TÉCNICO LISBOA

Spatiotemporal patterns of emergency prevalence and response in Portugal

Francisco Faria Barata

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisor: Prof. Rui Miguel Carrasqueiro Henriques

Examination Committee

Chairperson: Prof. José Carlos Martins Delgado
Supervisor: Prof. Rui Miguel Carrasqueiro Henriques
Member of the Committee: Prof. Cláudia Martins Antunes

October 2022

Acknowledgments

Quero agradecer em especial ao meu pai e a minha mae por sempre me terem dado a oportunidade de fazer formaçao no ensino superior. Ao meu irmao que sempre me inspirou e que muito me influenciou na pessoa que sou hoje.

Um obrigado gigante ao Prof. Rui pela paciencia, pelo carinho e pela mentoria, sem duvida uma pessoa que fez diferenca na minha vida pela forma como ve as todas as situacoes de forma positiva!

A todos os amigos e companheiros que fiz na faculdade (Rodrigo, Bonito, Ze Mickel, Marco, Maria, Carlota, Ramos, Esteves, Castro e tantos outros que vao para sempre estar comigo), pessoas que fizeram este percurso comigo e com quem construi grandes memorias!

Em ultimo lugar, um agradecimento tambem ao IST por esta oportunidade de formacao qualificada e distinta!

Abstract

The national-wide record of medical emergencies, monitored by Instituto Nacional de Emergência Médica (INEM), shows a notable increase on the number of emergencies along the last decade. Without ongoing reforms, we can reach the point of saturation in our emergency system and fail to give a response to citizens. Given the fact that it is impossible to have a perfect response for all the emergency cases and vehicles always available to be allocated, the management of the resources is viewed as an optimization problem. In this context, having more knowledge about the domain and the spatiotemporal distribution of emergencies can aid resource allocation and consequently yield a better emergency response. Given the importance of the addressed problem, research in this area can help to improve the success rate of rescue operations, saving more lives. The core purpose of this thesis is to discover if there is an underlying regional structure according to the patterns of medical emergency prevalence and response. With the desire to explore the possibility of a correlation between the location and its medical emergency features, we implemented partitioning and hierarchical time series clustering algorithms, using data aggregated by location. By using clustering methods, we were able to verify if the elements in nearby locations presented similar behaviour for medical emergency data. The clustering results showed that medical emergencies have a significant spatial correlation in terms of the number of emergencies, type of emergencies, and unit dispatch time.

Keywords

Spatiotemporal Data Mining · Medical Emergencies · Time Series · Clustering · Patterns

Resumo

O registo nacional de emergências médicas, monitorizado pelo Instituto Nacional de Emergência Médica (INEM), mostra que, ao longo da última década, o número de ocorrências de emergência continua a aumentar. Sem executar reformas, podemos chegar ao ponto de saturação do nosso sistema de emergência e deixar de produzir uma resposta adequada ao cidadão. Dado que é impossível ter uma resposta perfeita para todos os casos de emergência ou uma disponibilidade contínua dos veículos para alocação, a gestão dos recursos é encarada como um problema de otimização essencial. Neste contexto, ter mais conhecimento sobre o domínio, e como as emergências se comportam ao longo do tempo e do espaço, pode levar a uma melhor preparação e consequentemente uma melhor resposta no tempo. Dada a importância do problema abordado, pesquisas nesta área podem ajudar a melhorar a taxa de sucesso das operações de resgate, salvando mais vidas. O objetivo central desta tese é descobrir se existe uma estrutura regional subjacente de acordo com os padrões de prevalência e resposta a emergências médicas. Com o desejo de explorar a possibilidade de uma correlação entre o local e seus recursos de emergência médica, implementamos algoritmos de particionamento e agrupamento hierárquico sobre séries temporais, usando dados agregados por localização. Por meio de métodos de agrupamento, pudemos verificar se os elementos em localizações próximos apresentavam comportamento semelhante para dados de emergência médica. Os resultados de agrupamento mostraram que as emergências médicas têm correlação com o local, em termos de número de emergências, tipo de emergência e tempo de atendimento da unidade.

Palavras Chave

Prospecção de dados espaço-temporais · Emergências Médicas · Clustering · Séries Temporais · Padrões

Contents

List of Figures

List of Tables

Acronyms

ST	Spatiotemporal
EP	Emerging Pattern
KDD	Knowledge Discovery in Databases
GTS	Georeferenced Time Series
ILD	Inductive Loop Detectors
INEM	Instituto Nacional de Emergência Médica
DTW	Dynamic Time Warping
CODU	Centro de Orientação de Doentes Urgentes

Part I

Foundations

1

Introduction

Contents

One common denominator among humans is the fragility of our lives. Accidents and emergencies occur everywhere and at unpredictable times. To rescue people that are in a need, each country provides a 24 hour call service for emergency notifications and support.

In Portugal's continental territory, emergency medical services are coordinated by Instituto Nacional de Emergência Médica (INEM). In most cases, medical emergencies are reported to INEM through a phone call to the 112 number, where specialized medical staff classifies the emergency and dispatches the proper emergency vehicle (ambulance, helicopter, life-support vehicle, etc.), along with medical staff. Each vehicle is equipped to deal with different situations from light injuries to life-support. The operational productivity of INEM is crucial to Portugal and its actions have a huge societal impact. Hence, INEM needs to carefully plan its operations in order to minimize the overall response time to each emergency. In fact, the response time to a medical emergency is one of the key factors that determine the life or death of a person. There are two particularly relevant time periods in INEM's operations: (1) the amount of time to answer each 112 call and perform a diagnosis, and (2) the amount of time the emergency vehicle spends from dispatch until it reaches the emergency location.

The number of medical emergencies has been increasing every year (Figure 1.1). Given the observed growth of the number of medical emergencies, it is mandatory to perform a good optimization of our resources in order to guarantee efficient responses to emergencies. A first step to answer this task is to understand how the different Portuguese regions are organized with regards to the profile of occurring medical emergencies.

