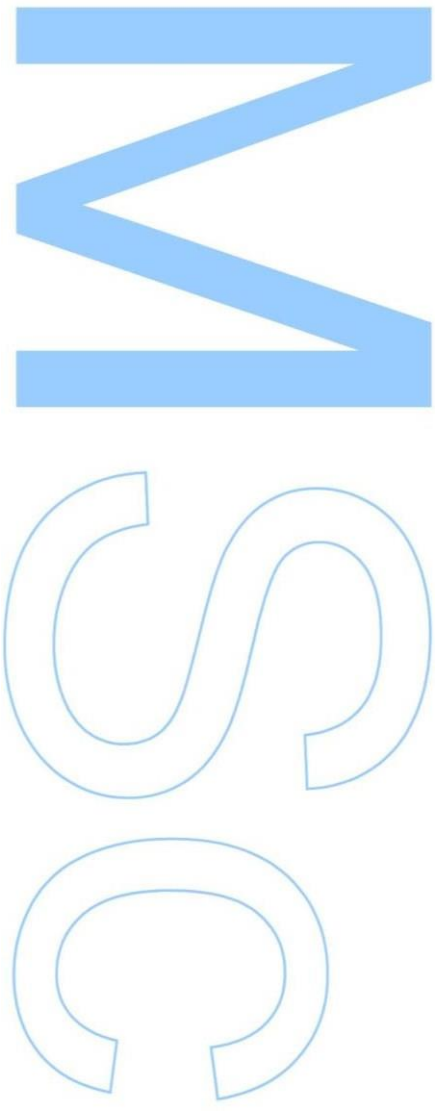
An exploratory data analysis of the TTR-FAP disease in Portugal

Rúben Xavier Correia Lôpo
Mestrado em Engenharia de
Redes e Sistemas Informáticos
Departamento de Ciência de Computadores
2022

Advisor
Alípio Mário Guedes Jorge, Associate Professor

Faculdade de Ciências da Universidade do Porto



All corrections determined by the Jury, and only those, were carried out.

The President of the Jury,

Porto,_____/_____/_____

To my mother Teresa Correia and my grandmother Ilda, the women of my life who sacrificed their own well-being for the sake of my education, to my cousin João Santos, to my close friends Rui Ramos and André Tse who have been with me since day one, to my advisor Alípio Jorge and Maria Pedroto who supported me during this project and to my close friends in my humble town of São João da Madeira. To them and to everyone I met in FCUP and helped me, because they are too many to name during these 5 years of joy.

Declaração de Honra

Eu, Rúben Xavier Correia Lôpo, inscrito(a) no Mestrado em Engenharia de Redes e Sistemas Informáticos da Faculdade de Ciências da Universidade do Porto declaro, nos termos do disposto na alínea a) do artigo 14.º do Código Ético de Conduta Académica da U.Porto, que o conteúdo da presente dissertação/relatório de estágio/projeto "An exploratory data analysis of the TTR-FAP disease in Portugal" reflete as perspetivas, o trabalho de investigação e as minhas interpretações no momento da sua entrega.

Ao entregar esta dissertação/relatório de estágio/projeto "An exploratory data analysis of the TTR-FAP disease in Portugal", declaro, ainda, que a mesma é resultado do meu próprio trabalho de investigação e contém contributos que não foram utilizados previamente noutros trabalhos apresentados a esta ou outra instituição.

Mais declaro que todas as referências a outros autores respeitam escrupulosamente as regras da atribuição, encontrando-se devidamente citadas no corpo do texto e identificadas na secção de referências bibliográficas. Não são divulgados na presente dissertação/relatório de estágio/projeto "An exploratory data analysis of the TTR-FAP disease in Portugal" quaisquer conteúdos cuja reprodução esteja vedada por direitos de autor.

Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Rúben Xavier Correia Lôpo

15 de Outubro de 2022

Acknowledgements

Throughout my life, my family has always been richer in values than in material goods or even in academic recognition. Before me, only one member of my vast family had the opportunity to study what they loved and to pursue a career.

First of all, and because this is the end of a long cycle of studies, I want to thank my closest family, which include my mother and grandmother. Despite having little to no school education, they nurtured in me the values of responsibility, altruism and commitment to make me become someone in life as they never had the opportunity to do so. To my father, and because life has not always been as we know it, my thanks for his concern in ensuring that I completed this milestone.

I would like to express my gratitude to Professor Alípio Jorge, for having accepted me in this project and for ensuring that I would complete it successfully. I would also like to thank him for the opportunity to join INESC TEC with a research scholarship at the Laboratory of Artificial Intelligence and Decision Support (LIAAD) in a dissertation project in an area that fascinates me a lot, which is Data Science. A special word to Maria Pedroto for her help, availability and friendship during a long year of work.

To finish an academic course is impossible without the existence of friendship. For that reason, my thanks to Diogo, Fábio, Daniel, Beatriz and Carolina for 5 years of memories built even with each one in his own personal journey. My thanks also to my course mates that I will take for all my life under a black cover full of unforgettable memories and that made me forget momentarily the difficulty of being a university student.

Five years ago I was about to give up on a dream because life is sometimes too hard to accept. So a thank you to myself 5 years ago for taking a difficult step that I now cherish so much.

My most heartfelt gratitude for 5 memorable years.

One of my great passions is writing. I was lucky enough a few years ago to find in poetry one of my talents. Therefore I want this work to start with the reading of a poem I wrote about science and what it meant to me when I chose it as my philosophy of life.

In science is the happening of the world

In science is the happening of the world,
The basis of the all-embracing foundation.
Since universality is people
With the profound knowledge.

The world is of the latter individual.
It grows when man persists,
When risk-taking is imminent,
To overcome the firm wrath.

In the true will to advance
The importance lies in knowing,
When in knowing one knows how to speak.

But in the act of making ourselves grow
It is important in the mind to know how to link,
Such sides of the valuable understanding.

Na ciência está o acontecer do mundo

Na ciência está o acontecer do mundo,
Base do fundamento abrangente.
Pois a universalidade é gente
Com o conhecimento profundo.

O mundo é do indivíduo segundo.
Cresce quando o homem é persistente,
Quando o arriscar é risco iminente,
Para vencer o firme iracundo.

No verdadeiro querer avançar
A importância reside no saber,
Quando no saber se o sabe falar.

Mas no ato de nos fazermos crescer
Importa em mente saber conjugar,
Tais lados do valoroso entender.

Abstract

Transthyretin-associated Familial Amyloid Polyneuropathy (TTR-FAP) is a fatal chronic disease with a significant number of cases in Portugal. The behaviour of the disease, its evolution and distribution of cases in the national landscape are public health issues that concern any endemic community. This is a disease with more than a century of history in the Portuguese community with a long record of clinical studies and that may benefit from the use of new concepts of data analysis and management.

Since this disease spreads from north to south of the country, it is important to provide professionals with a tool that allows a detailed geographical and territorial study. In this dissertation, we implemented a tool that applies concepts of geovisualisation with spatial data in order to understand the historical progression and how it is distributed among different territorial levels, from the whole country to districts and counties. This tool allows the choice of parameters to be considered, allowing the study of different subsets of users with different characteristics.

The visual comparison of different time periods of the disease can help health professionals to make more informed decisions in order to improve the quality of life, treatment and follow-up to patients. The tool is available online for data exploration and its code is available on GitHub for use in other geospatial scenarios. A paper on the processing of this data and geographic exploration of the disease was written and accepted at an international conference on data science.

Keywords: Transthyretin-associated Familial Amyloid Polyneuropathy, Geovisualisation, Spatial Data, Imputation

Resumo

Polineuropatia Amiloidótica Familiar (PAF), Paramiloidose ou Doença dos Pezinhos é uma doença crónica fatal com uma quantidade de casos significativa em Portugal. O comportamento da doença, a sua evolução e distribuição dos casos no panorama nacional são questões de saúde pública que se prendem a qualquer comunidade endémica. Esta é uma doença com mais de um século de história na comunidade portuguesa com um longo cadastro de estudo clínico e que poderá beneficiar da aplicação de conceitos novos de gestão e análise de dados.

Uma vez que esta doença se estende de norte a sul do país, é importante facultar os profissionais com uma ferramenta que permita um estudo geográfico e territorial detalhado. Nesta dissertação, foi implementada uma aplicação que reune conceitos de geovisualização com dados espaciais de forma a entender a progressão histórica e como esta se distribui por entre diferentes níveis territoriais, desde todo o país a distritos e concelhos. Esta ferramenta permite a escolha dos parâmetros a considerar, permitindo o estudo de diferentes subconjuntos de utilizadores com caraterísticas diferentes.

A comparação visual de diferentes períodos de tempo da doença pode ajudar os profissionais de saúde a tomarem decisões mais informadas de forma a melhorar a qualidade de vida, tratamento e acompanhamento dos pacientes. A ferramenta está disponível online para exploração de dados e o seu código está disponível no GitHub para uso em outros cenários geoespaciais. Um paper sobre o tratamento destes dados e exploração geográfica da doença com os mesmos foi escrito e aceite numa conferência internacional de ciência de dados.

Palavras-Chave: Polineuropatia Amiloidótica Familiar (PAF), Paramiloidose, Doença dos Pezinhos, Geovisualização, Dados Espaciais, Imputação

Contents

List of Tables

List of Figures

Listings

Acronyms

CSV Comma-separated Values

DM Data Mining

DS Data Science

GVis Geovisualisation

TTR-FAP Transthyretin-associated Familial Amyloid Polyneuropathy

Chapter 1

Introduction

1.1 Context and Motivation

Portugal is globally denoted as an epicentre [1] with regard to the incidence of cases of Transthyretin-associated Familial Amyloid Polyneuropathy (TTR-FAP). This disease is a rare and hereditary, which manifests itself in a progressive and neurodegenerative manner, has severe consequences over time and is clinically characterised as life-threatening. This motivates our contribution to providing a tool for the analysis of the disease that enables.

This condition represents a direct danger to its carriers and these patients have abnormal deposits of a protein called amyloid (amyloidosis) [2]. These deposits are mostly found in the peripheral nervous system, which includes the bodily nerves that sense pain. These deposits negatively affect the sensory capacity of these people especially in the extremities of the body. Due to the progressive characteristic of the disease, the central nervous system, which includes the brain and spinal cord, can be affected, putting vital organs at risk, like the heart [2].

More than 80 years ago, Corino de Andrade diagnosed the first TTR-FAP case that matched the characteristics described previously [3]. Analysing this data, if this diagnosis was made in 1939 and the patient was a 37-year-old woman, this generation already had information regarding the disease and it is possible to observe, roughly, at least 120 years of history due to the heredity of the disease.

As a result of the incidence of TTR-FAP in Portugal, it becomes evident that a study and analysis dedicated to the geographic behaviour of its identity and evolution is necessary. An analysis of geographical data, complemented by a statistical study that relates its background, may be relevant to understand the disease from an epidemiological perspective.

The medical analysis of this disorder is openly studied but, with regard to Portugal, the investigation and exploration of data from a territorial point of view is not the most detailed, comprehensive and complete, so this new data analysis may become essential for medical research in the short, medium and long term.

Due to the nature of the disease, patients and their families can be impacted economically and socially. Data Science (DS) tools can prove useful to understand behaviours and symptoms in automated procedures compared to traditional methods, with better accuracy and, more importantly, discover new information that, to the naked eye, would not be linkable.

As a result of this, through Data Mining (DM) algorithms and tools, we believe it will be possible to understand the behaviour of the disease in history and also how those generations relate in sub-groups relatively to the population, as well as giving the opportunity to healthcare professionals to formulate hypothesis themselves.

The work presented is carried out on real patient data, provided by Centro Hospitalar Universitário do Porto that includes Hospital Santo António, located in Porto, Portugal.

1.2 Main goals

To achieve success with this dissertation, objectives were set that, although flexible, are related to the motivation mentioned above.

- The main goal of this work is to contribute to the understanding of the epidemiological behaviour of TTR-FAP in the time and space dimensions.
- This objective was divided into 3 main parts: carrying out a spatio-temporal study with the historical data available, develop an interactive data control tool on visualisation to be used in the medical field that may enable tangible results and may explain geographical and demographic behaviour on the TTR-FAP data mainly in its spatial and temporal dimensions and making the tool available so that it can be applied and adapted to other data sets.

In addition to these objectives, the work shall contain essential actions in order to complete it and achieve the objectives mentioned above. These include:

- To study, understand, analyse and collect sources to build a State of the Art. This SotA should include an overview of visualisation theory and epidemiological visualisation while also reviewing studies on TTR-FAP and similar diseases.
- Develop tools for geographical visualisation of real data with spatio-temporal variables and apply data analysis tools regarding the age of onset as well as patient's locations.
- Find possible correlations in data that may be indirectly related to the geographical structure of patient subgroups and may explain their geography and location.
- To write a paper analysing the results obtained in order to enrich the information about the disease in the scientific community.

1.3 Dissertation layout

The dissertation is structured as follows. Chapter 1 presents the context of the problem, the motivation of the work and its objectives. Chapter 2 presents relevant concepts of data mining as its pipeline of actions and in specific the data preprocessing as well as an overview of spatial data, imputation and missing values. Chapter 3 presents a literature review related to the problem in 3 parts, studies related to TTR-FAP, visualisation theory and geovisualisation and a review of studies on endemic diseases and the techniques used. Chapter 4 discusses the data preparation, explaining the dataset used and its transformations including geocoding and imputation. Chapter 5 presents the primary exploration of GVis in the work that allowed the development of the techniques used later. Chapter 6 presents AmiVis, its prototype, features and architecture. Chapter 7 presents the information that has been gathered throughout the work resulting in demonstrations of spatial and temporal differences in the country. Chapter 8 concludes the document and presents possible future work. There are also 3 final appendices A, B and C that are used in the discussion of the results.

1.4 Contributions

This work resulted in contributions and tools to the scientific community:

- The online application AmiVis [4] and the availability of the code on GitHub [5].
- A paper was written about the process of data processing, creating visualisations, the application and future work that was also evaluated and improved. This paper was accepted to the Seventh Workshop on Data Science for Social Good, SoGood 2022 [6], held in conjunction with ECML PKDD 2022, in September 2022, at Grenoble, France.

Chapter 2

Relevant Concepts

In this chapter, we present the relevant concepts of Data Mining (DM) for this dissertation and introduce techniques and data tools addressed in various phases of the project.

2.1 Data Mining

The DM process is directly related to knowledge extraction through clean data collection and analysis. In modern times, society lives in a data age as the volumes of information produced are extremely high and originate from even more sources.

As Han et al. states [7], the technological evolution has created powerful data collection and storage tools. All aspects of modern life have become computable to some extent. For this reason, the challenge arises of drawing conclusions and facts that are modulated depending on whether the problem is human or not and how the data's subject limits the work to be done.

Consequently, it is important to note that most data goes through a pipeline of actions in order to make it possible to respond to a disparity of problems. Thanks to the evolution of the tools of this pipeline, problems that remain the same in time now obtain more advanced solutions and conclusions. In particular, as for the objective of this dissertation, which is the study and analysis of the behaviour of a disease and medical data, through a temporal and geographical visualisation of the territory, the problem remains the same as it was 100 years ago when the disease began to be studied. On an opposite, computational capabilities are extremely more advanced, allowing data to be analysed and treated differently, with modern procedures and with more accurate, credible and pleasurable hypothesis.

2.2 Data Mining Pipeline

The process of applying DM in the contemporary world is commonly planned according to a pipeline of operations. This set of operations and their order enable the transformation of information input into knowledge output.

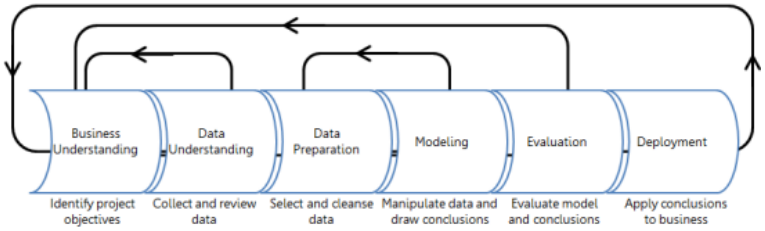


Figure 2.1: Cross Industry Standard Process for DM (CRISP-DM)[8].

As Juodyte states [8], the CRISP-DM model grew out of an earlier model named Knowledge Discovery in Databases (KDD). It is also important to mention that this pipeline is iterative and allows us to go back to previous steps and repeat them in order to strengthen the final output of data. Currently, the CRISP-DM model is followed by the community as an efficient and effective procedure that can be seen in the figure 2.1 and involves the following phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Firstly, in terms of Business Understanding, it is a phase that involves understanding the boundaries of the project and its objectives from a comprehensive and distribution point of view. Thus, first we define a DM problem, plan and how the outlined objectives will be reached through business success criteria. However, even in situations where the business is differentiated, as is the case in medicine, there will always be a problem, a plan to follow and conclusions to be drawn from the work, but it is up to the team working with the data to define this progress in coordination with the subject, never forgetting the resources, tools, techniques, risks and costs.

Secondly, when talking about Data Understanding, it covers the entire environment of activities related to the collection of data, as well as the activities that allow data scientists to become familiar with the peculiarities of the information, data quality problems and first approaches and hypotheses about what they have in hand. It is relatively common for data to be poorly organised and incorrect and it is used in the next phase.

Following the pipeline, Data Preparation involves all the activities required to obtain a final dataset ready to handle, including table, record and attribute selection as well as the application of other tools for data inclusion and exclusion named data selection. With the same care, the dataset is cleaned in a data cleaning phase from erroneous values and inconsistencies in data types and formats. The aim of this phase is to enable statistical analysis of the information using tools that require the data to be correctly formatted, normalised and regulated.

In the Modelling phase, the techniques and procedures to be applied to the data are selected and the parameters are adjusted to achieve the best results. The models can have different approaches, such as Anomaly Detection for detecting outliers and out-of-normal values, Association Rule Learning for modelling dependent relationships between variables, Clustering for discovering groups and subgroups of similar values, and Classification, which generalises information already known to be applied to new data and to classify it. On the other hand, in Regression, algorithms tend to infer the relationship of a dependent variable with independent variables while Summarization aims at a compact visualisation of the dataset.

Finally, in the Evaluation phase, a quality model must be tested in order to evaluate whether it meets the requirements imposed in the initial phases. Although Deployment is the last phase, the pipeline is always lively due to the fact that new knowledge and new data can always improve most hypothesis, making them applicable to modern standards.

While the introductory chapter 1 has already taken an overview of the Business Understanding and objectives of this dissertation, all the other phases are in the following chapters.

2.3 Data Preprocessing

Data preprocessing aims to evaluate and orchestrate factors that compromise data quality, such as accuracy, completeness, consistency, timeliness, believability, and interpretability, according to Han et al. [7]. Incorrect, incomplete, and inconsistent data are inseparable properties of the currently collected data that are part of real data databases and data warehouses and with very strong relationships to the daily life of their genesis. Complementarily, raw data is often in a form that is not the one needed for processing and this includes raw logs, documents and semi structured data.

The existence of incorrect data can be simply explained by human or computational error, but it is worthwhile introducing the concept of disguised missing data. Sometimes, for a variety of reasons as stated by Pearson [9], a user may choose to enter a default value or a wrong value in one of the information that concerns him. This is also applicable to the fields of medicine and medical data, since patients may, by personal choice, enter data that is different from what is real (such as the individual's secondary diseases or eating habits). Such data is fundamentally wrong, but for an automated system where it is entered according to its formatting rules, it is completely acceptable data. This type of data disrupts results and is naturally difficult to understand, detect and eliminate.

According to the factors, the data is not always complete because of lack of accessibility or misunderstanding when it is entered. The consistency of data collection methods also impact its quality. In the medical field, many patients are automatically assigned with data from their parents and relatives which translates into inconsistencies. It is also important to assess the data's believability which reflects how much the data are trusted by users and its interpretability which reflects how the data is perceived.

There are main tasks to take into account during data preprocessing, among which are data cleaning, data integration, data reduction and data transformation.

2.3.1 Data Cleaning and Data Integration

The data cleaning phase aims to solve problems such as missing values and noise in the data. The presence and identification of outliers, which may or may not be the objective of the work, and dealing with inconsistencies in the data are other purposes of this phase. Note that these procedures must be robust in order to avoid over fitting the data and making it difficult to obtain real results. Some of the problems and errors present in data collection are examples of:

- Technological problems in sensors and in their hardware limitations associated with information collection and transmission.
- Scanning errors due to lack of perfection of recognition techniques (e.g. speech-to-data).
- Refusal or preference to hide personal data for user privacy reasons, as mentioned earlier in disguised missing data.
- Manual errors and manually created data.
- The responsible entity may not be able to collect specific types of data due to their cost or consequences.

Regarding this step, when missing values are encountered, possible solutions involve ignoring the record or filling it with a manually entered value or a constant (e.g. Unknown). It is also possible to fill the value with an attribute trend calculation like the mean or median, which is calculated from the whole dataset or from classes or subsets of the dataset, and finally, the most likely value may be used. Still, the latter options introduce values that may not be correct and may introduce bias in the dataset, but still, the last technique is the most used. The process of estimating values is called imputation.

As for the outliers present in the data, these are possibly detected from clustering where values are grouped into groups called clusters that share the same characteristics. Records that differ from what is normal within these characteristics are considered outliers and an example of that can be seen in figure 2.2. The use of knowledge about the data itself, a priori, called metadata, can be beneficial in confirming the existence of outliers.

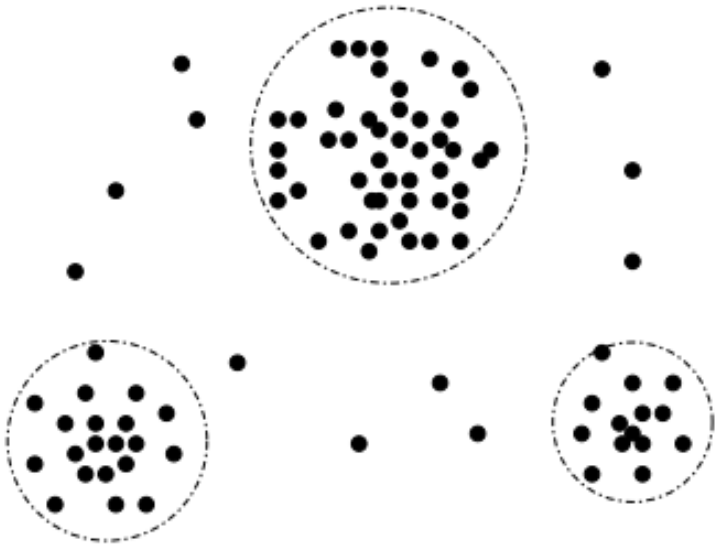


Figure 2.2: A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets[7].

Data Integration is a concept that involves cross-referencing data from various sources in order to gather information from different databases, cubes or files and represent a concept more generally. In addition, this process avoids redundancies and inconsistencies in the final dataset, increasing the speed of the subsequent DM process.

2.3.2 Data Reduction and Data Transformation

When you perform data reduction, you try to obtain a representative sample of the initial dataset that is smaller in volume but produces the same analytical results using techniques such as dimensionality reduction and numerosity reduction. On the one hand, when performing dimensionality reduction, one applies techniques that promise a compressed representation of the data but still in a representative way (e.g. removing irrelevant attributes, as shown by Gama et al. [10]). On the other hand, a numerosity reduction implies changing the data by a smaller and alternative data using parametric models (e.g., regression or log-linear models) or nonparametric models (e.g., histograms, clusters, sampling, or data aggregation). Regarding data compression, if it is possible to reconstruct the original data from the new subset, the data reduction is called lossless. If it is possible to reconstruct only an approximation, it is called lossy. In addition, there are algorithms that allow a more specific selection of relevant attributes for a final dataset. This final set of attributes must follow a distribution similar to the original to preserve the results, and most of these algorithms, according to Han [7], are greedy, since they choose the best answer that exists for a problem at a given time.

As far as data transformation is concerned, it comprehends the processes of:

- Smoothing with binning, regression and clustering techniques to remove noise.

- Attribute construction or feature engineering in which new attributes are created from existing attributes to help consecutive processes.
- Aggregation and summary operations to address datasets problems.
- Normalisation in which data is scaled in smaller ranges (e.g. -1.0 to 1.0).
- Discretisation in which numerical data are transformed into interval or concept labels (e.g. 0-10,11-20 or adult, senior). Through conversion methods between various types it is possible to use numeric data, this being the most widely used and simplest data type in DM algorithms, according to Aggarwal [11]. Discretisation of the data for this type can be done in a number of ways to achieve data granularity. It is also important to note that the data can have many different types of source and destination, which are shown in figure 2.3 as an example.

Source data type	Destination data type	Methods
Numeric	Categorical	Discretization
Categorical	Numeric	Binarization
Text	Numeric	Latent semantic analysis (*LSA*)
Time series	Discrete sequence	*SAX*
Time series	Numeric multidimensional	*DWT*, *DFT*
Discrete sequence	Numeric multidimensional	*DWT*, *DFT*
Spatial	Numeric multidimensional	2-d *DWT*
Graphs	Numeric multidimensional	*MDS*, spectral
Any type	Graphs	Similarity graph (Restricted applicability)

Figure 2.3: Portability of different data types [11].

- Concept hierarchy generation for nominal data in which the data is generalised to broader concepts such as a street to city or district.

In this context, it is important to mention the existence of data type portability because, due to the existence of multiple types, it is useful to make the information portable, transmissible and dynamic.

In short, the quality of the data tends to be less than desired, being inconsistent, incomplete and wrong. The techniques presented help to raise the quality of the data so that it is possible to extract the greatest amount of knowledge.

2.4 Spatial Data

Since the objective of this dissertation involves the processing of spatial data and since we need to transform this data to be able to apply specific techniques and procedures, we need to know the challenges to overcome. Spatial DM [7] is characterised by the ability to extract knowledge from quantities of data with spatial characteristics. Still, one of the main difficulties of dealing with this type of data is its nature of origin from different GIS (geographical information systems) [12].

Spatial DM is directly related to KDD (knowledge discovery in databases) defined as the discovery of implicit and previously unknown knowledge of considerable databases. Spatial data has logical and distance-related properties that are usually treated in a specific way. This kind of special data treatment involves different challenges and opportunities to explore data and reach conclusions and hypotheses.

Even with a Statistical Spatial Analysis [12] approach, there are some drawbacks considering that one has to pay attention to the statistical dependency between variables, since this method assumes independence and is not something one necessarily wants to show to the end user.

Koperski also talks about some primitives that should be taken into account when working directly with spatial data:

- Rules: Rules are not unique to spatial data but they are certainly a useful and applicable primitive for this type of data. A spatial characteristic rule can be, for instance, a rule that describes the general rental price of houses in a geographical area.
- Thematic Maps: This type of map allows for a visual representation of the distribution of one or more attributes but differs from generic maps in that the objective is to present the position of objects relative to other objects.
- Image Databases: Spatial databases composed of images and pictures generally used in remote sensing.

Koperski et al. [12] talk about two algorithms (SDD & NDD) that differ with respect to rules but which coincide on a very important point in the way spatial data queries are defined. Even without applying rules, it is possible to understand that different query schemes, namely in the attributes chosen, logically leads to different results.

If, on one hand, a Spatial-data-dominant generalisation algorithm extracts a set of explicit features relative to a map divided into areas for each previously delimited region, on the other hand, as with a Nonspatial-data-dominant generalisation algorithm, the focus can be on finding the region, beyond the pre-delimited area, that comprises a set of features. For this reason, it is useful to identify in the data the purpose to be found to better formulate hypotheses.

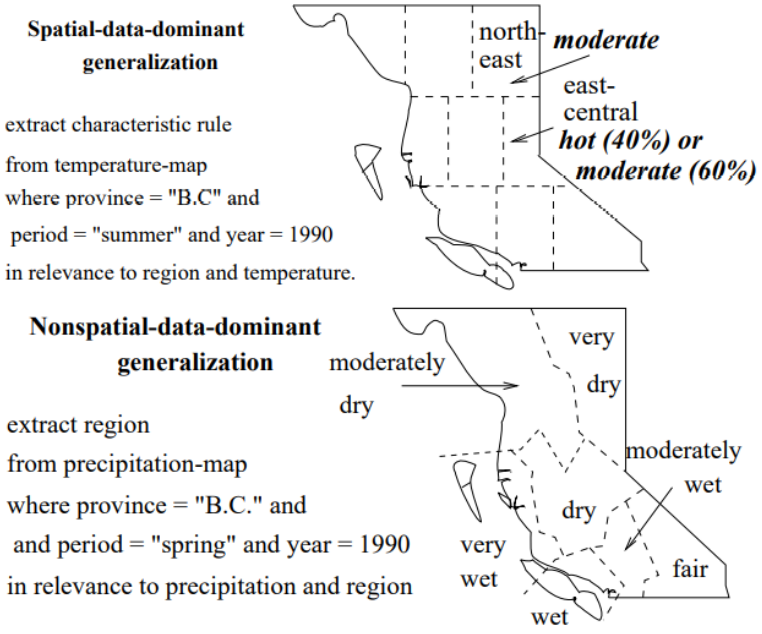


Figure 2.4: Example of a query and the result of execution of both spatial-data-dominant generalisation and nonspatial-data-dominant generalisation methods and their different outcomes[12].

2.5 Imputation and Missing Values

Value imputation is, first of all, directly related to the lack of attribute values. Real world databases often suffer from this problem and have a large part of their data incomplete or incorrectly entered as stated by [13]. Therefore, it is often convenient to capitalise on the ability to use and develop tools that not only provide solutions to this problem, but also maintain the quality of the data.

Depending on the problem and the data, it is possible to opt for simple or more complex data techniques. Still, one should track how biased the results become with the techniques as well as their complexity. In simple imputation, only one estimate is used while in multiple imputation, several estimates are used translating into uncertainty of the final results obtained.

In order to apply the imputation techniques, first one needs to know the types of data that are missing. Missing values can be classified in 3 different categories like [14] showed:

- MCAR - Missing Completely at Random - which defines missing values that are independent of observed or unobserved data and that do not include standards or norms. When data is in this category, the set of records without missing values is also a random sample of the population, which implies that the simplest imputation methods result in unbiased imputations. E.g. A laboratory value that could not be recorded because the blood tube was damaged.
- MAR - Missing at Random - The missing value is related to the observed data but not to the data not yet observed, and may include something systematic in its value assignment.

If the missing data falls into this category, there is relevant information that is lost and there is no universal method for dealing with the problem. E.g. The value of a person's income may depend on their age.

- MNAR - Missing Not at Random - The missing value is related to data not yet observed from the variable itself creating a missing problem which cannot be ignored and depends on the context. In this type of missing data category, case analysis is not based on random selection and there is probably inherent bias. E.g. Missing values on a form by patient choice.

Regarding solutions to solve these missing values, besides the possibility of removing observations and considering only complete records or ignoring these values in the analytic phase of the process of applying techniques and models, it is still possible to make estimates and impute values. The type of imputation and its evaluation should vary depending on the context of the problem and should be defined according to the purpose of the techniques to be used.

2.6 Summary

In this chapter, some DM concepts were briefly explained regarding the its pipeline with a extra focus on the data preprocessing phase that uses tools of data cleaning, data integration, data reduction and data transformation. We have also provided an introduction to concepts related to spatial data, imputation and missing values.

Chapter 3

State of the Art

In this chapter, we analyse information sources regarding the topic of the dissertation and in the first place, sources that focused on the disease, its behaviour and evolution. In the second place, information related to visualisation theory was analysed, so that the future work of this dissertation will be based on concepts resulting from a study on epidemiological visualisation. Finally, work applied to similar diseases was analysed in order to learn and understand how diseases are approached and how results are interpreted in similar situations.

3.1 Transthyretin-associated Familial Amyloid Polyneuropathy (TTR-FAP) clinical studies and geographic visualisations

3.1.1 TTR-FAP clinical studies

As a first contact with the disease, Andrade [3], in 1939, already had knowledge about a condition with incidence in the municipality of Póvoa do Varzim, belonging to the district of Porto in northern Portugal which was called "mal dos pésinhos" by the local inhabitants, which translates as foot disease. The author also mentions that the patients studied had the most derived family backgrounds, were from all social classes, had varied diets and had a high mortality rate (to date). The author also reveals that a large part of the known cases had rural jobs (such as fishermen) and reveals that the counties surrounding Póvoa do Varzim were affected to some extent (such as Vila do Conde).

In [1], the authors report, as of 2018, that after the age of onset (which defines the beginning of symptoms), patients have an expected survival of between 10 and 15 years. The results show that Portugal, and more specifically, the north of Portugal, has the highest prevalence rate on a global scale. Other results extracted from this study show that Portugal leads in the prevalence of TTR-FAP cases when it comes to small and medium population sizes (For n=5.5 and n=10, rounded values), per million inhabitants and not for prevalence in large values (n = 38), lead by China.

Moreover, Inês et al.[15] report that treatment for TTR-FAP includes liver transplant (since 1992) and tafamidis (since 2012 for patients followed at TTR-FAP specialised centres). Analysing the paper, there is a study on the prevalence and incidence of cases in Portugal. It is important to mention that the prevalence of a disease is the number of (total) cases in a defined population in a certain period of time, while incidence is determined by the number of new cases of the disease in a defined population during a period of time. It appears from this data that the incidence of the disease has been decreasing while prevalence has been increasing.

With regard to incidence, the number of yearly new TTR-FAP patients were 31% fewer between 2010 and 2016. There was also a peak in 2013 due to the transfer of patients to use tafamidis in specific treatment centres. A higher age of onset and late-onset were recorded in more recent years and male patients were found to be significantly younger than female patients. Geographically speaking, in this period 174 municipalities out of 278 (62.6%) were identified and in 19 exceeded the European threshold for rare disease (any disease affecting 5 or less in 10,000 persons in the community). Finally, one of the possible actions that may have had an impact on the number of cases is genetic counselling programmes to at-risk families.

In order to designing an approach that would allow the prediction of the year of onset of the symptoms of the disease with variation of input variables, Pedroto [16] looked into datasets of patients with a confirmed sick ancestor and with at least one relative followed in the Clinical Unit, patients with no known ancestor with the disease and no relatives in the unit and a mixed set of patients. Here the results are compared to the traditional method used by medical staff, without automation. From this study it is concluded that in the first dataset better results are obtained with Ridge and Decision Tree algorithms and the traditional model is worse. In the second dataset, the methods that stand out are Random Forest and Decision Tree. In the mixed dataset, Random Forest, Decision Tree and Ridge stand out. It is also possible to observe that the age of onset of patients with information about ancestors is determined more quickly than the opposite.

In [17], the authors surveyed the endemic situation in Europe. Comparing the evolution of the disease in the different countries, they show Portugal in two different representations. The first refers to the number of families taking into account adaptations, with special focus on the areas of Porto, Braga, Coimbra and Castelo Branco. The second indicates that Portugal is mostly (99%) constituted by a disease mutation (Val30Met), which indicates that patients will mostly be affected in the same way.

The article also states that the average age of onset of the disease is 33.5 years and that 87% of patients develop symptoms before the age of 40. Portugal is one of the countries with a faster diagnosis (with a shorter delay) where each patient only needs to consult 2 specialists. In 2012, according to the Ministry of Health, Portugal started to reimburse oral treatment by the Portuguese National Health System. The authors point out that a network approach can have an impact on the future of the diagnosis of the disease, starting with the awareness and education of professionals and the public, facilitated by a more competent and collaborative research in order to support more informed decisions. This type of information may be fundamental in explaining

changes in the data to be analysed and correlating the results.

