de linhas referente ao Top 8 atual das doenças com mais notificações no SINAVE.

Contudo, o gráfico de linhas tem a desvantagem de quantas mais linhas estiverem presentes no gráfico, mais confusa e difícil é a sua interpretação. Por isso, é possível selecionar apenas as doenças que se pretender ver, desmarcando as que não interessa, ocultando as linhas correspondentes no gráfico. Outro ponto importante que se deveria ser capaz de obter a partir deste gráfico era saber quando houve mais notificações de uma determinada doença do top 8 atual. Neste tipo de gráficos essa informação não é imediata, poderia ser no caso em que houvesse uma ou duas linhas, mas num gráfico onde estão representadas oito, essa tarefa é difícil. Por esse motivo, foi desenvolvido outro tipo de gráfico, o *heatmap*. O *heatmap* é uma representação a duas dimensões dos dados, nos quais os valores são representados por cores, providenciando uma forma visual imediata de sumarizar a informação nele contido. Por isso, utilizando exatamente os mesmos dados que suportam os gráficos de linhas, foram creados os gráficos “Heatmap - TOP8” e “Heatmap - TOP8 Atual”(figura 5.11). Assim, já é possível saber de forma imediata qual o dia que teve mais casos numa determinada doença, bastando para isso percorrer a linha correspondente à procura da cor mais quente (mais próxima de vermelho). Em ambos os casos, gráfico de linhas e heatmap, passando o cursores por cima dos dados, é exibido

uma *tooltip* com mais detalhes sobre os dados, nomeadamente quantas notificações ocorreram e qual a doença.

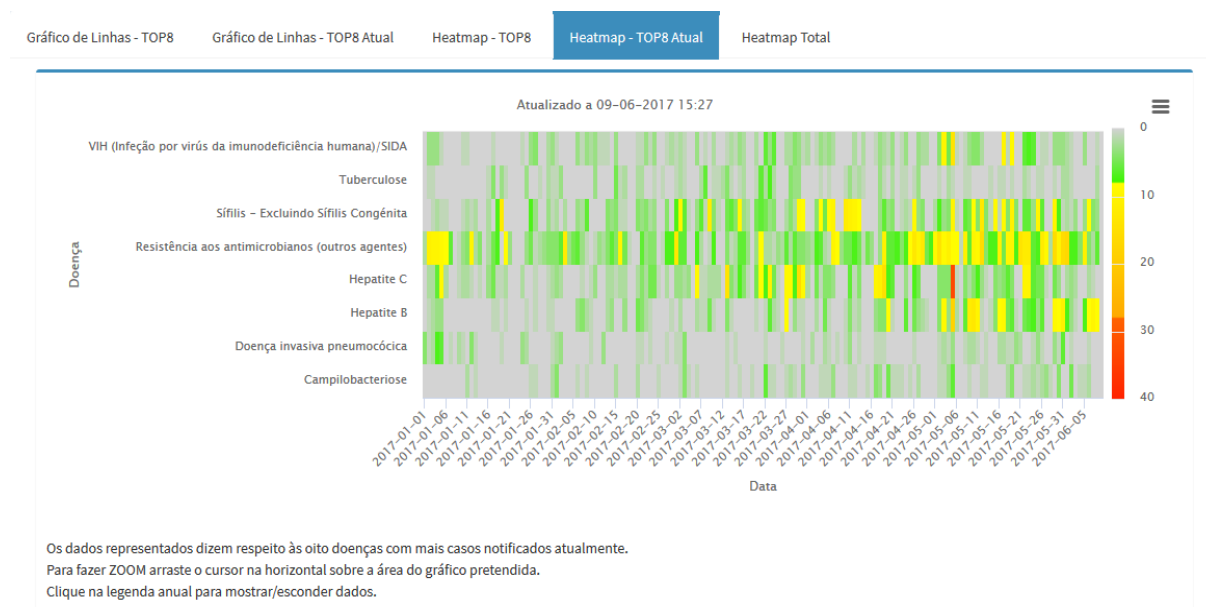


Figura 5.11: *Heatmap* referente ao Top 8 atual das doenças com mais notificações no SINAVE.

5.2.6 Visualização por distribuição geográfica

De forma a visualizar a incidência de notificações laboratoriais de doenças de declaração obrigatória em Portugal, foram utilizados mapas coloridos, tal como analisado na tabela 2.1, esta abordagem é uma das mais utilizadas em sistemas semelhantes. Os mapas não foram construídos de raiz, uma vez que o eVM também tem visualização de dados por área geográfica, e já possuem as divisões territoriais que se pretende. Por isso, apenas houve uma adaptação dos mapas de forma a ser possível representar os dados provenientes do SINAVElab. Ao contrário das outras visualizações, os mapas não foram desenvolvidos com a linguagem R e a biblioteca Highcharts porque não existe a possibilidade de representar mapas através dessa biblioteca, sendo por isso, excecionalmente desenvolvidos em JavaScript e D3. Através dos mapas é possível visualizar a incidência geral das notificações do SINAVElab mensal por ARS, NUTSIII, distrito e concelho, não estando discriminado por doença, isto significa que não é possível visualizar nos mapas onde incide cada doença notificada. No entanto, esses dados podem ser consultados através de tabelas que se encontram no mesmo painel onde constam os mapas, posicionando-se abaixo destes. Nessas tabelas é possível consultar, sobre o mapa correspondente, as notificações por doença, por dia, por mês e por ano. No mapa é ainda possível escolher o mês a visualizar através de uma timeline, arrastando o cursor ao longo da mesma. Caso se pretenda visualizar de forma automática a passagem do meses, também é possível fazê-lo, bastando que se accione o

botão de *play*. O mapa apresenta o número de notificações de doenças de declaração obrigatória em Portugal, por 100.000 habitantes (figura 5.12).