Currently, the allocation of resources and the forecasting of emergencies is done by relying on rules

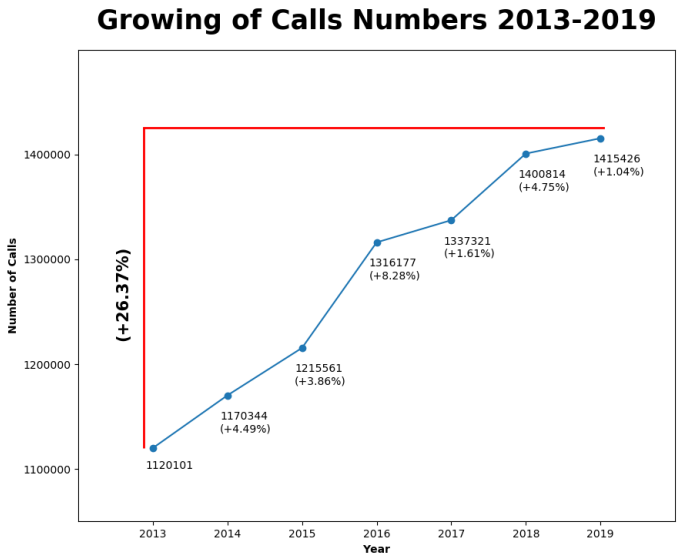


Figure 1.1: Growing of emergency calls registered by INEM in the last years. The percentage below each point is the variation relatively to the year before.

based on the experience of senior staff at INEM. For the proper allocation of resources for each region at all times of the year, there are important factors to be taken into account, including population mobility (e.g. a large number of people travel to the south of Portugal in summer, making the number of emergencies be much higher in those months) and external factors such as weather (some types of pathologies occur more in the hot or cold months of the year). Our objective is to detect if is there any underlying regional structure according to the patterns of medical emergency prevalence and response. For this purpose, we performed clustering methods over the medical emergency data agglomerated by region to investigate if the location of an emergency had a correlation with the medical emergency features. The possibility of discovering groups of regions with similarities in their medical emergencies data gives INEM the opportunity to plan the resources allocation in a more regional way, matching the needs of that region.

1.1 Major Contributions

The main contributions of this thesis:

1. a comprehensive exploratory analysis of medical emergencies in Portugal's continental territory over 2015–2019, with particular focus on: (1) studying emergency prevalence and response are spatiotemporally distributed, and (2) when differentiating the multiple types of pathologies, assessing type-specific spatiotemporal patterns of emergency;

2. testing the hypothesis if is there an underlying regional structure according to the patterns of medical emergency prevalence and response. The gathered results measure the spatial correlation among regions, showing how nearby locations synergistically compare to each other.

1.2 Organization of the Document

This thesis is organized in three parts, each part being divided into chapters.

Part I is divided into three chapters. Chapter 1 provides an introductory note on the thesis. Chapter 2 presents some essential concepts that are used throughout our work. Finally, Chapter 3 reviews related work in the emergency domain and spatiotemporal pattern mining.

Part II is divided into three chapters: Chapter 4 describes the data exploration over the case of study; Chapter 5 presents the clustering solution implemented; and Chapter 6 displays the results obtained from clustering stage.

Part III summarizes the achievements and conclusions of this work.

2

Background

Contents

Information systems are generally able to monitor the occurrence of events. Events can be characterized by one or more features and are often annotated with a specific georeference and timestamp. The presence of space and time originates a variety of Spatiotemporal (ST) data structures and representations, which leads to multiple ways of formulating ST Data Mining (STDM) problems and methods.

This chapter provides essential background regarding the properties and proposed methods used to explore this type of data, introducing: i) fundamental properties of ST data, some of which pose challenges for classical data mining algorithms (Section 2.1), ii) the different ST data structures developed to analyse distinct domains (Section 2.2), and iii) time series analysis and clustering (Section 2.3), and iv) key concepts of pattern discovery (Section 2.4).

2.1 Properties of Spatiotemporal Data

Shekhar et al. [19] explains that extracting interesting and useful patterns from ST datasets presents obstacles that do not exist in traditional numeric and categorical data. The reason why this takes place is the complexity of ST data structures and relationships between ST instances. Atluri et al. [4] described two generic properties of ST data: *auto-correlation* and *heterogeneity*.

The auto-correlation property occurs since the observations made at nearby locations and timestamps are not independent, instead, they are *correlated* with each other. According to Tobler's first law of geography [15] "Everything is related to everything else, but near things are more related than distant things.". In spatial statistics, this form of spatial dependence is called the spatial auto-correlation effect. This auto-correlation property can be found in ST datasets resulting in a coherence of spatial observations (e.g., surface temperature values are consistent at nearby locations) and smoothness in temporal observations (e.g., changes in traffic activity occur smoothly over time). The fact that ST datasets are embedded in continuous space and time presents a challenge for many classical data mining techniques, turning them less effective. Also, standard evaluation schemes such as cross-validation may become invalid in the presence of ST data, because when generating training and test sets randomly it is possible that they are correlated with each other. Hence, the auto-correlation of observations in ST data has to be taken into account in order to build models and algorithms that present useful and clear information about the data.

The heterogeneity property can be observed both in space and time in varying ways and levels. An example of this property is how different spatial regions of the brain perform different functions and hence show varying physiological responses to stimulus. Spatial heterogeneity and temporal non-stationarity have a huge impact on ST datasets. Shekhar et al. [19] proposed that ST data samples do not follow an identical distribution across the entire space and overall time. Instead, different geographical regions and temporal periods may have distinct distributions. This heterogeneity in space and time requires

the learning of different models for varying ST regions since classic data mining formulations make the assumption of homogeneity (or stationary) between observations. That implies that every observation belongs to the same population and is identically distributed. Modifiable Area Unit Problem (MAUP) or multi-scale effect is another challenge since the result of spatial analysis depends on a choice of appropriate spatial and temporal scales.

2.2 Spatiotemporal Data Structures

There are several types of ST data structures in different real-world applications. This divergence occurs in the process of data collection and representation, changing the way space and time are stored to be used afterwards. For this reason, in each ST problem, it is mandatory to understand the ST data structures available to make the most effective use of STDM methods.

The most common categories of ST data, according to Atluri et al. [4], are the following four: (i) event data, which consists in discrete events occurring at specific locations and times (e.g., road accidents or incidences of crime); (ii) trajectory data, where trajectories of moving bodies are being measured (e.g., taxi route); (iii) point reference data, where a continuous ST field is being measured at moving ST reference sites (e.g., measurements of surface temperature collected using weather balloons); (iv) raster data, where observations of an ST field are spatial and temporally grouped at fixed cells in a grid (e.g., mapping several attributes in landscapes). Having these four ST data structures, the first two (events and trajectories) considerably differ from the last two (point reference and raster data). Events and trajectories record observations associated with discrete events and objects. Point reference and rasters capture information of continuous or discrete ST fields. A mapping between ST data structures and ST data instances is presented in Figure 2.1.

Given the available type of data, collections of Events were the structure used in this project.

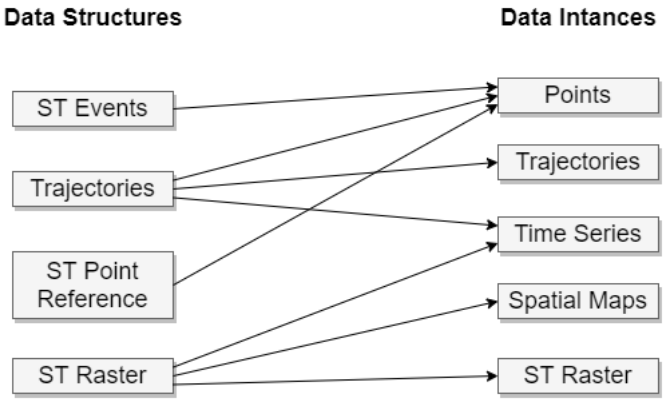


Figure 2.1: Mapping showing the different categories of ST data instances that can be build from ST data structures.