Since TTR-FAP can have different mutations, it is important that the same one as the Portuguese majority of patients have is analysed and, in paper [18], the landscape of Majorca, Balearic Islands, Spain, is scrutinised. The data analysis was performed with patients between 2001 and 2012. Regarding asymptomatic carriers, the gender difference is not extraordinary, but there are many more men with symptoms than women (65% to 35%). Statistically significant differences were found in the mean age at diagnosis for asymptomatic and symptomatic patients (54 to 44 years, $p < 0.05$) and regarding their employment status (75.8% vs. 50.0%; $p < 0.01$). The age of onset in this study was much older (49.8 years) than in Portugal (33.5 years).

Such as the previously reference in [15], there is also a discussion of treatment methods and therapies for the condition in article [19]. Without going into excessive detail, two of the greatest changes, expressed in the figure 3.1, with respect to this topic relate to the:

- appearance of liver transplantation around 1990 which revolutionised what would hitherto have been a methodology of no therapy.
- introduction, around 2012 in Europe, to tafamidis and further pill treatments that have changed the outlook of the hegemony of liver transplantation.

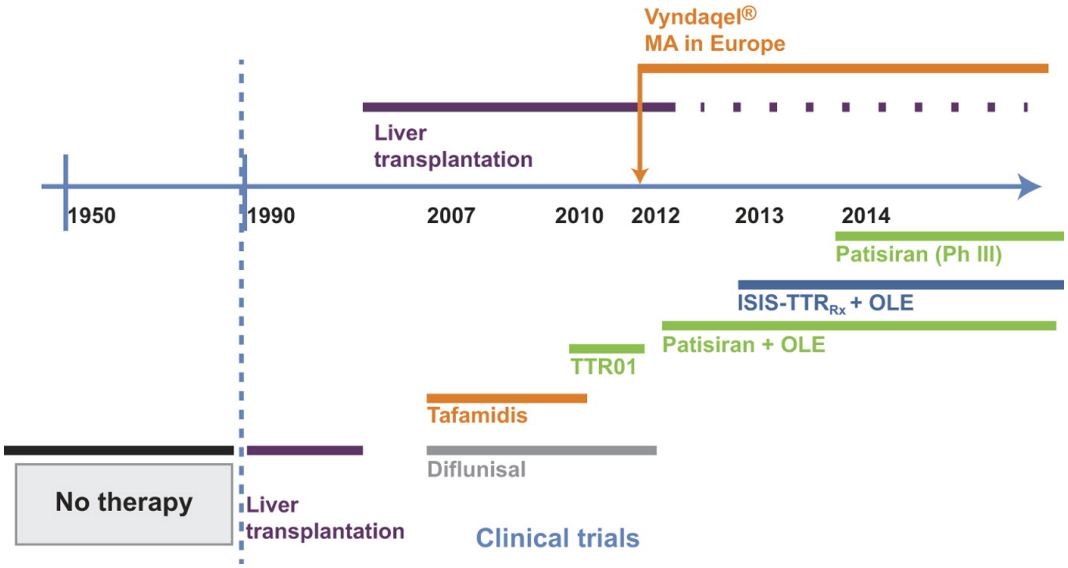


Figure 3.1: Timeline of access to anti-amyloid therapies for patients with TTR-FAP (Vyndaqel is a trade name for tafamidis) [19].

These historical changes may explain some of the variations in the information present in this work.

In article [20], there is a study on the history of TTR-FAP and treatment effects in patients with the most common variant in Portugal, with data from two Portuguese reference centres. Patients were studied taking into account the treatment carried out, or not, and whether this was by liver transplant or with tafamidis.

As this study was applied to Portuguese patients, the results will be directly related to the data to be worked on. Both genders are affected and age of onset is more common in young adults. The authors refer that untreated affected patients up to 1 year have better chances of survival, and this may be due to better medical conditions. They also note that the low survival rate when liver transplantation was more common may reflect the clinical and social characteristics of the time, including a later disease stage, poor nutritional status, health illiteracy and low awareness of disease treatment.

Nevertheless, Coelho et al.[20] report that liver transplantation, in particular, has been extremely important for the prognosis of TTR-FAP in the last 25 years prior to 2016, making the disease a more chronic condition with delayed treatment than a previously known fatal disease. This information is supported by a 60% reduction of SMR (Standardised mortality ratio) from 1991 to 2016.

3.1.2 TTR-FAP Geovisualisation (GVis)

Mazzeo et al. [21] studied the endemic area of Sicily in Italy over a 20-year period from 1995 to 2015. In addition to the analysis of the family background, the manifestations of the disease and treatment were investigated and, more precisely, demographic details and the geographical distribution of the patients' residence and origin.

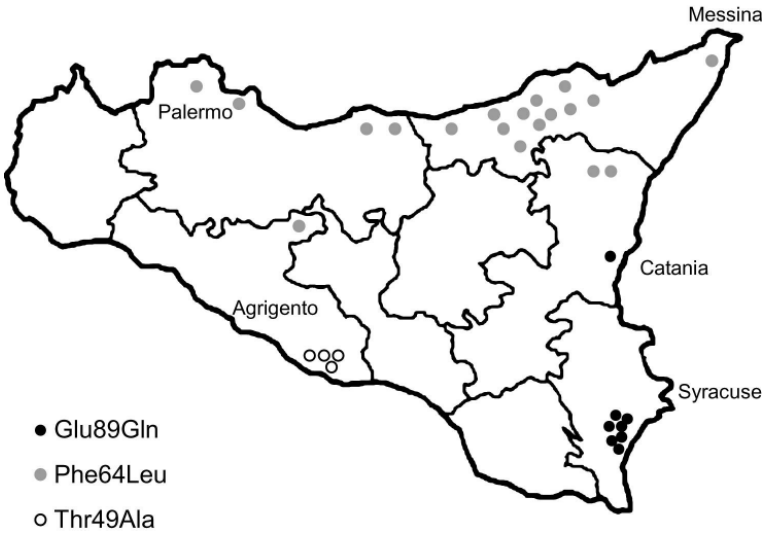


Figure 3.2: Geographical distribution of TTR-FAP patients families in Sicily according to different mutations [21].

As far as geographical visualisation is concerned, the authors present a map of Sicily, in figure 3.2 with the individuals presented as points scattered across their municipalities. They group these dots as families related to the presence of the disease and its mutations. The authors further statistically analyse each mutation relative to the families denoted on the map. They state in their discussion that the prevalence in the Sicilian region of 8.8/1M inhabitants is significant, although lower than in the Portuguese endemic regions.

They conclude that there is a correlation between patients from more than 30 families geographically distributed in 3 of the largest areas of the island (Messina, Agrigento and Syracuse, according to the clusters of points on the map).

A very interesting insight about the cases of TTR-FAP in the regions of Póvoa do Varzim and Vila do Conde, in the north of Portugal, is presented in [22] which are, to date, pioneer locations for known cases of the disease.

In 1952, the authors report that 84% of the patients belonged to these areas. Another interesting fact is that at the time of writing (1995), 35% of the patients originate from these two locations. 76% of them had lived in this area until the diagnosis of the disease. A peculiar fact of this study reveals that the fertility of affected women is on average 3.7 children. They also refer that this population, until the beginning of the medical analysis of the disease, was by habit quite reserved and showed a predisposition to marriages of similar social classes (in this case, fishermen, who have already been mentioned in [3] as one of the primary professions of the disease at the time).

The authors also present figure 3.3 , which analyses the area of origin and residence between 1939 and 1992, in relation to the total number of patients per municipality.

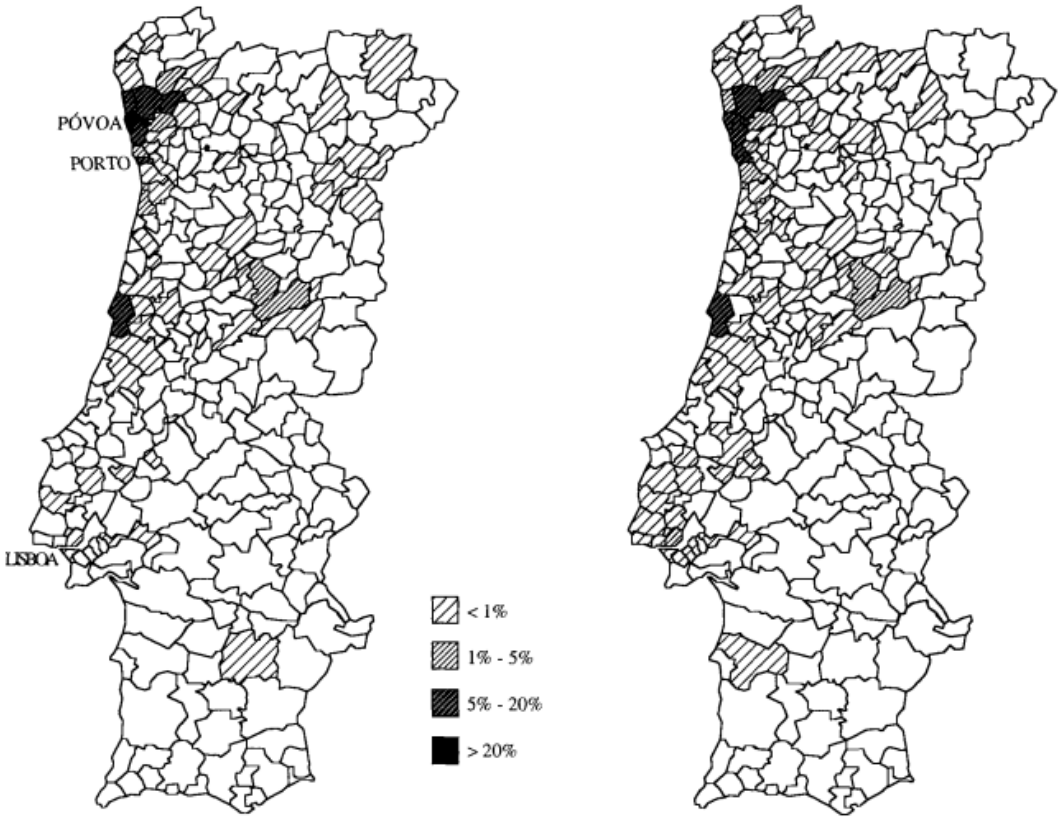


Figure 3.3: Area of origin (left) and area of residence (right) of all patients diagnosed at CEP between 1939 and 1992 [22].

In these maps it is already possible to understand that the great concentration of cases is in

the area of Porto and Póvoa do Varzim, with some numbers still in Figueira da Foz (Coimbra), Covilhã (Castelo Branco) and Seia (Guarda). Visually, the maps look quite identical and perhaps the biggest difference is that the districts further north of Porto are no longer the places of residence of the patients, who prefer municipalities further south.

The study reveals that due to the symptoms of the disease containing impotence, the fertility in a higher number of women may be a consequence of the lower impact this symptom has in patients with a higher age of onset.

In study [23], the authors elucidate on the current epidemiological situation of TTR-FAP in Japan, one of the 3 countries with the highest focus of cases. The study was conducted between 2003 and 2005 with data collected annually.

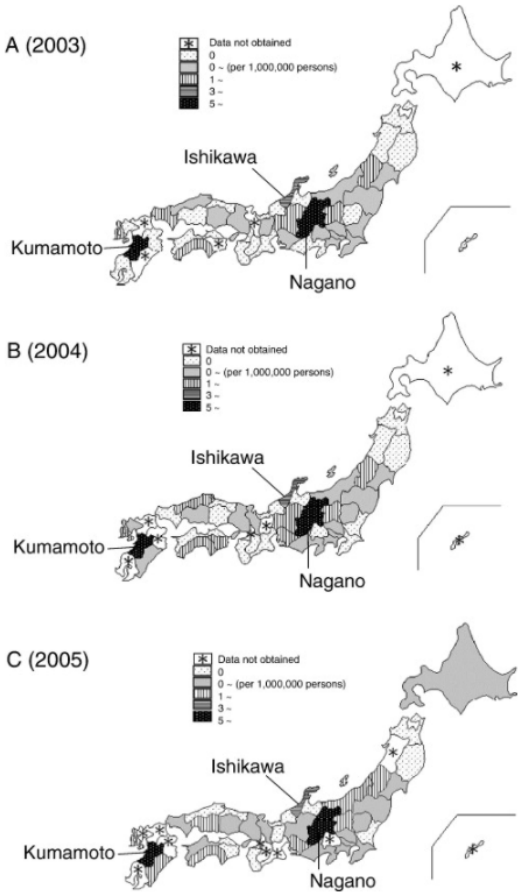


Figure 3.4: The prevalence (per 1,000,000 persons) of familial amyloidosis corrected by data availability during the period from 2003 to 2005 in Japan [23]

In the demographic results obtained by the authors, there was an outbreak in three distinct areas of Japan and the prevalence of the disease changed little or not at all over the years, with Nagano, Kumamoto and Ishikawa being the endemic foci, as seen in figure 3.4 . Even so, the country showed a wide distribution throughout its territory.

The authors presented the data visualisation by prefectures regarding the prevalence of the

disease, annually. Patients in Japan had the same characteristics as in Portugal in 2 out of 3 foci.

Finally, as in study done by Mazzeo el al. [21] and the Italian geographical representation, the authors in [24] also did a clinical study on the geographical distribution of TTR-FAP in South Korea between 1995 and 2014.

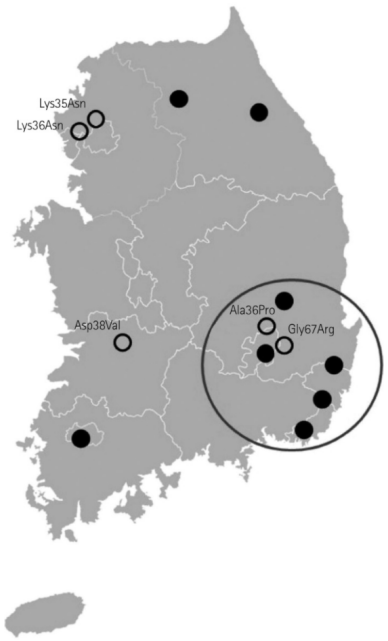


Figure 3.5: Distribution of TTR-FAP gene mutations in Korea (Total 13 families) [24].

The study was mostly medical but the authors present a geographical visualisation of the mutations with the points scattered in space, as shown in the figure 3.5, where the largest cluster is in south-east Korea.

After the analysis of these articles, it was possible to conclude that the clinical status of TTR-FAP is under constant review and study, but an in-depth investigation of the geographical distribution and evolution of the disease has not been done. Besides this, the demographic representation and exploration of different variables in all the articles is done in a shallow perspective.

Therefore, a possible exploration of the data in a geographical manner may prove to be fundamental and innovative in the research fields of the disease, correlating clinical with statistical data.

3.2 Research and theory regarding geographical and cartographic visualisation

This section aims to clarify the study done on the theory of visualisation from a geographical point of view. Thus, by analysing articles and bibliographic references on the subject, it will be possible to make decisions in the future data analysis presented with properly identifiable arguments. Some of these concepts may prove to be crucial in the use of certain methods, tools and visualisations so that their presentation is duly justified and appropriately portrayed.

In [25], the author makes a very detailed analysis of the concepts of geographic visualisation. Firstly, he refers to a definition provided by the 2001 research agenda of the International Cartographic Association Association (ICA) Commission on Visualisation and Virtual Environment which reads as follows: "Geo-visualisation integrates approaches from visualisation in scientific computing (ViSC), cartography, image analysis, information visualisation, exploratory data analysis (EDA), and geographic information systems (GISystems) to provide theory, methods and tools for visual exploration analysis, synthesis, and presentation of geospatial data". Secondly, it mentions other less technical references but with another separate point of view. These other definitions seek to state that the use of visualisation tools should be directed to a problem in order to find solutions, with special attention to the human capacity to interpret and extract knowledge from its results.

After this, a motto is presented to be followed when the objective is the search for knowledge through visualisations: exploration, analysis, synthesis, and presentation. This methodology defines the timeline of the work, exploration of the data and its analysis, summary of its characteristics and presentation to the public. Following this methodology, it is said that there are 3 "driving forces" commanding the evolution of geospatial visualisation and they are the technological and graphic development, the exorbitant amount of spatial data collected today and the dissemination that the Internet causes in the expansion of this scientific field.

Nöllenburg explains two concepts relevant to the mental processing of information. The first, which he calls Visual Thinking, specifies that the goal should not be the generation of images with high computational can, but the focus on the generation of images capable of generating new ideas by the people who are interpreted. This method implies that it is possible to capture patterns in visualisations that can be analysed by a human. The second, Graphic Variables, which recalls how location, size, density/size of texture elements, orientation and colour hue, saturation, shape and value can be different means of communication with the map reader.

In the following chapters, we are introduced to different visualisation methods and techniques that will introduce some of the differences and purposes of each in order to understand their place in the catalogue of visualisations. The types discussed involve the processing of Geo-spatial Data that is characterised by its georeferencing nature. This information is divided into 2 dimensions (latitude and longitude) or into 3 dimensions (the previous ones including altitude). These are:

- 2D Cartographic Visualisation - This is the most common method of representing geospatial data where a map relating to a defined area is presented through its spatial variables. On the one hand, the author mentions the importance of representation in cartograms that allow geographic representation of a space in which the proportion of its spatial subdivisions are relative to the data. These cartograms demonstrate a greater facility for the map reader to be able to focus on relevant locations regardless of their territorial dimension, resulting in better conclusions about social, economic and political problems.

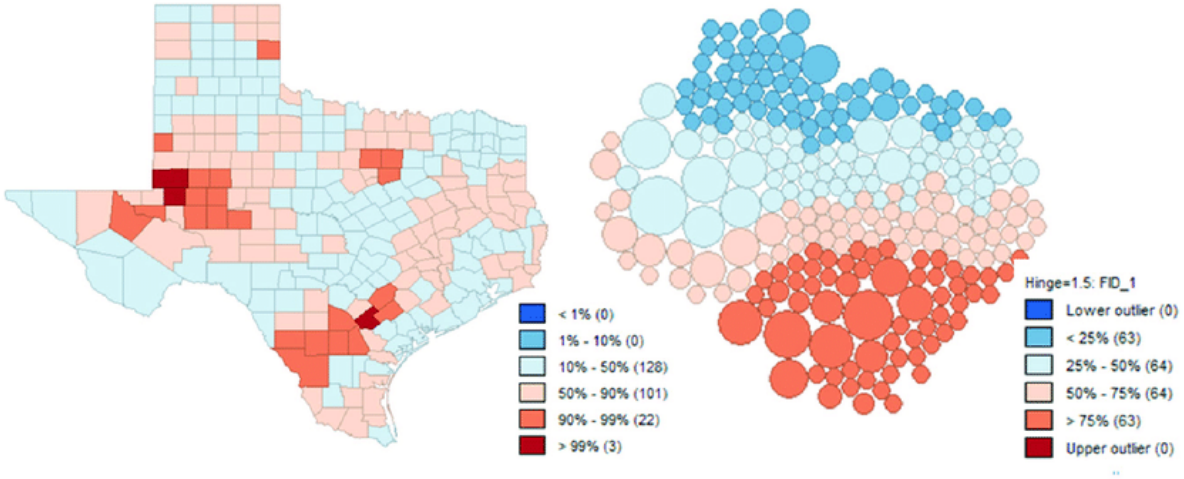


Figure 3.6: Choropleth and Cartogram maps of numbers of fracking sites by county, Texas. [26].

On the other hand, he mentions the importance of Choropleth maps which, through the manipulation of colours and textures, represent peripheries, districts and other territorial divisions. For the attribution of these values, it also highlights the importance of the ability to order the data.If the data is sortable (whether nominal, logical or numeric), then the colour systems must demonstrate the sorting explicitly. Both views are shown in figure 3.6.

- 3D Cartographic Visualisation - The use of more advanced technologies is the next step to evolve GVis into another dimension. The document highlights that human perception, living in a realistic 3-dimensional environment, adapts and stimulates much more easily to representations of this type. Nevertheless, there is a trade off in representation with these new technologies: while on the one side more delicate and detailed representations are obtained, on the other side the capacity for abstraction that distinguishes map readers and, consequently, the unique and distinct conclusions of each one are lost.

  As an example, the author mentions that sometimes the use of only two temporal variables and one time variable are more beneficial than three spatial variables. One of the adversities of this type of representations is the ease with which the user of these softwares can lose orientation, needing to re-orientate himself too often in his spatial analysis.

- Visual Data Mining (DM) Tools - Visual DM, as a synonym of Exploratory Data Analysis (EDA), emerges as the concept of establishing a connection of DM techniques, such as

algorithms, to detect knowledge and patterns. The author divides these 3 categories, two of which seemed to be of greater relevance:

- Geometric Techniques - Such as scatter plots and parallel coordinate plots (PCP) . Scatter plots are of a popular use, while the others are more infrequent and allow to display high-dimensional data in a single representation. With this type of visualisation it is possible to identify behaviours and patterns by linking variables and make comparisons between similar entities. One problem with this type of representation is the over plotting of data that makes it eligible due to the amount of information and variables. Note that the order of the variables presented may affect their interpretation.
- Graph-Drawing Techniques - Expressing data relationships through links between elements and nodes can be advantageous. The author gives very enlightening examples as the representation by graphs for GVis, such as the representation of a metro line or streets of a city through a graph that remains in the memory of the reader. The author also presents a representation of this type of graphs in the figure 3.7, using the nodes and edges.

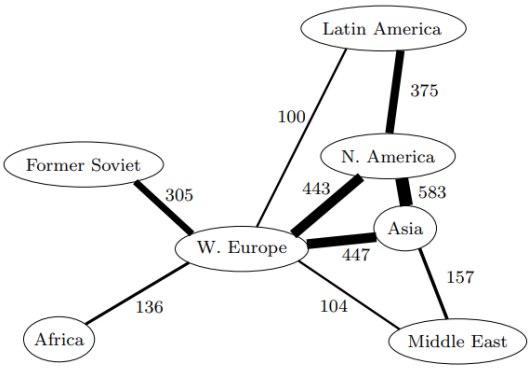


Figure 3.7: A graph showing trade relationships. Edges are weighted by trade volume and drawn shorter and thicker with increasing weight. Only edges with at least 100 billion dollars trade volume are shown. [25].

- Animation - The use of animation allows the use of the third dimension (time) to create geographical visualisations such as map movies and animated maps. This kind of new paradigm adds a special aspect to the visualisation: time differences are seen clearly. For example, the moment of an earthquake may become clearer in a video due to the frames. One of the disadvantages of animation is the possible loss of details in the passage of time, which escape human perception. The control of variables is said to be one of the greatest benefits to be added to this type of representation.
- Spatio-Temporal Visualisation - The author states that spatio-temporal data should be treated according to its characteristics in which are the appearance and disappearance of attributes, differences in location, shape and size, as well as qualitative and quantitative differences in the data. The author gives examples of different visualisations used to

demonstrate data in 3D form such as the space-time cube and shows, as in figure 3.8, how data can be modelled on a screen depending on its trajectory, a feature that is not possible on 2D maps as there is no perception of velocity.

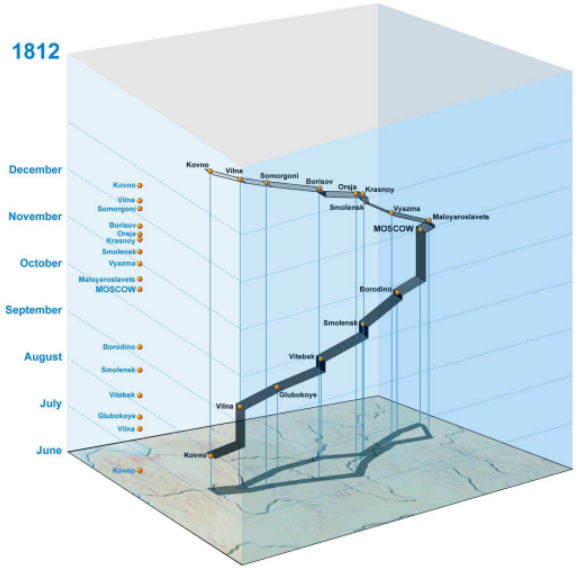


Figure 3.8: A space-time cube visualisation of Napoleon's march in Russia, Image courtesy of M.-J. Kraak. [25].

The author also gives examples of Change Maps, which use Choropleth maps at different points in time. He also warns that instead of presenting a map that covers all data as a whole, the strategic division between time series makes it feasible, as much as possible, to represent the data temporally, without ever forgetting that there is always a loss of scrupulous information in the creation of time intervals. Finally, one strategy was to divide the land into territorial divisions, so that each division was assigned another data visualisation over time that referred only to that part of the land, as in figure 3.9.

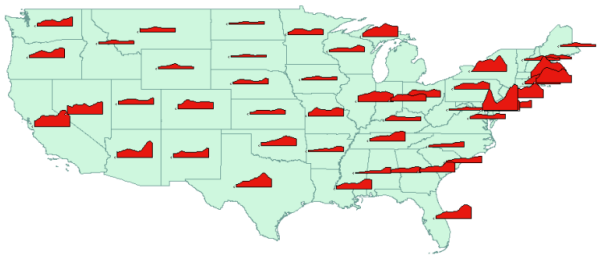


Figure 3.9: Cartographic representation of the spatial distribution of the burglary rates in the USA. Image courtesy of G. Andrienko [25].

- Interactive User Interfaces - Here the author explains how important it is that, when developing a tool for user interaction with visualisation, it should have the same mechanisms as a camera i.e. the ability to manage perspective, magnitude and detail. In a complementary way, it should be possible to control the dimensionality of the attributes,

when it is meaningful.

The author, besides reinforcing the colour palette used to express data patterns, mentions that the potential of GVis lies in the representation of multiple views of the same data in the same dashboard, for comparison purposes. This means that changing the dimensionality or attributes of one view also affects another and generates a multiple impact that can lead to the generation of patterns and conclusions.

- Combining Visual and Computational Exploration - Finally, the author's words are that DM and machine learning are able to analyse large volumes of data in order to automate knowledge extraction, while other tools are optimised for data visualisation and that both should complement each other in order to solve the problem with their distinctly interconnected tools. Among these collaborations, k-means clustering together with self-organising maps (SOM) are mentioned.

In the end, the author ends with a presentation of some geographic visualisation tools already developed among which the author highlights ArcView, XGobi, Cartographic Data Visualizer, CommonGIS, GeoVISTA Studio.

In article [27], the authors focus on describing GVis as something broader that involves all methods related to large spatial datasets and where this data is not so specifically caged. The authors refer to the importance of distinguishing GVis from cartography, in what is the critical need for prior investigation, be this theoretical, conceptual or methodological.

After analysis, the authors share the conclusion that due to technological developments, this is reflected in a greater amount of data being available. As a result, there is more information to be visualised, more methods, tools and methodologies to visualise it as well as increased technological capabilities needed by data scientists. Obviously, more preparation for data handling will translate into better results.

Finally, the authors list some of the challenges that threaten to arise in geographic visualisation research, in the short, medium and long term.

In the short term, the authors list the interactivity between the users and the data to be represented in the visualisation, cognitive problems related to the need for knowledge about the problem, the computer interface design and its importance with the user, since it is the main contact point between the developers and the viewer. The authors also warn about the importance of considering the type of format of the visualisation, the relationship between different users for the same visualisations (in their goals and differences), as well as the dimensionality of the visualisation, dynamism and animation.

When obtaining some visualisations, in the medium term other challenges arise to consider. This involves, for example, considering different sensory levels, as well as their limitations and variations, such as in colour and sound. In addition to this, there is the challenge of maintaining the effectiveness of the visualisation with new data values or more or fewer pixels while still

producing real or abstract visualisations. In the long term, there is the challenge of creating new creative visualisations as well as the creation of automatic visualisation systems.

Complementarily, to bring together concepts, methods and tools applied to problems solved with visualisations was the goal for authors in [28]. Initially, they start by mentioning how the difference between static and animated maps can bring advantages namely in the discovery of "'peculiar oscillation', in one of the studies they discuss, and how animation, by being able to incorporate the time variable, facilitates its natural human perception. They also criticise timeless maps for being unable to incorporate patterns temporally.

The authors also refer to four types of time: universe time which is absolute and linear, cyclic time which represents, for example, diurnal patterns, ordinal time which concerns the natural order of events time as distance using a spatio-temporal dimension. This type of temporal analysis is applied directly to the representations. It refers, further on, to how world time (days, years, centuries, typically) is converted into animation time (usually seconds). Examples of temporal animation include the growth of populations and the spread of diseases.

Throughout the document, it is pointed out that animated maps, also known as movie maps or change maps, are mostly used to demonstrate geographical changes and patterns. This animated form of visualisation, by increasing or decreasing the execution time, directly changes the amount of data to be displayed. If it is too long, the greater the detail but the more difficult it is for the user to understand all the frames. If it is too short, little knowledge will be extracted from the visualisation. They also mention the importance of the existence of a time scale, as in figure 3.10 in parallel to the spatial scale on static maps.

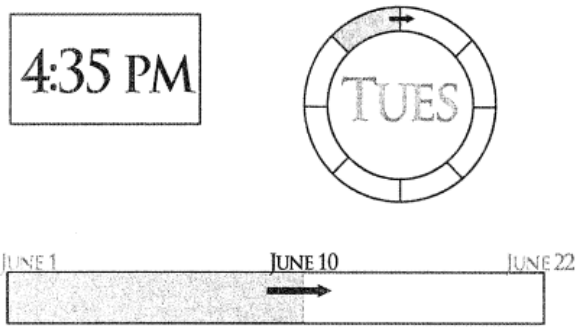


Figure 3.10: The three most common kinds of temporal legend - digital clock, cyclical and bar. The graphical cyclic and bar legend can communicate at a glance both the specific instance and the relation of that moment to the whole (source: M. Harrower)[28].

On the flip side, the disadvantages of using this type of visualisation are also mentioned. The use of animation is not always justified, especially when the process of changing data is instantaneous and not over time (e.g. buying a house). Most important events at short time intervals will be better represented in data tables or static maps. Although chloropleth maps work very well up to +- 7 chunks of information at the same time, one of the open questions is how many data frames per time unit should be displayed in order for the user to retain it. The

authors also explore in detail this topic, namely in the aspect of animation comparison between different views and have a discussion about the overload of micro interactions in animation. Finally, they conclude that visualisations should, above all, have the user's cognitive aspect in mind.

In [29], the authors define GVis, the concept that extends from traditional cartographic in two ways: through the way it builds knowledge (in addition to the way it presents it) and in the way it enables the dynamic display of data and its changes. The authors reinforce the idea that the animation chosen for the map, a multivariate representation that allows dynamic display in "small multiples" (small distinct visualisations in the same panel) and an interaction with the user through the adjustment of variables. An example of small multiples can be seen in figure 3.11 and there was already a representation of it in [23], when studying the TTR-FAP behaviours in Japan, as in figure 3.4.

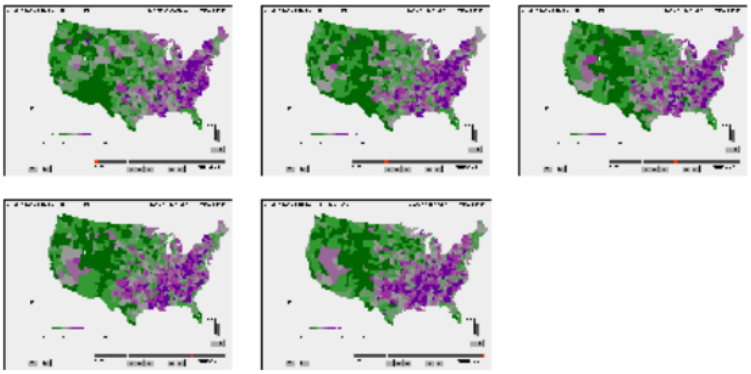


Figure 3.11: Small multiples depicting the 5 time steps for heart disease mortality rates [29].

The authors developed a prototype data mapping visualisation. Throughout their analysis they prototyped it in such a way that it was possible to study various subsets of data over time as well as examine temporal sequences and make comparisons through time. The authors conducted a study for the application of the tool related to cardiac problems. In their conclusions, there were professionals who were able to extract knowledge from the animated visualisation that others could only extract later. An example of visualisation through their panels can be seen in the figure 3.12.

Since exploring the evolution of digital cartography and the steps towards a better historical understanding of its paradigms is important in order to ensure a better use of knowledge, Maceachren and Kraak present a good discussion in [30]. They eventually introduce to the idea that visualisation is a balance between visual thinking, knowledge construction, communication as information transfer, the public nature of data and the level of interaction. All of these narrow down into a presentation of knowns and revelation of unknowns.

The focus of the speech is on the importance that, different datasets, when assessed independently, are easily understood but their relationship is not so easily perceived, hence the work of visualisation and its environment. The quality of pattern extraction in its presentation is what will answer the questions and solve the problems. This quality is directly influenced by

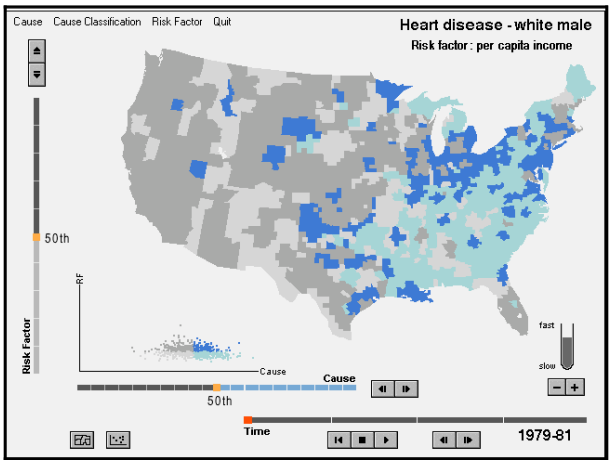


Figure 3.12: A sample display layout illustrating the full set of available controls applied to a bivariate "cross" map. On each cross map blue is used to indicate higher mortality rates and dark shades depict higher values for the risk factor [29].

tools for filtering, detailing and noise removal. Moreover, it proves more useful to use less detail which translates into higher abstraction and effectiveness.

The authors also elucidate the directions to follow in cartographic visualisation, which involve:

- Investigating the implications of changes in the use of tools for visualisation.
- Designing a conceptual model for tools, visualisation and data quality.
- Investigating the implications of changes in the use of tools for visualisation. Designing a conceptual model for tools, visualisation and data quality.
- Understanding decision making through the outputs of the tools in which GIS (Geographic Information System) fits and understanding which visualisations serve each problem.
- Investigate different solutions, such as 3D, for solving the same problems, as well as other methods and tools.

3.3 Studies on other diseases with epidemiological behaviour

This section aims to analyse and understand which methods, tools and procedures have been used by other researchers in the epidemiological study of similar diseases. Since TTR-FAP is only one of many active diseases in humankind, there will certainly be data processing mechanisms that will serve as a basis for the future management of the disease data in this dissertation.