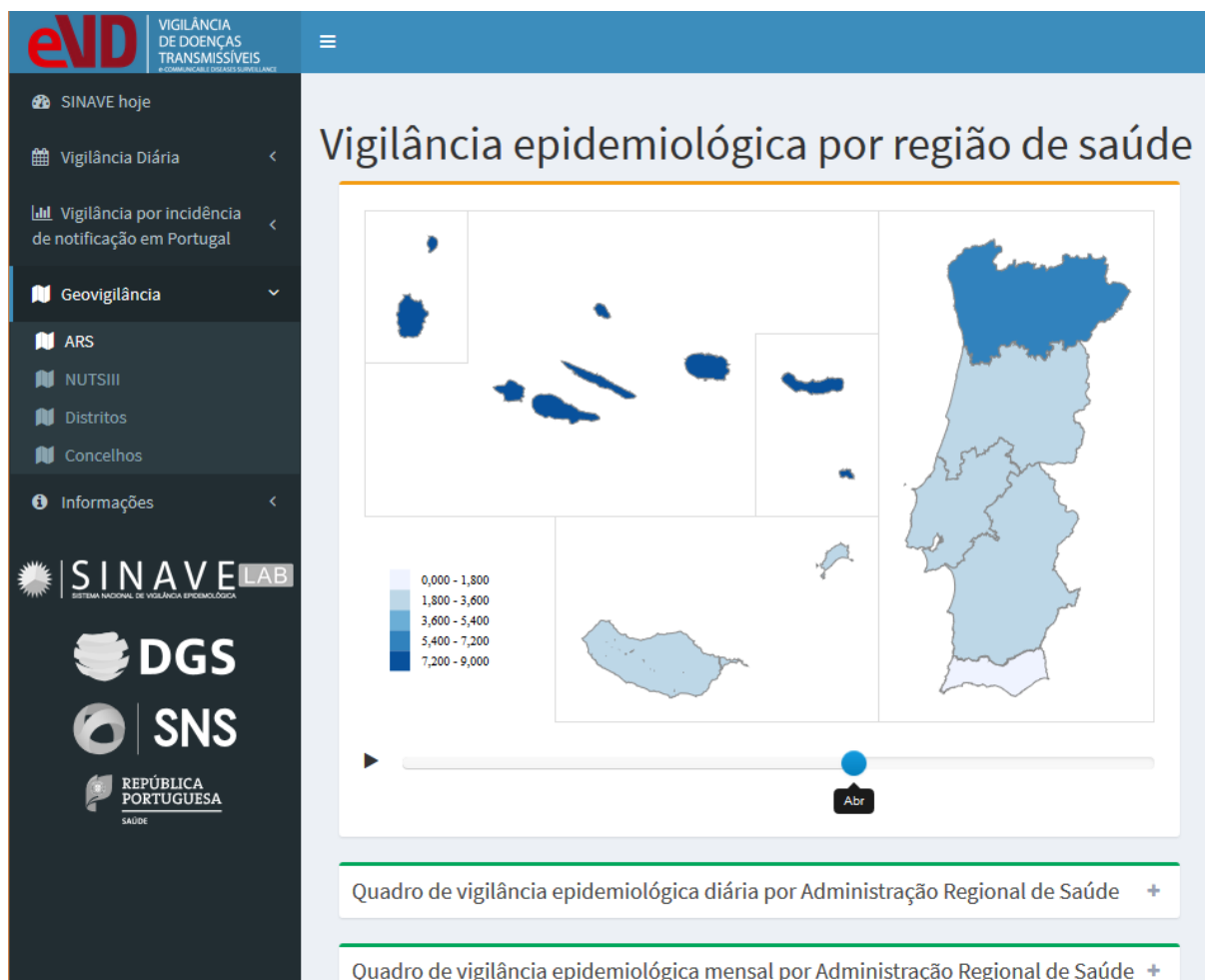


Figura 5.12: Mapa com dados referentes à incidência de notificação de doenças de declaração obrigatória por ARS.

5.3 Demonstração do potencial

De forma a mostrar o potencial do sistema, encontram-se disponíveis algumas figuras representativas do mesmo, tendo sido escolhida como exemplo a doença de sífilis excluindo sífilis congénita. Na figura 5.13 é possível ver os dados referentes às notificações laboratoriais da doença, enquanto que na figura 5.14 é possível ver informações sobre o grupo etário e género dessas mesmas notificações. É também demonstrada na figura 5.15 a distribuição territorial por NUTSIII, averiguando no mês de Abril qual a região com maior incidência de notificações laboratoriais por 100.000 habitantes, confirmando as notificações de sífilis excluindo sífilis congénita nesse mês para essa região.