2.2.1 Event Data Structure

According to Atluri et al. [4], ST events are a common structure in real-world applications such as criminology (incidence of crime and related events), epidemiology (disease outbreak events), transportation (road accidents), Earth science (land cover change events like forest fires and insect disease) and social media (Twitter activity or Google search requests).

An ST event can generally be represented by a *point* instance, as shown in Figure 2.1, with an exact location and time, which describe where and when the event occurred, respectively. Besides those key attributes, location and time, a ST event may also contain non-ST attributes, known as *marked attributes*, that provide additional information about the event.

In this context, an ST event can be represented, as illustrated by Neves et al. [16], as a tuple (**x**, *s*, *t*), where:

- **x** $= (x_1, \cdots, x_m)$ is the observation, either *univariate* (m=1) or *multivariate* (m>1) depending on the number of monitored marked attributes.
- *s* is the *spatial extent* of the observation **x**. The spatial extent *s* can be any spatial representation associated with the event, such as a geographic coordinate or a trajectory.
- *t* is the *temporal extent* of the observation **x**, either given by a time instant or a time interval.

In this project, emergency calls are seen as ST events. Given the amount of timestamps recorded in each emergency call, we extended the representation of an event to a tuple (**x**, *s*, **t**). Our dataset is composed by a collection of events, E = $(e_1, \cdots, e_n)$, being formally classified as a *ST event dataset*. Each event $e_i = (\mathbf{x}_i,\ s_i,\ \mathbf{t}_i)$ is an emergency call that occurred in the location s_i (string with district and county), with a set of timestamps $\mathbf{t}_i$ recorded (creation of the occurrence, response unit dispatched, arrive of the dispatched unit, close of occurrence) and a set of observations $\mathbf{x}_i$ that contains emergency information (type of unit assigned, priority, emergency nature, etc.).

2.3 Time Series Data Analysis

Time series have a relevant role in diverse fields including medicine, aerospace, finance, business, meteorology and human activity. Time series data are sequences of measurements over time describing the behaviour of systems [2].

A time series is an ordered set of observations $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$, each observation $\mathbf{x}_t$ being recorded at a specific time point *t*, with $1 \leq t \leq T$. Time series are referred as *univariate* when only one attribute is recorded, $\mathbf{x}_t \in \mathbb{R}$, or *multivariate*, $\mathbf{x}_t \in \mathbb{R}^m$, where $m > 1$ is the multivariate order, i.e., number of attributes recorded. In our project, we used both univariate and multivariate time series.

Time series can be decomposed into *trend*, *seasonal*, *cyclical*, and *irregular components* using additive or multiplicative models, as explained by Brockwell et al. [6]. Classical approaches for time series

analysis and forecasting generally rely on statistical principles, including *auto-regression*, *differencing* and *exponential smoothing operations* [20]. An example of a time series is presented in Figure 2.2, where is visible the upward trend in the number of emergency calls between 2013-2019.

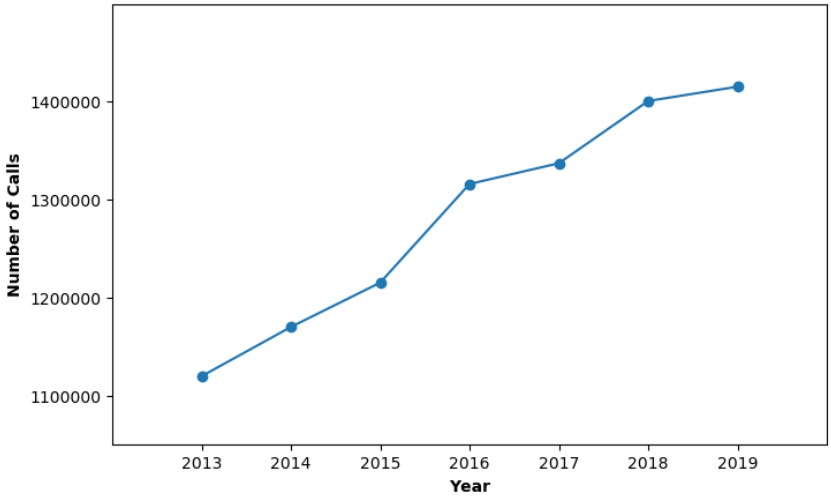


Figure 2.2: Evolution of the emergency occurrences number in Portugal (continental territory) recorded by INEM between 2013-2019.

Zhao and Bhowmick [23] suggested four kinds of patterns that can be obtained from time series data: *trend analysis*, *similarity search*, *sequential patterns* and *periodical patterns*.

Trend analysis is the discovery of evolution patterns over time. They can be long-term trend movements, cyclic movements or variations, seasonal movements and irregular/random movements.

Similarity search aims to find slightly different sequences. The similarity among two time series, $\mathbf{x}_1$ and $\mathbf{x}_2$, can be computed using proximity measures. One possibility is the use of Minkowski distance that consider one-to-one correspondence between the elements of the two arrays, as shown in equation 2.1.

However, sometimes it is the case that two similar time series are not exactly aligned with one another but show the same pattern of activity over time. Measures such as Dynamic Time Warping (DTW) and Fréchet distance are able to capture such forms of similarity among time series, as demonstrated by Atluri et al. [4]. Based on the length of the time series that we are trying to match, can be classified as: *subsequence matching* and *whole sequence matching*.

Sequential patterns, as already explained in section 2.4.1, are relationships between occurrences of sequential events, to discover if there exists any specific order for the occurrences.

Periodical patterns are recurring patterns that occur periodically in the time series. Periodicity can be daily, weekly, monthly, seasonal and yearly. Periodical patterns can be viewed as sequential pattern mining by taking the periodical subsets of the time series as a set of sequences.

$$D_M(\mathbf{x}_1, \mathbf{x}_2) = \sqrt[q]{\sum_{i=1}^{m}(\mathbf{x}_{1,i} - \mathbf{x}_{2,i})^q}, \tag{2.1}$$

where *m* is the multivariate order and *q* is a positive integer. The most used values are $q = 1$, $q = 2$ and $q = \infty$, corresponding to Manhattan, Euclidean and Chebyshev distance, respectively.

2.3.1 Georeferenced Time Series

Time series describing evolving behaviour of one or more features recorded at fixed locations and uniform intervals are referred as *georeferenced* [22]. Neves et al. [16] formulates that Georeferenced Time Series (GTS) can be represented as a tuple $(\phi, \mathbf{x})$, where ϕ is a pair (*latitude*, *longitude*) describing the location in which the series **x** is being recorded.