Vindenes and other investigators [31] focused on one type of tuberculosis with data from 2012 to 2015 and aimed at a demographic exploration of the data through a clustering criteria. They combined genetic and spatiotemporal data to predict whether individuals were more likely to be in a cluster. This disease has a high impact on the socio-economic life of those affected and the data analysed varies from a wide range of personal information, from generic data such as age and gender to data such as employment drug use. These authors defined clusters as cases with less than 3 years of detection between them and within a 50km radius.

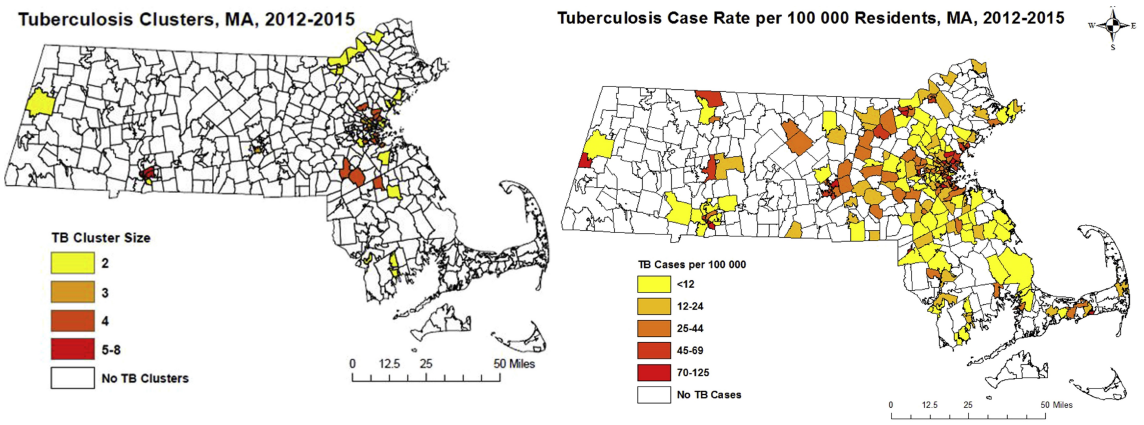


Figure 3.13: TB clusters by residence at diagnosis in Massachusetts, 2012–2015. Two or more cases with identical genotypes, which were close in time (3 years) and geographic space (<=50km [km]), were defined as TB clusters (left).Tuberculosis case rate per 100,000 Massachusetts residents, 2012–2015 (right)[31].

In addition to the geographical exploration in territorial areas in which the number of cases per 100,000 inhabitants was evaluated over time, they also present the geographical area with the number of clusters by territorial division, as in figure 3.13 as well as a record adjudication diagram seen in figure 3.14 (also called study flowchart), the number of clusters by size, and statistics on cases by race (which ultimately has an impact on the conclusion).

In general, from this epidemiological study on tuberculosis, three important concepts can be mentioned: the study and correlation with secondary information that at first would not be related to the data but leads to interesting conclusions, the use of an initial diagram to explain, expose and describe the cases analysed from the raw data until its treatment and the clustering of the data in the dataset into groups that have similar characteristics.

First and foremost, as an example of the cross-referencing of medical data with secondary

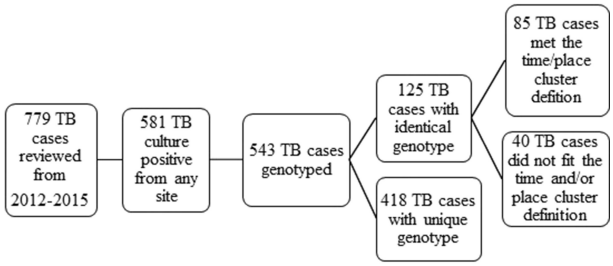


Figure 3.14: Overview of the TB cases reviewed for identification of clusters in Massachusetts, 2012–2015 [31].

data, Prof Jean Gaudart and the other researchers [32] analysed the first wave of cases of COVID-19, the most recent pandemic on a global scale, in relation to the French population and how the disease behaved epedimologically. The data, from the months of March to May 2020, was cross-referenced with French hospitals and climate data and, for each of the territorial divisions, the incidence, mortality and case fatality rate were studied. They used, as have other studies, Moran's I statistic [33–35] and Gaussian kringing smoother to correlate geographic coordinates, multivariate generalised additive models using DAGs (directed acyclic graphs) to relate secondary data and adjusted Moran I statistic with Pearson residuals (all using R and some of its packages).

Of particular relevance is the wide variety of small multiples that are presented in figure 3.15 showing cases and deaths per 100,000 population, deaths per 100 cases, age ranges, climate classes, nature of urbanisation, economic status, health services, lag between first cases, as well as beds and pharmaceuticals per 100,000 population.

As another example of data correlation, the paper [34] assesses the risk of Cutaneous Leishmaniasis in a state of Brazil. The study applied statistical and visualisation methods and correlated data on seasons, characteristics of infected people and deforestation. Clusters were identified and the non-homogeneity of the disease in its distribution was concluded.

This endemic disease was also assessed taking into account the presence of health centres and basic sanitation in the vicinity. The authors used two types of techniques: Local Moran's Index and Kernel density estimator to explain the distribution patterns of cases in areas up to 300 meters which resulted in case density maps, shown in figure 3.17. One of the statistical visualisations that is presented is an overlay of the precipitation and clusters existing during the year, as well as deforestation, as in figure 3.16.

Other instance regarding data correlation was done by Angelou, Kioutsioukis and Stilianakis [36], were they focused on a West Nile Virus in Greece and its climate-related epidemiological transmission risk. They applied models to study the number of infected humans and the weeks of incidence of transmission, with relevance to variables such as precipitation, temperature, presence of parasites, humidity, soil water and wind speed and their impact on the prevalence of the disease. Case outbreaks were also studied and were done at the municipality scale of the

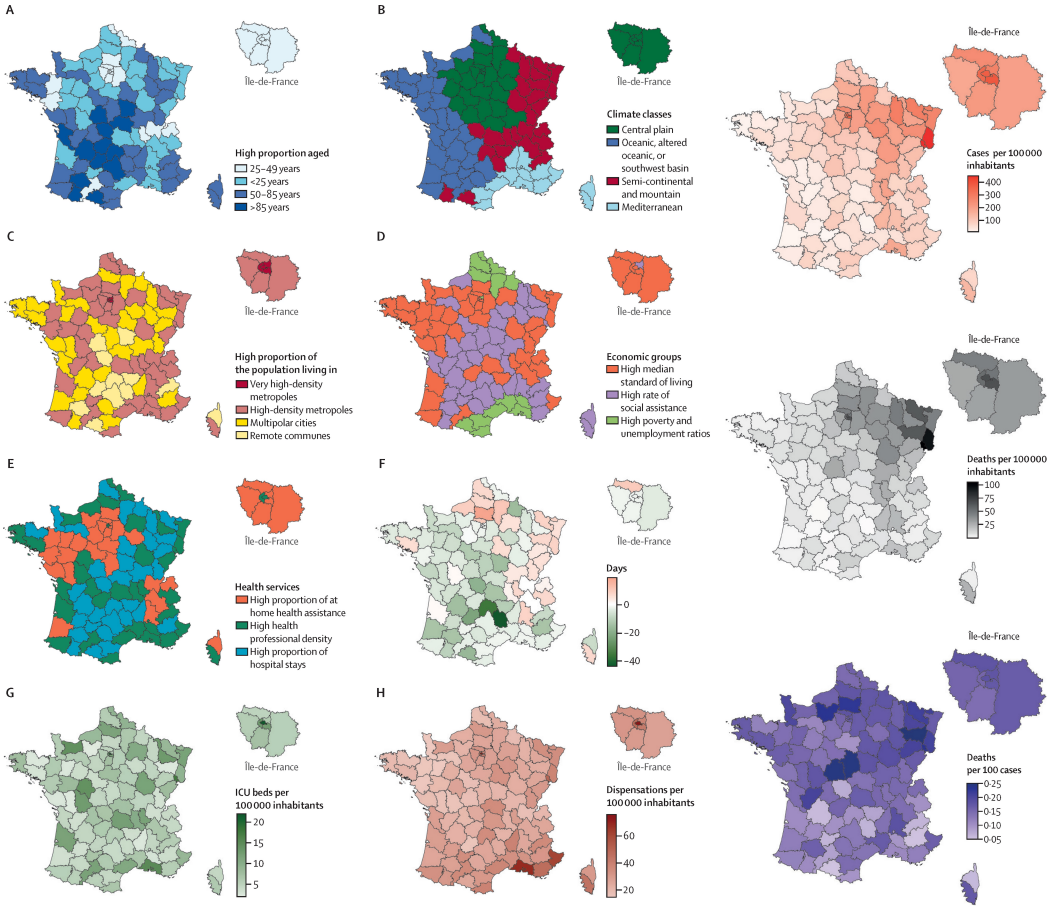


Figure 3.15: Maps of covariates, showing population age structure (A), climate classes (B), urbanisation (C), economic profile (D), population health and health-care services (E), the lag between the first COVID-19-associated death and lockdown (F), baseline intensive care capacity (G), and chloroquine and hydroxychloroquine dispensations in pharmacies (H) and Spatial heterogeneity of COVID-19 in France, showing cumulative in-hospital incidence, in-hospital mortality rate, and in-hospital case fatality rate (last column)[32].

territorial area. The spatial data of each municipality was taken from Google Maps and its API. Among the visualisations presented, it is possible to highlight the mapping of municipalities by years with observed cases and also by range of number of cases.

Secondly, the record adjudication diagram, which has already been mentioned in [1], is a common component in various epidemiological analysis papers, in order to portray the entirety of the data to make clear the exclusions, removals and disregards of some its elements, as is the case in [31], [33] and [37].

The article [37] shows a spatial and epidemiological analysis of visceral leishmaniasis in northern Brazil and intended to study the impact of improvements in education, house conditions and nutrition on the incidence of the disease, between 2007 and 2017. The data on the general population of the municipalities was taken from the official Brazilian database and the count of cases is presented in a record adjudication diagram. The most important visualisations presented

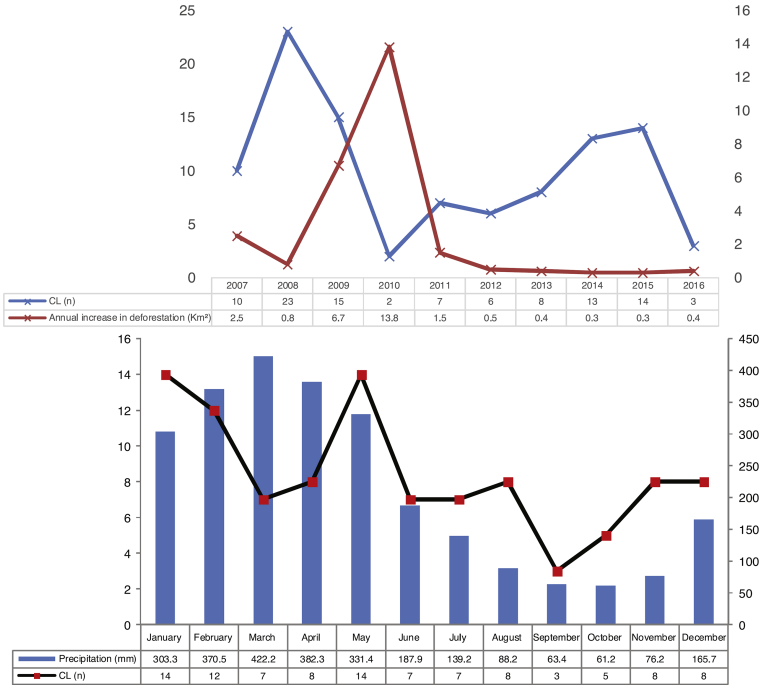


Figure 3.16: Annual analysis of American Tegumentary Leishmaniasis cases in relation to the rate of deforestation, 2007 to 2016, Cametá, Pará, Brazil (above) and Monthly analysis of cases of cutaneous leishmaniasis related to precipitation level, 2007 to 2016, Cametá, Pará, Brazil (below)[34].

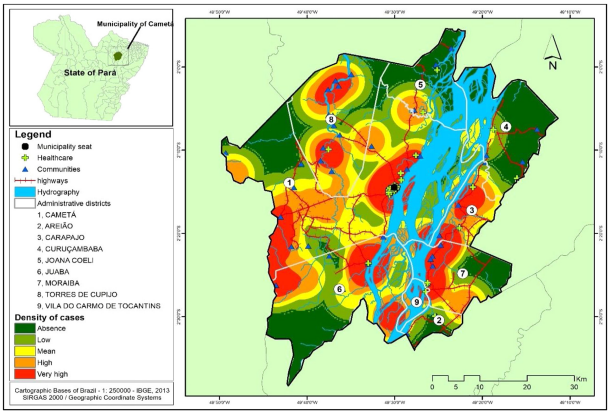


Figure 3.17: Case Density of Cutaneous Leishmaniasis, 2007 to 2016, Cametá, Pará, Brazil [34].

in this study show the odds ratio and the incidence per 100 thousand inhabitants, as well as the lethality rate. This research showed a classical view of the clinical study of the disease and a simplistic exploration of mapped statistical data, with some emphasis on the local analysis of cases.

Thirdly, and regarding clustering, Yang and the other authors [33] aggregated spatial, genomic and epidemiological data to study the behaviour of tuberculosis in Shanghai, China. They proceeded to identify clusters of the disease that share the same patterns, the relationships of proximity of residence and risk of contracting it, and the correlation with the migrants of the

cities.

After studying the background of the disease and analysing it, they suggest that the results are explained by the chain of transmissibility. The number of cases is again presented through a record adjudication diagram as in figure 3.18. Along with this, a geographical map of the city of Songjiang in Shanghai is presented, with a circular graph overlapping each territorial division to represent the presence of positive cases in natural and non-natural persons from the city. This type of representation is identical to figure 3.9, where information on specific territorial areas is expressed in a single image, in a global overview of a larger territory, despite not using time.

Some of the methods used for predictions include multi variable logistic regression to test hypothesised associations of spatial proximity and genetic similarity. The authors considered a minimum number of 4 cases to visualise clusters in their dimensions and one of the most relevant visualisations contains the mean notification rate and kernel density estimation of cases by territory for migrants and residents. Finally, there is also a representation of the clusters on the map in figure 3.19, with at least 4 elements, which clearly translates the location of the data on the map. This paper combined two of the points mentioned earlier by Vindenes in [31], the use of the diagram and use of clustering.

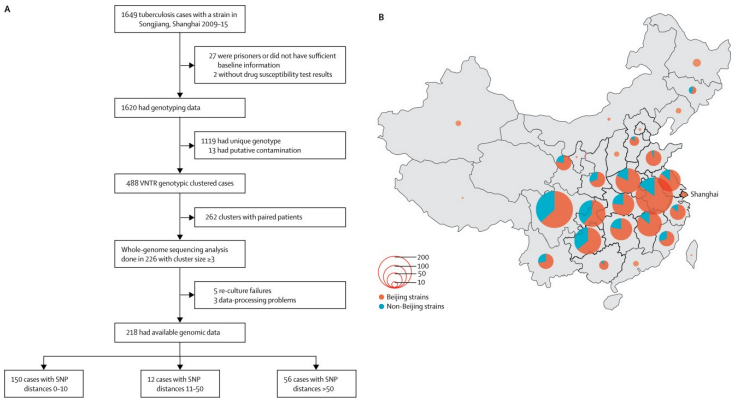


Figure 3.18: Study flowchart (A) and distribution of provinces of origin for migrant patients with tuberculosis in Songjiang, Shanghai, 2009–15 (B)[33].

Sequentially, Jing and other reasearchers [38] worked with patient data from 2004 to 2018 regarding Acute Hemorrhagic Conjuctivitis in China. The work used descriptive statistical methods to characterise the disease and epidemiology. The incidence of cases at the district level was analysed, spatial correlation was studied to identify hot spot and cluster regions and spatiotemporal patterns, and it was concluded that the northern region was the epicentre of clusters.

The authors performed a statistical and demographic analysis of the data in order to characterise the disease epidemiologically, an annual visualisation of the incidence of cases through geographic maps and identification of peak cases, the study of spatial auto correlation with the Global Moran's I statistic and High/Low clustering, as in figure 3.20 and, finally, the representation of the main areas of clusters.

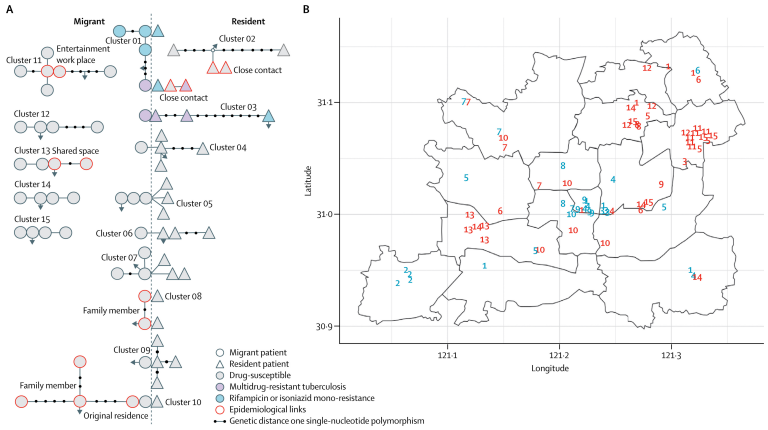


Figure 3.19: Genetic clusters with at least four patients (A), and spatial distribution of genetic clusters (B)[33].

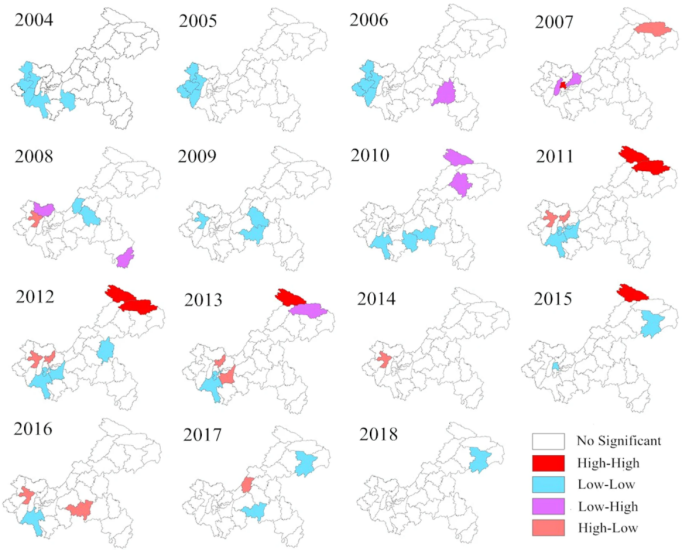


Figure 3.20: Annual local spatial auto correlation of AHC in Chongqing from 2004 to 2018 [38].

Given this, it is necessary to clarify that a correlation between two variables is a statistical relationship that represents their association. On the other hand, auto correlation is a correlation between two values of the same variable in two or more temporal instances. Smith [39] defines auto correlation as a "mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals". Spatial auto correlation is a broader term that involves a greater number of variables due to the use of spatial, temporal and statistical data.

The article also points to the importance of the methods used: profiling of the location regarding climatic and geographical conditions, the analysis of spatial auto correlation through GISA and LISA, Global and Local [40] Indicators of spatial association and the use of Kulldorff's method for Scan Statistics. It is important to mention that LISA Maps use important terms used by the authors:

- high-high (high-incidence regions surrounded by high-incidence regions, which are highly epidemic regions)
- low-low (low-incidence regions surrounded by low-incidence regions, which are lowly epidemic regions)
- high-low (high-incidence regions surrounded by low-incidence regions)
- low-high (low-incidence regions surrounded by high-incidence regions)

Likewise, Pordanjani and the other investigators [35] decided to analyse the data from 2006 to 2014, of Acute Lymphoblastic leukemia disease in Iran, and understand its epidemiological environment. The work performed focused on spatial auto correlation and clustering, identification of disease hot spots and cold spots, understanding the impact of geoclimatic conditions on disease incidence and comparison between local and global regression models applied.

To study this type of paediatric cancer, the authors mention the importance of considering spatial epidemiology and find that observations are not simply individual, but are a result of dependencies on other spatially auto correlated observations. This suggests that the values from spatially proximate observations are more identical than distant observations. Still, it warns that each region has a different type of dependencies and a global association may not explain all the data, as spatial data are largely heterogeneous. In order to keep these concepts in mind, they used the GWPR, geographically weighted Poisson regression model. The study also applied Global Moran's I to its auto correlation procedures and used OHSA, optimised hot spot analysis, for cluster detection.

Among the exposed visualisations, small multiples of cumulative incidence rate and cold spots/hot spots (correlated with altitude and longitude) stand out, as shown in figure 3.21.

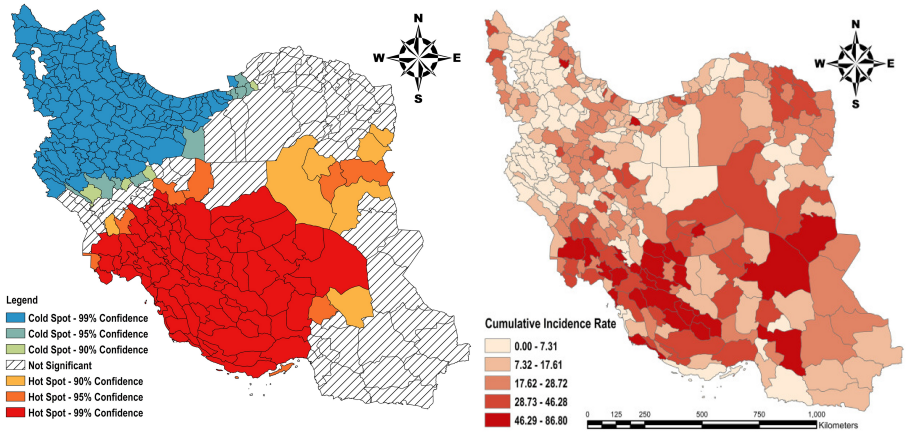


Figure 3.21: Hotspots and coldspots of ALL incidence in Iran at 90%, 95%, and 99% confidence level (left) and Cumulative incidence rate of ALL by counties in Iran during 2006–2014 (right)[35].

3.4 Summary

In this chapter we reviewed a lot of information related to three different topics. The first part gave an overview of the clinical and geographic situation in Portugal and abroad related to TTR-FAP. The second part was a literature review that extended the fundamental concepts of GVis and what to have in mind when using tools that are related to spatial data. Finally, we also had a literature review about other diseases that used correlation of data, visualisations and spatial data in their work. All of this information, allowed this dissertation to be more educated in what are the concepts used in this kind of work.

Chapter 4

Data Preparation and Preprocessing

As previously stated, an aim of this dissertation is to develop Geovisualisation (GVis) techniques and tools for the study of Transthyretin-associated Familial Amyloid Polyneuropathy (TTR-FAP) disease in the national environment. The data came from the specialised unit of the Hospital Santo António, located in Porto, Portugal, which provided an anonymised dataset of patients and relatives related to these patients, all belonging to their family trees, providing a lot of correlated data. This chapter is an overview of the work plan and shows the preprocessing of the data to make it clean, as complete as possible and usable for a greater efficiency in the application of techniques.

4.1 Primary Dataset Insights

The health unit from which the data comes from is also a centre specialising in TTR-FAP, which highlights that the data with which results will be shown has a wide credibility. The unit provided a Comma-separated Values (CSV) with information on patients and, in addition, information on directly related individuals in the patients' families. The information in the CSV can be found in detail in the tables 4.1 and 4.2, where the attributes and values of each record are summarised.

Despite the information in the table, there are some attributes for which some considerations should be made. First, there is no clear unique identifier for each record and for that reason it has been added in order to treat each record individually in future operations and data processing techniques.

Secondly, it is important to mention the heterogeneous property of values present in the attribute genetic_symbols . Since it represents different possible states in which a record can be found in the database, it takes different forms. The individual may be affected with TTR-FAP or may be clear which indicates that he/she does not carry the disease genes. On the other hand, he/she may be unaffected which means that he/she is asymptomatic or a carrier of the genes. A heterozygous individual is a carrier who received the gene from both parents. People classified as possibly affected are considered to be possibly symptomatic, while all other values are considered

Table 4.1: Initial CSV information 1/2

Attribute name	Type	Example	Insight
id	Double	73	The id identifies the position of the individual in the family tree to which it belongs.
fam_file_path	Character	fams\100-199\100.FAM	This attribute shows the file path of the family of each register.
date_of_birth	Character	1957000	Date of birth for each individual with year (4 numbers) and day count (3 numbers).
date_of_death	Character	unknown	Date of death for each individual with year (4 numbers) and day count (3 numbers).
sex	Character	female	Sex of each individual. Male, female or unkown.
mother_id	Double	63	This identifies the mother of the individual by her id in the family tree.
father_id	Double	22	This identifies the father of the individual by her id in the family tree.
number_of_individuals	Double	1	Number of individuals the record refers to. Sometimes when there is no correct information about birth and sex of different individuals, they are grouped in the same record.
genetic_symbols	Character	affected	Last known state of the individuals (affected, clear, unaffected, carrier, heterozygous, possibly affected or errors.
adoption_type	Character	adopted_out_of_family	Adoption status of the individual.
proband	Double/Binary	0	This binary attribute identifies if the individual was the first patient of the family to be registered on the Medical Unit.

Table 4.2: Initial CSV information 2/2

Attribute name	Type	Example	Insight
annotation_1	Character	hearsay, "9"	Tree related specific annotations.
origem_desc	Character	AVEIRO AROUCA	District and county of origin for each record.
res_desc	Character	LISBOA SINTRA	District and county of residence for each record.
sitdoenca	Double	4.0	Classification used by the healthcare professionals to explain the clinical situation of the patients
anoinicio	Double	2000	Year of onset of symptoms registered.

to be errors for this work.

Thirdly, the attribute sitdoenca is meant to internally classify individuals with respect to their symptoms. Its numerical value can be assigned to symptomatic patients, patients with some symptoms but not enough to be considered clinically symptomatic or asymptomatic patients. Values considered between 4 and 5 represent these values. For this work, all other values are considered errors.

4.2 Data Cleaning and Feature Engineering

The data preprocessing phase that had to be done unconditionally involved a lot of operations since the data had a lot of inconsistencies, from missing data to wrong data, even if belonging to the type of attributes.

The work was carried out in R language, complemented by its packages that allowed the use of specific operations to solve problems. The IDE (Integrated Development Environment) used was RStudio, an open-source user interface that facilitates the handling of R functionalities with menus to maintain the organisation of work. With this tool, all the scripts done and files were easily accessed in a single application.

The primary CSV was introduced in the IDE, as well as other secondary datasets used in more advanced stages of the work with the help of the package readr [41]. With this package the data was introduced in WINDOWS-1252 encoding, single-byte character encoding of the Latin alphabet, since the initial data contained characters specific to the Portuguese alphabet, like the tilde or the character "ç". On the other hand, in order to facilitate future operations and since many packages do not include functions for this type of encoding, the package stringi [42] was

used for character string processing facilities. With the help of the stri_trans_general function, a conversion to Latin-ASCII encoding was implemented.

Working with a dataset with thousands of records may be time-consuming and complex, so there were packages used that proved to be fundamental to solve eventual challenges. One of the most important packages, dplyr [43], was used throughout the work due to its enormous usefulness. This package is a grammar that allows the control and manipulation of data, providing verbs to solve common problems in record management. There are 5 operations that, combined with group_by() allow you to apply various queries:

- mutate() which adds new variables as functions of other variables.
- select() allows variables to be sorted by name.
- filter() allows variables to be sorted by value.
- summarise() which reduces the values of multiple variables to a single summary.
- arrange() which changes the order of the records.

The datasets were used in a tibble structure which allows a better formatting, also more restricted, of the data frames. Allied to this package, the stringr package [44] allows other string operations that, in the case of this dataset, will be very useful since there are many erroneous values. Among the solved problems, the following stand out:

- The fam_file_path attribute, where the families' file paths are included, contained a backlash. This backlash directly affected the encoding in strings, since there are special characters from this backlash together with the consecutive character. For that reason, we proceeded to two solutions. First, the backlash was replaced by another more simplistic character, like the character "*". From this change, it was possible to consider substrings of the attribute and create a new attribute, called famN, which considers only the numerical identifier of the string. An example is the string "fams\1-99\13.FAM" to which the attribute "13" will correspond. This is one example of feature engineering in preprocessing.
- We changed values of the start year attribute from -1 and 0 to NA since this is the default value in R used in most functions for missing values.
- The value of date_of_birth and date_of_death were changed from Character type to a numeric type with NAs. Negative dates, dates of birth or death above the current year, dates of birth from unrealistic years (e.g. year 1000) and records where the year of death was greater than the year of birth were classified as NA.
- An imputation was applied, explained in subsection 4.4 of this chapter, in order to populate the records with missing information concerning the location of residence and origin, as well as the next section regarding dates of onset.

- Another example of feature engineering is the creation of distinct columns in the locations. By default, the attributes origin_desc and res_desc contain, respectively, the district and municipality of origin and the district and municipality of residence of the same individual in the same string. For that reason, combined functions from the dplyr package [43] with the stringr package [44] were applied and 4 new attributes were created. With all of them being composed by characters, it becomes easier to access the district of origin, county of origin, district of residence and county of residence (district_ori, county_ori, district_res, county_res).

- Specific columns were created for the year of birth and year of death. The set of characters for the dates are in a String (e.g. "1963000") format. The first four digits being for the year and the last three for the days of the year (from 1 to 365). Most years have a default of days of 0 ("000" in the last three digits) so the information on years was highlighted for future use.

4.3 Geocoding

Geocoding is the process of converting location information, such as a street or address, into geographic coordinates, such as latitude and longitude. These coordinates can then be used in practical mapping applications. The ggmap package [45], which provides a collection of functions for visualising spatial data and modelling this data on top of static maps from different online sources was used together with the tidyverse package [46], which shares several functions for data structures in Data Science (DS), allowing geocoding to be applied to the data.

As explained above, the spatial data are strings related to the district and county of origin and residence of each record. This means that there is enough information to apply geocoding to the data, since we want to obtain the latitude and longitude of each record to proceed with the application of GVis tools.

Firstly, a unique dataframe was created that gathered the union of unique origins and unique residences. This way, we obtain all the district-county pairs existing in the primary dataset. There are a total of 117 locations of origin and a total of 159 locations of residence which culminate in a total of 174 unique district-county pairs in this dataset out of a total of 278 pairs at the national level.

To gather the spatial data of longitude and latitude, the ggmap [45] API was used which worked directly with the Google Cloud Platform [47]. GCP allows you to solve cloud-related problems through Google, including data management, hybrid & multi-cloud, and AI & ML. Through this platform, it was possible to enable Geocoding related APIs like Geolocation API, Maps Static API or Places API. After selecting which APIs the project fits into, a unique key was generated to access the cloud information in ggmap [45] operations.

After having access to the cities data frame present in the primary dataset and access to

google APIs for geocoding, it was necessary to run functions like mutate_geocode to gather all latitudes and longitudes in the cities data frame and transform it also in a tibble. Using another dplyr [43] tool, this tibble was joined with the cities' spatial information and attributes. Finally, 4 new attributes were obtained: the latitude and longitude of origin (ori_late ori_long) and the latitude and longitude of residence (res_lat and res_long).

4.4 Imputation of values

4.4.1 Imputation of Locations

As previously stated, there is a considerable lack of values with regard to the area of location, both residential and of origin of patients. Although not all records are representative of patients with the TTR-FAP condition, we know that all records belong to a family which is entered in the health unit because someone in a close generation has the disease. Therefore, every record has a connection to an affected patient that can result in some data to be retrieved (a family can have connections between grandparents, parents and sons despite not all of them being a carrier). For this reason, it is justified to apply an imputation to records without these values. This process involves solving the problem of missing data by replacing it with estimates that are computed by different techniques and methods.

One of the problems that may affect the quality of this spatial data is due to the data input itself. It is regular procedure for recent records to inherit the location of origin and residence of parents and often the value is not even entered into the database. The fact that this happens together with the real lack of values for reasons such as the patient's willingness not to share their data, culminates in a set of missing data that could be avoided.

Among the 174 unique district-county pairs, it makes sense that all records that have these values are used in imputation. Since we have information on the family numeric identifier and therefore have information on the family subgroup to which each record belongs, it makes sense to use all patients, non-patients and known relatives that have spatial information to determine the value of each patient location.

The imputation was performed, together with the study of the performance of the methods and the analysis of the results, in a separate script and considered only the attributes related to the unique id, the id in the family tree, the family, the date of birth and sex, as well as the id of the father and mother and the locations of origin and residence. Packages such as caret [48] for tools related to classification or regression training and arsenal package [49] that provides a pack of R functions for Large-Scale Statistical Summaries assisted this imputation work.

Although there were 34654 records registered in the dataset, only those with existing data regarding the location of origin and residence were considered, creating two distinct datasets, the first with 6472 for existing records with origin and the second with 6098 records with available

residence information. It is important to note that the primary dataset available that contains all data, includes records of affected, clean, unaffected, carrier, heterozygous (carrier who received the gene from both parents) and possibly affected individuals. Since records are organised by families and this information is key to this study but their medical condition is not relevant for the imputation of values, the performance study considers individuals that did not develop any form of the disease yet but have valuable geographical data that we can use because they are related to patients that in fact have TTR-FAP.

Two algorithms/techniques were applied to the data: future generation parenting in 1 and family mode in 2 both for origin location and for residence. The mode is a self-explanatory algorithm, since the records will inherit the mode from the family location. Future generation parenting is a process of trying to relate the spatial data of an individual's origin and residence to his/her parents and grandparents, which are the most recent previous generations. Both take as input all the records of the dataset and produce a modified dataset with new values for patients' origin and residence,

Algorithm 1 Future Generation Parenting

```
for generation = 1, 2, ... do
    for individual = 1, 2, ... do
        if i_location exists then
            continue
        end if
        Consider only the subset of the i individual's family F
        if i_motherid > 0 then
            Use individual's mother id to find her location l.
            i_location ← l
            continue
        end if
        if i_fatherid > 0 then
            Use individual's father id to find his location l.
            i_location ← l
        end if
    end for
end for
```

The Future Generation Parenting algorithm is carried out as many times as there are generations desired. In other words, if for a record the location is to be evaluated, considering only 1 generation, only the parents of that record are considered, whereas for 2 generations the parents and grandparents are considered. The algorithm is applied to all individuals. If the location already exists, the procedure is to move on to the next record. If no location exists, a subset of records equivalent to the individual's family identified by the family number is considered. If there is a mother of the individual with a location, this location is used for the individual who did not have one. If there is no mother but there is a father of the individual

with location, use this location for the individual who did not have it.

Algorithm 2 Family Mode

```
for individual = 1, 2, ... do
    if i_location exists then
        continue
    end if
    Consider only the subset of the i individual's family F
    Calculate mode M of the family F
    i_location ← M
end for
```

The concept of mode applied to the dataset assigns, for all individuals, the mode of the family subset to which they belong identified by family number. This happens for all individuals who effectively do not have a location.

It is important to note that in the following results, numbers considering 2 generations were analysed in the Future Generation Parenting algorithm. The application of 1,2 and 3 generations was evaluated, but the difference for the 3rd generation no longer justified the tradeoff between solving more cases or repeating too often the location of an older relative. If we applied the algorithm at the origin, the difference of solved cases considering one generation and two generations would go up by 20%, but the difference when considering the 3rd generation would only improve by 0.02%. As for the residence, applying the second generation relative to the first generation solved 17.6% more cases, but the difference when considering 3 generations only solved 0.02% more cases.