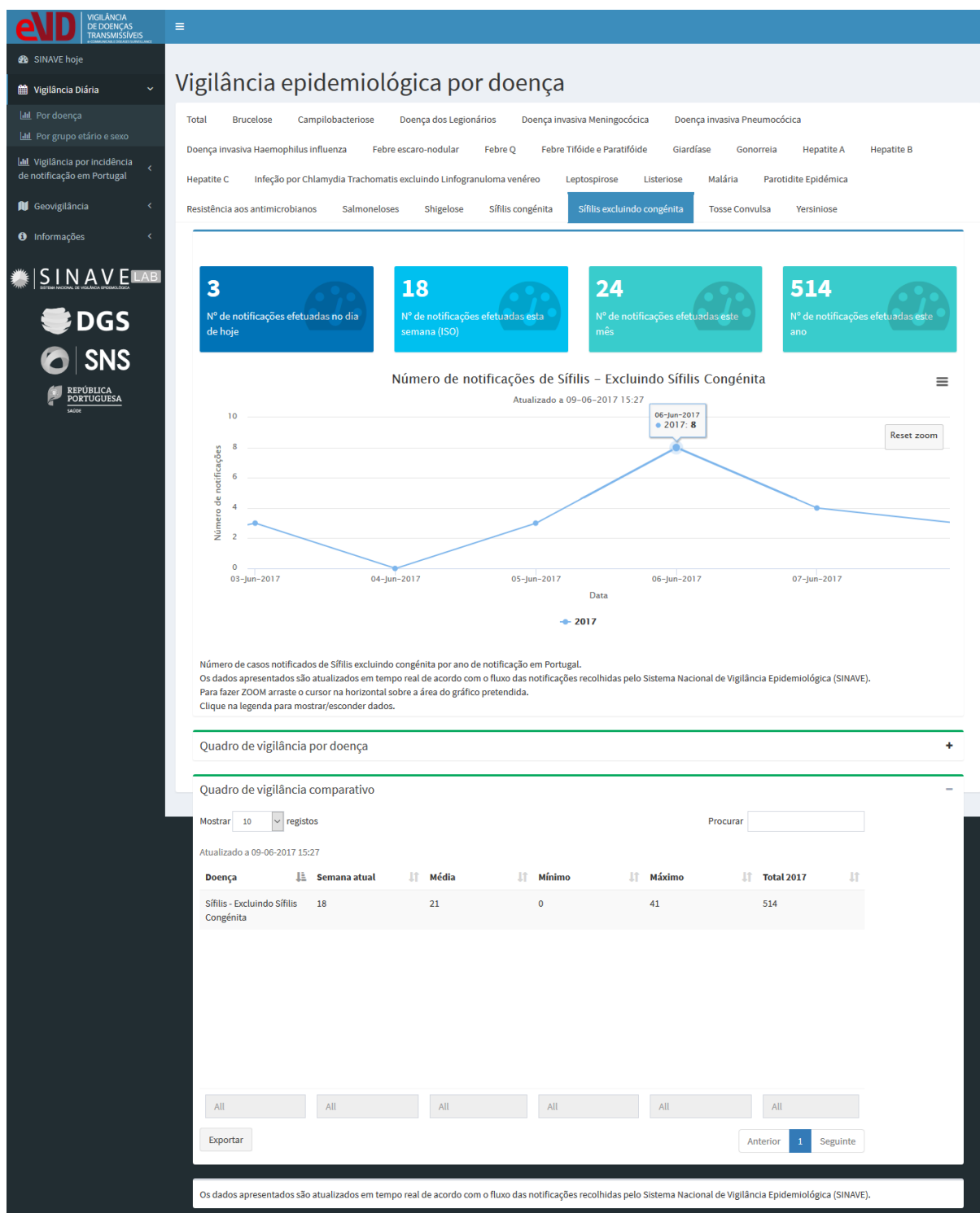


Figura 5.13: Painel de vigilância diária, por doença, estando selecionada a doença de Sífilis excluindo sífilis congénita.

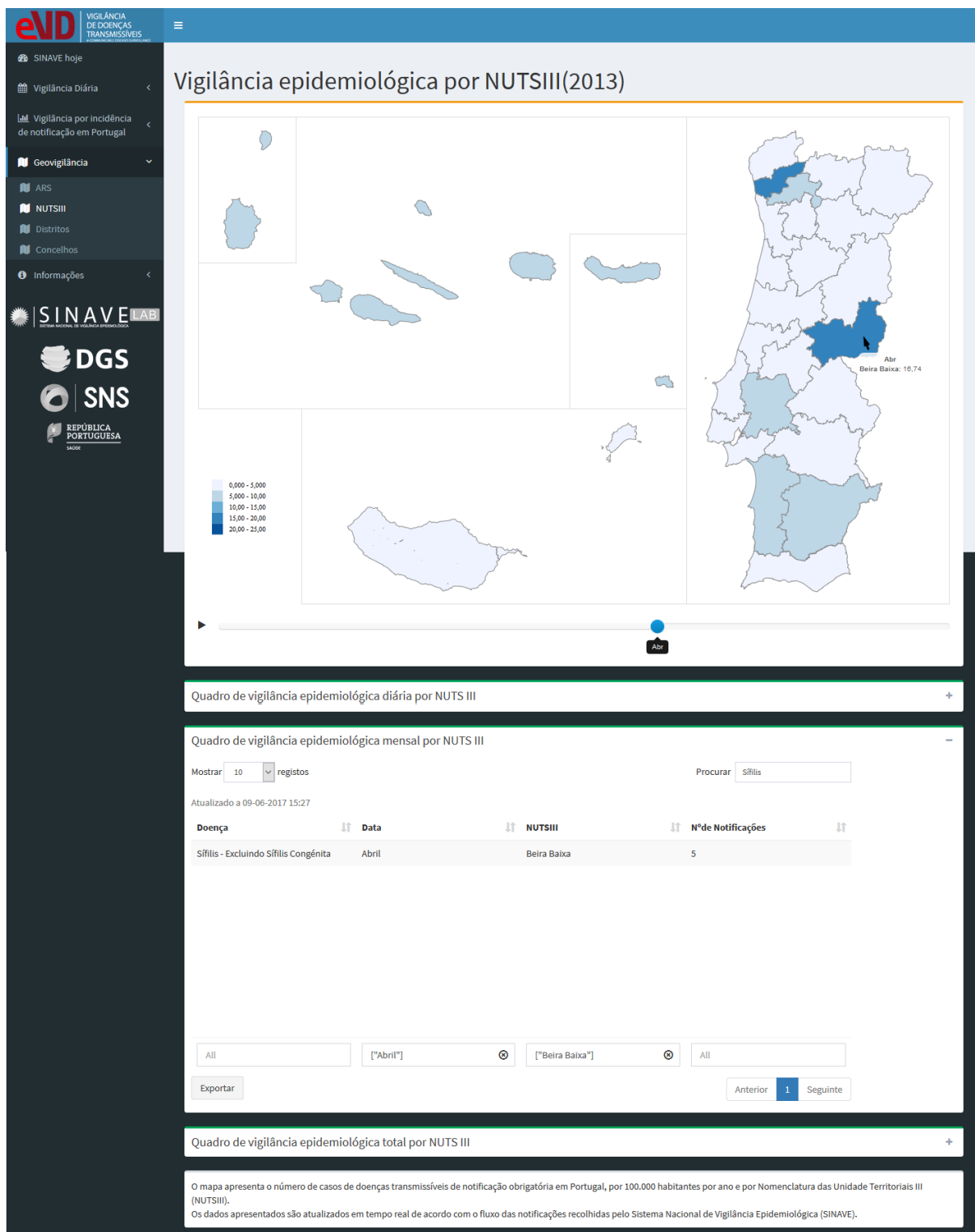


Figura 5.15: Painel de vigilância por divisão territorial, por NUTSIII, no mês de Abril, averiguando na região com maior incidência de notificações por 100.000 habitantes, quantas notificações de sífilis excluindo sífilis congénita ocorreram.

6

Avaliação

Conteúdo

6.1	Características da avaliação	67
6.2	Testes com utilizadores	67
6.3	Cenários de teste com utilizadores da DGS	73
6.4	Discussão	74

6.1 Características da avaliação

A avaliação consistiu em duas componentes: testes com utilizadores e cenários de teste. Os testes com utilizadores tinham como objetivo avaliar o sistema a nível de interatividade e usabilidade, sendo os utilizadores voluntários sem conhecimento do domínio. Nos cenários de teste o objetivo era a avaliação por parte de potenciais utilizadores da DGS, simulando tarefas que fazem atualmente sem o sistema, percebendo se estava funcional, intuitivo e com boa usabilidade.