In the context of this project, medical emergencies from the same region can be aggregated to provide views on the cumulative numbers or average values of interesting features (number of occurrences, the priority of the emergency, dispatch time of the vehicle, etc.). Figure 2.3 shows the registered number of occurrences for the Lisbon district over the last years, where is clearly observable an upward trend. Due to the multiplicity of features being recorded, the target GTS generally have high multivariate order.

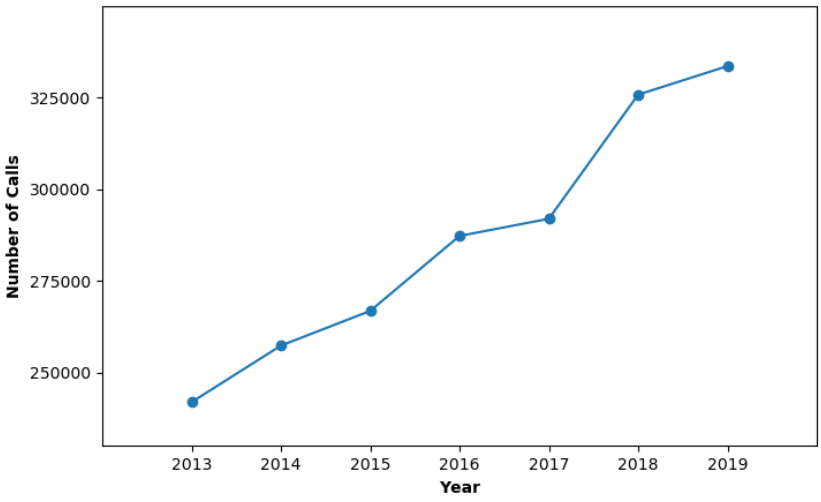


Figure 2.3: Evolution of the emergency calls in Lisbon district between 2013-2019.

2.3.2 Clustering Time Series

In the presence of a significant number of time series, clustering is a technique that can be applied in order to form homogeneous groups with similar behaviours. Given a set with *N* observations, $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, a cluster $\mathcal{C}$ is a subset of the original space, where $\mathcal{C} \subseteq \mathcal{X}$. Clustering task aims to find a set of clusters that satisfy specific intracluster and intercluster criteria of similarity, i.e., similarity of elements inside the same cluster is maximized while minimizing similarity between elements from different clusters.

Clustering algorithms work fundamentally in three steps, where two repeat in a cycle until the algorithm converges, as represented in Figure 2.4.

1. **Representation:** Consists in the definition of parameters such as the number of classes and scale of features. This step usually includes feature selection (i.e. the identification of ”key” features that have impact on the observation class) and/or feature extraction (i.e. at least one transformation is made on the input features).

2. **Similarity Computation:** This step uses a function to measure the similarity between observations. . This step repeats after Step 3, recalculating the similarity between the new groups formed.

3. **Grouping:** This step groups the observations using the similarities computed in the previous step. The two main grouping processes are hierarchical and partitioning clustering.

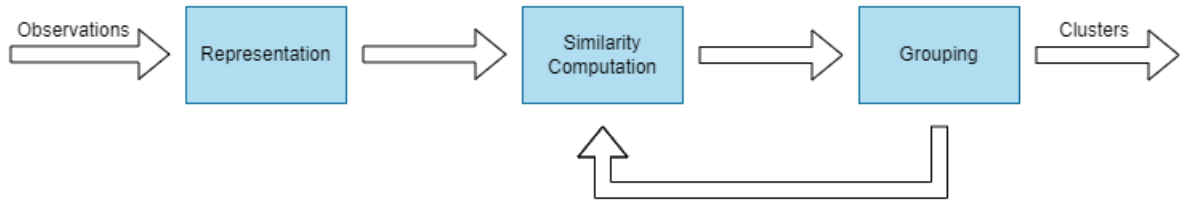


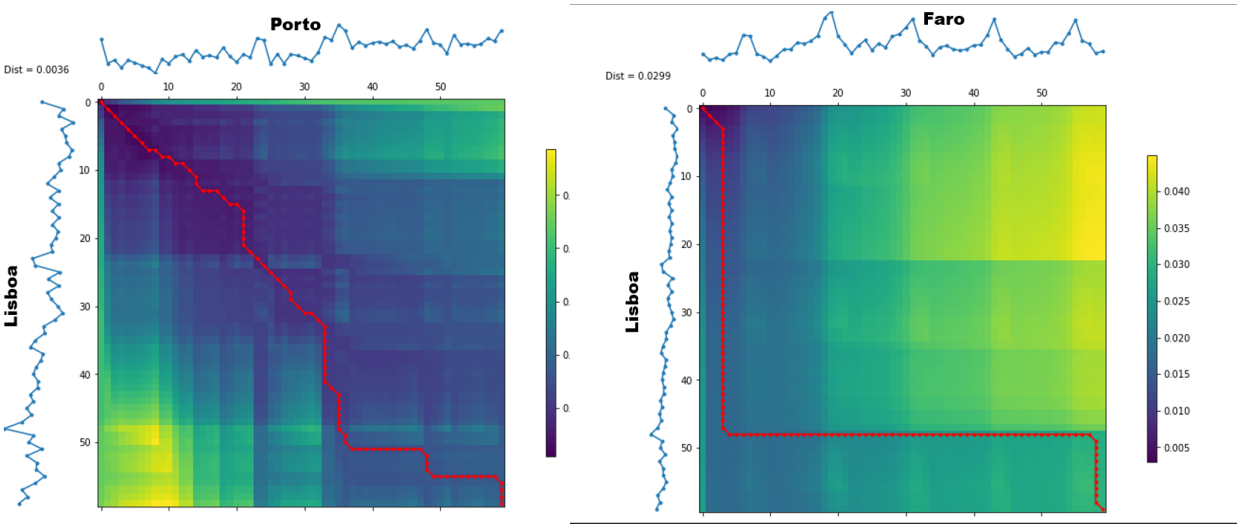
Figure 2.4: Overview of the clustering process

As shown by Giorgino [9], the adaptability and efficiency of DTW lead to be the commonly most used method when computing the similarity between time series (Step 2). DTW is an algorithm that builds a distance matrix by comparing every pair of elements of two time series. To calculate the distance between points is often used the Euclidean distance, obtained by replacing *q* with 2 in equation 2.1.

Two examples from our domain are given in Image 2.5, where 2.5a shows the distance matrix (represented as a heatmap), computed when comparing the emergency calls from Lisboa and Porto. In 2.5b is represented the comparison between Lisboa and Faro. The red line in both images represents the best wrapping path for that pair of time series, i.e. the path with the lower total distance.

A conclusion that can be observed here is that Lisboa and Porto have similar behaviours (distance values in matrix are low and the wrapping path moves almost in a diagonal) while Lisboa and Faro do

not present that relation (distance values in matrix are higher and the wrapping path moves in straight lines, showing difficulty in match values across the time series).