Another question concerns the assessment of mothers or fathers in the first place. This difference is almost irrelevant in the case of places of origin, since the difference between considering one or the other first impacts 0.002% cases on average. The case is not much more significant in the case of residence, with the difference being 0.036% cases on average. Even so, the dataset only gives us information about the existence of father or mother in the same database. Information such as the separation of an individual's parents or the sharing of a house at the current moment of registration in the dataset could be beneficial for considering the location of the father or mother in the first place. By default and because it is biologically more natural for humans to reside with their mother in their early years until they become independent, the mother is evaluated first.

The values in the tables result from an application of imputation to a training and test by random splitting the data in 70/30. Cross-validation, a resampling procedure used to evaluate machine learning models on a limited data sample, was applied. Ten folds were used during the evaluation of the methods compared to the use of Future Generation Parenting and mode.

There are irresolvable records that result from the data split itself, since entire families can be entirely both in the training and the test meaning that they are impossible to predict. These records are expressed in table 4.3 for both locations. It is also important to mention that since

Table 4.3: Unique families in the origin and residence study, using a K-Fold = 10 and using up to 2 generations in the parenting values.

K-F=10, P=2	Origin	Residence
All Unique Families	821	824
Avg Train Unique Families	773,3	770,1
Avg Test Unique Families	579,7	564,3
Avg Unique Families Test Parenting w/ at least 1 NA record	444,8	438,4
Avg Unique Families Test Parenting+Mode w/ at least 1 NA record	47,7	53,9

Table 4.4: Evaluation values for mode and parenting of future generations plus mode for origin and residence, using a K-Fold = 10 and using up to 2 generations in the parenting values. The first half are values taken from applying Parenting and the second half are those in comparison with and against Mode.

K-F=10, P=2	Origin	Residence
Available	6472	6098
Train 0.7	4590	4343
Test 0.3	1882	1755
Avg Irresolvable records	1092,8	1046,9
Avg Correctly Predicted records	780,6	473,4
Avg Wrong Predicted records	8,6	234,7
Precision (Parenting) for predicted records	0,989	0,669
Avg Parenting + Mode Irresolvable records	58,1	70,3
Avg Different values Parenting vs Mode (for predicted by Parenting)	3,8	140,5
Avg Wrong predicted Parenting (for predicted by Parenting)	8,6	234,7
Avg Wrong predicted Mode (for predicted by Parenting)	6,4	167,3
Avg Precision (Mode) for predicted values by Parenting	0,992	0,764

only records that have location values are considered, the families will be smaller than they were actually recorded in the original dataset. Even so, as it is possible to verify in the table, the number of families that it is not possible to predict at least 1 record is about 1/20 of the existing total. Here it is possible to verify that the numbers of families evaluated are quite similar for both locations and that a considerable number are used in Train and Test splits (more than 90% of families in Train 773 out of 821 for origin and 770 out of 824 for residence a and almost 70% of families in Test 580 out of 821 for origin and 564 out of 824 for residence).

The table 4.4 is divided into two parts. The first part refers to the application of Parenting, to the records that it can predict the location and to those that it is not possible. The second part shows data concerning the combination of both techniques and the comparison between the records.

- Avg Parenting + Mode Irresolvable records - shows the number of records that are on

average impossible to resolve by applying Parenting and then the mode.

- Avg Different values Parenting vs Mode (for predicted by Parenting) - shows the average amount of different values of location predicted by both techniques, for the values that Parenting was able to predict.
- Avg Wrong predicted Parenting (for predicted by Parenting) - shows the average amount of wrong predicted values for the values that Parenting was able to predict.
- Avg Wrong predicted Mode (for predicted by Parenting) - shows the average amount of values wrongly predicted by Mode for the values that Parenting was able to predict.
- Precision (Mode) for predicted values by Parenting - average precision of mode values for values predicted by Parenting.

From table 4.4, it can also be seen that Parenting successfully predicts most of the values it tried to predict (Avg Correctly Predicted vs Avg Wrong Predicted) when it comes to the origin of the patients but, when it comes to residence, these values are one third lower (0.99 vs 0.67). These values show that patients of more recent generations tend to have the same origin as their relatives but live in different locations. When comparing the Parenting values with the mode, regarding the values that Parenting was able to predict, both correctly and incorrectly, in both cases the mode is the procedure with better results, and although it is not very different in origin, in residence these values are 10% more accurate.

With these results, it was possible to apply the mode as the imputation of locations to the data, which resulted in two distinct datasets that will be used in the remaining visualisation work. Considering only affected, carrier, heterozygous and possibly affected patients, the final datasets regarding all data entries containing origin and residence locations each have 5782 and 5762 records.

4.4.2 Imputation of Dates

In addition to imputation of locations, in order to complement the study with regard to the time intervals of the disease and the date of onset of symptoms, values were imputed from the year of birth. The average number of years required for symptom detection in the dataset was calculated from the data that had both year of birth and year of symptoms information and this value was 36 years. Previously, a study done by Parman et al. [17] pointed out that this number would be 33.5 years and mostly develop symptoms before the age of 40.

Given that this dataset has a subset of patient values that are registered in a time window of a specified medical unit and, of course, are data coming from the same specialists, the calculated mean years of 36 years was used, which does not diverge so much from the other study and both are in a similar patient life span, as well as being under the 40 year threshold.

4.5 Record Adjudication Diagram

A record adjudication diagram is a common procedure done in [31, 33, 36, 37] is presented in image 4.1:

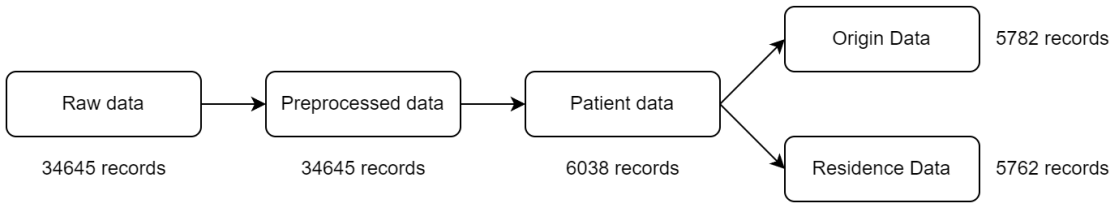


Figure 4.1: Record Adjudication Diagram of the TTR-FAP data.

Raw data cotains all records including family members related to patients with the disease. Preprocessed data changed attributes' values but only a subset of records regarding only affected, carrier, heterozygous and possibly affected patients were considered. The final subset of users considered were the ones with information on origin or residence.

4.6 Summary

This chapter introduced the data used during the dissertation, as well as the data cleaning procedures, the geocoding tools used and the imputation process. There were already some conclusions about the data explained on the imputation tables that were discussed. At the end of this chapter, data was finally ready to be uses in the next development phases.

Chapter 5

Exploratory Geovisualisation

After obtaining the imputed datasets, we proceeded to create primary Geovisualisation (GVis). Using multiple tools, this chapter contains explanations of what type of visualisations the data was able to produce as well as some of the methodologies used during this work. Images in this chapter are for reference since most of the data visualisations are to be discussed on the Results chapter 7.

5.1 Interactive Maps

With the data finally available with a greater completeness of locations, the first application of this data was with the use of the package sf [50] that allows a standardised format to encode spatial vector data along with the package mapview [51] that allows to conveniently create interactive visualisations of spatial data. This type of visualisation allowed the unique identification of points in space and, at the same time, allowed the user to obtain information about the locations, namely the attribute values, by hovering the mouse at any location. In image 5.1 it is possible to see some information regarding one of the patients using this interactive visualisation.

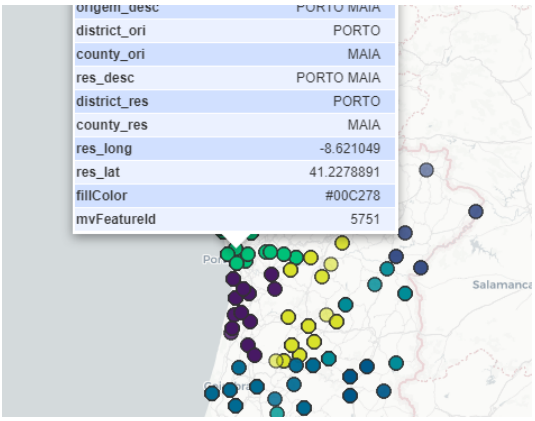


Figure 5.1: Map viewer overview of attributes when hovering the mouse.

5.2 Static Maps

The static visualisations performed involved the use of several packages for the use of the spatial data. The sp package [52] allows the use of classes and methods for spatial data, the ggplot2 package [53] allows the creation of elegant data visualisations with a grammar for graphs combined with the ggspatial package [54] that adds to it a framework for spatial data and the osmdata package [55] imports 'OpenStreetMap' data as simple spatial objects.

With these tools, it is possible to visualise the data over a static map obtained from the coordinates of Portugal. The static image where the data is drawn is obtained from a google maps image with a changeable editing criterion since it is possible to consider different map elements such as terrain elevations or geographical divisions. One of the examples of this type of application, can be seen in the figure 5.2 where all the unique locations present in the dataset in mainland Portugal are represented, in a similar way to the interactive map.

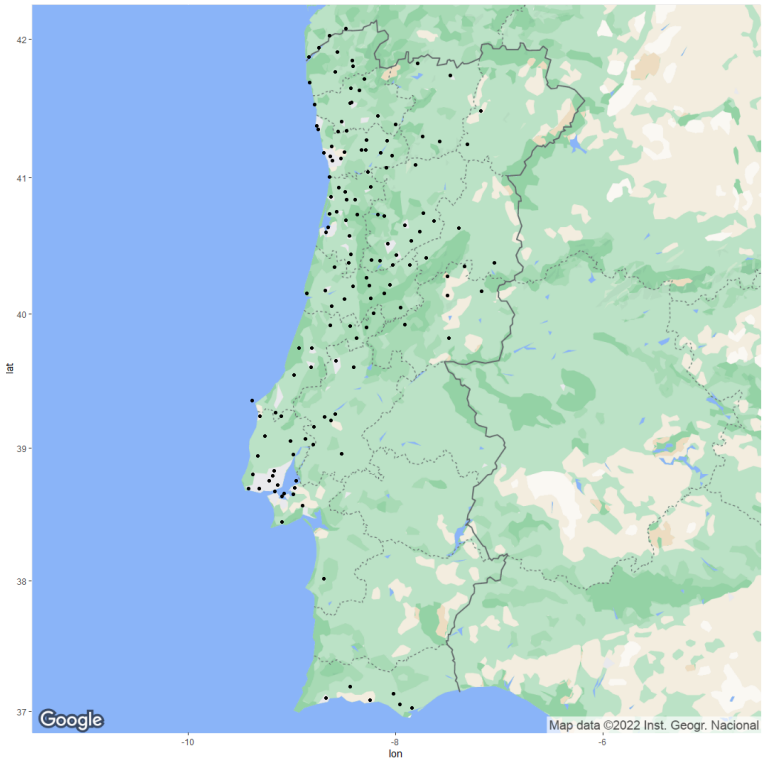


Figure 5.2: Unique locations of TTR-FAP individuals by residence.

Most visualisations follow a problem solving pipeline. This means that a goal or problem to solve with a visualisation is defined and then a subset of records or attribute values of those records are considered. After this selection, GVis tools are applied which results in one or more images. As an example, we define the goal of creating a visualisation that represents all the locations of origin of patients in the dataset, where we can visualise the incidence of each district and county. Then, using dplyr, an example of code would be as follows in listing 5.1:

1

```
per_district_ori <- pafTibble_noNA_ori %>%
  group_by(district_ori) %>%
  group_by(origem_desc) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

geo_per_district_ori_withNames <- left_join(per_district_ori,
                                            locations_tibble_withNames,
                                            by = c("origem_desc" = "cities"))
```

Listing 5.1: Example of data structures for GVis.

These are data compositions that represent the grouping of data regarding the district and county of origin of all patients. This way we will have easily accessible data structures for spatial data methods. To these are added the data relating to the geocoding previously done, obtaining data structures that in addition to counting cases also have spatial and nominal information. The listing 5.1 is an example of two of those data structures.

Using the methodology for creating visualisations, 3 other elements are added to the static map: the points referring to the origin locations, the legend that will include information for reading the image and aesthetic elements. In this listing example 5.2, clusters of similar locations are created in red and black. The black ones represent all locations that are below the 90th percentile of case counts for all locations, while the red ones are the locations with the highest incidence. The percentile and consequently the number of cases are adjusted by taking into account the subset case counts. The labels of the locations are adjusted using the ggrepel package [56], which allows for greater control and customisation of the visualisations. This type of GVis results in the following image 5.3 that will be later discussed.

```
quantileNamed <- round(quantile(geo_per_district_ori$count, 0.90),digits = 0)

ggmap(background_map) +

  geom_point(data = geo_per_district_ori,
             aes(x = lon, y = lat, size = count),
             color = ifelse(geo_per_district_ori$count <quantileNamed,'black',NA),
             alpha = 0.4) +

  geom_point(data = geo_per_district_ori,
             aes(x = lon, y = lat, size = count),
             color = ifelse(geo_per_district_ori$count >= quantileNamed,'red',NA),
             alpha = 0.4) +

  geom_point(aes(color = paste("Incidence above", quantileNamed, "cases"))) + #b
  geom_point(aes(color = paste("Incidence below", quantileNamed, "cases"))) + #r

  geom_text_repel(data = geo_per_district_ori_withNames,
                  aes(x = lon, y = lat,
                  label = ifelse(count >= quantileNamed, county, '')), size = 2) +

  labs(title = "Incidence of affected and possibly affected individuals, carriers
  and heterozygous for TTR-FAP, by municipality and origin, in Portugal",
       size = "Incidence (N)", color = "Incidence above percentile 90") +

  scale_color_manual(values = c("red", "black"))
```

Listing 5.2: Example of one exploratory Gvis.

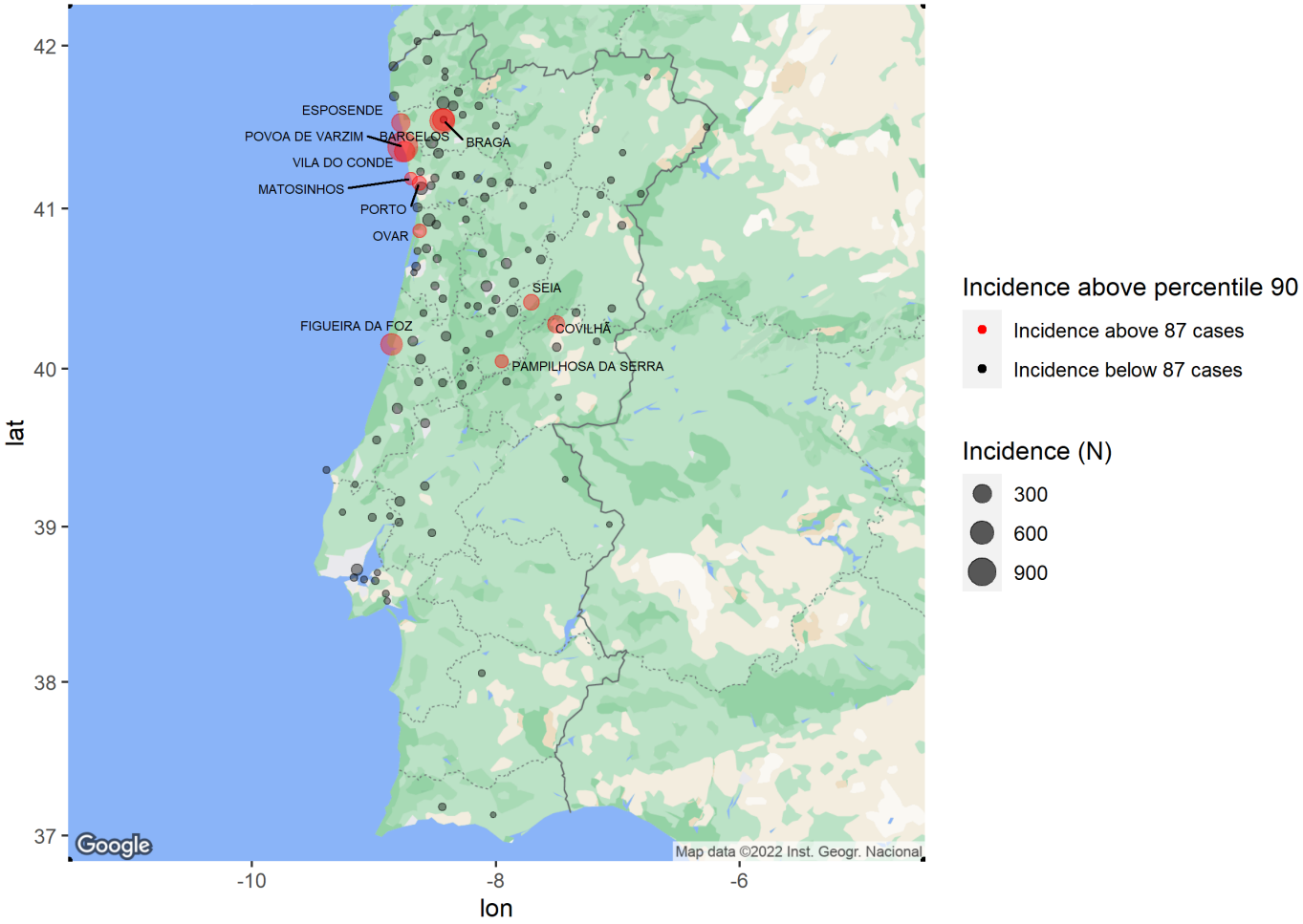


Figure 5.3: Incidence of affected and possibly affected individuals, carriers and heterozygous for TTR-FAP, by county and origin, in Portugal.

Using the concept of small multiples referenced earlier in [29], problem pipelines were defined to solve with GVis. As usual, an objective was first defined, the subset of information was selected and this was introduced to obtain the visual representation.

The first problem involved being able to observe the uniqueness of disease locations over time. Both for origin and residence may suffer from spatial variations over the time window of the data. For this reason, a 10-year interval was defined which resulted in visualisations similar to the example in figure 5.4. This way, it is possible to observe the spatio-temporal changes of patients in a simple and straightforward manner. Another example of this time interval problematic is the spatial incidence by time. If the problem is to evaluate which localities have the highest incidence in a certain period of time, images like the figure 5.5 give this possibility.

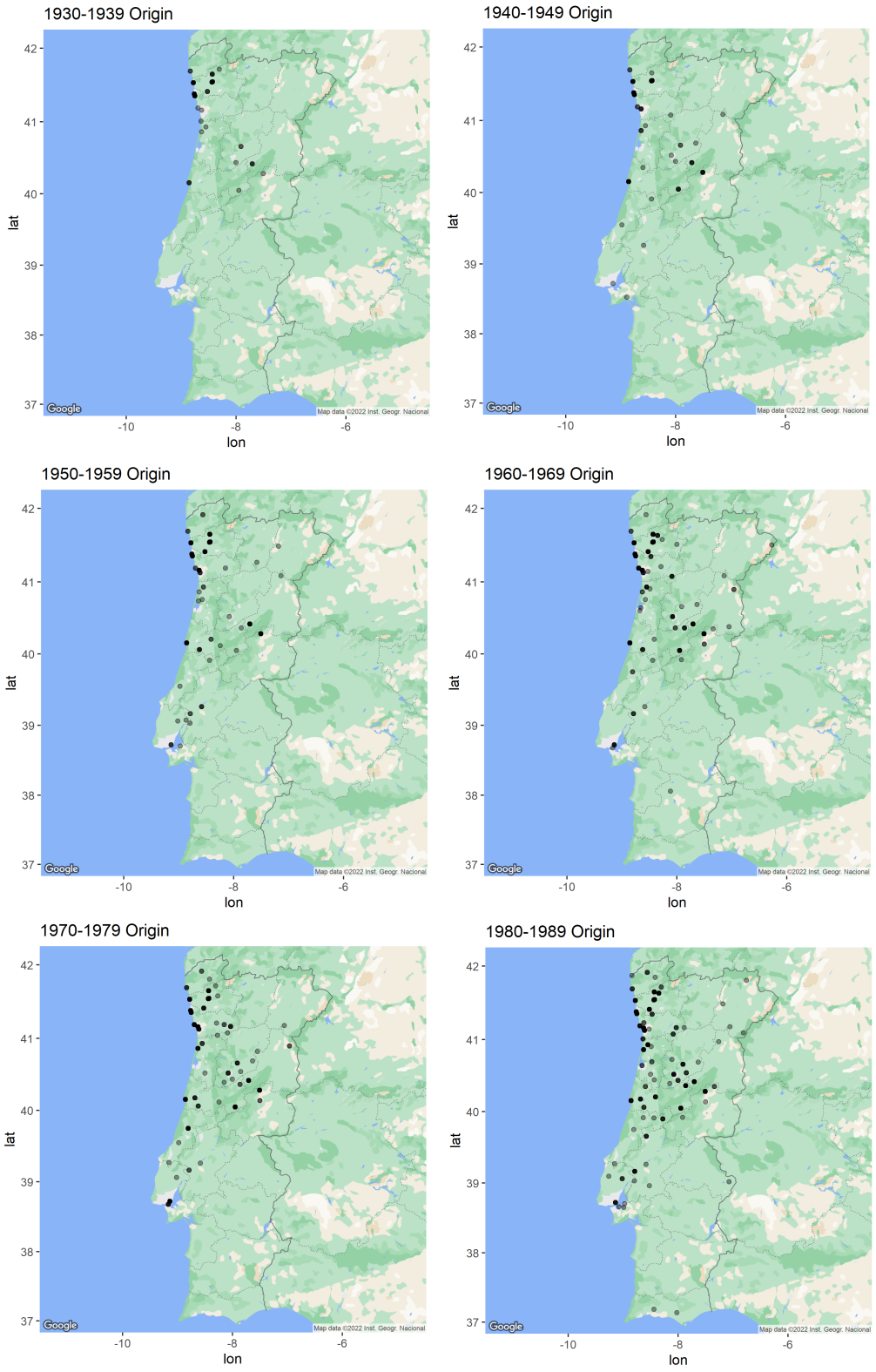


Figure 5.4: Small Multiples of unique locations of TTR-FAP individuals in 6 different decades, by origin.

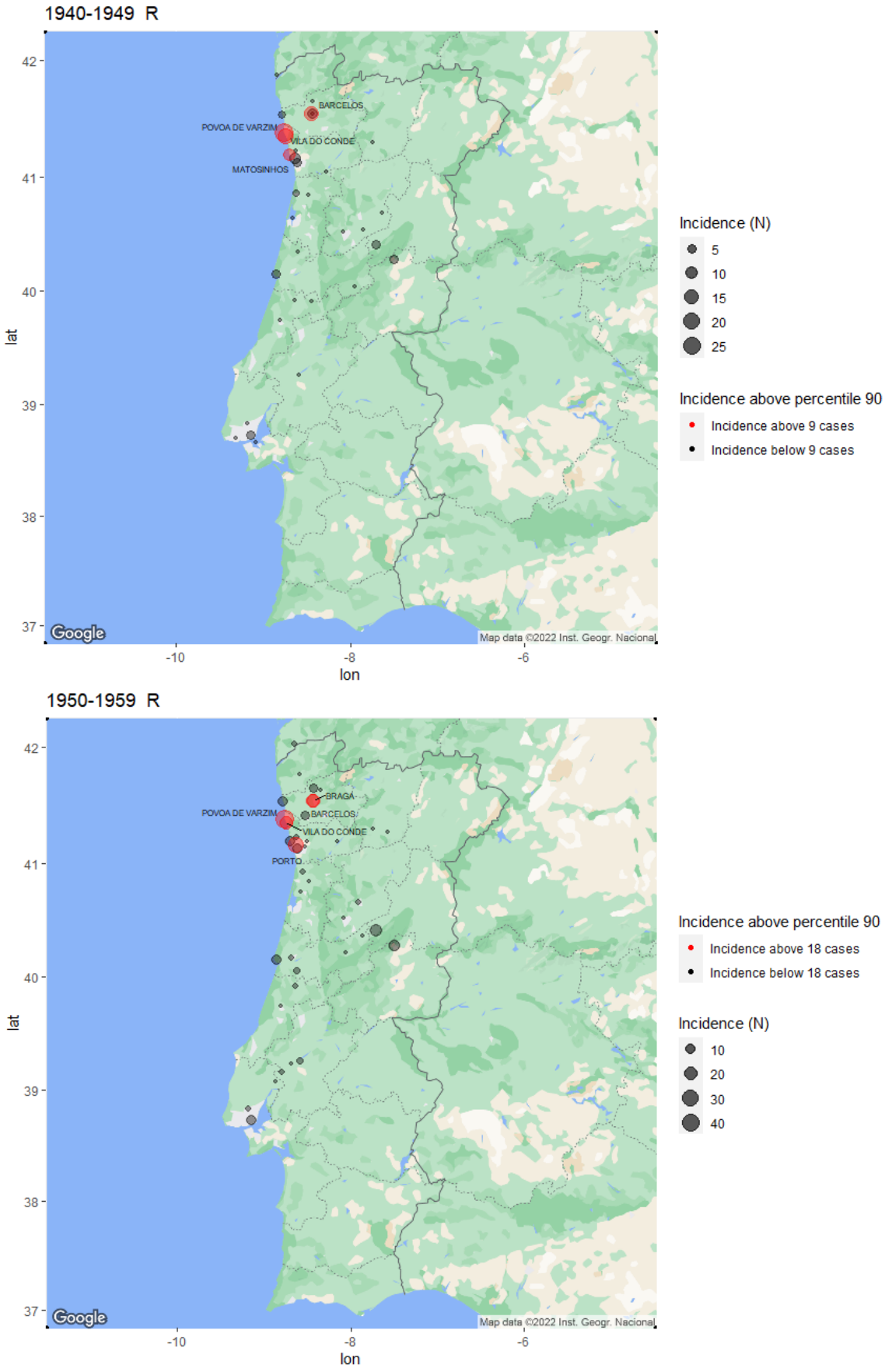


Figure 5.5: Small Multiples of the locations' incidence of TTR-FAP individuals in 2 different decades, by residence.

5.3 Territorial Maps

Another approach that broadened the way of visualising data and that we can mention was widely used in the work of [21–24], was the ability to visualise the locations of patients taking into account their geographical and territorial area directly related to the border limits of the countries in which they are located. For this reason and because it would make it easier to understand the most affected regions, tools were implemented that would provide this territorial division together with the data.

In a first exploration of how the data structures would work, the packages maps [57] and mapdata [58] allowed to create expectations of how the data would be organised in sets of information vectors at various levels and belonging to different regions.

Still, the GADMTools package [59] was the most impactful in this area. This package allows the creation, manipulation and export of maps belonging to GADM. GADM is a repository of databases that map the administrative areas of all countries, with all the available territorial sub-divisions. These data have high spatial resolution with unique identifiable attributes for access. With this information, it is possible to create a variety of unique map visualisations in order to transcribe the data into the most realistic environment possible.

GADM allows the use of the national territory as a whole, of 20 first level subdivisions corresponding to the 20 districts, 18 in the mainland and the 2 in the archipelagos. It also allows the use of all the second level subdivisions that correspond to the counties of each district and all the third level subdivisions that correspond to the municipalities or parishes of each county.

In addition to this type of detailed information about the outline of the country and all its subdivisions, GADM also provides information such as Elevation above sea level (m), Average annual temperature (°C), Total annual precipitation (mm) and data concerning Nighlight (nocturnal activity). An example of the country outlines can be seen in the figure 5.6.

As the available data is only up to the second level of Portuguese territorial subdivision, the county level, we only mapped this data discarding the municipalities. The GADM files are in .rds, a type of data file that is usually smaller than a text file and therefore frees up storage space. This file type preserves data types and classes, such as factors and dates, which makes the file easier to load.

Different tools were used to deal with problems related to districts and municipalities. We have changed the data formatting and encoding of the .rds files themselves so that, when new data is added in the same format as the data we have available, it will be similar and compatible. In what concerns formatting, all data has been considered in capital letters, with a different encoding that becomes compatible with the encoding of the original dataset and even more specific problems like for example the district of "CASTELO BRANCO" in a dataset is called "C BRANCO" in the other. Only districts and municipalities in mainland Portugal were considered, excluding the two archipelagos and their information.

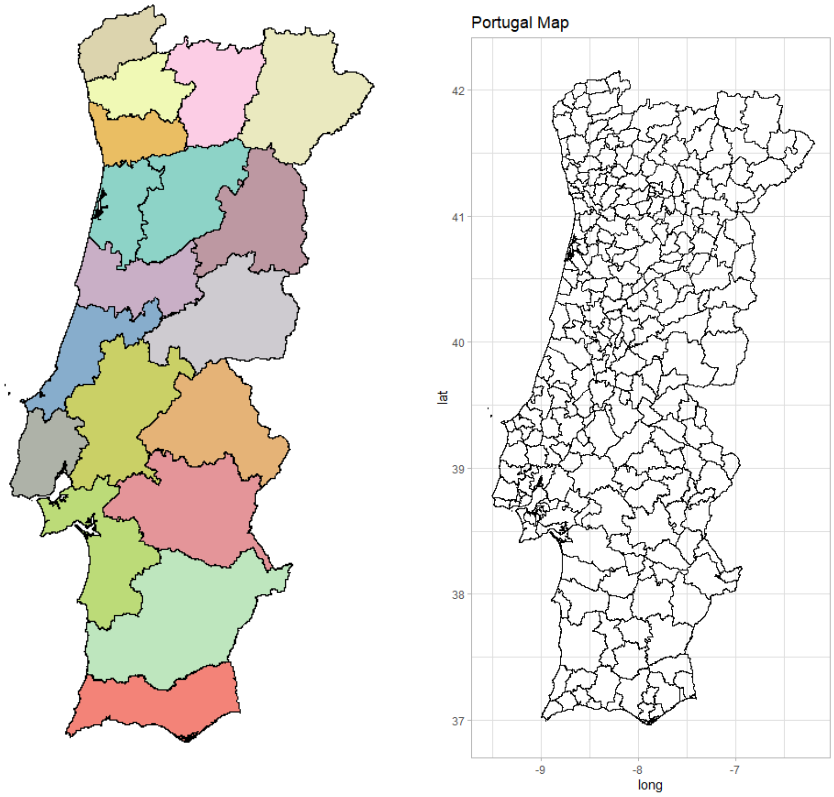


Figure 5.6: Portugal's district and county outlines using the GADM data.

A set of functions belonging to the scales package [60] was used to create colour palettes to colour the territorial subdivisions. By creating a red colour palette, where each colour, with a more or less blacker tone, corresponds to one of the percentiles related to the amount of patients per geographical area. The 10 colours obtained, from white to blacker red, correspond to the 10 percentiles (from 0 to 1 with intervals of 0.1).

The same problem solving pipeline is used, choosing a subset of data, applying them to the GVis techniques but this time, in order to preserve the high quality of the spatial data and because these visualisations involve different formats, the tiff package [61] was used, which allows working with images of high pixel density. The grid package [62] allowed to join Small Multiples in a grid in order to be possible to compare differences.

While districts are a small number of unique cases in the dataset, there are 278 counties in mainland Portugal. For this reason and because not all the counties are written in the same way in both the GADM data and the patient dataset, we made this an automatic process. Using the stringdist package [63] and creating lists of the counties in the dataset and in the GADM, it was possible to compute the distances between the string values of both. This way, the list present in GADM was replaced by the dataset values of the patients that had the smallest distance for each location. This means that similar strings that differed by a few characters are now compatible between the two datasets. The color palette and pipeline was used in the same way as in the

districts and it generated images like the figure 5.7.

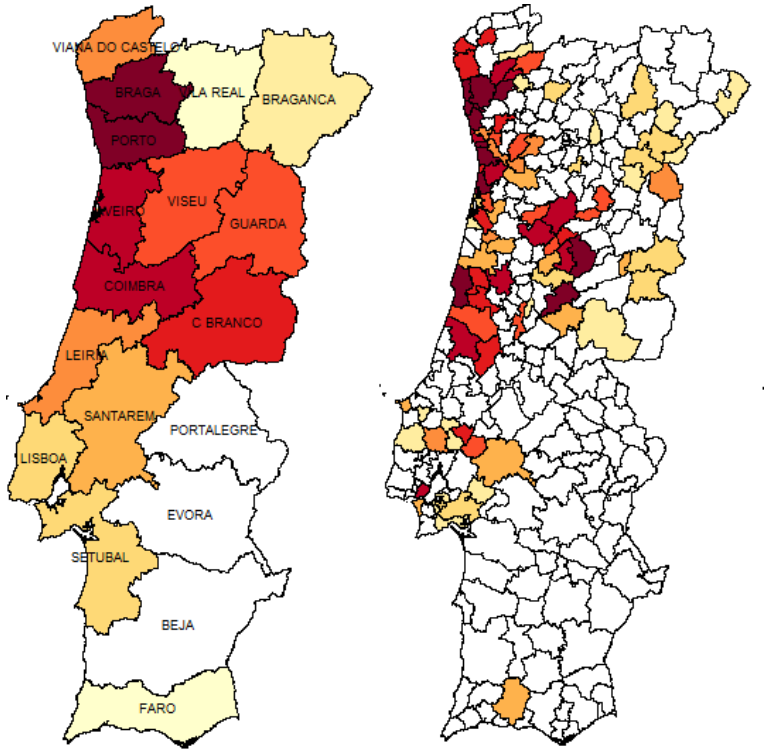


Figure 5.7: Incidence of affected and possibly affected individuals , carriers and heterozygous for TTR-FAP, by district, county and origin, in Portugal.

5.4 Summary

This chapter showed multiple types of GVis, using different tools and techniques, that were created using the data available. Interactive, Static and Overview maps were most of these types and their development was also discussed briefly.

Chapter 6

AmiVis: a GeoVisualisation app

In this chapter we introduce AmiVis, the tool designed to give an user the ability to control the Geovisualisation (GVis) that they want to view, created with the purpose to visualise spatial data in the country. First, we discuss a prototype done in Adobe XD and then AmiVis's features and architecture.

6.1 App prototype

Firstly, a prototype application was designed using Adobe XD software [64]. This design tool allows to create a vector-based user environment for web and mobile applications. By doing so, it is possible to build a mock-up of the future visualisation tool for the Transthyretin-associated Familial Amyloid Polyneuropathy (TTR-FAP) data without compromise that can demonstrate how the initial ideas can be applied and used by the public.

This prototype has been idealised with the construction of panels and buttons that would allow user navigation using a set of connections that made a specific type of configuration. In this initial prototyping, different features were designed. An example menu was outlined to choose parameters such as the choice of location, time interval, region at district or county level, as well as the unique location map. Static and overtime visualisation panels, statistics, comparisons with external factors and the records' table were also outlined. All this exploratory work of idea brainstorm can be seen in the figure 6.1 that resulted in a starting point for the development of the AmiVis tool.

Figure 6.1: AmiVis prototype panels for Static/Overtime GVis, Statistics, External Factors and Table in Adobe XD.