6.2 Testes com utilizadores

De forma a se perceber se o sistema permite recolher informações precisas sobre o estado atual da vigilância epidemiológica em Portugal, como o número de casos de cada doença de declaração obrigatória, grupo etário, sexo e região dos pacientes e se o mesmo é de fácil utilização, perceptível, atrativo e de agrado do utilizador, foram realizados testes com utilizadores. Sendo o público alvo do sistema, o público em geral, o sistema foi testado através de voluntários.

O teste com utilizadores foram realizados com 20 voluntários entre os dias 12 e 26 de Junho, tendo lugar na sala técnica do departamento de engenharia informática 0.07, do Pavilhão de Informática III, no campus da Alameda do Instituto Superior Técnico e incidia principalmente em cinco tarefas. As tarefas exercitavam as várias partes do sistema, abrangendo todas as funcionalidades requeridas aquando da definição de requisitos a cumprir e não era necessário que o utilizador levasse qualquer material para a sessão de testes.

As cinco tarefas, que foram realizadas pelos utilizadores por ordem aleatória, são:

1. Saber o número de notificações totais do SINAVElab no dia de hoje.
2. Saber qual a faixa etária com maior incidência de Hepatite B.
3. Saber qual a doença com mais notificações no SINAVElab no dia 5 de Maio de 2017.
4. Saber qual a ARS com maior incidência de doenças de declaração obrigatória no mês de Maio e qual a doença com mais notificações nesse mês para essa ARS.
5. Saber se nesta semana, a doença com mais notificações da NUTSIII com maior incidência de doenças de declaração obrigatória no mês de Fevereiro, está acima ou abaixo da média das semanas do ano.

A sessão de testes dividiu-se em quatro fases:

1. A primeira, com duração de cinco minutos, consistiu numa breve explicação do sistema a testar e âmbito do teste. Foi também lembrado ao utilizador que iria testar o sistema, que não era o

utilizador que iria ser testado mas sim o sistema, não tendo que se sentir constrangido nem ter receios de errar.

2. De seguida foi dada liberdade total ao utilizador para explorar e usar livremente o sistema, tendo como duração cinco minutos.
3. Após um primeiro contato com o sistema, o utilizador realizou o conjunto das cinco tarefas por ordem aleatória, assegurando assim a independência intra-tarefas. Esta fase teve o tempo necessário para que o utilizador pudesse concluir as cinco tarefas, sendo que inicialmente estava previsto que esta fase tivesse como duração máximo entre os 10 e os 15 minutos, facto que se veio a comprovar.
4. Concluída a realização das tarefas, foi pedido ao utilizador que preenchesse um questionário de satisfação para aferir alguns aspetos sobre o sistema avaliado. Esta fase teve a duração de cinco minutos.

A fase de realização dos testes consistia na atribuição aleatória das tarefas a realizar, solicitando ao utilizador qual a ordem pela qual desejava realizar as tarefas. Antes de iniciar qualquer uma das tarefas, o ponto inicial era a página principal do sistema, menu "SINAVE hoje". Cada tarefa era considerada como tendo um fim correto quando o utilizador respondia acertadamente à questão implícita na tarefa, não possuindo qualquer ajuda na sua realização, exceto se o utilizador "bloquear", permanecendo muito tempo parado.

6.2.1 Os utilizadores

Os 20 utilizadores que testaram o sistema forma voluntários, estando distribuídos por dois grupos etários: 90% dos utilizadores entre os 18 e 30 anos e 10% dos utilizadores entre os 41 e 60 anos, sendo na sua maioria do sexo masculino, 70% dos utilizadores. Em relação ao grau de instrução completo (figura 6.1), os utilizadores estão distribuídos principalmente por Ensino Secundário, 50% dos utilizadores e Ensino Secundário - Licenciatura, 45% dos utilizadores.

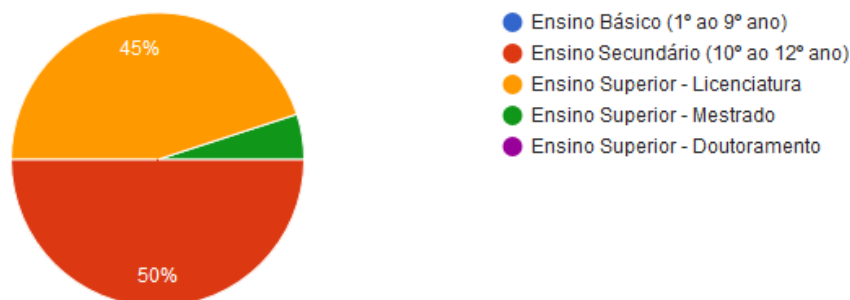


Figura 6.1: Grau de instrução completo dos 20 utilizadores.

6.2.2 Dados recolhidos

Durante a realização dos testes por parte do utilizador, foram recolhidos dados como o tempo de realização de cada tarefa assim como o número de erros, sendo considerado um erro quando o utilizador percorria uma tabela, gráfico ou menu desnecessário para a concretização da tarefa, estando os dados recolhidos representados na tabela 6.1.

Utilizador	Tarefas									
	1		2		3		4		5	
1	11	0	45	0	45	2	53	0	300	4
2	24	1	103	1	55	3	92	1	355	4
3	19	0	195	2	31	0	74	1	621	7
4	1	0	163	0	35	0	44	0	344	2
5	1	0	26	0	112	3	93	0	363	4
6	2	0	30	0	64	1	51	0	222	2
7	62	1	35	1	18	0	78	0	77	1
8	25	0	60	0	25	0	51	0	100	0
9	23	0	56	0	53	0	55	0	104	0
10	1	0	23	0	90	1	60	0	134	0
11	1	0	31	0	39	0	64	0	184	0
12	1	0	53	0	26	0	88	0	245	0
13	1	0	25	0	19	0	53	0	64	0
14	2	0	41	0	36	0	58	0	97	0
15	7	0	31	0	62	0	64	0	71	0
16	7	0	48	0	71	0	69	0	74	0
17	22	1	73	0	47	0	102	0	147	0
18	1	0	22	0	54	0	55	0	132	0
19	7	0	50	1	98	0	112	0	156	0
20	1	0	19	0	36	0	62	0	82	0
Média	10,95	0,15	56,45	0,25	50,8	0,50	68,9	0,10	193,6	1,20

Tabela 6.1: Tempos, em segundos, e número de erros cometidos por utilizador em cada uma das tarefas realizadas na sessão de testes.