(a) DTW matrix comparing Lisboa and Porto emergency calls numbers.

(b) DTW matrix comparing Lisboa and Faro emergency calls numbers.

Figure 2.5: Examples of DTW matrices computed comparing time series from Lisboa-Porto and Lisboa-Faro. The red line shows the best alignment found (best wrapping path).

For our project a central objective when clustering time series from ST data was to find spatially coherent groups of locations with similar temporal activity [4] (e.g. districts that are connected with the same patterns).

2.4 Pattern Discovery

In each domain, new interesting patterns (non-trivial, implicit, previously unknown, and potentially useful) can be identified. According to Chen et al. [7], this process has the name of *pattern discovery*. The purpose of pattern discovery is to find the relations among attributes and/or among their values. Knowledge Discovery in Databases (KDD) is an iterative process to explore patterns for either descriptive or predictive ends. During the search, evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get interesting and more valuable results, as shown in Figure 2.6. Patterns can be used to augment the knowledge acquired from clustering solutions to answer real-world problems. In the context of this work, their discovery is suggested as a relevant direction to complete the proposed methodological principles, grounded primarily on clustering principles.

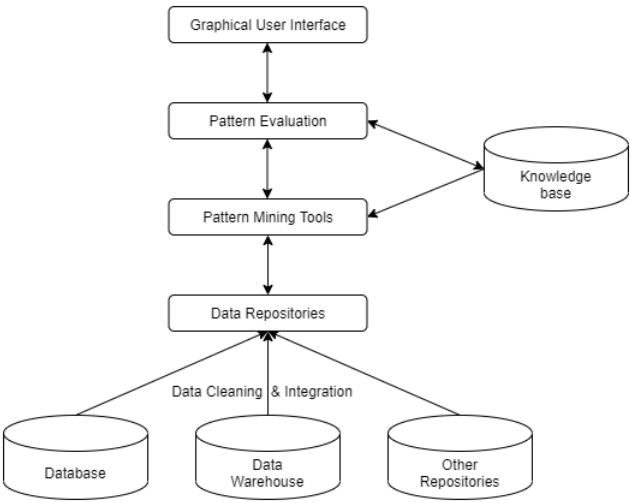


Figure 2.6: Knowledge Discovery in Database process.

Wong and Wang [21] formulated that a dataset $\mathcal{D}$ has a set of N observations, such as $\mathcal{X} = (X_1, \cdots, X_N)$, with M attributes. Let $\mathcal{Y} = \{Y_1, \cdots, Y_M\}$ represent the attribute set. Then each attribute Y_j, $1 \leq j \leq M$, can be seen as a random variable taking values from its domain $\mathbb{Y}_j$. For a continuous attribute, $\mathbb{Y}_j = \mathbb{R}$ and for a categorical attribute, $\mathbb{Y}_j = \{\alpha_j^1, \cdots, \alpha_j^m\}$, where *m* is the cardinality of the alphabet.

Having $\mathcal{A}$ and $\mathcal{B}$ as a subset of $\mathcal{Y}$, where $\mathcal{A} = \{A_1, \cdots, A_a\}$, $\mathcal{B} = \{B_1, \cdots, B_b\}$, $a, b < M$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$, different measures can be calculated in order to evaluated the *association rule* $\mathcal{A} \Rightarrow \mathcal{B}$. Each element of $\mathcal{A}$, A_j, is a subset of the existing values in $\mathbb{Y}_j$, $A_j \subseteq \mathbb{Y}_j$. This implication represents that $\mathcal{A}$ occurred and triggered $\mathcal{B}$.

The *support* measure for the rule $\mathcal{A} \Rightarrow \mathcal{B}$, shown by Han et al. [11], is the percentage of observations in $\mathcal{D}$ that contain both $\mathcal{A}$ and $\mathcal{B}$,

$$support(\mathcal{A} \Rightarrow \mathcal{B}) = P(\mathcal{A} \cup \mathcal{B}) = \frac{\#Observations\ containing\ both\ \mathcal{A}\ and\ \mathcal{B}}{\mid \mathcal{D} \mid}. \tag{2.2}$$

Goethals [10] defines that a rule is considered *frequent* if its support is no less than a given *minimal support threshold* θ, with $0 \leq \theta \leq 1$.

The *confidence* measure for the rule $\mathcal{A} \Rightarrow \mathcal{B}$, as demonstrated by Han et al. [11], is the percentage of observations in $\mathcal{D}$ containing $\mathcal{A}$ that also contain $\mathcal{B}$. In other terms, confidence corresponds to the conditional probability $P(\mathcal{B} \mid \mathcal{A})$,

$$confidence(\mathcal{A} \Rightarrow \mathcal{B}) = P(\mathcal{B} \mid \mathcal{A}) = \frac{support(\mathcal{A} \cup \mathcal{B})}{support(\mathcal{A})} = \frac{\#Observations\ containing\ both\ \mathcal{A}\ and\ \mathcal{B}}{\#Observations\ containing\ \mathcal{A}}. \tag{2.3}$$

Similarly to the support, it is possible to choose a value for the *minimal confidence threshold* γ, with $0 \leq \gamma \leq 1$.

Rules that satisfy both minimal support threshold and minimal confidence threshold are called *strong* [11]. It is also important to acknowledge that not all strong rules are interesting. Therefore, the support-confidence framework should be augmented with a pattern evaluation measure, which promotes the mining of *interesting* rules. Examples of those types of measures are: *lift*, $\mathcal{X}^2$, *all confidence*, *max confidence*, *Kulczynski* and *cosine*. All these measures use support and confidence values in their calculations. For example, lift value of rule $\mathcal{A} \Rightarrow \mathcal{B}$,

$$lift(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{P(\mathcal{A} \cup \mathcal{B})}{P(\mathcal{A})P(\mathcal{B})} = \frac{confidence(\mathcal{A} \Rightarrow \mathcal{B})}{support(\mathcal{B})}. \tag{2.4}$$

Given a ST dataset, as described by Neves et al. [16], a pattern is a spatially correlated set of frequent, periodic and coherently changing observations over time. For each pattern, several criteria of interest can be measured: (i) *pattern support*, the number of observations satisfying the pattern; (ii) *pattern length*, the multivariate order and ST extension of the given pattern; (iii) *pattern strength*, the association strength among the elements composing a pattern.

2.4.1 Frequent Pattern Discovery

According to Atluri et al. [4], a pattern occurs frequently over multiple observations in a dataset (e.g., frequently bought groups of items in a store). Given the variety of data structures and instances in ST applications, there are several categories of frequent pattern mining problems that can be formulated in the presence of ST components. As described in section 2.2.1, in the context of this project each observation corresponds to an event that is represented by a point instance. Given this type of instance, Atluri et al. [4] suggested the exploration of two frequent patterns: *co-occurrence patterns* and *sequential patterns*.