6.2 AmiVis Features

The AmiVis application has features which are described in this section:

- A menu for choosing parameters. These parameters are applied to the visualisations relative to the primary dataset and allow different images to be visualised in time and space. It is possible to obtain subsets for the location of origin or residence, the sex to be considered, the district and municipality, and the range of symptoms, as well as to consider records without dates that impact the year of symptom onset. Below these parameters there is a map that expresses all the unique locations present in the primary dataset.
- An Introduction panel contains an explanation of most of the features and their panels. The next panel, Load Files, allows the upload of files by source and residence as well as the download of the Comma-separated Values (CSV) with the filtered data for later use.
- There is a panel dedicated to static GVis where it is possible to see the comparisons of the primary dataset that contains all the records directly with the dataset filtered through the parameters. There are three secondary panels, one for the visualisation of the geographic map of incidences by location and the other two for the territorial division of the country into districts and municipalities. Figure 6.2 shows this panel.

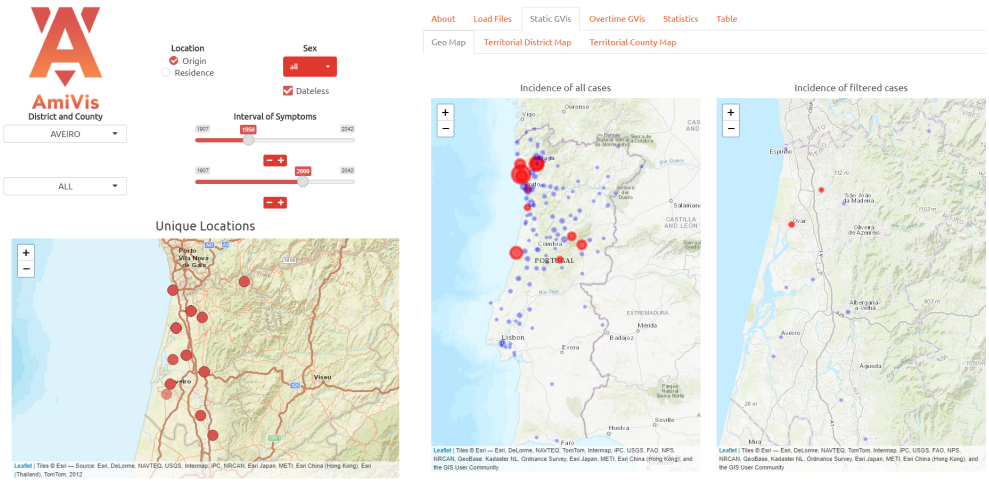


Figure 6.2: AmiVis panel for Static GVis.

- The panel, Overtime GVis, gives the user the possibility to control the space of years to visualise. This way, it is possible to define an interval of, for example, 10 years, and visualise the evolution of cases every 10 years from the beginning to the end of the total period of time. This feature is available for incidence maps and territorial maps of districts and municipalities. Figure 6.3 shows this panel.
- There is a panel for statistics on both the primary dataset and the parameter-filtered dataset. It is possible to access a time series comparison of cases and a count of cases per

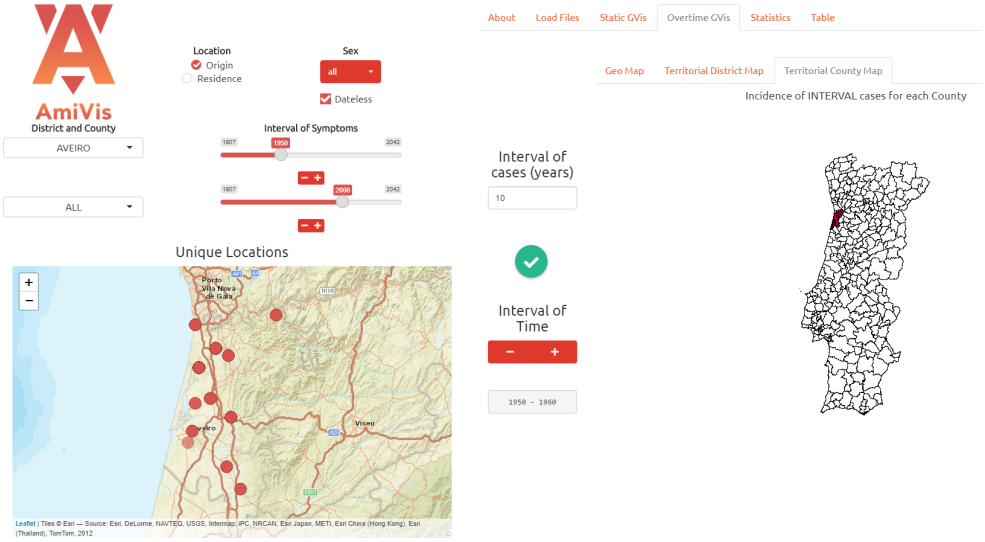


Figure 6.3: AmiVis panel for Overtime GVis.

decade as well as cases per district and county. Figure 6.4 shows this panel.

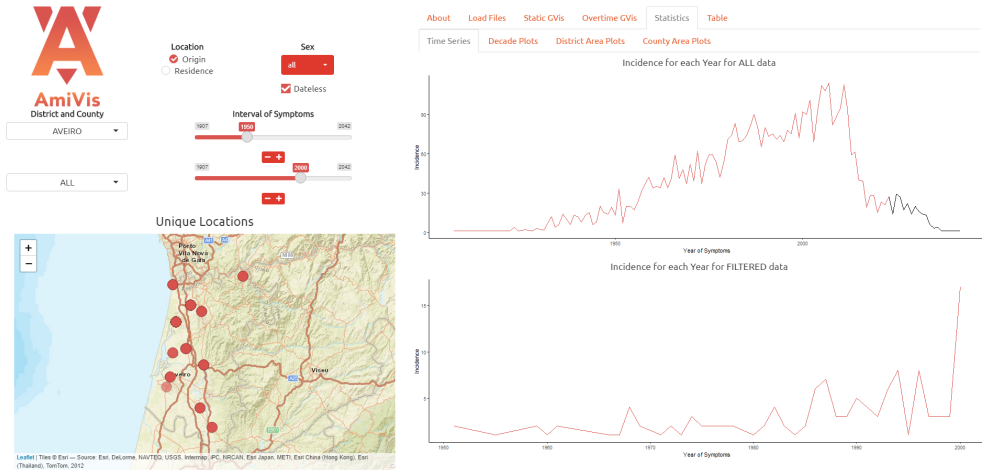


Figure 6.4: AmiVis panel for Statistics.

- The Table panel allows a detailed visualisation of the filtered dataset as well as a more customized search of any attribute and its ordering. Figure 6.5 shows this panel.

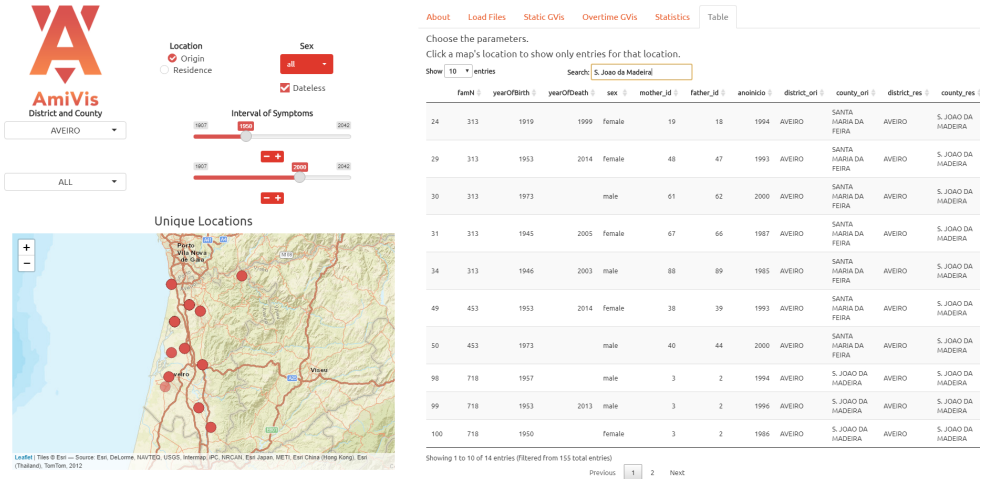


Figure 6.5: AmiVis panel for Table records.

6.3 App architecture

Since the initial explorations of GVis were carried out in R language as well as the preprocessing of the data, it seemed logical to use this type of technology for the development of the tool. The shiny package [65] was used, a package that allows the construction of an interactive web application environment from the R language. As well as allowing the application to be hosted on web pages, it also allows them to be embedded in R Markdown documents or build dashboards. Finally, there is compatibility between Shiny and CSS themes, htmlwidgets, and JavaScript actions.

A Shiny application has two main components, a user interface object and a server function which are connected through the shinyApp function that takes them as arguments and creates the Shiny app object from this UI/server pair. Image 6.6 shows AmiVis file structure. Although these are the essential components, the available file contains other elements that are part of the application:

- The file folders contain non R script files that are used during the uptime of the application. The data folder contains the rds files for districts and municipalities that are used in the construction of territorial maps and contains a CSV Dummy with random data for exploratory use of the application. The www folder contains other types of files such as the application logo image.
- The run.R script contains the instructions for the host, directory and port to use in the application.
- The R script init.R contains code for installing the packages required for using the application if they are not already installed. There are required packages that were already mentioned like are dplyr [43], sf [50], ggplot2 [53], maps [57], mapdata [58], sp [52], tiff [61] and GADMTools [59] and shiny [65]. There are also new packages such as

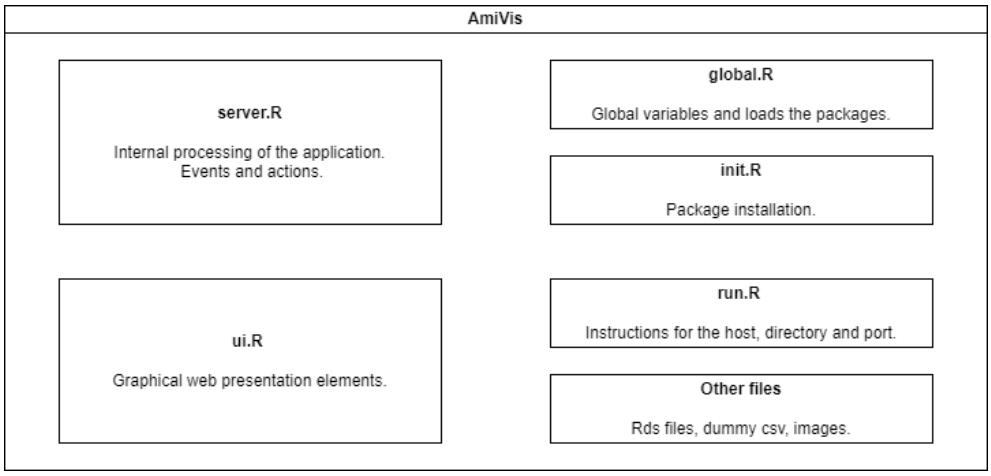


Figure 6.6: AmiVis file structure.

shinyWidgets [66] that enables the use of a collection of custom input controls and UI components for Shiny apps, shinythemes [67] that enable custom themes for the overall Shiny app, DT [68] that enables data objects in R to be rendered as HTML tables using JavaScript, tools [69] that helps producing human-readable summary statistic from raw data and other macros, leaflet [70] that enables the creation and customisation of interactive maps using the 'Leaflet' JavaScript library, tidyr [71] that helps manipulating data and RColorBrewer [72] that provides color schemes for maps.

- The R script global.R contains global variables accessible in the various scrips, mainly server and UI. It is also the script that loads the packages.
- The R script ui.R contains the graphical web presentation elements.
- The R script server.R contains functions regarding the internal processing of the application as well as the events related to the actions performed in the UI.

6.3.1 AmiVis ui.R

The UI is composed of a fluidPage. This type of pagination divides space into a grid layout that is also divisible. A Fluid Grid Systemm was used which uses 12 columns that can be flexibly subdivided into rows and columns using fluidPage() and fluidRow() functions.

The application code is divided into 2 main parts. The first part comprises the page header as well as the customisation of the widgets used in the development of the tool. In this set of operations, there is embedded HTML and CSS code in order to customize widgets that don't have that option. For this reason, it is necessary to define styles applied to elements of the widgets and redefine their behaviour and visualisation. Examples of this type of embedded code are changing the colour of the year range bars, the font and the arrangement of the buttons relative to the columns they are in.

On the other hand, all the rest of the UI is divided into rows and columns with the elements distributed by them. As each row uses 12 columns, there was a distinct planning and execution so that the size of the window altered the space distribution of the columns in a reasonable way, making the whole page responsive. Each widget is identified by an Id and a value that are accessed from the server in order to interact with what the user sees.

6.3.2 AmiVis server.R

The Server is made up of two different sets of operations, these being the auxiliary functions for handling and display of the visualisations and the events that are triggered by actions directly related to the widgets present in the UI.

An integral part of the auxiliary functions are those which create the incidence and territory maps with total and filtered datasets, as well as the functions for creating statistics. There are also functions that link the change of the UI with the server and take care of applying those changes back to the UI. Finally, there are functions that directly change the initial dataset in order to apply feature engineering to it, one such example being the decades and symptom start dates.

Part of the events handled in the application is the upload of the origin and residence files as well as their choice in the menu, the choice of sex, district, municipality, year interval and location. Also part of these events are buttons present in Overtime GVis and the option to download the filtered CSV.

6.4 Summary

In this chapter we gave an overview of how the early prototype of the AmiVis application was designed, a brief summary of its features and, more importantly, how the app is designed in different files.

The application uses concepts already discussed previously in the State of the Art chapter 3 such as 2D Cartographic Visualisation , Spatio-Temporal Visualisation and Interactive User Interfaces [25] . As stated in [27] it tries to expand the application to different objectives and dimensionalities in an automatic approach, whose quality of the results is a direct consequence of the tools for filtering, detailing and noise removal, as referenced in [29].

AmiVis is a combination of an UI and a Server, complemented by other files that make the app run in R language with different packages. The app is online at shinyapps [4] and its code is available on GitHub [5].

Chapter 7

Results

In this chapter we discuss results that are possible to extract from the analysis of the general dataset and also from the datasets on the origin and residence of the patients. We also relate the data to other previous studies on the disease to provide a comparison and to introduce what could be a methodology for analysing Geovisualisation (GVis) and to understand the presence of the disease in the national area.

7.1 Overall Data Analysis

Firstly, data on the totality of cases from affected and possibly affected, carriers and heterozygous patients will be the one considered. An overview of case numbers can be seen in the picture 7.1, which shows the total incidence of cases per year. As mentioned before in [3], the first known case was in 1939, marked by the green line. From this point onwards, and considering also the half decade before, there is a clear progressive increase of the disease.

The brown line marks the introduction of liver transplant treatment (1992) while the black line marks the introduction of the use of tafamidis (2012) [15]. It is possible to note that the number of existing cases in the decade preceding liver transplantation is similar to the number a decade later. Furthermore, the introduction of tafamidis treatment in patients coincides quite nicely with the decrease in cases taking into account the incidence. In 2012 Portugal also started to reimburse oral treatment by the Portuguese National Health System [17], so the registration of cases in treatment units may also have been affected by this policy decision, as they increased in number.

In the chapter 4 on data preparation it was mentioned that this dataset contained a total of 117 locations of origin and a total of 159 locations of residence which culminate in a total of 174 unique district-county pairs in this dataset out of a total of 278 pairs at the national level. In [15] it was noted that there were a total of 174 district-county pairs out of 278 (62.6%) from 2010 to 2016, the period which contains the highest peak incidence of these cases. Still, with regard to all records containing either geographical information of origin or residence or both,

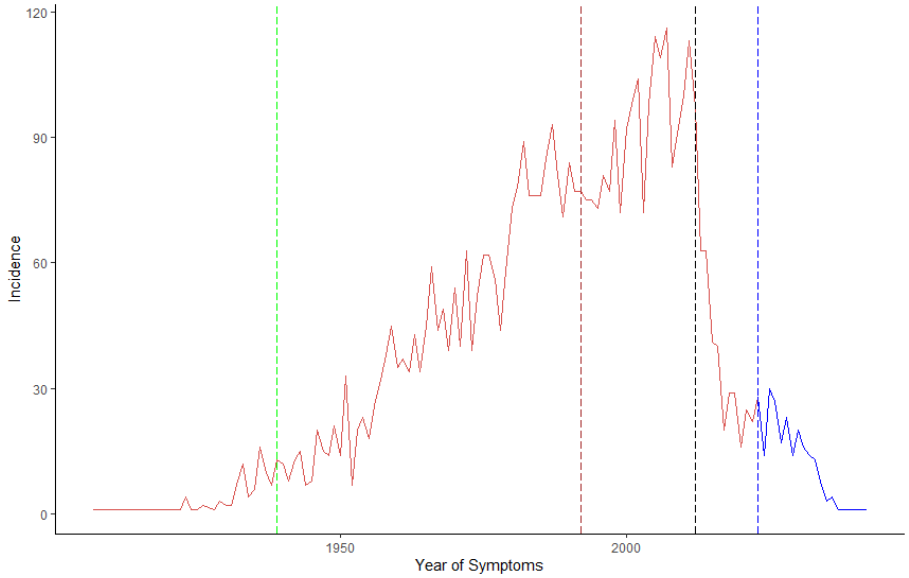


Figure 7.1: Incidence of affected and possibly affected individuals, carriers and heterozygous for TTR-FAP, by year, in Portugal. Green line marks 1939 (first known case), brown line marks 1992 (liver transplant treatment), black line marks 2012 (tafamidis) and blue line marks 2023.

the number of unique counties in this subset is 163 (58.6%). In this period it reports a decrease in the number of cases which also occurs in this dataset in this period of time

Other studies like [1, 15] take into account the prevalence of the disease. This type of representation takes into account the number of cases of patients in a certain period of time. For this it is necessary to know how many patients are alive during that period. Even so, there are records that do not have data related to the date of death of the patients. For these patients, an average life expectancy of 60 years was considered, according to the study of [73], which shares results for the same mutation in Portugal, Val30Met, which normally expresses symptoms up until the age of 40 and causes death up to 20 years later from symptom-related causes. Thus, in the figure 7.2 is the prevalence curve of the disease. The vertical lines have the same meaning as in the previous figure. There is an evolution of prevalence in these data until the end of the 20th century. The decrease of cases with the advent of treatments seems to correlate with the curve, but the decrease is more accentuated. This can possibly be explained by prevention and awareness measures regarding the disease, symptoms and treatments. More recent data is also scarcer which may explain a more abrupt decrease. These data would also be more accurate if more information existed regarding the dates of the patients.

In [18] , little gender difference is reported considering asymptomatic carriers, while in symptomatic carriers, men have more signs than women (65% vs 35%) in a study of the same Portuguese mutation but in Spain. In this dataset of 6038 records, there are 2970 women (49.1%), 3059 men (50.6%) and 9 unknowns. If we consider only asymptomatic patients (with value 4 in the sitdoenca attribute) then 581 patients are considered and there are 375 women (64.5%) and 206 men (35.5%). For values of 4.5, which represent patients with some symptoms but

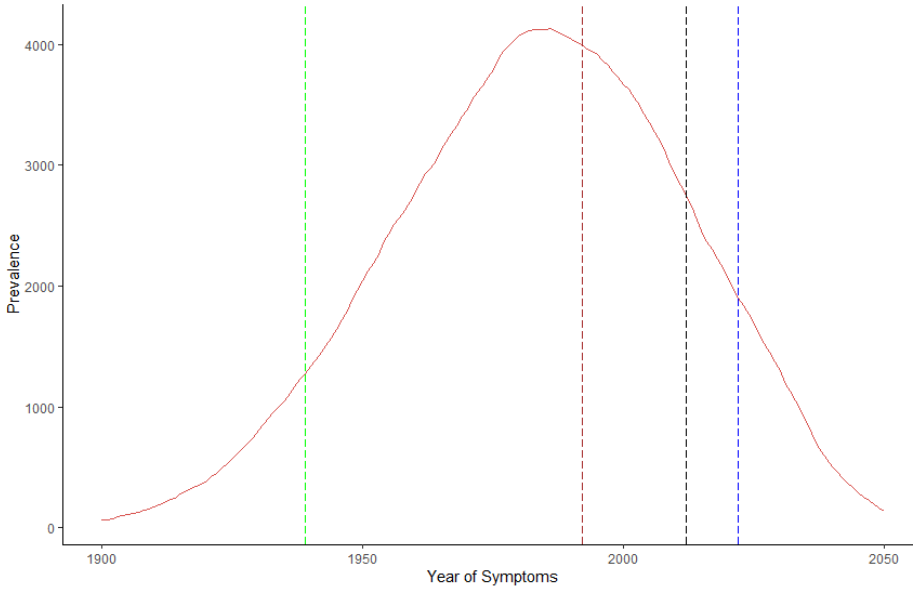


Figure 7.2: Prevalence of affected and possibly affected individuals, carriers and heterozygous for TTR-FAP, by year, in Portugal. Green line marks 1939 (first known case), brown line marks 1992 (liver transplant treatment), black line marks 2012 (tafamidis) and blue line marks 2023

who cannot be considered clinically as symptomatic, there are 97 records, 65 women (67%) and 32 men (33%). For values equal to 5, representing 2694 symptomatic patients, there are 1218 women (45.2%) and 1476 men (54.8%). These values show a great disparity in relation to values for asymptomatic patients and with some symptoms and a smaller disparity in relation to values for symptomatic patients, meaning that the data from these patients are contrary to the values from the spanish study.

7.2 Geographic Data Analysis

There are 3 appendices and all of them show data from a subset of records belonging to affected and possibly affected individuals, carriers and heterozygous for TTR-FAP, without records that are considered Dateless (that don't have information of date of birth and therefore no age of onset of symptoms). These appendices put together what could be an informative data sheet on a group of records for the purpose of medical analysis:

- The first subset's appendix A brings together all the national data in the dataset divided into data with origin and residence information. (2x4 pages).
- The other two subsets appendix B and C contain all the data of the districts of Porto and Lisbon of the primary dataset divided into data with origin and residence information (2x2x4 pages).

Each of the 3 data sheets is divided into 5 parts for such subset.

- 1 - Static GVis of the incidence of TTR-FAP patients with geographical and territorial maps by district and county.
- 2 - Time series and cases per decade.
- 3 - Overtime GVis with geographic map for all periods of 25 years after the 1st case in the dataset in 1907.
- 4 - Comparison of visualisations for men (above) and women (below).
- 5 - Incidence for each district and county.

7.2.1 AmiVis Data Analysis of mainland Portugal

In what concerns appendix A, regarding the general situation of mainland Portugal, it is possible to see a similar concentration of cases in terms of origin and residence in the districts of Braga, Porto and Coimbra. Something to note are the differences in the significant districts but not at the top of the incidence values. The district of Aveiro is common as a district with considerable values but it is joined by the district of Castelo Branco in the origin while in relation to the residence of the patients, it is joined by the district of Lisbon, while Castelo Branco drops in priority in terms of the residence of the individuals. Most of these districts are mentioned in [17] as Portugal's disease focus.

There are fewer people living than being originally from Guarda, in the case of the counties of Figueira de Castelo Rodrigo and Aguiar da Beira and this phenomenon is identical in the counties of Castelo Branco, Fundão and Oleiros of the Castelo Branco district. On the other hand, there are many more people living in the counties Oeiras, Amadora, Loures, Sintra and Cascais of the Lisbon district and Almada, Seixal, Palmela and Montijo of the Setúbal district.

With regard to the evolution of the locations of origin every 25 years, it is notable the appearance of cases in the North of the country, which extends to the central coastline and the northern interior of the country. In relation to the locations of residences, these also have a high incidence in the north of the country and in the coastal and central interior. On the other hand, the incidence in the south coastal area, and in Lisbon and Setúbal in particular, is much more notorious and progressive.

As for the difference between genders, regarding the origin, men have a fewer incidence of cases in the north centre, in Viseu and Aveiro, while the existence of affected women is greater in the south and Algarve. As for residence, there is a higher incidence of men in the Viseu and Guarda districts, but there are more women living in the districts of Leiria, Castelo Branco and Faro.

In [22] the authors refer that in 1952, 84% of the cases belonged to individuals from Vila do Conde and Póvoa de Varzim, districts of Porto. In 1995, 35% would have originated from these locations. From the data provided, it is possible to see that, in the totality of recorded cases

over time, the two largest counties in terms of incidence are Póvoa de Varzim and Barcelos (of Braga), followed by Vila do Conde. As for these two counties of Porto, in 1952 they represent 36% of the cases and in 1995, 29%, by origin. These numbers are 36% and 24% for residence. These numbers account for all cases up to and including that year and show a decrease in cases of origin and residence in these counties, although they are problematic in the same proportion.

7.2.2 AmiVis Data Analysis of Porto, the most affected region

Regarding appendix B, it gathers all the data concerning the district of Porto, which is the most affected district in mainland Portugal. Its peak of registered cases is in the 90's and 00's. Among the differences in the territorial incidence of the municipalities, on the coast there is only the difference in the residence of the cases in Matosinhos and Maia. On the other hand, there is much greater residence in the interim counties of the district, namely the counties of Santo Tirso, Valongo, Gondomar and Amarante. In the map of residential incidence it is notable the presence of unique locations within the district that are not included in the register of origins.

The evolution every 25 years ends up stabilising in what concerns the difference of locations, starting on the northern coast of the district and spreading to the centre and southern coast of the area.

With regard to the difference between genders residency, there are more men in the municipalities of Paredes, Penafiel and Marco de Canaveses, while the rest of the municipalities are similar with regard to residence. Regarding the origin of these individuals, there are more men originating from the opposite municipalities of Vila Nova de Gaia, Marco de Canavezes and Baião.

7.2.3 AmiVis Analysis of Lisbon

Considering that Lisbon is the capital of Portugal and there is a clear shift of residential cases to this district, appendix C contains the subset of cases belonging to this region. As far as the origin of cases in Lisbon is concerned, they mostly belong to the northern counties of Torres Vedras, Alenquer and Azambuja with the exception of the county of Lisbon which has the most cases. Regarding the residence of the patients, there is a big difference because most of the counties have an incidence of cases but this is greater in the southern districts of the Lisbon district with the counties of Lisbon and Sintra having the majority of affected residents.

The time series of Lisbon cases is very different from that of Porto which is similar to the general cases in the country. In the case of Lisbon, the origin of cases is much more noticeable from 1970 to 2000 in relation to other decades while the residence of patients in quantity of cases was gradually increasing from 1960 to 2000 with a considerable jump in the 1950 decade.

In periods of 25 years, it is notable the presence of cases in Alenquer that move southward in terms of origin, while after 25 years it is already possible to observe the existence of residential

cases in the capital.

The origin of the men is Alenquer and Lisbon while the women also come from adjacent municipalities such as Torres Vedras and Azambuja. As for the residence of the patients, most municipalities are occupied by men affected by the disease with more notoriety in the south while women are much less numerous in the north of Lisbon. An interesting data is that only 4 of the 16 municipalities of Lisbon originate individuals.

7.3 Summary

This chapter described the results of the work carried out in this dissertation, focusing on the comparison and analysis of the data for the whole country, the district of Porto and Lisbon. A paper was also written and accepted at a conference that brings together Data Science (DS) work for community areas such as healthcare.

Chapter 8

Conclusion

In this dissertation, Data Science (DS) and Geovisualisation (GVis) topics were addressed, culminating in the analysis of data relating to Transthyretin-associated Familial Amyloid Polyneuropathy (TTR-FAP) patients. Literature regarding other studies of the disease was analysed as well as visualisation theory for spatial data applied to endemic diseases.

Data transformation techniques were discussed and applied resulting in final datasets of previously available data ready to work with and dissect. This data was fed into the primary visualisations and more importantly into the application designed and built during this period. AmiVis allows the users themselves to hypothesise the data they want to see and, as stated in [29], when assembling the geovisualisations into animated GIFs with image equivalent frames, each user can extract different information that others will only conclude later.

The final application allows the control of different variables and allowed to gather relevant information for the construction of data sheets present in the appendices that show an overall distribution of people in time and space for this disease.

Finally, this work concludes with the accreditation in the form of acceptance and publication of a paper [6] in a workshop of an international conference of DS.

8.1 Future Work

This work will benefit greatly from future applications that can extend its functionality. First, the web app can use different attributes to achieve even different levels of subset of users, such as symptomatic men who died in the 1930s and were adopted. In this dissertation we have chosen the attributes that are, in our opinion, the most impactful and urgent. These attributes can also make other types of GVis possible. This work can be upgraded with features like revealing whether a localisation was imputed or not and consider even family tracking.

Secondly, the patient data can be correlated with different national data such as Doctors per speciality (related to TTR-FAP, if obtainable), Doctors per number of inhabitants, health

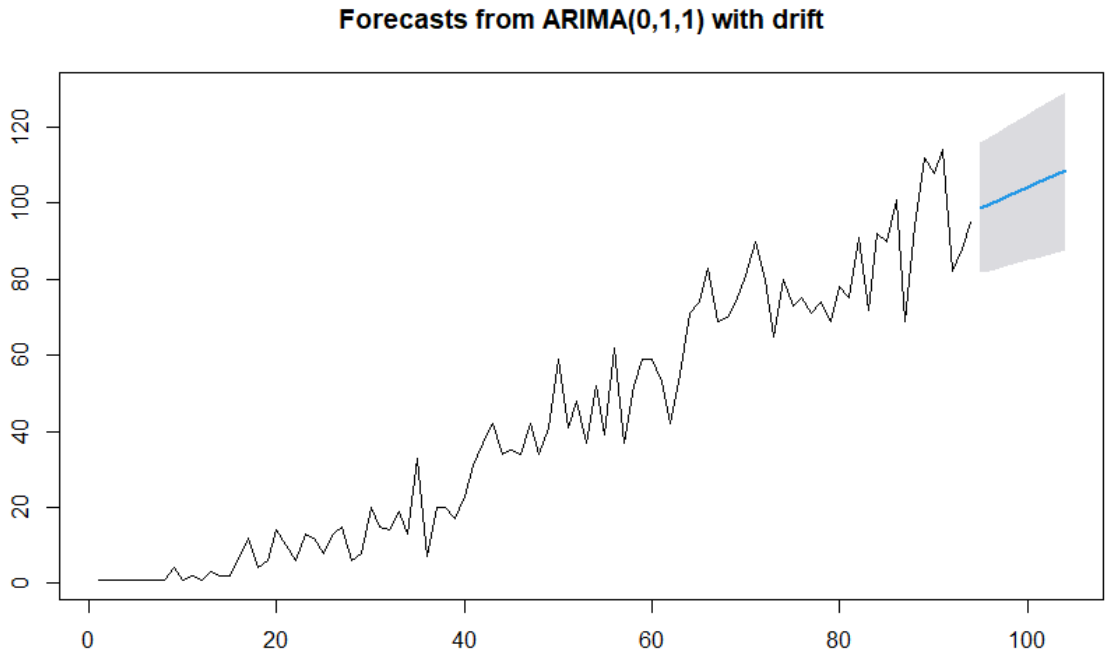


Figure 8.1: Example of forecasting for the number of affected and possibly affected individuals, carriers and heterozygous for TTR-FAP in Portugal for each year with onset symptoms for 2011-2015

facilities per region, welfare index, scientific publications on TTR-FAP, number of Researchers (FTE) in research and development activities (R&D), internal migration balance according to censuses, and Hospitals/consultations/Health Centers per inhabitant. Much of this information is difficult to access publicly or, if it exists, is scarce in its temporal granularity.

Thirdly, imputation of values can be improved by designing more complex procedures that would result in a higher accuracy of predictions, specially for residence taking in account the results gathered. These methods could also be introduced in the application and users would be able to see the different outcomes of applying different imputations.

Future work also involves creating models that are able to predict the incidence of the disease over time and that can be adapted to the location chosen in the application. To do this, a symptom year prediction model needs to be adapted so that more patients with recent years of symptom onset can be correctly used. Examples like the figure 8.1 using the package forecast [74] show forecasting with ARIMA modelling for the years after 2010, considering the worsening of cases while the figure 8.2 shows what would be an incidence growth curve with the incidence package [75]. Although these methods work relatively well for value estimation and even prediction for a 95% confidence interval, more detailed individual forecasting is still needed to ensure higher data quality over the last decade.

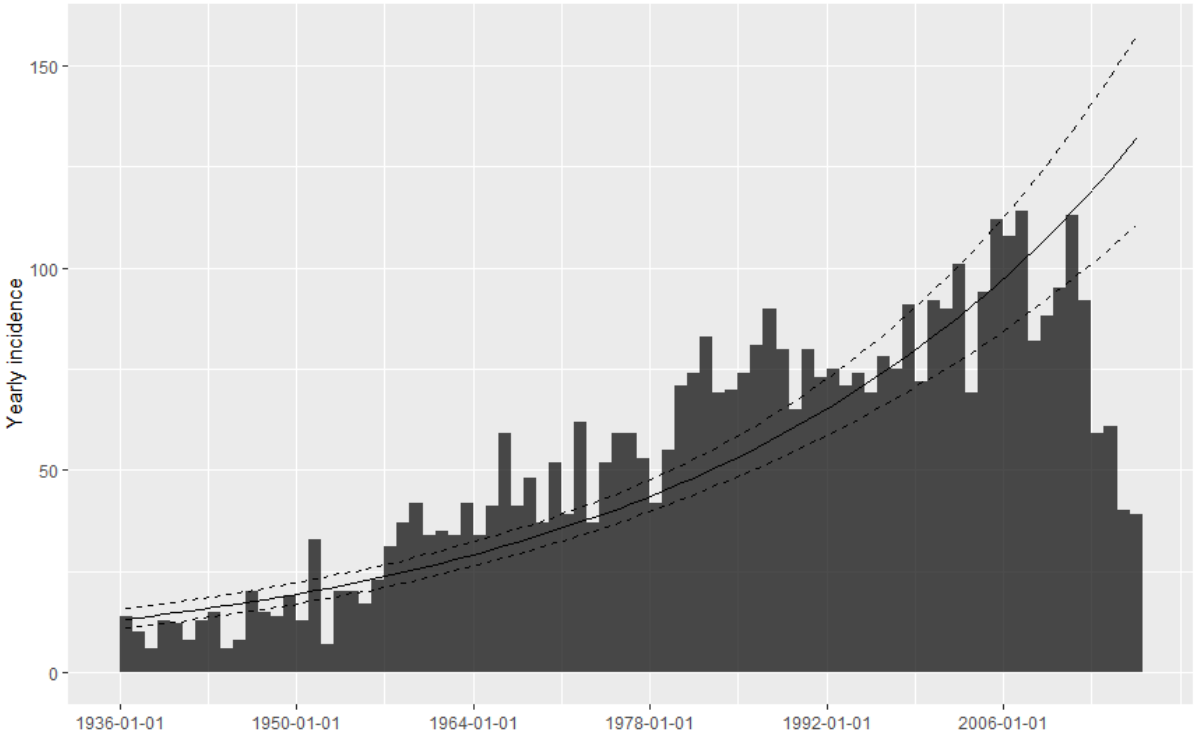


Figure 8.2: Incidence of affected and possibly affected individuals , carriers and heterozygous for TTR-FAP, by district, county and origin, in Portugal for each year with onset symptoms between 1936 and 2006 with curve by early fitting.