Observando a tabela 6.1 e as figuras 6.2 e 6.3, constata-se que o tempo médio para a realização das tarefas aumenta da primeira para a última, facto expectável, uma vez que a complexidade das mesmas também é crescente, sendo a quinta tarefa a mais complexa. Na primeira tarefa, a que tinha um menor grau de complexidade teve como tempo médio 10,95 segundos. Para a realização desta tarefa era apenas necessário observar o painel informativo da página inicial, encontrando-se a resposta sem qualquer interação sendo visível num painel com o valor da resposta. Nesse sentido, o tempo médio está dentro do esperado, pois os utilizadores apenas tinham que procurar o local da informação. Atendendo ao número de erros cometidos, média de 0,15, também está dentro do esperado dada a simplicidade da tarefa. Os erros cometidos consistiram em encontrar exatamente a mesma resposta mas noutro menu menos imediato. As tarefas 2, 3 e 4, têm complexidade semelhantes, exigindo que o utilizador navegue pelos menus do sistema. Os tempos médios são parecidos, 56,45; 50,8 e 68,9

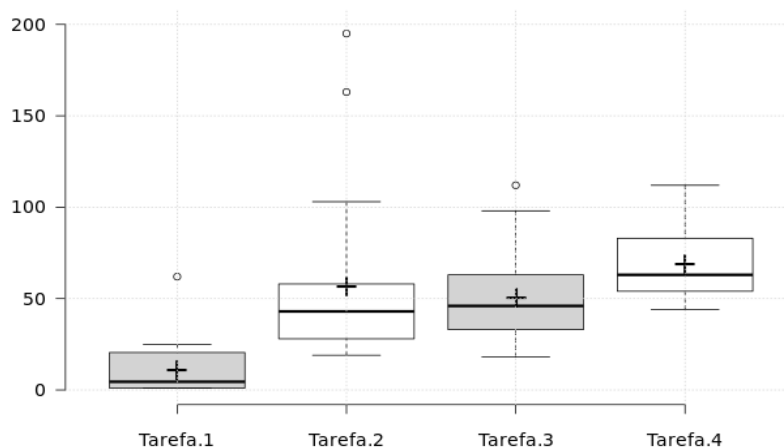


Figura 6.2: Boxplot dos tempos de realização das tarefas 1, 2, 3 e 4.

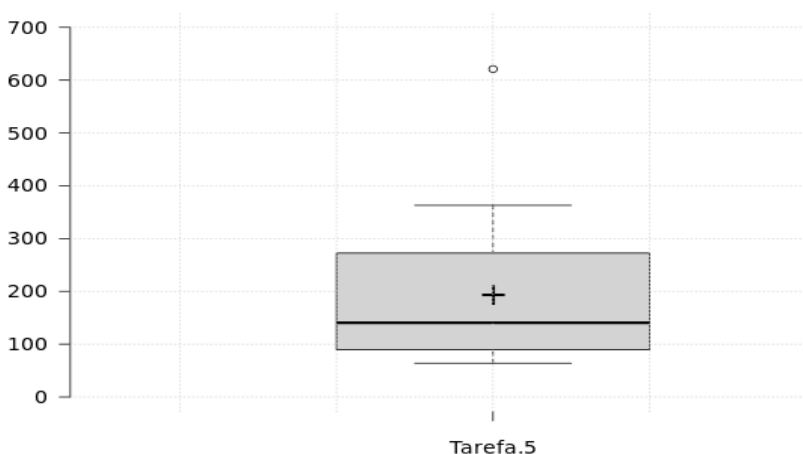


Figura 6.3: Boxplot dos tempos de realização da tarefa 5.

segundos para as tarefas 2,3 e 4 respetivamente. Estes tempos podem ser considerados como bons uma vez que requeriam navegação entre menus, e interpretação de gráfico de barras com necessidade de interação para se obter na *tooltip* a resposta na tarefa 2, assim como interpretação de um mapa e consequente leitura numa tabela filtrando e comparando os valores nela apresentados nas tarefas 3 e 4. O número de erros também é positivo pois em nenhuma das tarefas o número médio é superior a 0,50, e consistiram na leitura errada da faixa etária na *tooltip* do gráfico de barras na tarefa 2 e na análise de tabela diárias em vez de tabelas mensais nas tarefas 3 e 4. A última tarefa requeria a navegação num primeiro menu com interpretação de um mapa e consequente leitura, filtragem e comparação de valores na tabela mensal para se obter a primeira parte da resposta. Obtendo essa resposta, a doença com maior incidência no mês de Fevereiro na NUTSIII com maior número de notificações, requeria ainda a navegação noutra menu e análise atenta de uma tabela com valores comparativos. Esta tarefa demorou, em média, 193,6 segundos, o que equivale a 3 minutos e 14 segundos. Este valor

é positivo pois a tarefa requer interação com o sistema e a resposta está segmentada em duas fases, relembrando o pouco tempo de utilização prévia do sistema por parte dos utilizadores que o testaram, sendo esperado uma melhoria à medida que a utilização seja mais frequente. O número médio de erros foi de 1,20, o que é positivo pois envolvia a navegação entre vários menus e consulta de mapa e tabelas, consistindo na sua maioria em leitura de tabelas diárias em vez de mensais. Atentando aos valores globais, pode-se considerar que o sistema teve um bom resultado pois o número de erros é baixo, indicando que o sistema é intuitivo. E ainda, em menos de cinco minutos é possível obter informações, sem a necessidade de obter qualquer autorização por parte da DGS, que atualmente pode levar dias.

6.2.3 Questionário de satisfação

O questionário de satisfação consistia num questionário de 15 perguntas de escolha múltipla, sendo as respostas possíveis dadas numa escala de 1 a 5, em que 1 corresponde a "Discordo Muito" e 5 corresponde a "Concordo Muito", e ainda 1 pergunta de resposta aberta. Das 15 perguntas de escolha múltipla, 10 dizem respeito à avaliação por parte do utilizador das afirmações da escala de usabilidade de sistemas (do inglês, SUS) e que são:

1. Gostaria de utilizar este sistema frequentemente.
2. Achei que o sistema era desnecessariamente complexo.
3. Achei o sistema fácil de usar.
4. Penso que iria precisar de apoio técnico para usar o sistema.
5. Acho que as várias funções do sistema estão muito bem integradas.
6. Penso que havia demasiadas inconsistências no sistema.
7. Imagino que a maioria das pessoas aprenda rapidamente a usar o sistema.
8. Achei o sistema muito complicado de usar.
9. Senti-me muito confiante a usar o sistema.
10. Preciso de aprender antes de poder usar este sistema.