A co-occurrence pattern, as explained by Aydin et al. [5], is a subset of attributes whose observations frequently co-occur in close spatial and temporal proximity to each other. In the presence of ST events of varying types, a co-occurrence pattern can also be a subset of ST events, as described by Atluri et al. [4]. For example, given a dataset of crime and other events in a city, it is interesting to find ST events that occur together (e.g., bar closing and drunk driving).

A sequential pattern is an order of events that occurs frequently [23]. In a dataset, storing observations with the order in which they occurred or having, in each observation, a timestamp that identifies when it took place makes possible the search for this type of patterns. For example, a customer in a store that buys a computer, later buys a mouse and then a printer ($Buy\_computer \rightarrow Buy\_mouse \rightarrow Buy\_printer$). A sequential pattern can be found when the occurrence of an ST event can trigger other ST events, explained by Atluri et al. [4]. The goal is to discover ordered lists of events types such as $e_1 \rightarrow e_2 \rightarrow \cdots \rightarrow e_n$, where events belonging to type e_1 trigger events of type e_2, that further triggers

events of type e_3. This process continues until the final event in the chain (e_n) occurs. When identified, these sequences offer users the opportunity to make more accurate and useful predictions, supporting real-world decisions.

2.4.2 Emerging Pattern Discovery

The concept of Emerging Pattern (EP) was introduced by Dong and Li [8] in the context of multivariate observations collected from two periods/datasets. An EP was defined as a multivariate pattern whose support suffered a significant change between two given timestamps. Neves et al. [16] extends this notion of EP to encompass an arbitrary number of time periods and to further incorporate spatial information. In the presence of a ST dataset, an EP is a set of spatially correlated observations where specific *growth*, *fitness* and *support* criteria are satisfied.

The *growth* criterion defines the rate at which observations change over time. For example, given a location and periodicity, a growth rate of 1% indicates that the values of a given observation increase 1% in the observed period.

Having a specific growth rate, the *fitness* (error) criterion is defined by comparing the values of the observations with the given expectations. For instance, fluctuations of the observed values around the expected values produce residues that can be used to characterize the fitness of a given EP. EPs that present an accuracy value below a given threshold should be discarded based on these criteria.

The *support* criterion defines the number of observations satisfying the given growth and accuracy criteria.

EP discovery can be applied with specific growth, fitness and support thresholds. As described by Neves et al. [16], in order to guarantee pattern quality, other properties should also be satisfied by EPs, in addition to the three already presented: (i) *non-triviality* (novelty) and *actionability* (support decisions); (ii) *robustness*, bounded noise tolerance; (iii) *statistical significance*, excluded spurious patterns that occur by chance; (iv) *interpretability*; (v) *coverage*, complete solutions spanning different geographies and time periods; (vi) *efficiency*, with respect to the pattern retrieval process.

Complementarily to clustering solutions, the discovery of frequent and emerging patterns from spatiotemporal data are generally a relevant step for knowledge acquisition. In the context of this work, their discovery is suggested as a relevant direction to complete the proposed methodological principles, grounded primarily on clustering principles.

3

Related work

Contents

This chapter compiles related work that provides an import basis for the methodologies and concepts used for this study. This work aims at exploring, modeling, and clustering the ST distribution of medical emergency calls over the last years (2015-2019). In this context, relevant previous studies are divided into two parts. Section 3.1 covers works in the same domain of our project, i.e., medical emergency. Section 3.2 presents advances in the research area of mining patterns from ST data.

3.1 Emergency domain

Lamine et al. [13] studied the distribution of the incoming calls of Emergency Medical Assistance Centre (EMAC), a 24h assistance service for emergency response, and proposed possible ways of improving the management of this service. An approach based on the combination of process mining and discrete event simulation techniques is suggested. Process mining allowed to discover the control flow of the incoming call processing and Key Performance Indicators (KPI) such as the answer speed and call duration. Discrete event simulation was used to reproduce the dynamic behaviour of the EMAC in order to subject them to predictive experimentation, without risk or disturbance of the real-life system.

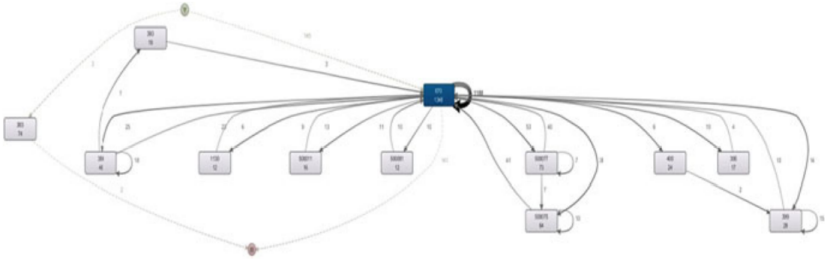
The study case was conducted with data gathered at the EMAC located in Albi, France. Based on the sequence and timing of the activities in the event log dataset, process mining was performed using Disco software. The development of the simulation model is carried out with the software package Witness. Witness is a discrete event simulation software where it is possible to describe the several existing states and the conditions that determine the change of state. In this study, the simulation was used to create an ”As-Is” model, where each call is processed until the end of the communication, even if another call is in the queue, and compare it with a ”To-Be” model, where possible improvements (optimizing business rules, effective staff deployment, etc.) can be performed.

The pilot testing was carried out on the most critical day for the EMAC. With the ”As-Is” model, all the cases are treated according to the first-come first-served discipline. The ”To-Be” model was tested with a new rule: rather than process a call until the end, Lamine et al. [13] proposed to put it on hold in order to quickly evaluate the severity of the new call. If the new call has more priority than the current one, it is processed first and then the call putted on hold continues to be handled. The results of this simulation study showed that applying this new rule could be the first step to improve the performance of this call centre by taking into account the characteristics of an incoming call in the queue.

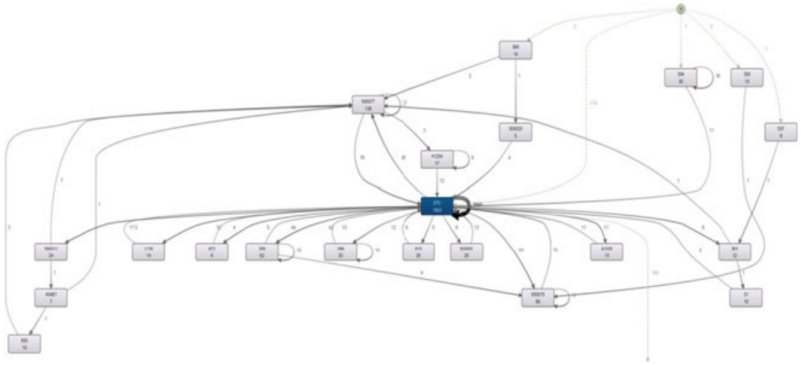
Antonelli and Bruno [3] adapted tools, from other research fields, to obtain a model of healthcare network (hospitals, medical centres, drugstores, etc.) and then perform a meaningful analysis from the data produced. In this project, a data warehouse that merged data from several types of healthcare organizations, collecting the patient movements inside the healthcare network, is proposed. With this data warehouse, the medical history of each patient becomes available, allowing both physicians to express meaningful analyses at patient level (e.g., searching for similar patients based on their medical history or predicting future events in care pathway) and the systems manager to operate at organization level (e.g., discovering which are the most accessed resources or which are the anomalous managements of patients). Antonelli and Bruno [3] use process mining applied to healthcare data to identify the processes and derive meaningful insights from complex temporal relationships existing between activities and resources involved in the process.