Finally, a final iteration of this work could be the prediction of cases by geographic area that relates adjacent areas and population movement over the years in order to understand the next steps of the disease.

Most of this future work will benefit from the datasets that originated in this dissertation as well as being inserted in the application as new functionalities for users to access.

Bibliography

[1] H. H. Schmidt, M. Waddington-Cruz, M. F. Botteman, J. A. Carter, A. S. Chopra, M. Hopps, M. Stewart, S. Fallet, and L. Amass, "Estimating the global prevalence of transthyretin familial amyloid polyneuropathy," *Muscle Nerve*, vol. 57,5, pp. 829–837, 2018 May.

[2] Medlineplus, "Transthyretin amyloidosis." URL, Last accessed 23 December 2021.

[3] A. Corino, "A peculiar form of peripheral neuropathy, familiar atypical generalized amyloidosis with special involvement of the peripheral nerves," *Brain*, vol. 75,3, pp. 408–27., 1952.

[4] R. X. Lôpo and A. M. Jorge, "Amivis - demo." URL, Last accessed 15 August 2022.

[5] R. X. Lôpo and A. M. Jorge, "Amivis - github." URL, Last accessed 15 August 2022.

[6] SoGood2022, "Seventh workshop on data science for social good." URL, Last accessed 15 August 2022.

[7] J. Han, M. Kamber, and J. Pei, "Data mining: Concepts and techniques," Third Edition, 2012.

[8] M. Juodyte, "Overview: Data mining pipeline," 2017.

[9] R. K. Pearson, "The problem of disguised missing data," *SIGKDD Explor. Newsl.*, vol. 8, p. 83–92, jun 2018.

[10] J. Gama, A. P. d. L. Carvalho, K. Faceli, A. C. Lorena, and M. Oliveira, *Extração de Conhecimentos de Dados*. No. 3, Edições Sílabo, September 2017.

[11] C. C. Aggarwal, "Data mining, the textbook," 2015.

[12] K. Koperski, J. Adhikary, and J. Han, "Spatial data mining: progress and challenges survey paper," *Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada.*

[13] K. Lakshminarayan, S. A. Harp, R. Goldman, and T. Samad, "Imputation of missing data using machine learning techniques," p. 140–145, 1996.

[14] A. R. T. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons, "Review: A gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[15] M. Inês, T. Coelho, I. Conceição, F. Duarte-Ramos, M. de Carvalho, and J. Costa, "Epidemiology of transthyretin familial amyloid polyneuropathy in portugal: A nationwide study," *Neuroepidemiology*, vol. 51(3-4), p. 177–182, 2018.

[16] M. Pedroto, A. Jorge, J. Mendes-Moreira, and T. Coelho, "Predicting age of onset in ttr-fap patients with genealogical features," pp. 199–204, 2018.

[17] Y. Parman, D. Adams, L. Obici, L. Galán, V. Guergueltcheva, Suhr, O. B., and T. Coelho, "Sixty years of transthyretin familial amyloid polyneuropathy (ttr-fap) in europe: where are we now? a european network approach to defining the epidemiology and management patterns for ttr-fap," *Current opinion in neurology*, vol. 29 Suppl 1(Suppl 1), pp. S3–S13, 2016.

[18] J. Reinés, T. Vera, and M. Martín, "Epidemiology of transthyretin-associated familial amyloid polyneuropathy in the majorcan area: Son llàtzer hospital descriptive study," *Orphanet J Rare Dis 9*, vol. 29, 2014.

[19] P. N. Hawkins, Y. Ando, A. Dispenzeri, A. Gonzalez-Duarte, D. Adams, and O. B. Suhr, "Evolving landscape in the management of transthyretin amyloidosis," *Annals of Medicine*, vol. 47:8, pp. 625–638, 2015.

[20] T. Coelho, M. Inês, Conceição, I., M. Soares, M. de Carvalho, and J. Costa, "Natural history and survival in stage 1 val30met transthyretin familial amyloid polyneuropathy," *Neurology*, vol. 91(21), 2018.

[21] A. Mazzeo, G. Russo, M. Di Bella, F. Minutoli, C. Stancanelli, L. Gentile, S. Baldari, S. Carerj, A. Toscano, and G. Vita, "Transthyretin-related familial amyloid polyneuropathy (ttr-fap): A single-center experience in sicily, an italian endemic area," *Journal of neuromuscular diseases*, vol. 2(s2), p. S39–S48, 2015.

[22] A. Sousa, T. Coelho, J. Barros, and J. Sequeiros, "Genetic epidemiology of familial amyloidotic polyneuropathy (fap)-type i in povoa do varzim and vila do conde (north of portugal)," *American Journal of Medical Genetics (Neuropsychiatric Genetics)*, vol. 60, pp. 512–521, 1995.

[23] Y. Kato-Motozaki, K. Ono, K. Shima, A. Morinaga, T. Machiya, I. Nozaki, A. Shibata-Hamaguchi, Y. Furukawa, D. Yanase, C. Ishida, K. Sakajiri, and M. Yamada, "Epidemiology of familial amyloid polyneuropathy in japan: Identification of a novel endemic focus," *Journal of the neurological sciences*, vol. 270(1-2), p. 133–140, 2008.

[24] K. Choi, J. Seok, B. Kim, Y. Choi, H. Shin, I. Sunwoo, D. Kim, J. Sung, G. Lee, E. Jeon, N. Kim, J. Min, and J. Oh, "Characteristics of south korean patients with hereditary

transthyretin amyloidosis," *Journal of Clinical Neurology (Korea)*, vol. 14, pp. 537–541, Oct. 2018. Publisher Copyright: © 2018 Korean Neurological Association.

[25] M. Nöllenburg, *Human-Centered Visualization Environments*. Springer, 2006.

[26] C. Hupy, R. Weichelt, C. Wilson, and J. Hupy, *Extending into STEM: The Geospatial Education Initiative.*, pp. 95–106. 09 2016.

[27] A. Buckley, M. Gahegan, and K. Clarke, "Geographic visualization," January 2001.

[28] M. Harrower and S. I. Fabrikant, *Geographic Visualization: Concepts, Tools and Applications; The Role of Map Animation for Geographic Visualization*. May 2008.

[29] A. MacEachren, F. Boscoe, D. Haug, and L. Pickle, "Geographic visualization: designing manipulable maps for exploring temporally varying georeferenced statistics," pp. 87–94, 1998.

[30] A. Maceachren and M. Kraak, "Exploratory cartographic visualization advancing the agenda," *Computers and Geosciences*, vol. 23, pp. 335–343, 1997.

[31] T. Vindenes, M. Jordan, A. Tibbs, T. Stopka, D. Johnson, and J. Cochran, "A genotypic and spatial epidemiologic analysis of massachusetts' mycobacterium tuberculosis cases from 2012 to 2015," *Tuberculosis*, vol. 112, pp. 20–26, 2018.

[32] J. Gaudart, J. Landier, L. Huiart, E. Legendre, L. Lehot, M. K. Bendiane, L. Chiche, A. Petitjean, E. Mosnier, F. Kirakoya-Samadoulougou, J. Demongeot, R. Piarroux, and S. Rebaudet, "Factors associated with the spatial heterogeneity of the first wave of covid-19 in france: a nationwide geo-epidemiological study," *The Lancet Public Health*, vol. 6, no. 4, pp. e222–e231, 2021.

[33] C. Yang, L. Lu, J. L. Warren, J. Wu, Q. Jiang, T. Zuo, M. Gan, M. Liu, Q. Liu, K. DeRiemer, J. Hong, X. Shen, C. Colijn, X. Guo, Q. Gao, and T. Cohen, "Internal migration and transmission dynamics of tuberculosis in shanghai, china: an epidemiological, spatial, genomic analysis," *The Lancet Infectious Diseases*, vol. 18, no. 7, pp. 788–795, 2018.

[34] A. da Silva Sousa Júnior, N. V. Gonçalves, C. do Socorro Carvalho Miranda, B. de Oliveira Santos, R. de Oliveira, R. da Costa, S. K. da Trindade Noguchi, J. S. de Sousa Oliveira, E. Matsumura, and V. R. da Cunha Menezes Palácios, "Cutaneous leishmaniasis spatial distribution and epidemiological and environmental risk factors in cametá, state of pará, brazil," *The Brazilian journal of infectious diseases : an official publication of the Brazilian Society of Infectious Diseases*, vol. 24(4), p. 330–336, 2020.

[35] S. R. Pordanjani, A. Kavousi, B. Mirbagheri, A. Shahsavani, and K. Etemad, "Spatial analysis and geoclimatic factors associated with the incidence of acute lymphoblastic leukemia in iran during 2006–2014: An environmental epidemiological study," *Environmental Research*, vol. 202, p. 111662, 2021.

[36] A. Angelou, I. Kioutsioukis, and N. I. Stilianakis, "A climate-dependent spatial epidemiological model for the transmission risk of west nile virus at local scale," *One Health*, vol. 13, p. 100330, 2021.

[37] C. A. L. Machado, A. da Paixão Sevá, F. Dantas-Torres, and M. C. Horta, "Spatial analysis and epidemiological profile of visceral leishmaniasis, northeastern brazil: A cross-sectional study," *Acta Tropica*, vol. 208, p. 105520, 2020.

[38] D. Jing, H. Zhao, and R. Ou, "Epidemiological characteristics and spatiotemporal analysis of acute hemorrhagic conjunctivitis from 2004 to 2018 in chongqing, china," *Sci Rep*, vol. 10 9286, 2020.

[39] T. Smith, "Autocorrelation." URL, Last accessed 19 January 2022.

[40] Esri, "How local outlier analysis works." URL, Last accessed 15 January 2022.

[41] H. Wickham, "Package read 2.1.2." URL, Last accessed 23 July 2022.

[42] M. Gagolewski, "Package stringi 1.7.6." URL, Last accessed 23 July 2022.

[43] H. Wickham, R. François, L. Henry, and K. Müller, "Package dplyr 1.0.9." URL, Last accessed 23 July 2022.

[44] H. Wickham, "Package stringr 1.4.0." URL, Last accessed 23 July 2022.

[45] D. Kahle, H. Wickham, and S. Jackson, "Package ggmap 3.0.0." URL, Last accessed 23 July 2022.

[46] H. Wickham, "Package tidyverse 1.3.1." URL, Last accessed 23 July 2022.

[47] Google, "Google cloud platform." URL, Last accessed 24 July 2022.

[48] M. Kuhn, "Package caret 6.0-92." URL, Last accessed 24 July 2022.

[49] E. Heinzen, J. Sinnwell, E. Atkinson, T. Gunderson, and G. Dougherty, "Package arsenal 3.6.3." URL, Last accessed 24 July 2022.

[50] E. Pebesma, "Package sf 1.0-7." URL, Last accessed 27 July 2022.

[51] T. Appelhans, F. Detsch, C. Reudenbach, and S. Woellauer, "Package mapview 2.11.0." URL, Last accessed 27 July 2022.

[52] E. Pebesma and R. Bivand, "Package sp 1.4-7." URL, Last accessed 27 July 2022.

[53] H. Wickham, W. Chang, L. Henry, T. L. Pedersen, K. Takahashi, C. Wilke, K. Woo, H. Yutani, and D. Dunnington, "Package ggplot2 3.3.6." URL, Last accessed 27 July 2022.

[54] D. Dunnington, "Package ggspatial 1.1.5." URL, Last accessed 28 July 2022.

[55] M. Padgham, B. Rudis, R. Lovelace, and M. Salmon, "Package osmdata 0.1.9." URL, Last accessed 28 July 2022.

[56] K. Slowikowski, "Package ggrepel 0.9.1." URL, Last accessed 28 July 2022.

[57] R. A. Becker, A. R. Wilks, R. Brownrigg, T. P. Minka, and A. Deckmyn, "Package maps 3.4.0." URL, Last accessed 28 July 2022.

[58] R. A. Becker, A. R. Wilks, and R. Brownrigg, "Package mapdata 2.3.0." URL, Last accessed 28 July 2022.

[59] J. P. Decorps, "Package gadmtools 3.9-1." URL, Last accessed 28 July 2022.

[60] H. Wickham, "Package scales 1.2.0." URL, Last accessed 28 July 2022.

[61] S. Urbanek, "Package tiff 0.1-11." URL, Last accessed 28 July 2022.

[62] P. Murrell, "Package grid 4.2.0." URL, Last accessed 28 July 2022.

[63] M. van der Loo, "Package stringdist 0.9.8." URL, Last accessed 28 July 2022.

[64] Adobe, "Package shinywidgets 0.7.0." URL, Last accessed 4 August 2022.

[65] W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges, "Package shiny 1.7.1." URL, Last accessed 4 August 2022.

[66] V. Perrier, F. Meyer, and D. Granjon, "Package shinywidgets 0.7.0." URL, Last accessed 4 August 2022.

[67] W. Chang, "Package shinythemes 1.2.0." URL, Last accessed 4 August 2022.

[68] Y. Xie, J. Cheng, and X. Tan, "Package dt 0.23." URL, Last accessed 4 August 2022.

[69] R. C. Team, "Package tools 3.6.2." URL, Last accessed 4 August 2022.

[70] J. Cheng, B. Karambelkar, and Y. Xie, "Package leaflet 2.1.1." URL, Last accessed 4 August 2022.

[71] H. Wickham and M. Girlich, "Package tidyr 1.2.0." URL, Last accessed 4 August 2022.

[72] E. Neuwirth, "Package rcolorbrewer 1.1-3." URL, Last accessed 4 August 2022.

[73] Y. Ando, T. Coelho, and J. Berk, "Guideline of transthyretin-related hereditary amyloidosis for clinicians.," *Orphanet J Rare Dis*, vol. 8, 31, 2013.

[74] R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, K. Kuroptev, M. O'Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, and F. Yasmeen, "Package forecast 1.1-3." URL, Last accessed 4 August 2022.

[75] T. Jombart, Z. N. Kamvar, and R. FitzJohn, "Package incidence 1.1-3." URL, Last accessed 4 August 2022.

Appendix A

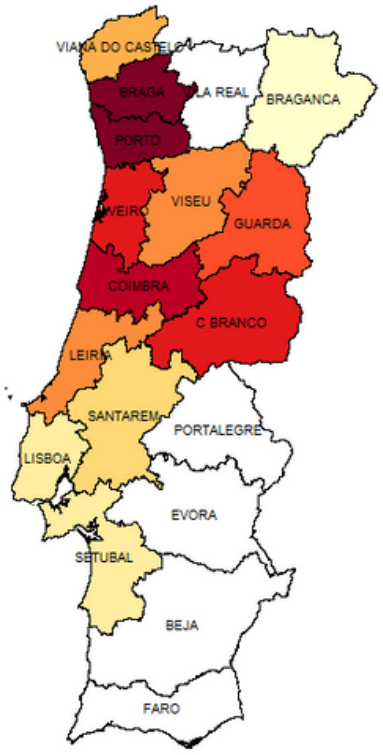
Portugal TTR-FAP GVis by Origin and Residence

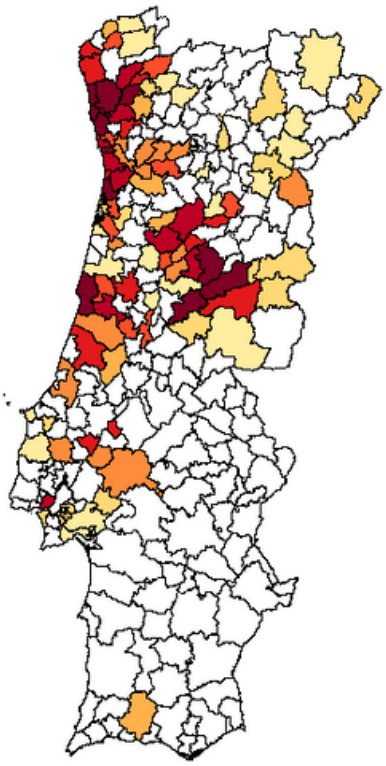
´

AmiVis Sheet

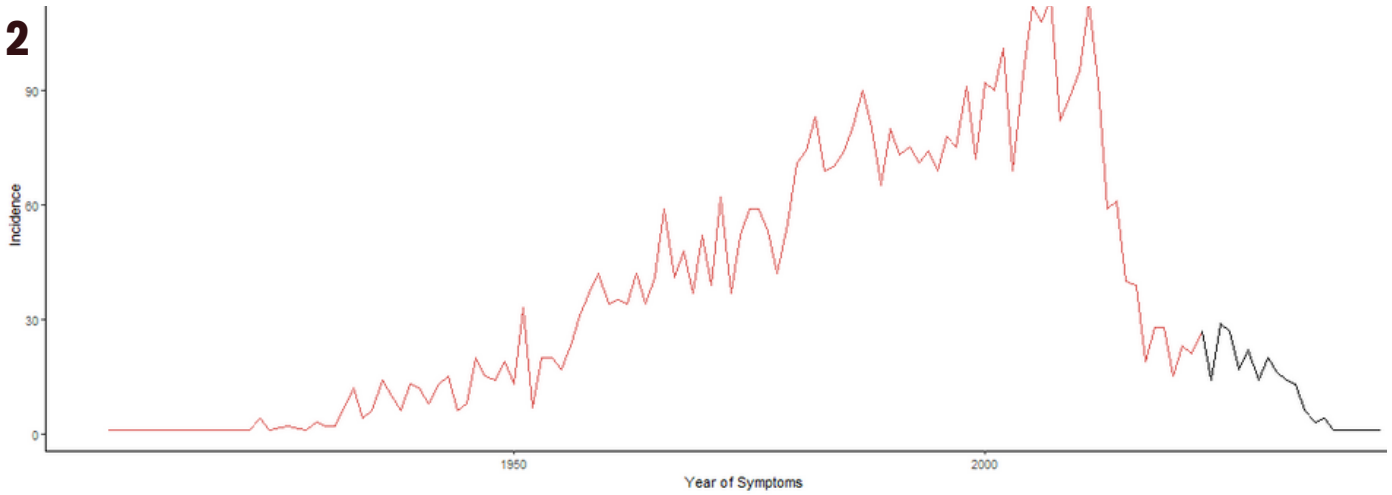
Portugal TTR-FAP GeoVisualisation by Origin

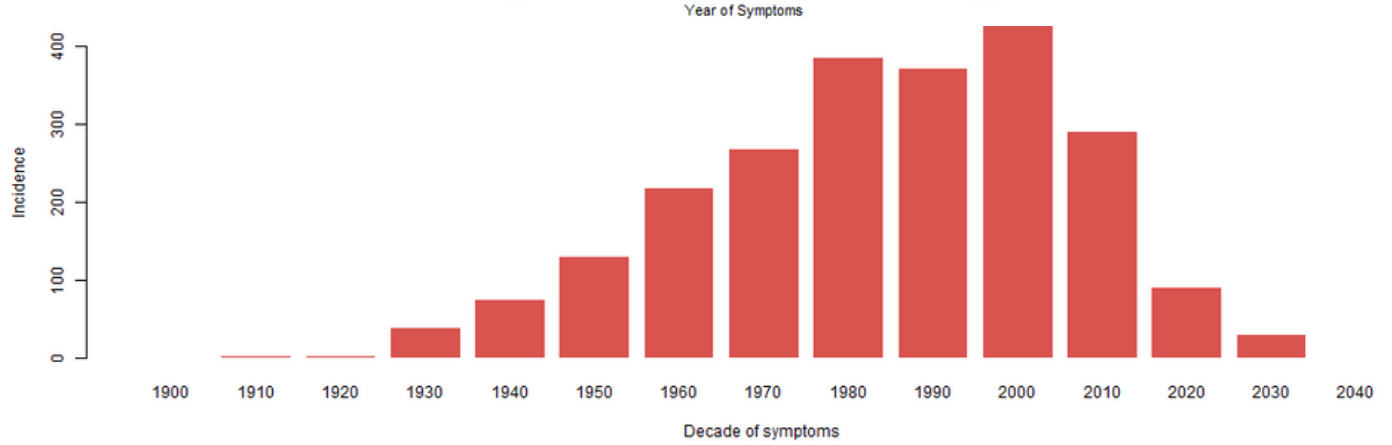
1





2

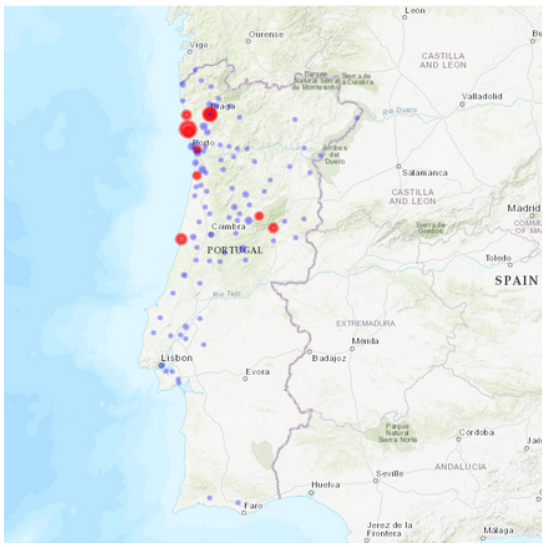


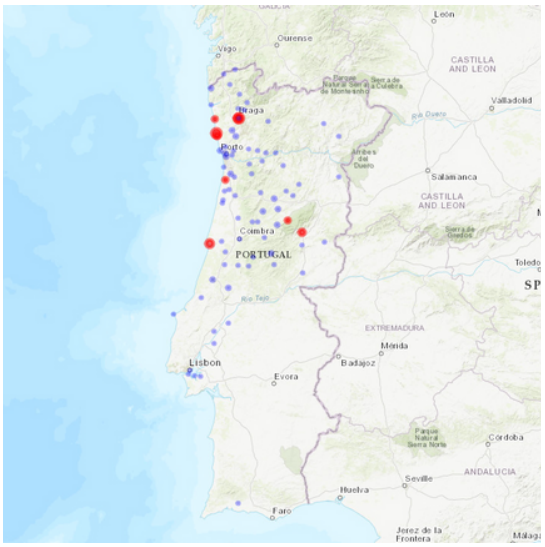


AmiVis Sheet

Portugal TTR-FAP GeoVisualisation by Origin

3

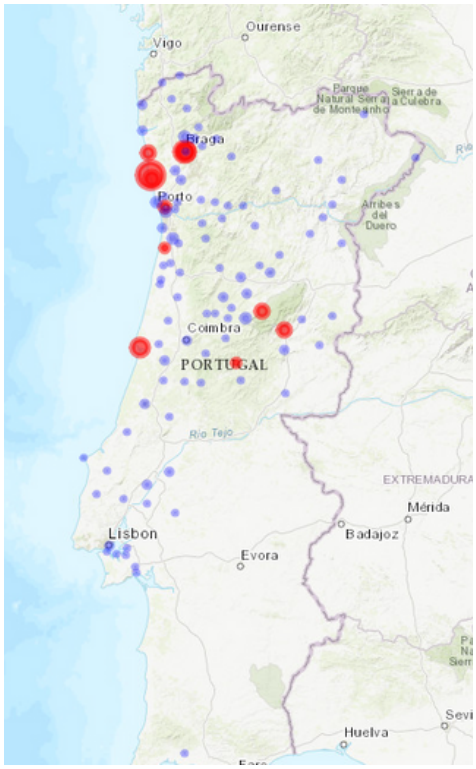


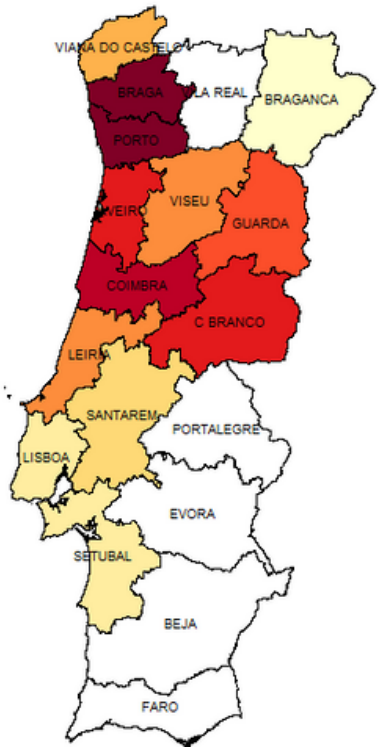


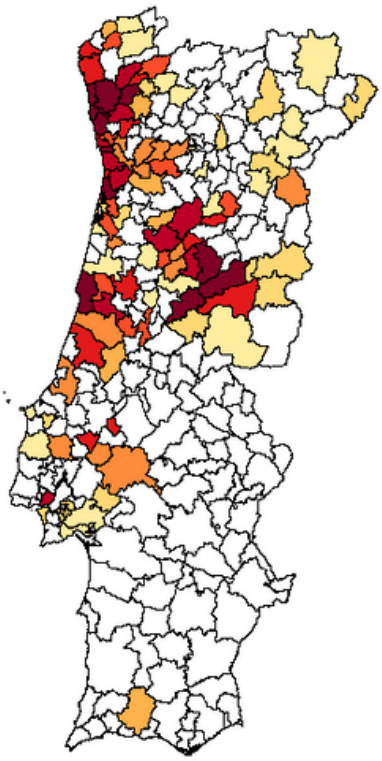
AmiVis Sheet

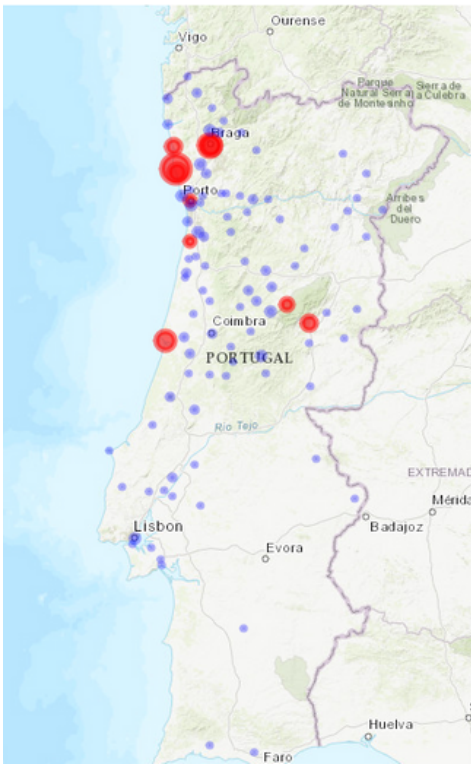
Portugal TTR-FAP GeoVisualisation by Origin

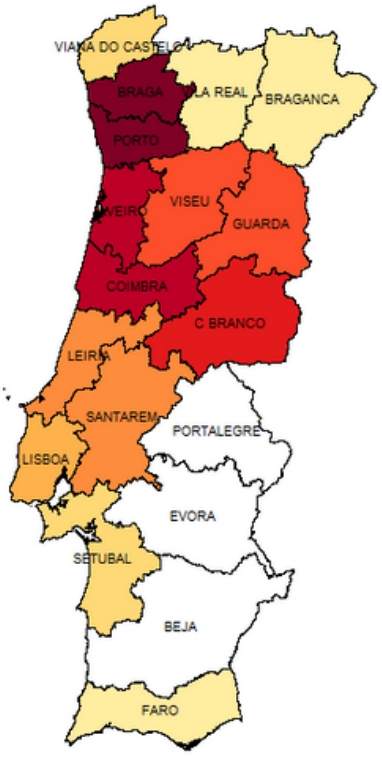
4

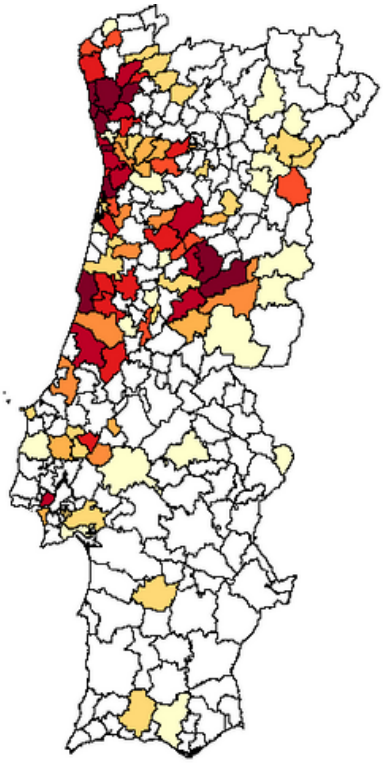








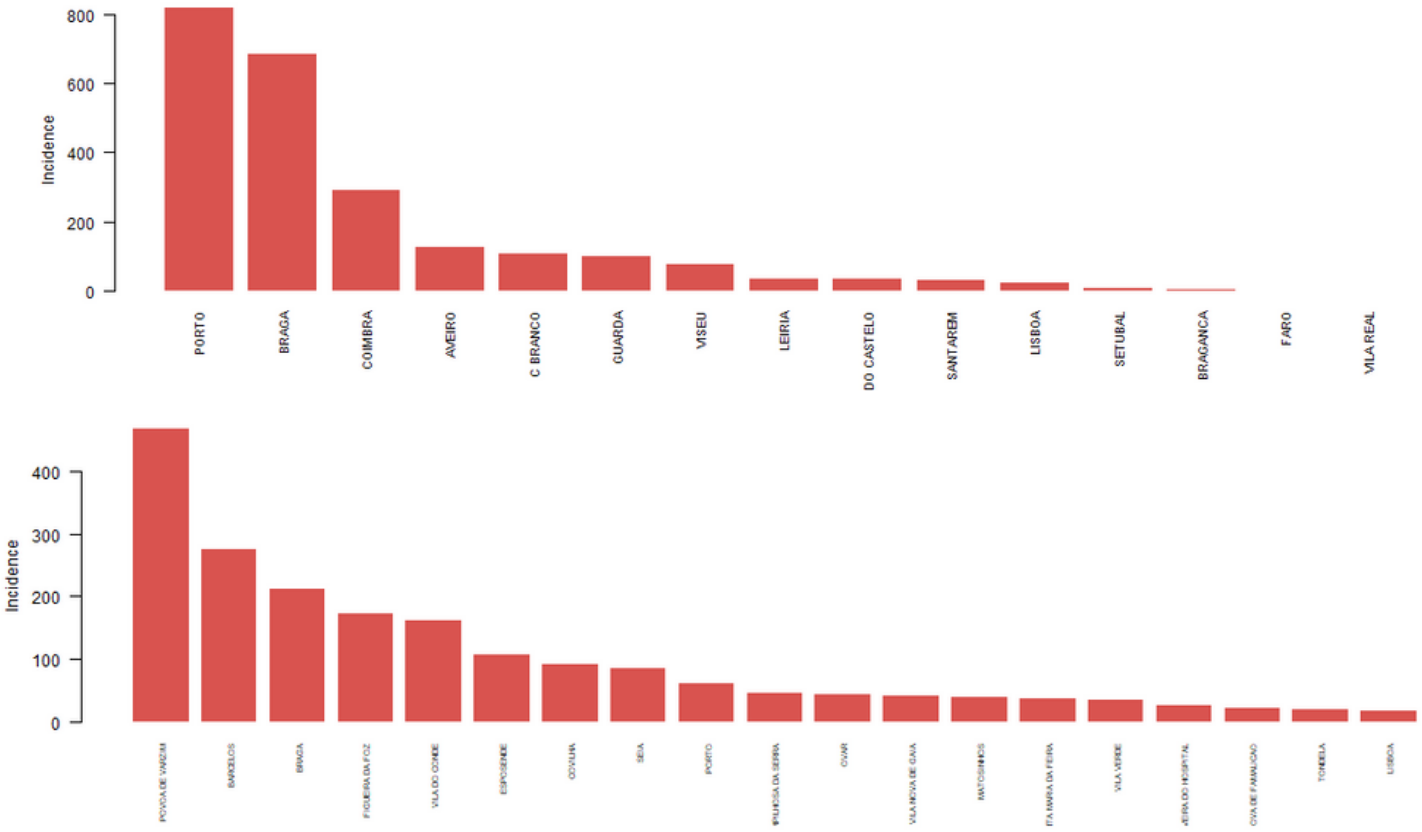




AmiVis Sheet

Portugal TTR-FAP GeoVisualisation by Origin

5



Captions:

1 - Static GeoVisualisations of the incidence of TTR-FAP patients with geographical and territorial maps by district and county.

2 - Timeseries and cases per decade.

3 - Overtime Geovisualisation with geographic map for all periods of 25 years after the 1st case in the dataset in 1907.

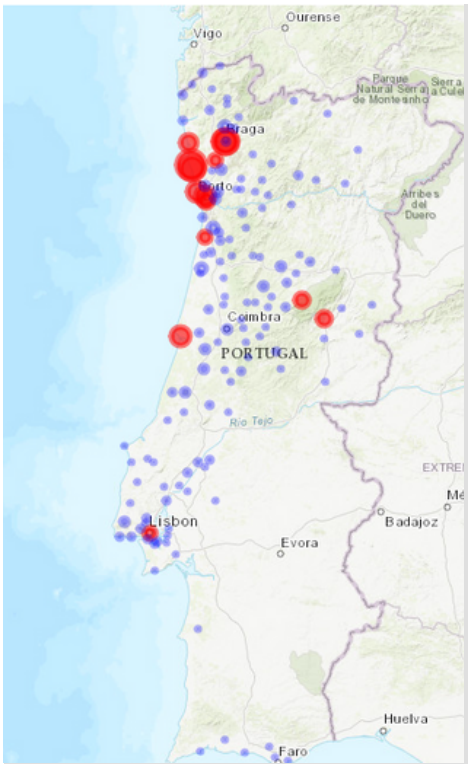
4 - Comparison of visualisations for men (above) and women (below).

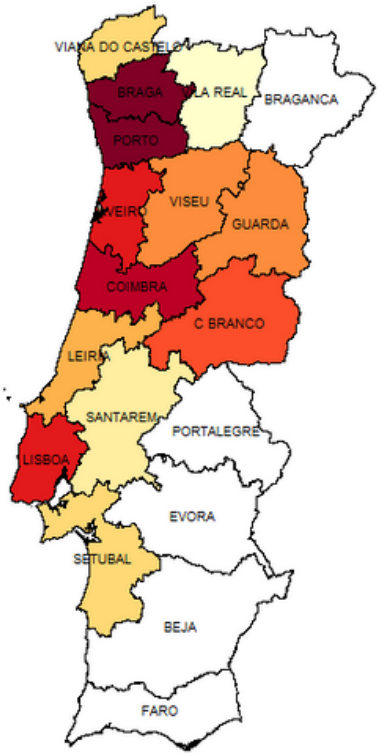
5 - Incidence for each district and county.