A escala de usabilidade de sistemas (SUS) é um questionário de 10 perguntas de escolha múltipla, com respostas entre 1 e 5, sendo 1, "Discordo Muito" e 5, "Concordo Muito", que permite avaliar de forma rápida e de baixo custo a usabilidade de sistemas (<http://measuringu.com/sus>).

As respostas são posteriormente convertidas numa pontuação de SUS, sendo que para isso é aplicado o critério:

- Às perguntas pares (2,4,6,8 e 10), subtrai-se 1 à resposta do utilizador.
- Às perguntas ímpares (1,3,5,7 e 9), subtrai-se a resposta do utilizador a 5.
- Soma-se o total das respostas do utilizador já convertidas e multiplica-se por 2,5. Desta forma, a pontuação SUS por utilizador varia entre 0 e 100.
- Por último, calcula-se a pontuação SUS média.

Através da pontuação média SUS é possível comparar com a pontuação que se considera padrão de um bom sistema em termos de usabilidade, sendo essa pontuação de 68 pontos. Sistemas que tenham pontuação de SUS média abaixo de 68 pontos consideram-se sistemas que têm de ser fortemente melhorados a nível de usabilidade, enquanto que acima de 68 pontos consideram-se sistemas que têm uma usabilidade acima da média. A pontuação padrão de referência pode ser mais precisa, considerando-se que sistemas com pontuações médias acima de 74 têm usabilidade muito boas e são do agrado do utilizador. Pontuações médias acima de 80.3 são considerados sistemas de ótima usabilidade, sendo totalmente do agrado do utilizador, recomendando o sistema (<http://measuringu.com/sus>). A avaliação por parte dos utilizadores já convertida em pontuação SUS, encontra-se na tabela 6.2.

Utilizador	Perguntas										Total	Total x 2.5
	1	2	3	4	5	6	7	8	9	10		
1	3	4	4	4	4	4	4	4	3	2	36	90
2	1	3	3	4	3	3	3	2	2	0	24	60
3	3	2	2	1	3	3	3	3	2	0	22	55
4	1	3	3	4	3	4	3	4	3	4	32	80
5	2	4	3	4	3	4	4	4	4	3	35	87,5
6	1	3	1	1	3	2	3	3	2	1	20	50
7	3	4	4	4	4	4	4	4	4	3	38	95
8	4	4	3	4	4	4	1	3	2	3	32	80
9	3	4	3	4	3	3	3	3	3	3	32	80
10	3	3	3	2	3	4	3	4	3	1	29	72,5
11	4	4	4	3	4	4	4	3	4	4	38	95
12	3	3	3	2	4	3	2	2	2	2	26	65
13	3	4	4	4	4	3	3	4	3	1	33	82,5
14	3	3	3	4	4	4	2	3	3	4	33	82,5
15	4	1	3	1	3	3	3	3	0	2	23	57,5
16	4	4	4	4	4	4	4	4	4	4	40	100
17	3	3	3	4	4	4	3	4	3	3	34	85
18	3	3	3	3	3	4	3	3	3	4	32	80
19	2	4	3	1	3	2	3	4	3	1	26	65
20	3	3	4	3	4	3	4	4	4	2	34	85
Média	2,8	3,3	3,15	3,05	3,5	3,45	3,1	3,4	2,85	2,35	30,95	77,375

Tabela 6.2: Pontuação SUS das 10 perguntas.

A média da pontuação SUS é de 77.375 pontos, o que significa que está acima da média dos 68 e consequentemente se pode considerar que o sistema tem uma boa usabilidade. Quando comparado com a pontuação padrão de referência mais precisa, fica aquém dos 80.3, valor a partir do qual se considera uma usabilidade excelente, contudo é um valor alto, permitindo aferir que são necessárias poucas alterações para tornar a usabilidade do sistema excelente. Nas questões 1, 9 e 10, a pontuação média é inferior a 3. Se considerarmos o teor dessas questões, constatamos que podem estar interligadas uma vez que abordam a confiança (questão 9), a necessidade de aprendizagem (questão 10) e a frequência na utilização do sistema (questão 1). Uma possível conclusão que se pode retirar é que o facto dos utilizadores acharem que não têm conhecimento suficiente do domínio do sistema faz com que se sintam menos confiantes na sua utilização e sintam necessidade de aprender antes de o utilizarem. Para melhorar estes aspetos deve-se tornar o sistema mais intuitivo e apelativo durante a descrição de certos termos técnicos para que o utilizador se sinta mais confiante e não sinta necessidade de aprender a utilizá-lo, proporcionando assim uma maior frequência de utilização.

6.3 Cenários de teste com utilizadores da DGS

Para além dos testes com utilizadores voluntários, o sistema foi também submetido a uma avaliação por parte de duas técnicas superiores da DGS que irão utilizar o sistema. Esse teste consistia na navegação simulando tarefas que virão a realizar na utilização do sistema. O processo de desenvolvimento da tarefa consistia na realização da tarefa por parte do utilizador, seguindo a abordagem de dizer em voz alta o que está a pensar e o que pretende fazer, comentando as interações que ia tendo com o sistema.

Uma das tarefas foi a identificação do número de casos de Sarampo ocorridos em Portugal em 2017. Partido do menu principal, "SINAVE hoje", ambos os utilizadores acharam pouco intuitivo a divisão das doenças, isto é, as doenças com baixa incidência, doenças com menos de 30 casos nos últimos 4 anos, estão separadas das restantes doenças de declaração obrigatória, mas um utilizador com pouco conhecimento do domínio não sabe a distinção de quais as doenças, navegando por isso para o menu das doenças com mais notificações que se encontra no menu "Vigilância Diária". Para tornar esta divisão mais clara, foi sugerido pelos utilizadores que se acrescentasse uma legenda a explicar a divisão das doenças no menu das doenças com mais notificações. Esta divisão existe devido à diferença de representação dos dados, visto que as doenças com baixa incidência têm poucos casos, pelo que não seria adequado utilizar a mesma abordagem das restantes doenças, que à priori, têm muitos mais casos. Também comentaram que, apesar do ponto referido anteriormente, o sistema estava bem concretizado.