Data collected by an Italian Healthcare Territorial Agency (HTA) between 2007 and 2012 was the case study. Lamine et al. [13] performed two alternative segmentations of the dataset, one based on gender and the other based on age. In each segmentation, process mining techniques were applied using Disco software to reconstruct the actual movements of patients among the providers of a HTA. Disco exploits the Fuzzy Miner algorithm for process mining, which uses significance/correlation metrics to interactively simplify the process model at the desired level of abstraction. This tool automatically discovers process maps by interpreting the sequences of activities.

One of the results observed was found using gender segmentation and a specific pathology, patients suffering from asthma. The mobility of male patients (Figure 3.1a) and female patients (Figure 3.1b) leads to analysis where is possible to discover similarities and differences. Both graphs present the same most accessed centre (dark blue) since the majority of examinations are performed in this centre. Another similarity is that processes usually do not involve more than two different centres. The major difference is the number of visited centres: male patients visit 12 different centres, while female patients visit 21 different centres, showing higher mobility of female patients. The preliminary results obtained proved the applicability and the usefulness of the proposed approach. It becomes easier for healthcare managers to clearly be informed about the status of services under their responsibility, and to suggest improvements to system inefficiencies.



(a) Process flow for male asthma patients.



(b) Process flow for female asthma patients.

Figure 3.1: Process flow extracted by Disco on a) male asthma patients and b) female asthma patients [13].

Li et al. [14] discuss the problem of finding risk patterns in medical data. Risk patterns are defined, in this paper, by a statistical metric, relative risk, which has been widely used in epidemiological research. For example, if pattern $\mathcal{S}$ is smoking and c is having lung cancer, it is possible to ascertain the impact of such behaviour,

$$RR(S) = \frac{P(c \mid S)}{P(c \mid \neg S)} = \frac{\frac{support(S \cup c)}{support(S)}}{\frac{support(\neg S \cup c)}{support(\neg S)}}. \tag{3.1}$$

A $RR = 3.0$ means that people who smoke are three times more likely to get lung cancer than those who do not. The problem of mining risk patterns was characterised as an optimal rule discovery. They developed an efficient method to exhaustively find all high-risk patterns in high-level interactions and present understandable results to medical practitioners. They used the anti-monotone property to efficiently prune searching space, distinguishing it from other association rule mining algorithms that use post-prune techniques.

The implemented method has been applied to a real-world project of detecting adverse drug reactions. The dataset used connects data from hospitals, pharmaceuticals and medical services. In particular, this study focus was to determine how the Angiotensin-converting enzyme (ACE) inhibitors usage is associated with Angioedema (swelling beneath the skin) and identify types of patients at risk.

The risk patterns returned were verified by domain experts that considered them of great interest. Statistical evaluations concluded that the proposed method is able to find statistically significant patterns from large and skewed datasets.

Jasso et al. [12] implemented an algorithm to automatically detect medium to large-scale emergency events, such as earthquakes or fires, based on an analysis of the ST characteristics of 911 call activity. Obtaining an early indicator of the presence, location, and spatial extent of the medium to large-scale event can lead to an informed response and consequently achieve better results in rescuing citizens. The algorithm created aims at detecting hotspots of emergency calls linked with a catastrophe while avoiding calls that occur close in time and space but are not related to each other. Hotspots were defined as instances where a large number of calls happen within a short distance and a short time interval. Any two calls that happen within the minimum inter-call distance as well as within the minimum inter-call time from each other are clustered together. In the end, after all clusters are calculated, only those that present a size bigger than the minimum number of calls are selected as hotspots.

Data used in this study was collected for the San Francisco Bay Area, from 01/09/2004 to 30/06/2007, and for San Diego County, from 01/11/2005 to 30/06/2007. The developed algorithm was applied to the collected data. They explored the effect of different parameter settings (minimum inter-call distance, minimum inter-call time and minimum number of calls per cluster) in order to detected medium to large-scale emergency events reported in the news. It was possible to observe that only people within visual or audio range of an emergency event tend to call to report it, thus the location of the detected clusters correspond closely to the location of the emergency event.

An example of a detected event, a gasoline pipeline explosion in California, is presented in Figure 3.2. With this plot, the areas where the event had an impact on citizens are clearly observable. A reasonable set of parameters was suggested suitable for effective detection of the location and extent of major emergency events.

Using data mining techniques, Selvam and Thivakaran [18] developed a model to identify patterns based on an analysis of the characteristics of 911 call activity. This study's goal was to help allocate emergency responders and help strategy makers implement proactive strategies to provide quick responses. Two steps were performed in this study. First, it was conducted a hotspot analysis to determine areas of high/low call volume. Second, they performed a spatial analysis to explore factors that contribute to the high/low call volumes.

This study analyses 911 emergency call data compiled from Montgomery County in Pennsylvania, between 12/10/2015 and 10/08/2016. The dataset was loaded and preprocessed. After being preprocessed, data was clustered, where K-Means clustering was the technique selected. In the end, they

Figure 3.2: Pipeline explosion in California. Star indicates the location of the explosion. A Red circle indicates that the call is part of a detected cluster [18].

performed a hotspot analysis using Arc GIS software, that implements Getis-Ord Gi* algorithm. The outcome of this tool is a Z score for each feature. For high Z scores more acute the clustering of high values (hot spot) and respectively, for low Z scores more acute the clustering of low values (cold spot).

The gathered results helped to identify areas within Montgomery County where the concentration of 911 calls is high and low, as presented in Figure 3.3. Due to the fact that collected data only corresponds to a 10 month period, it is possible that the gathered results are not representative. Analysis like the one realized in this work can improve our strategy making and lead to an efficient planning, where resources are allocated on a need basis. When allocated in an optimized way, response times will decrease and more lives will be possible to save.