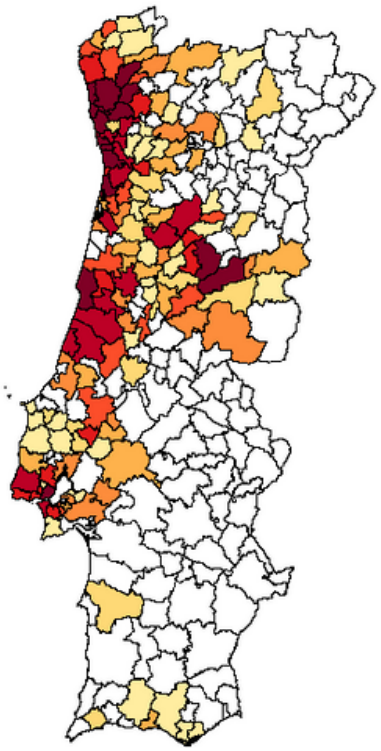
AmiVis Sheet

PORTUGAL TTR-FAP GEOVISUALISATION BY RESIDENCE

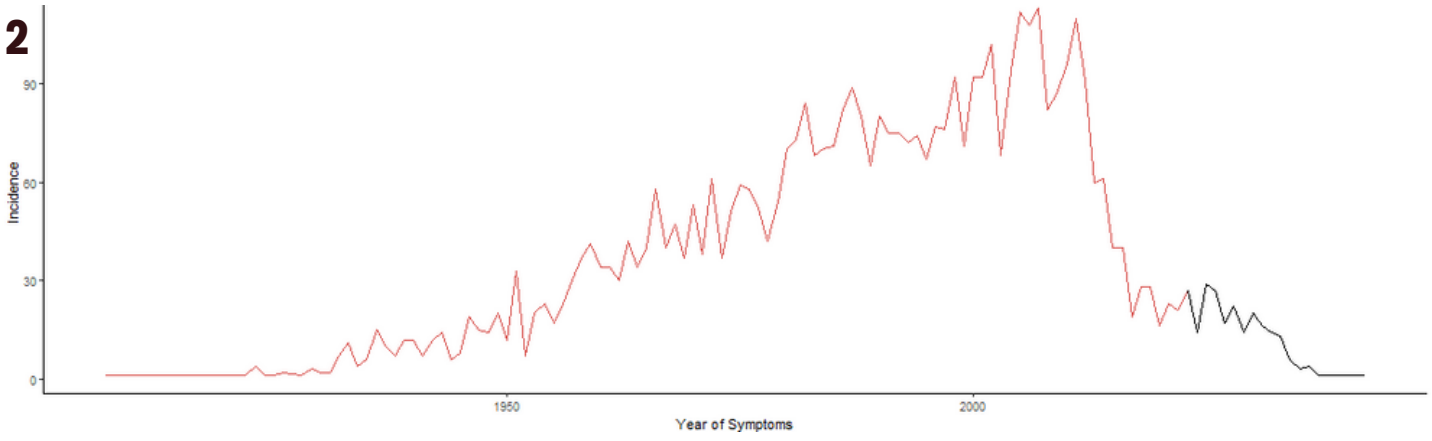
1

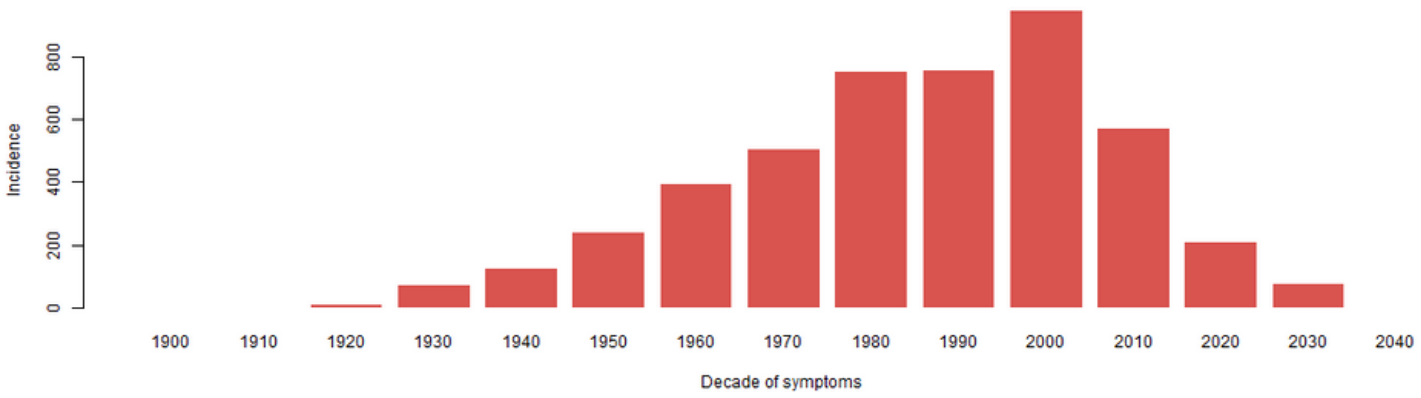






2

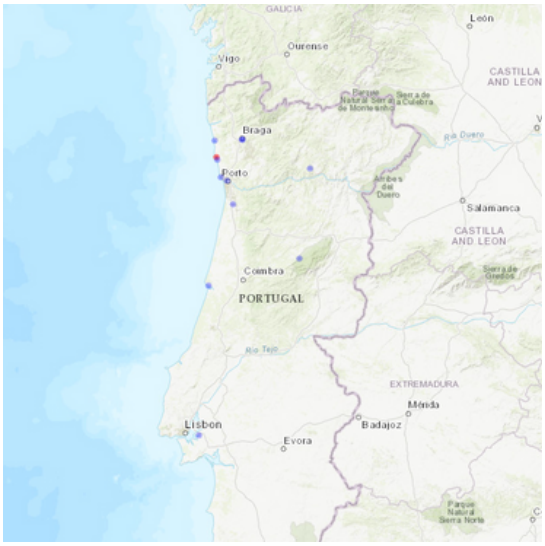




AmiVis Sheet

Portugal TTR-FAP GeoVisualisation by Residence

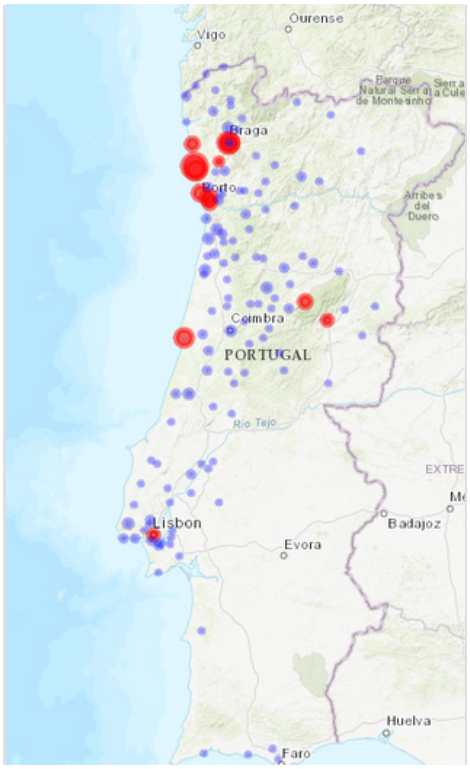
3

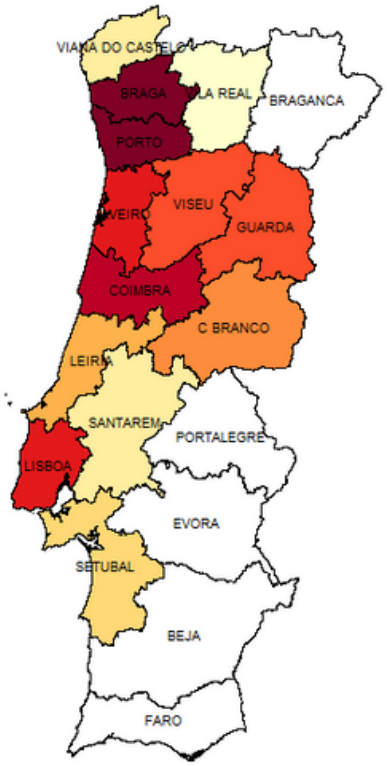


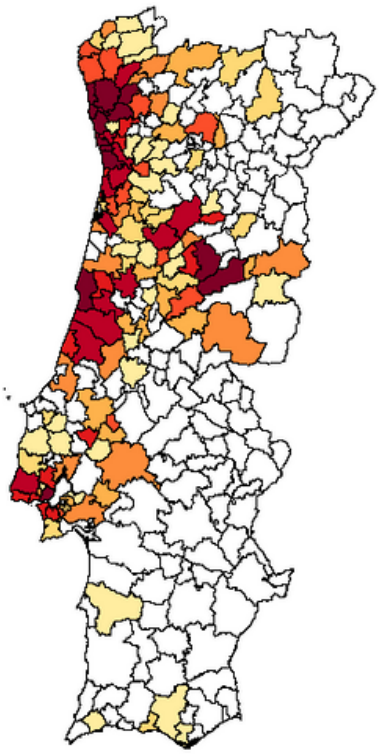
AmiVis Sheet

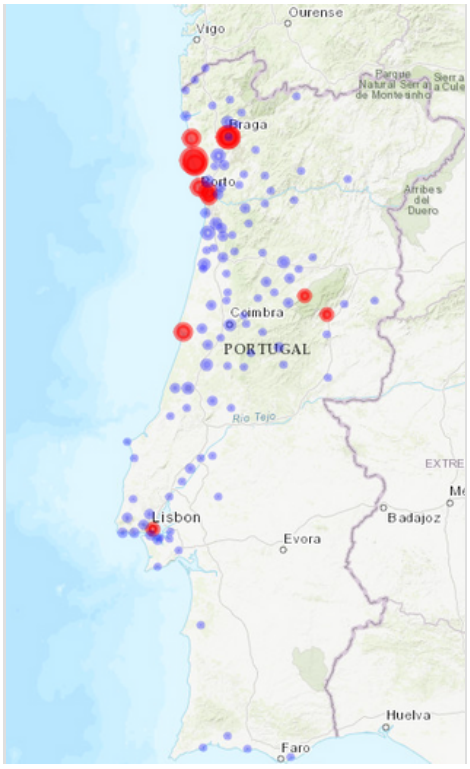
Portugal TTR-FAP GeoVisualisation by Residence

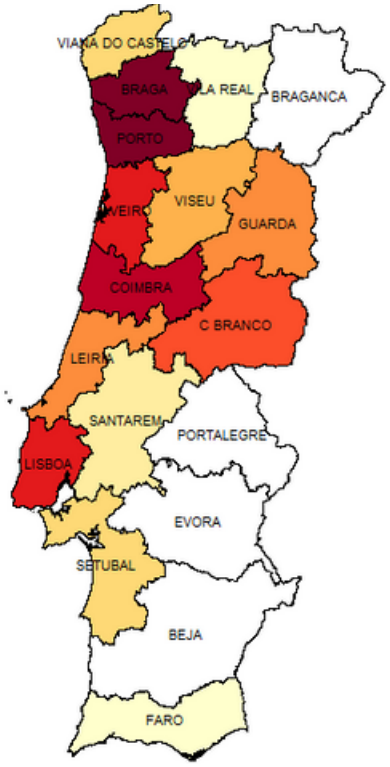
4

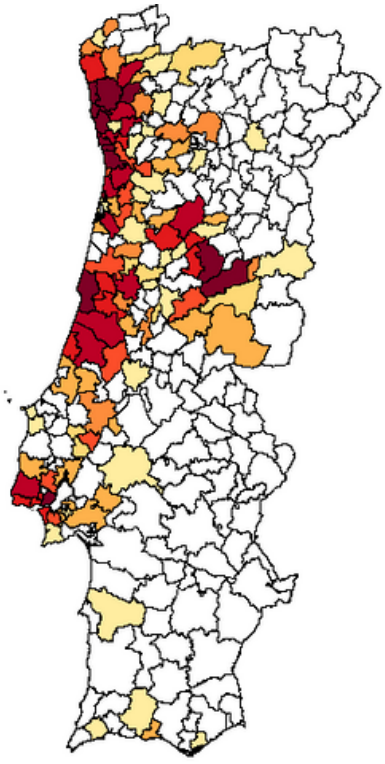






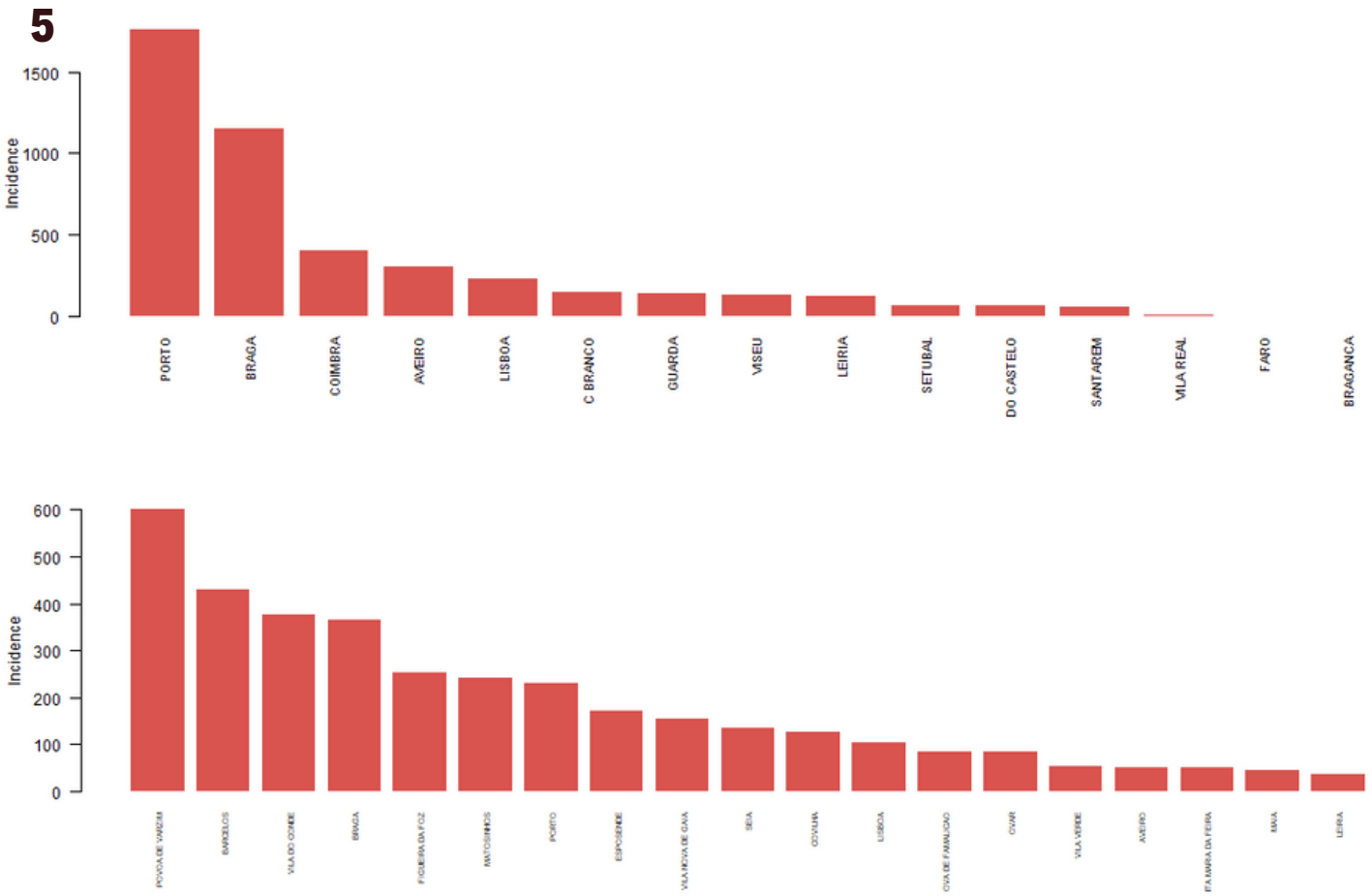






AmiVis Sheet

Portugal TTR-FAP GeoVisualisation by residence



Captions:

1 - Static GeoVisualisations of the incidence of TTR-FAP patients with geographical and territorial maps by district and county.

2 - Timeseries and cases per decade.

3 - Overtime Geovisualisation with geographic map for all periods of 25 years after the 1st case in the dataset in 1907.

4 - Comparison of visualisations for men (above) and women (below).

5 - Incidence for each district and county.

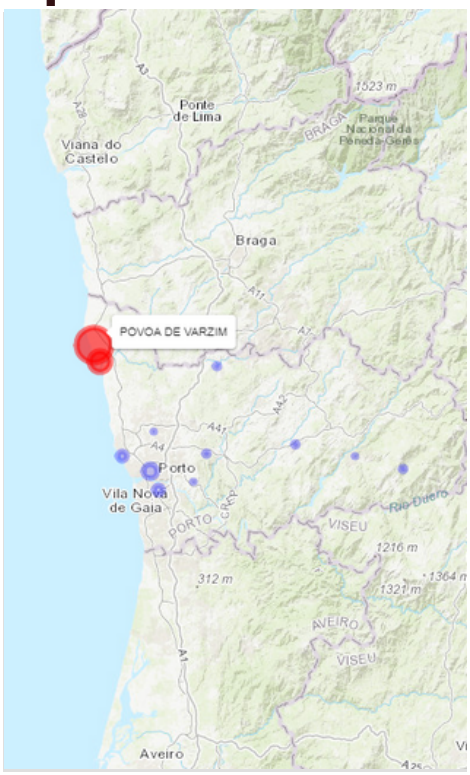
Appendix B

Porto District TTR-FAP GVis by Origin and Residence

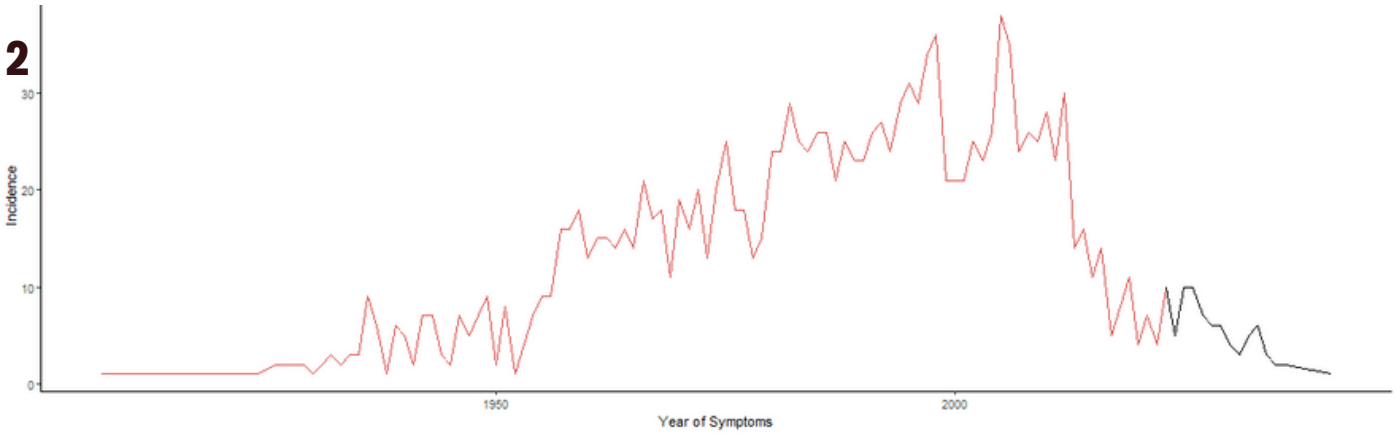
AmiVis Sheet

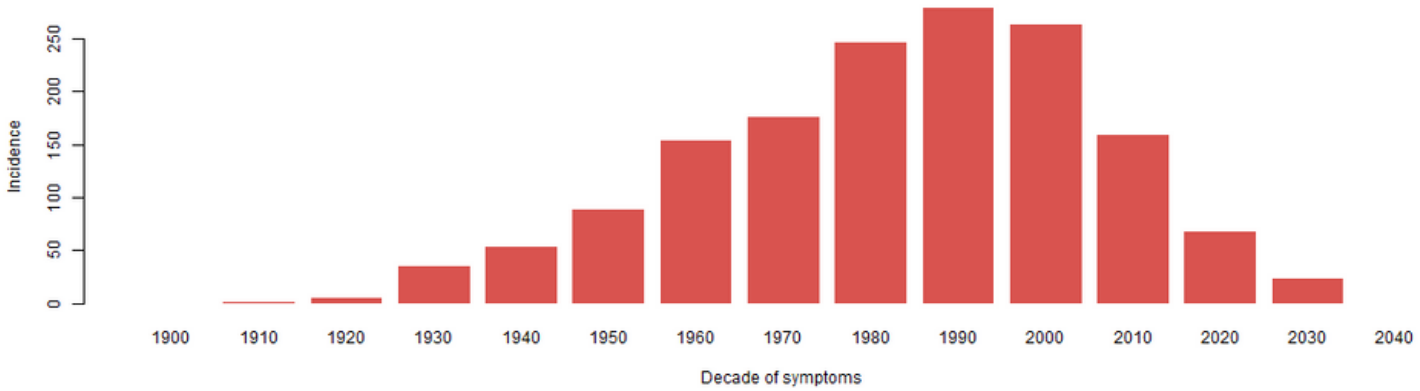
Porto District TTR-FAP GeoVisualisation by Origin

1



2

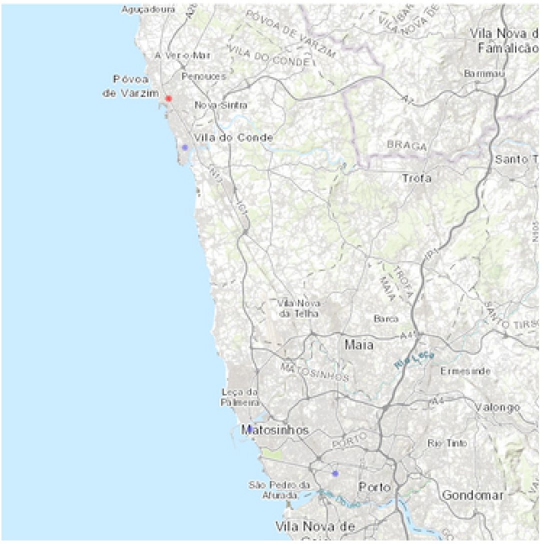


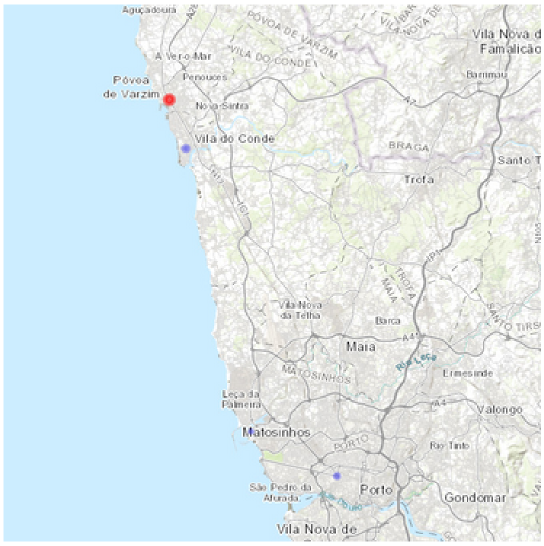


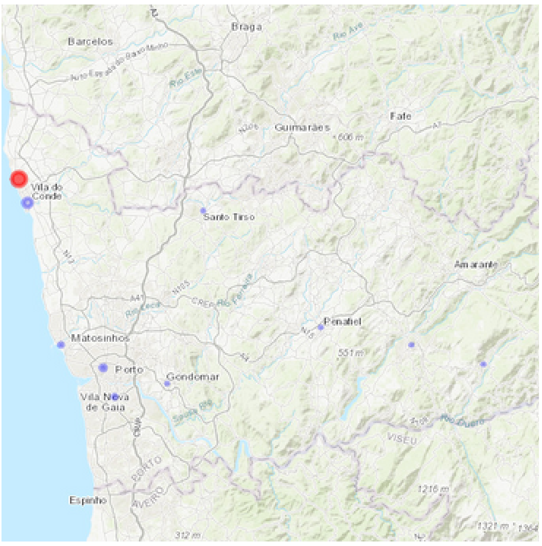
AmiVis Sheet

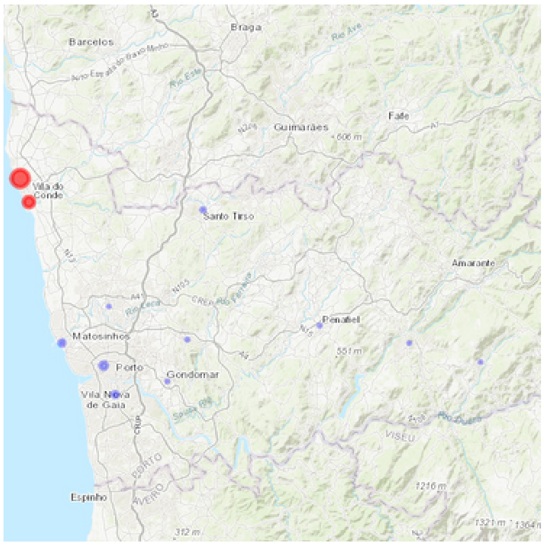
Porto District TTR-FAP GeoVisualisation by Origin

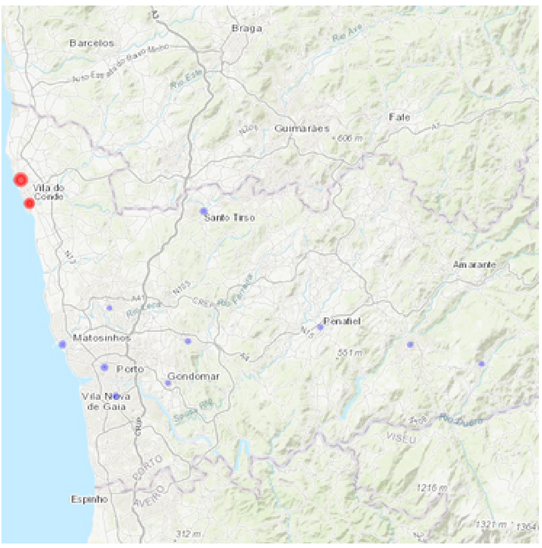
3







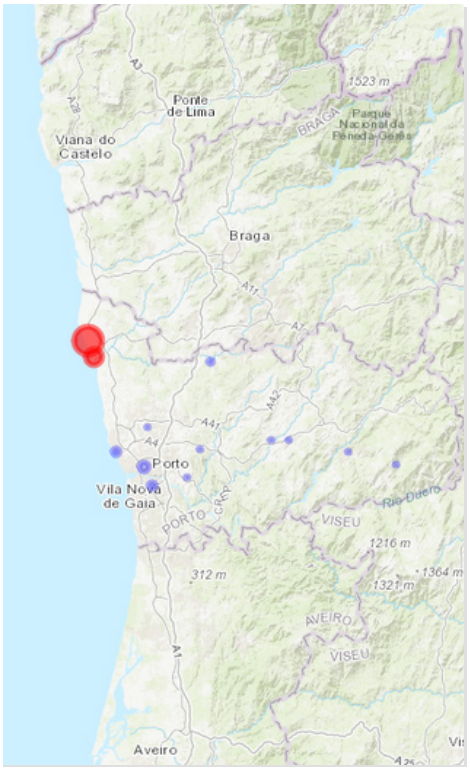


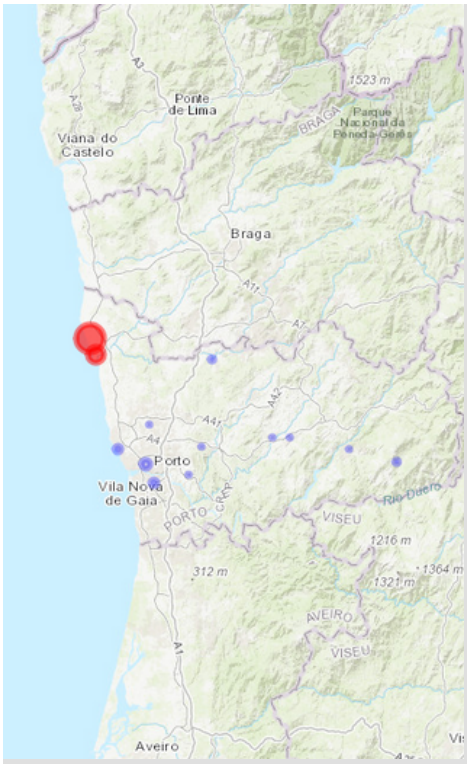


AmiVis Sheet

Porto District TTR-FAP GeoVisualisation by Origin

4

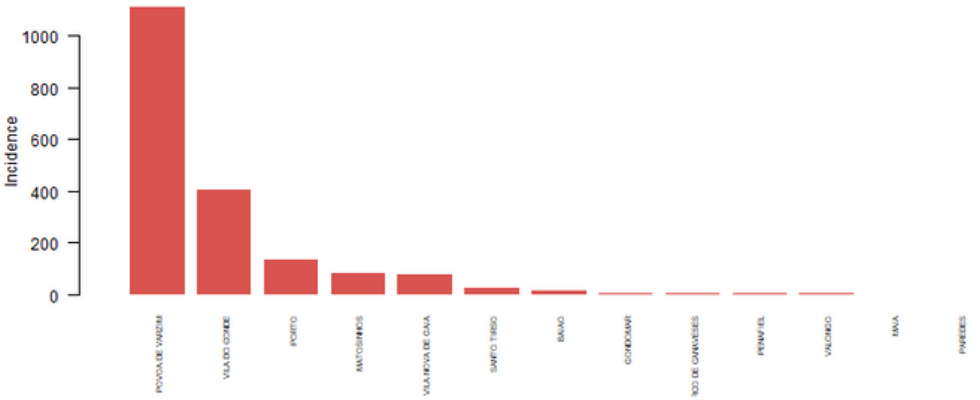




AmiVis Sheet

Porto District TTR-FAP GeoVisualisation by Origin

5



Captions:

1 - Static GeoVisualisations of the incidence of TTR-FAP patients with geographical and territorial maps by district and county.

2 - Timeseries and cases per decade.

3 - Overtime Geovisualisation with geographic map for all periods of 25 years after the 1st case in the dataset in 1907.

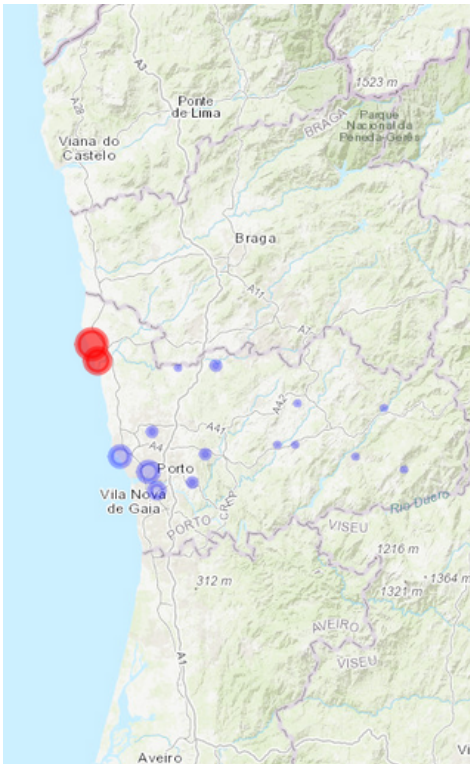
4 - Comparison of visualisations for men (above) and women (below).

5 - Incidence for each district and county.