A outra tarefa que os utilizadores escolheram realizar foi a identificação de grupos etários e sexo

de várias doenças. Nesta tarefa os comentários foram todos positivos, sendo que a navegação estava intuitiva e o gráfico apresentado era útil e bastante perceptível.

Na restante navegação é de salientar que foi referido o facto dos mapas acrescentarem pouca informação, uma vez que era mais útil para eles um mapa onde fosse possível ver a distribuição geográfica por doença ao invés de ver a incidência sem discriminação de doenças. No entanto, este caso não figurava nos requisitos inicialmente definidos, pelo que não foi alvo de desenvolvimento, aliando ao facto de mais uma vez, não se querer mostrar ao público em geral todos os detalhes possíveis para não alarmar a população.

6.4 Discussão

Analisando os resultados dos testes e a lista de requisitos, pode-se considerar que o sistema foi bem conseguido e que os objetivos foram alcançados. A sessão de testes com utilizadores mostra que o sistema permite uma rápida percepção dos dados sobre a vigilância epidemiológica em Portugal e que a interação é intuitiva, uma vez que nos testes realizados o número de erros é baixo, não esquecendo do facto que o utilizador apenas interagiu com o sistema na própria sessão, o que perspectiva que ao voltar a utilizar o sistema terá um melhor desempenho, e ainda o tempo médio de realização ser adequado à complexidade das tarefas exigidas. É também possível concluir que o sistema tem uma boa usabilidade, alcançando de média 77.375 pontos na SUS, ficando acima da média dos 68 pontos de referência e ficando pouco abaixo da média dos 80.3 pontos a partir dos quais os sistemas são considerados como tendo excelente usabilidade.

Através dos testes com utilizadores da DGS constata-se que a análise realizada de forma automática e presente no sistema, está funcional e abrange todos os pontos solicitados, pelo que cumprem-se os requisitos inicialmente propostos, mostrando-se ainda agradados com o sistema desenvolvido.

7

Conclusão

Conteúdo

7.1 Contribuição	77
7.2 Trabalho futuro	77

O sistema desenvolvido tinha como objetivo principal fornecer um meio de vigilância de doenças de declaração obrigatória em Portugal, em tempo real. Esse objetivo foi alcançado, sendo agora possível utilizar o eVD Lab para obter informações em tempo real sobre o estado atual das notificações laboratoriais de doenças de declaração obrigatória em Portugal.

Em termos dos requisitos, foram todos alcançados, uma vez que é possível visualizar em tempo real informações sobre a incidência de doenças de declaração obrigatória por diversos factores como geográficos, grupos etários e género. O sistema também permite ver tendências e evoluções ao longo do tempo. É também possível exportar os dados para análise na DGS e ou por entidades externas. Para além dos requisitos e objetivos propostos, foram ainda acrescentadas as seguintes funcionalidades:

- Saber as 8 doenças de declaração obrigatória com mais notificações desde o início do ano em tempo real.
- Saber o número de notificações laboratoriais de cada doença, exceptuando as de baixa incidência, discriminado por dia, semana, mês e ano atual.
- Comparar o número de notificações laboratoriais, discriminado por doença, da semana atual com a média semanal.

Para além dos requisitos terem sido cumpridos, através da avaliação com utilizadores constata-se que o sistema tem uma boa usabilidade, assegurando uma boa interação dos utilizadores com o mesmo.

7.1 Contribuição

Com o eVD Lab é agora possível visualizar diversas informações em tempo real sobre a notificação de doenças de declaração obrigatória em Portugal, tarefa que até à existência deste sistema, não era possível. O desenvolvimento deste sistema inovador contribui para a modernização dos meios de controlo e prevenção de doenças transmissíveis, objetivos a curto prazo da CDC e da ECDC, colocando Portugal como um dos pioneiros na disponibilização pública de um sistema de visualização de dados sobre doenças de notificação obrigatória.

7.2 Trabalho futuro

O sistema desenvolvido, o eVD Lab, é apenas a primeira versão, pelo que pode ser melhorado em vários aspetos futuramente. Ao se adquirir novos conhecimentos sobre a prevenção de doenças, o eVD Lab, pode e deve acompanhar essa evolução, adicionando novas análises aos dados, assegurando que o sistema se mantém atualizado, preciso e objetivo na análise feita.

Em novas versões do eVD Lab, podem também ser adicionadas novas visualizações de forma a assegurar que o sistema possibilite recolher novas informações necessárias e mais detalhadas do que as atuais.

A versão atual foi desenvolvida a pensar no público em geral, sendo que não é necessário grande conhecimento do domínio para se perceber os dados apresentados, contudo, na minha opinião, seria bastante útil para a DGS possuir uma versão privada do eVD Lab. Isto significa que poderia ser considerada a hipótese de haver um sistema de autenticação que permita ter uma apresentação dos dados privada para que a DGS possa obter análises mais aprofundadas sobre os dados do SINAVElab, analisando-os com mais detalhe e atentando a mais atributos das notificações realizadas pelas clínicas laboratoriais. Nessa versão privada, que seria acedida através de uma autenticação, poderia possuir também visualizações mais precisas e detalhadas que não podem ser divulgadas publicamente, como a distribuição geográfica a nível de freguesias, e ainda essa mesma distribuição geográfica mas discriminada por doença.

Outro ponto que pode ser melhorado em futuras versões do eVD Lab é a tecnologia utilizada, principalmente na camada de visualização. Esta camada ao ter sido desenvolvida em R, apesar de corresponder ao pretendido, limitou o potencial de visualizações utilizadas. Poderia ser considerada a hipótese de melhorar essa camada recorrendo a JavaScript pois permite uma maior liberdade na criação de visualizações, possibilitando uma maior interatividade na própria visualização e entre visualizações, algo que com o R e a biblioteca de Highcharts ficou inexistente.

Bibliografia

- Alonso, W. J. and McCormick, B. J. J. (2012). EPIPOL: a user-friendly analytical tool for the extraction and visualization of temporal parameters from epidemiological time series. *BMC public health*, 12(1):982.
- Avruskin, G. a., Jacquez, G. M., Meliker, J. R., Slotnick, M. J., Kaufmann, A. M., and Nriagu, J. O. (2004). Visualization and exploratory analysis of epidemiologic data using a novel space time information system. *International journal of health geographics*, 3(1):26.
- Blevins, M., Wehbe, F. H., Rebeiro, P. F., Caro-Vega, Y., McGowan, C. C., and Shepherd, B. E. (2016). Interactive data visualization for HIV cohorts: Leveraging data exchange standards to share and reuse research tools. *PLoS ONE*, 11(3):1–10.
- Carroll, L. N., Au, A. P., Detwiler, L. T., chieh Fu, T., Painter, I. S., and Abernethy, N. F. (2014). Visualization and analytics tools for infectious disease epidemiology: A systematic review. *Journal of Biomedical Informatics*, 51:287–298.
- Chui, K. K. H., Wenger, J. B., Cohen, S. A., and Naumova, E. N. (2011). Visual analytics for epidemiologists: Understanding the interactions between age, time, and disease with multi-panel graphs. *PLoS ONE*, 6(2).
- Deodhar, S., Chen, J., Wilson, M., Bisset, K., Barrett, C., and Marathe, M. (2015). EpiCaster: An Integrated Web Application for Situation Assessment and Forecasting of Global Epidemics. *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 156–165.
- Dunne, C., Muller, M., Perra, N., and Martino, M. (2015). VoroGraph: Visualization Tools for Epidemic Analysis. *Extended Abstracts of the ACM CHI'15 Conference on Human Factors in Computing Systems*, 2:255–258.
- Edberg, S. C. (2005). Global Infectious Diseases and Epidemiology Network (GIDEON): a world wide Web-based program for diagnosis and informatics in infectious diseases. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 40(1):123–6.

- Gesteland, P. H., Livnat, Y., Galli, N., Samore, M. H., and Gundlapalli, A. V. (2012). The EpiCanvas infectious disease weather map: an interactive visual exploration of temporal and spatial correlations. *Journal of the American Medical Informatics Association : JAMIA*, 19(6):954–9.
- Karlsson, D., Ekberg, J., Spreco, A., Eriksson, H., and Timpka, T. (2013). Visualization of infectious disease outbreaks in routine practice. In *Studies in Health Technology and Informatics*, volume 192, pages 697–701.
- Livnat, Y., Rhyne, T. M., and Samore, M. (2012). Epinome: A visual-analytics workbench for epidemiology data. *IEEE Computer Graphics and Applications*, 32(2):89–95.
- Maciejewski, R., Livengood, P., Rudolph, S., Collins, T. F., Ebert, D. S., Brigantic, R. T., Corley, C. D., Muller, G. A., and Sanders, S. W. (2011). A pandemic influenza modeling and visualization tool. *Journal of Visual Languages and Computing*, 22(4):268–278.
- Nguyen, H. T., Tran, P. V., Ngo, H. T., Tran, T. T., Tuan, L., and Le, D. (2013). Visualization of epidemic data. In *Proceedings of 5th International Conference on healthGIS*, page 6.
- Plaza-Rodríguez, C., Appel, B., Kaesbohrer, A., and Filter, M. (2016). Discussing State-of-the-Art Spatial Visualization Techniques Applicable for the Epidemiological Surveillance Data on the Example of *Campylobacter* spp. in Raw Chicken Meat. *Zoonoses and Public Health*, 63(5):358–369.
- Ran, R., Zhao, C., Xu, X., and Yao, G. (2013). A Case Study on Epidemic Disease Cartography Using Geographic Information. 7798:232–243.
- Weng, J., Xu, Y., and Sharma, A. R. (2012). EPIDEMIC ANALYSIS AND VISUALIZATION BASED ON DIGITAL EARTH SPATIO- TEMPORAL FRAMEWORK 2 . State Key Laboratory of Remote Sensing Science , Institute of Remote Sensing Applications of Chinese Academy of Sciences , Beijing , 100101 , China * Corresponding a. pages 7220–7223.



Lista de doenças transmissíveis de notificação obrigatória

Lista de doenças transmissíveis de notificação obrigatória
(Despacho nº 15385-A/2016 de 21 de dezembro)

Doenças de Declaração Obrigatória				
Botulismo	Doença Invasiva Pneumocócica	Hepatite B	Paralisia Flácida Aguda	Tétano, excluindo Tétano Neonatal
Brucelose	Doença Invasiva por <i>Haemophilus influenzae</i>	Hepatite C	Parotidite Epidémica	Tétano Neonatal
Campilobacteriose	Ébola	Hepatite E	Peste	Tosse Convulsa
Cólera	Equinococose/Hidatidose	Infeção pelo novo Coronavírus (MERS-CoV)	Poliomielite Aguda	Toxoplasmose Congénita
Criptosporidiose	Febre amarela	Infeção por <i>Bacillus anthracis</i>	Raiva	Triquinelose
Dengue	Febre Escaro-Nodular (<i>Rickettsiose</i>)	Infeção por <i>Chlamydia trachomatis</i> , Incluindo Linfogranuloma venéreo	Rubéola Congénita	Tuberculose
Difteria	Febre Q	Infeção por <i>Escherichia coli</i> produtora de Toxina Shiga ou Vero (<i>Stec/Vtec</i>)	Rubéola, excluindo congénita	Tularémia
Doença de Creutzfeldt-Jakob (DCJ)	Febre Tifoide e Febre Paratifoide	Infeção por vírus do Nilo Ocidental	Salmoneloses não <i>Typhi</i> e não <i>Paratyphi</i>	Variola
Doença de Creutzfeldt-Jakob variante (vDCJ)	Febres hemorrágicas virais e febres por arbovírus	Infeção por vírus ZIKA	Sarampo	VIH (Infeção pelo vírus da imunodeficiência humana) /SIDA
Doença de Hansen (Lepra)	Giardíase	Leishmaniose Visceral	Shigelose	Yersiniose
Doença de Lyme (Borreliose)	Gonorreia	Leptospirose	Sífilis Congénita	Resistências aos antimicrobianos
Doença dos Legionários	Gripe Não Sazonal	Listeriose	Sífilis, excluindo Sífilis congénita	
Doença Invasiva Meningocócica	Hepatite A	Malária	Síndrome Respiratória Aguda - SARS	