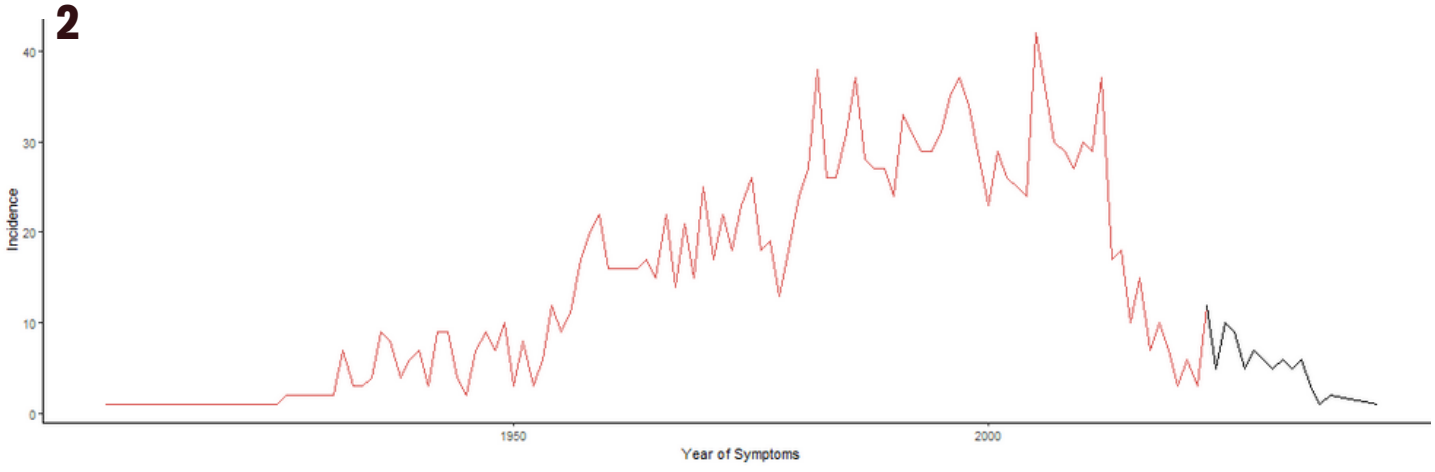
AmiVis Sheet

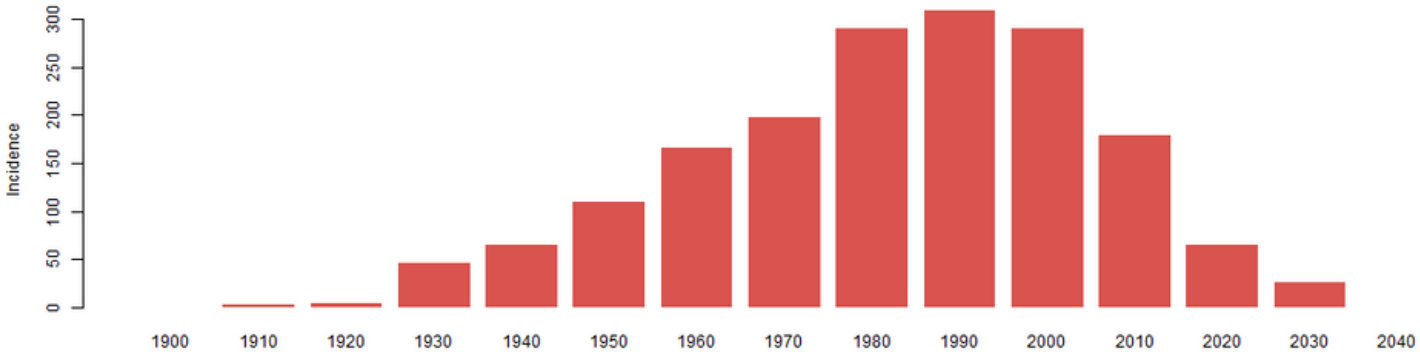
Porto District TTR-FAP GeoVisualisation by Residence

1



2

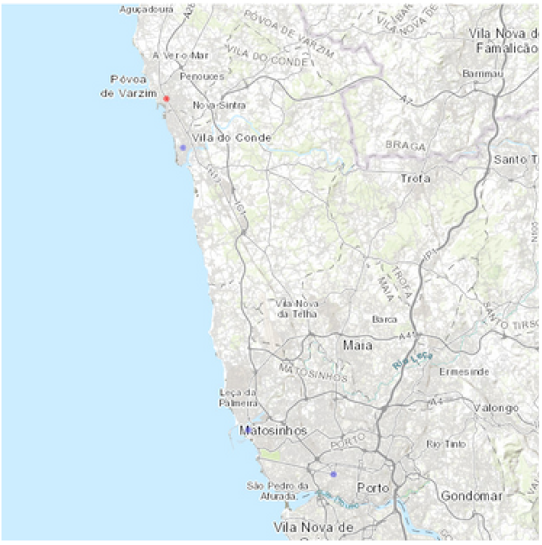


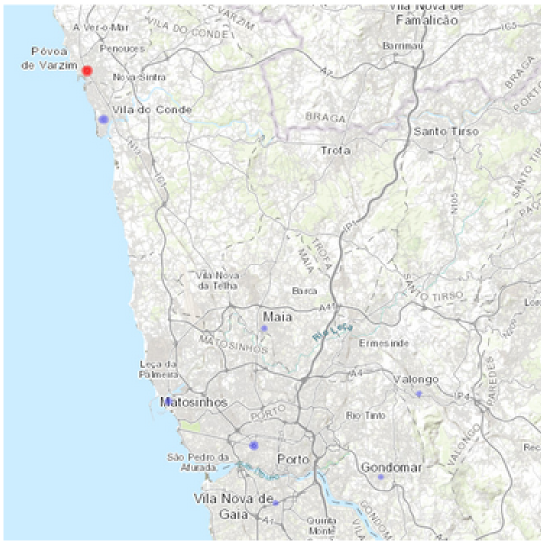


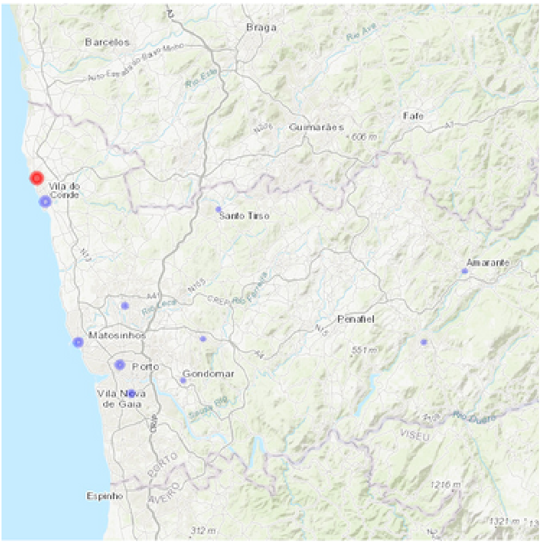
AmiVis Sheet

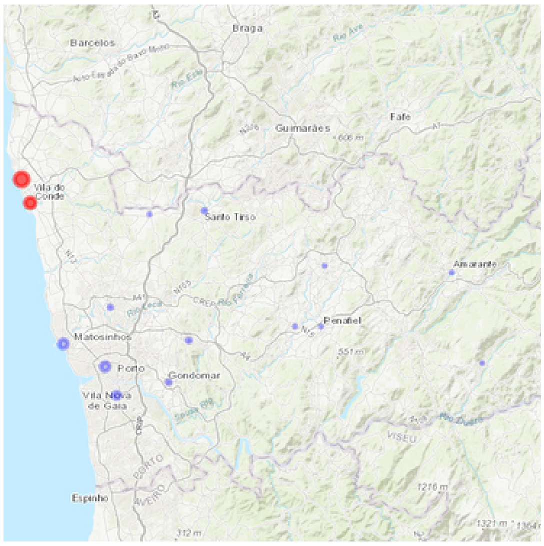
Porto District TTR-FAP GeoVisualisation by Residence

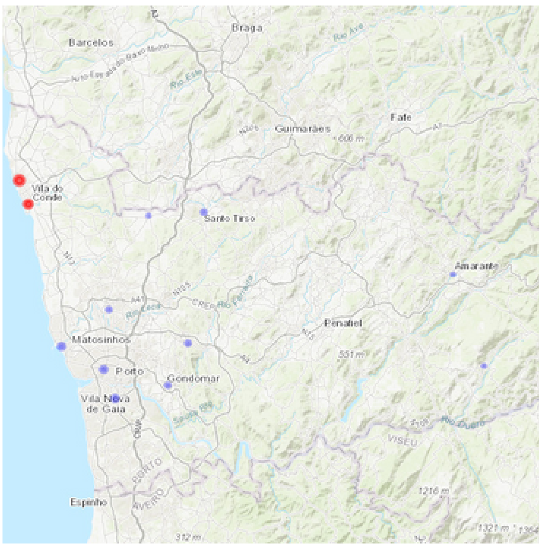
3







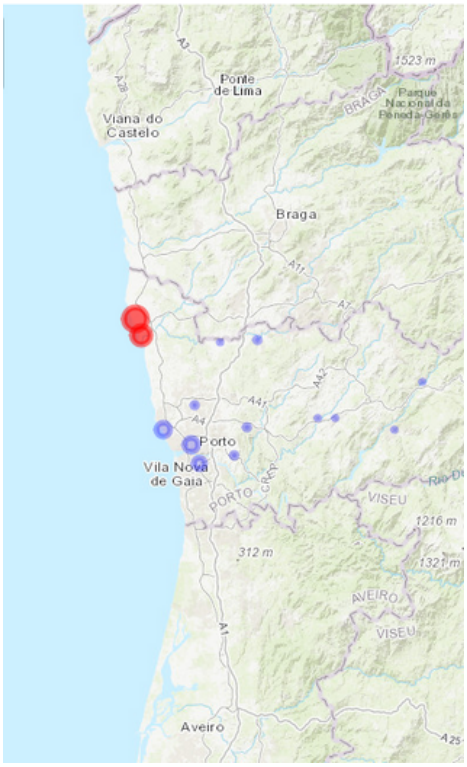


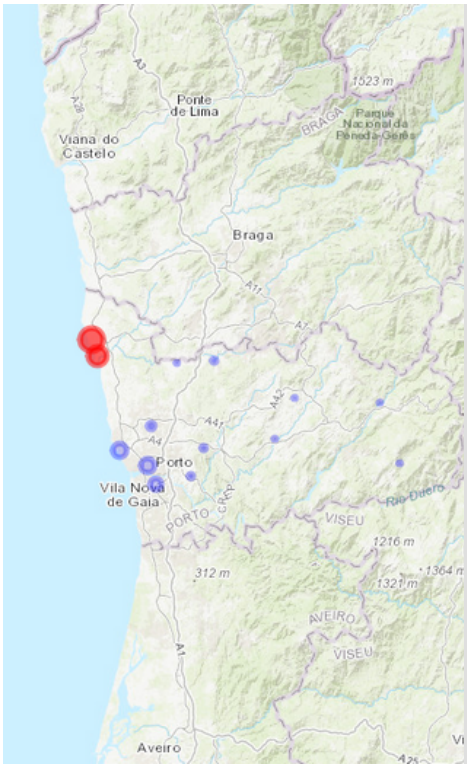


AmiVis Sheet

Porto District TTR-FAP GeoVisualisation by Residence

4

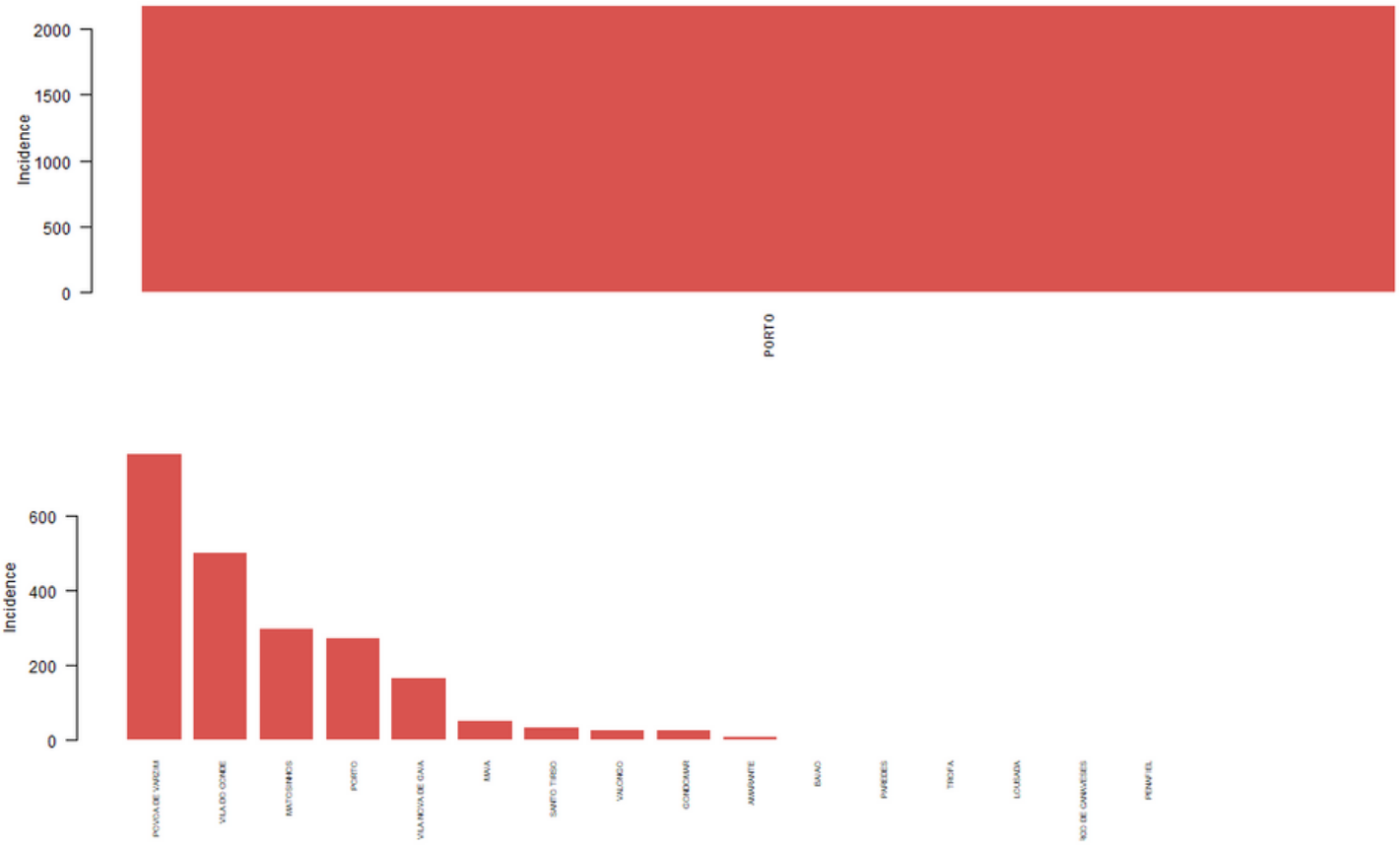




AmiVis Sheet

Porto District TTR-FAP GeoVisualisation by Residence

5



Captions:

1 - Static GeoVisualisations of the incidence of TTR-FAP patients with geographical and territorial maps by district and county.

2 - Timeseries and cases per decade.

3 - Overtime Geovisualisation with geographic map for all periods of 25 years after the 1st case in the dataset in 1907.

4 - Comparison of visualisations for men (above) and women (below).

5 - Incidence for each district and county.

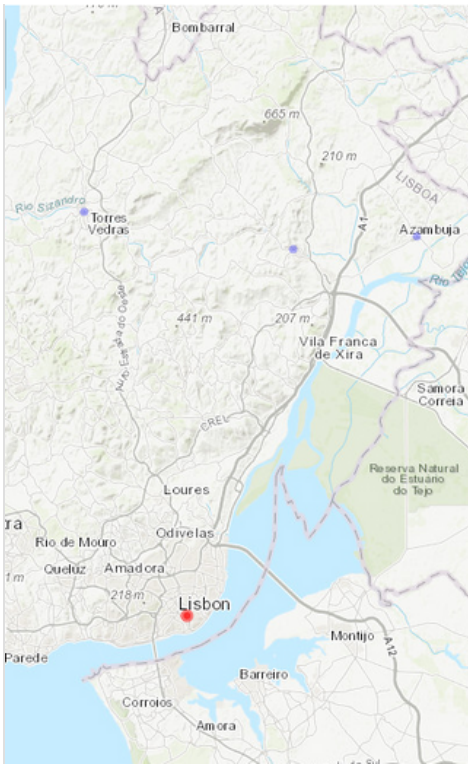
Appendix C

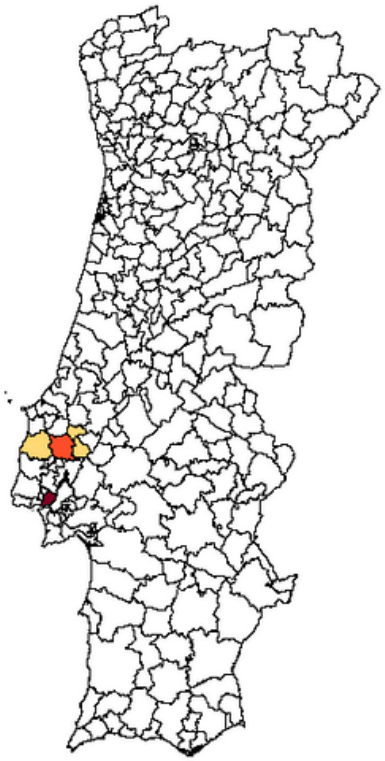
Lisbon District TTR-FAP GVis by Origin and Residence

AmiVis Sheet

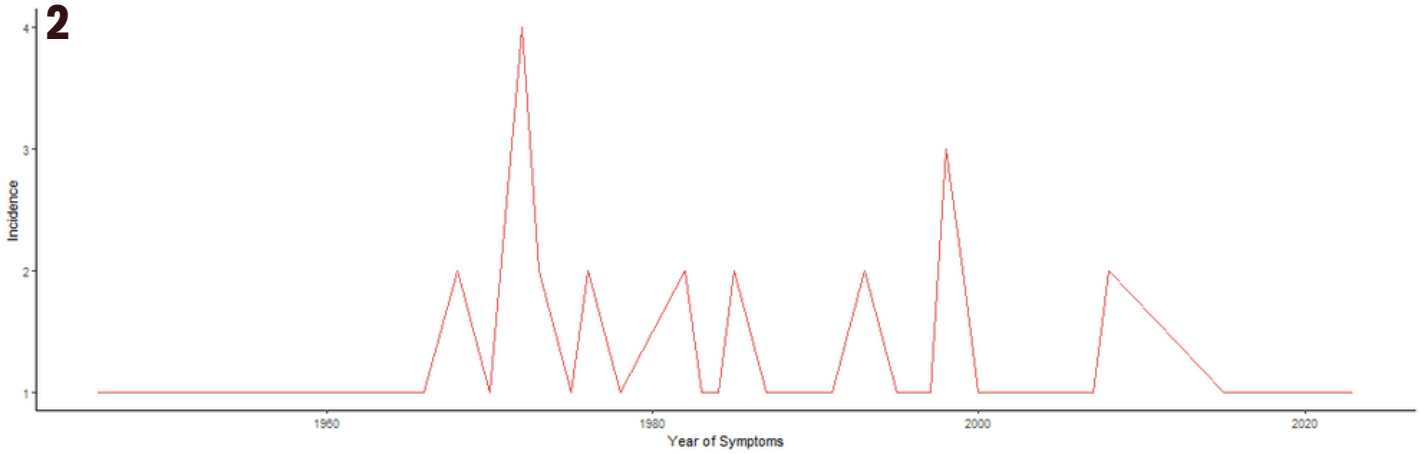
Lisboa District TTR-FAP GeoVisualisation by Origin

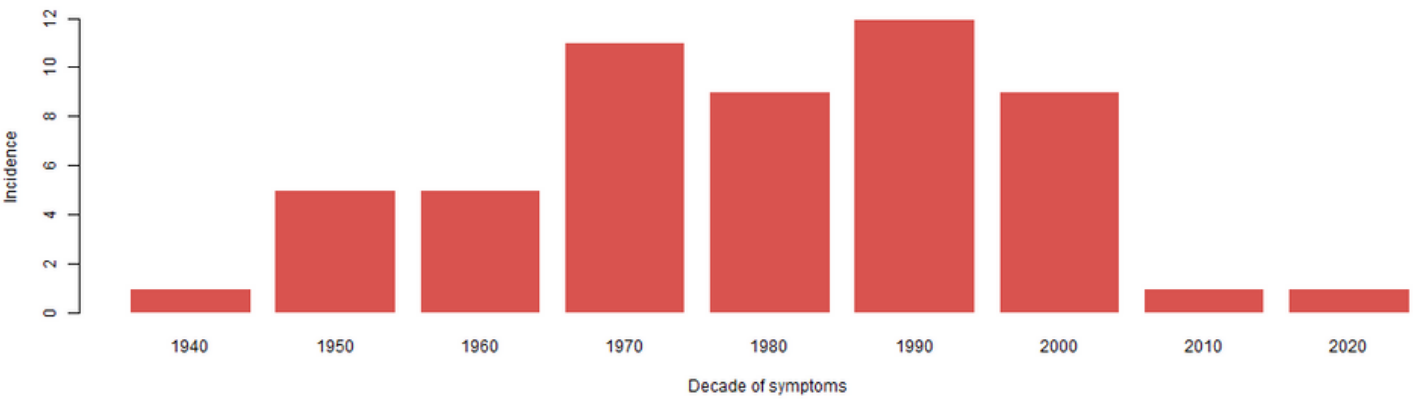
1





2





AmiVis Sheet

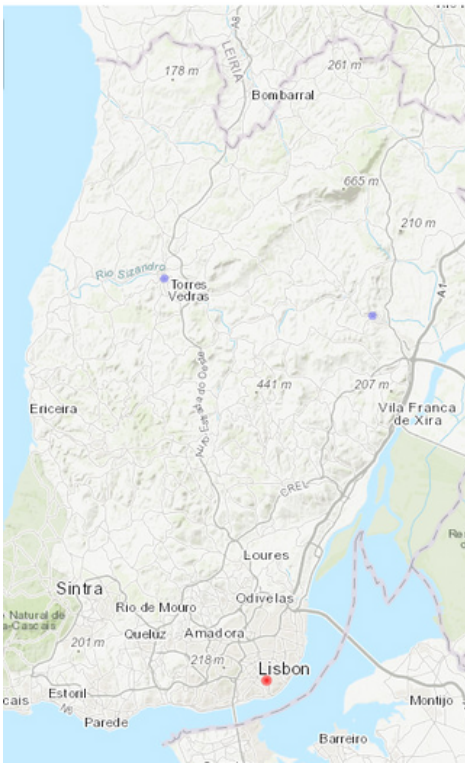
Lisboa District TTR-FAP GeoVisualisation by Origin

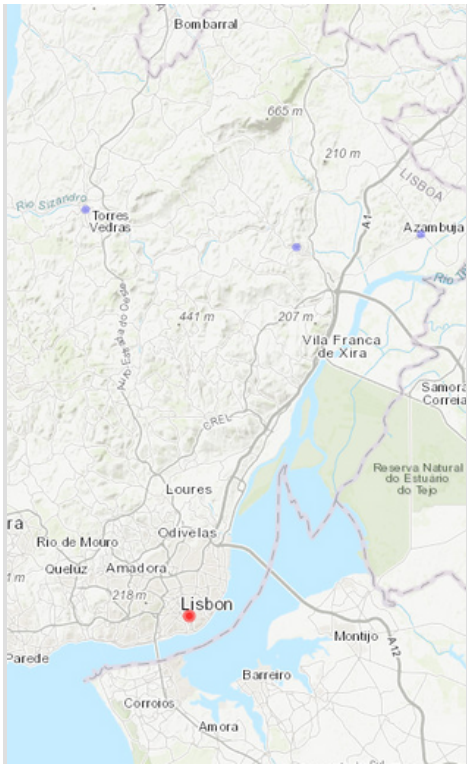
3

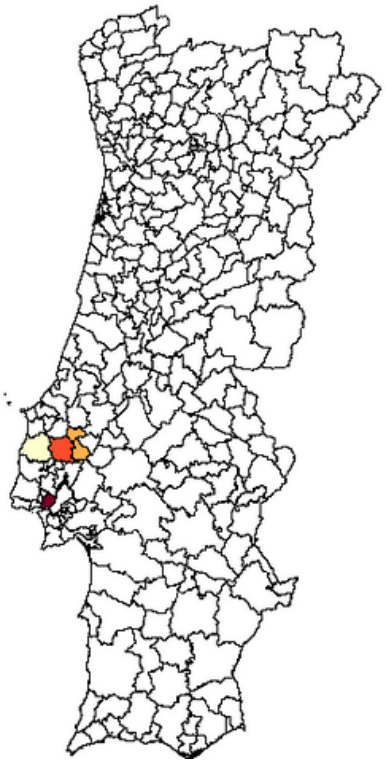
AmiVis Sheet

Lisboa District TTR-FAP GeoVisualisation by Origin

4



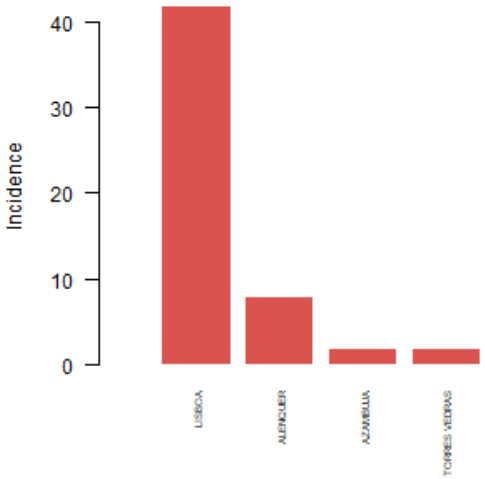




AmiVis Sheet

Lisboa District TTR-FAP GeoVisualisation by Origin

5



Captions:

1 - Static GeoVisualisations of the incidence of TTR-FAP patients with geographical and territorial maps by district and county.

2 - Timeseries and cases per decade.

3 - Overtime Geovisualisation with geographic map for all periods of 25 years after the 1st case in the dataset in 1907.

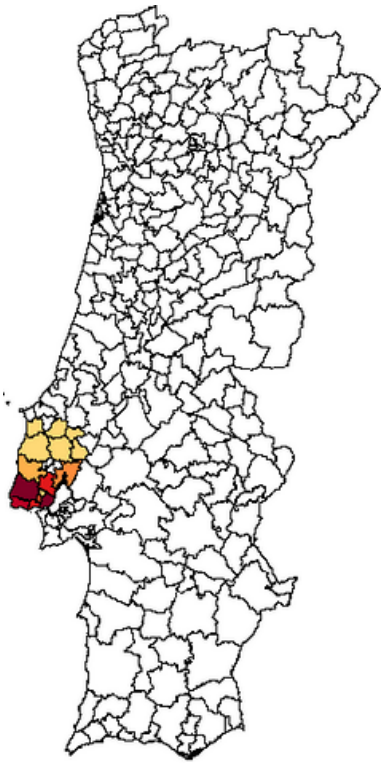
4 - Comparison of visualisations for men (above) and women (below).

5 - Incidence for each district and county.

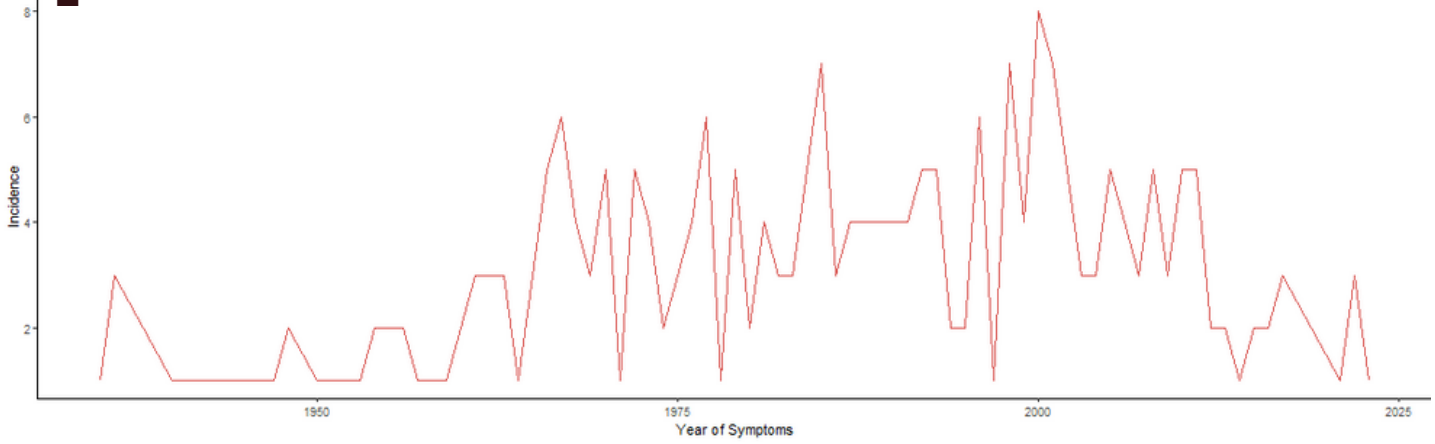
AmiVis Sheet

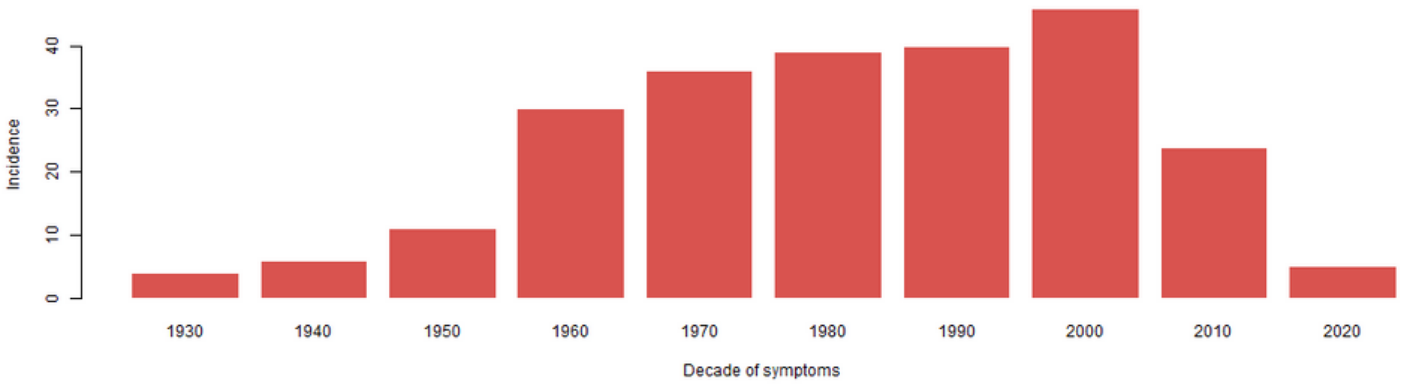
Lisboa District TTR-FAP GeoVisualisation by Residence

1



2

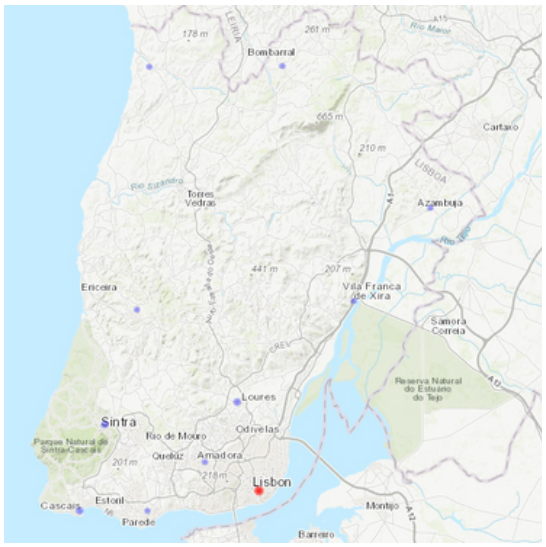




AmiVis Sheet

Lisboa District TTR-FAP GeoVisualisation by Residence

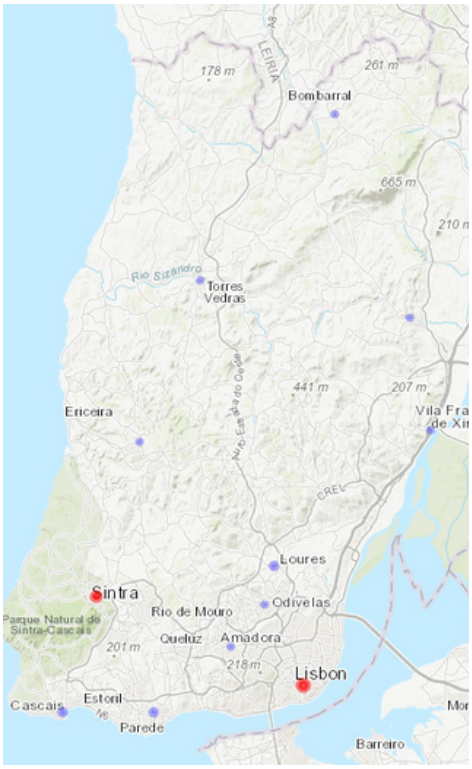
3

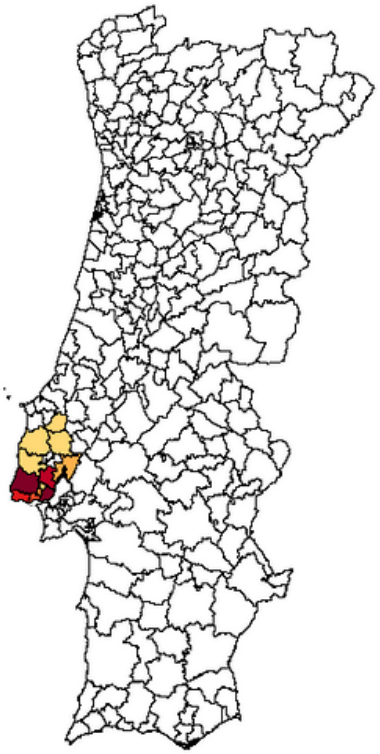


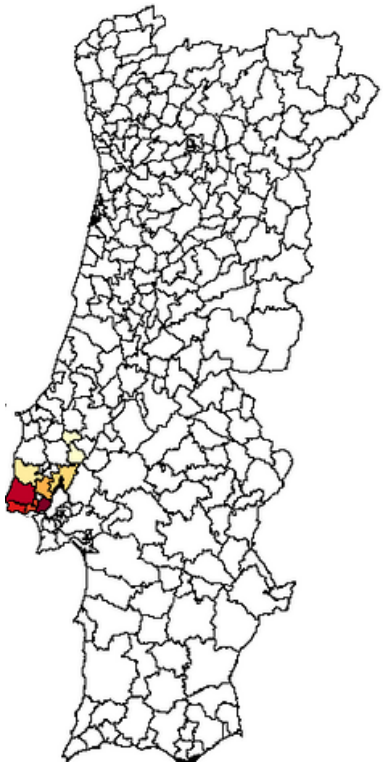
AmiVis Sheet

Lisboa District TTR-FAP GeoVisualisation by Residence

4



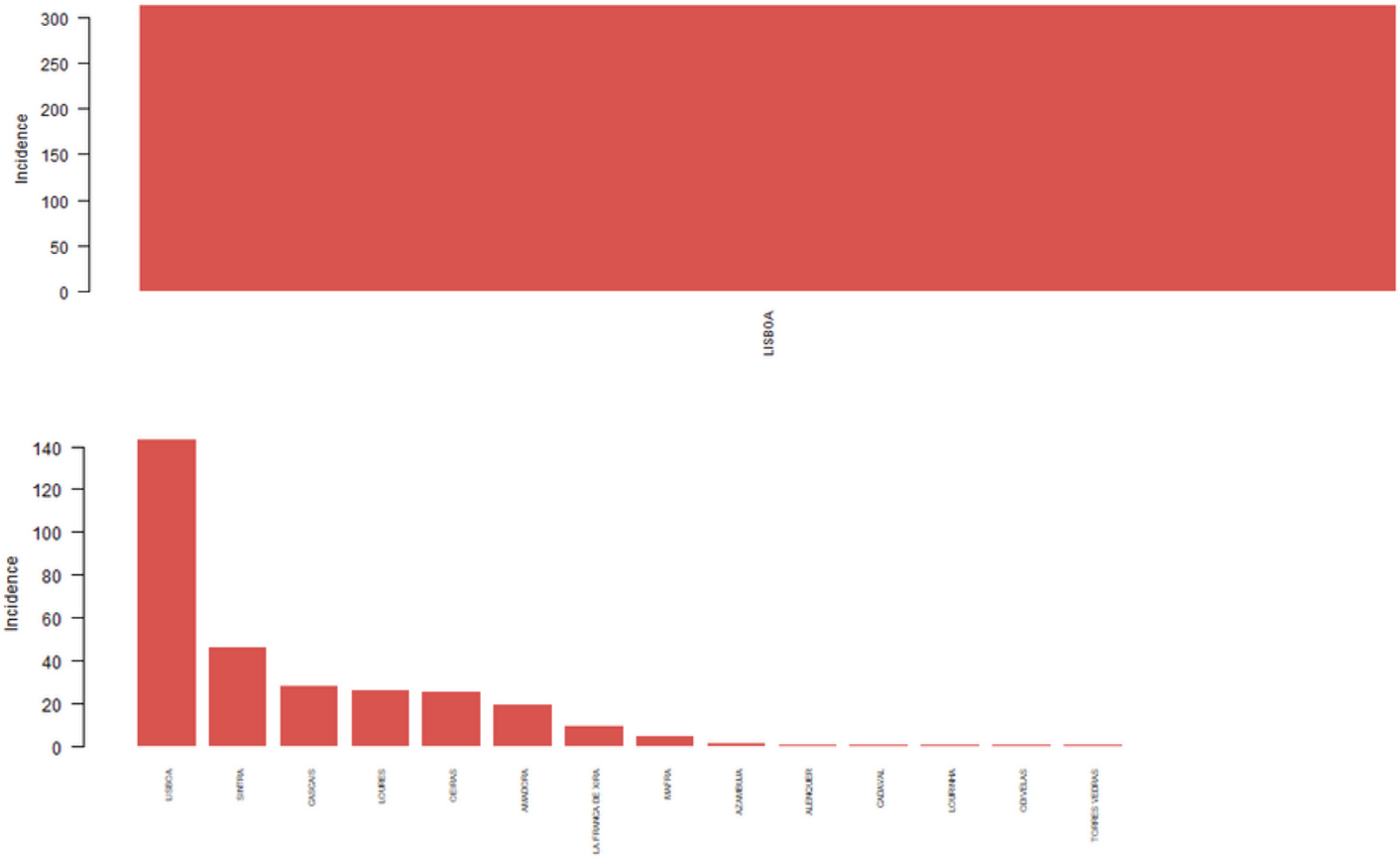




AmiVis Sheet

Lisboa District TTR-FAP GeoVisualisation by Residence

5



Captions:

1 - Static GeoVisualisations of the incidence of TTR-FAP patients with geographical and territorial maps by district and county.

2 - Timeseries and cases per decade.

3 - Overtime Geovisualisation with geographic map for all periods of 25 years after the 1st case in the dataset in 1907.

4 - Comparison of visualisations for men (above) and women (below).

5 - Incidence for each district and county.