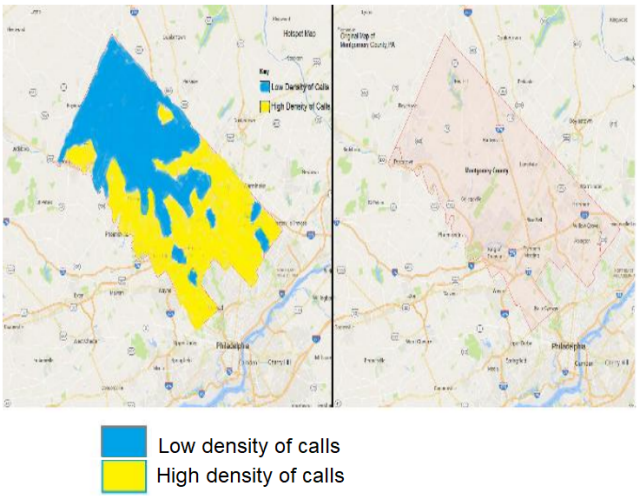


Figure 3.3: Map of Montgomery County and its respective 911 hotspot detected [18].

3.2 Advances on spatiotemporal pattern mining

Neves et al. [16] proposed a scalable method to comprehensively detect EPs from heterogeneous sources of ST data generated by large sensor networks. Emerging Event PATtern miner (E2PAT) is a linear-time method formed by two steps. First, transformation procedures are applied to consolidate the original ST data sources and map them into new data structures. Second, EPs are discovered from the transformed data by combining three principles: (i) spatial intersection and time windowing operations, for the comprehensive traversal of search space; (ii) combined use of time series differencing operations with linear regressors; (iii) integrative scoring to measure the relevance of EPs and control the amount of false positive and false negative discoveries.

The study case considered was Lisbon's road traffic monitoring system, a large scale network of mobile and fixed sensors. Detecting new EPs at an early stage offers urban planners the opportunity to make the necessary provisions to urban mobility. The data used by Neves et al. [16] was collected from two different sources: GTS data produced by Inductive Loop Detectors (ILD), which are fixed sensors placed in city junctions to measure the number, speed and type of vehicle passages over time, and timestamped trajectory data produced by vehicles with mobile sensors (GPS), provided by WAZE .

A visualization tool, shown in Figure 3.4, was developed to support the analysis and guide the navigation throughout the outputted pattern solutions. The tool provides an interface for querying the desirable area, sensors, and time granularity. Results of the discovered congestion patterns (Figure 3.5a) and decongestion patterns (Figure 3.5b) are presented.

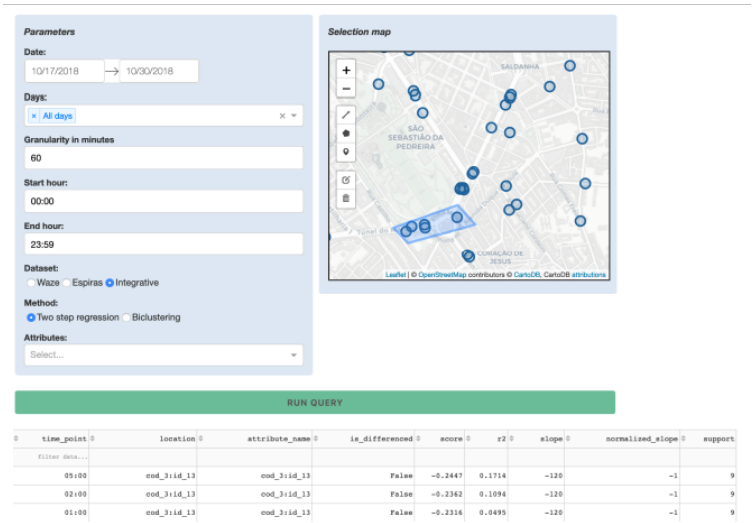


Figure 3.4: Overview of the user dashboard for querying the road data sources [16].

(a) Congestion patterns.

(b) Decongestion patterns.

Figure 3.5: Map visualization of the found patterns from both ILD-WAZE data sources using score-based coloring of point-based and trajectory-based EPs [16].

In an alternative study, [17] addressed the problem of mining actionable patterns of road mobility from heterogeneous sources of traffic data. A two-step methodology was proposed: (i) mapping the original data sources into new data structures using transformation procedures and (ii) perform pattern-based biclustering over the new data structures to discover traffic patterns. Biclustering aims at finding subsets of observations with values correlated on a subset of variables. Given a dataset defined by a set of observations $\mathcal{X} = \{x_1, \cdots, x_n\}$ and variables $\mathcal{Y} = \{y_1, \cdots, y_m\}$, an element $a_{ij} \in \mathbb{R}$ corresponds to the value of attribute y_j for the observation x_i. A bicluster $B = (\mathcal{I}, \mathcal{J})$ is a n x m subspace, where $\mathcal{I} = (i_1, \cdots, i_n) \subseteq \mathcal{X}$ is a subset of observations and $\mathcal{J} = (j_1, \cdots, j_m) \subseteq \mathcal{Y}$ is a subset of variables. Specific criteria like *homogeneity*, *statistical significance* and *dissimilarity* must be satisfied by the identified biclusters.

Homogeneity criteria are commonly guaranteed through the use of a merit function, such as the variance of the values in a bicluster. The homogeneity determines the coherence, quality and structure of a biclustering solution. The coherence of a bicluster is determined by the observed form of correlation among its elements (coherence assumption) and by the allowed value deviations from perfect correlation (coherence strength). The quality of a bicluster is defined by the type and amount of accommodated noise. The structure of biclustering solution is defined by the number, size, shape and positioning of biclusters. An element e_{ij} within a bicluster have coherence across variables if $e_{ij} = c_j + \gamma_i + \eta$, where c_j is the expected value of variable y_j, γ_i is the adjustment for the observation x_i, and η is the noise factor. A bicluster has a *constant coherence* when $\gamma_i = 0$, and *additive coherence* otherwise, $\gamma_i \neq 0$. A bicluster $B = (\mathcal{I}, \mathcal{J})$ satisfies the *order-preserving coherence* assumption if the values for each observation in $\mathcal{I}$ follow the same ordering along the same subset of variables in $\mathcal{J}$. Figure 3.6 illustrates biclusters with

constant, additive and order-preserving coherence (right) found in real-valued data (left).

Statistical significance criteria, in addition to homogeneity, guarantee that the probability of a bicluster's occurrence deviates from the expectations.

Finally, dissimilarity criteria can be further placed to guarantee the comprehensive discovery of non-redundant biclusters.

Lisbon's city was the study case considered and data was acquired from two heterogeneous sources: GTS data from ILDs and multivariate event collections from GPS sensors, provided by WAZE. BicPAMS (Biclustering based on PAttern Mining Software) was the selected biclustering approach as it combines state-of-the-art principles on pattern-based biclustering.

Gathered results confirm the relevance of biclustering to unravel non-trivial, meaningful, actionable and statistically significant patterns able to combine heterogeneous road traffic aspects. A considerably high number of dissimilar and statistically significant patterns were discovered.

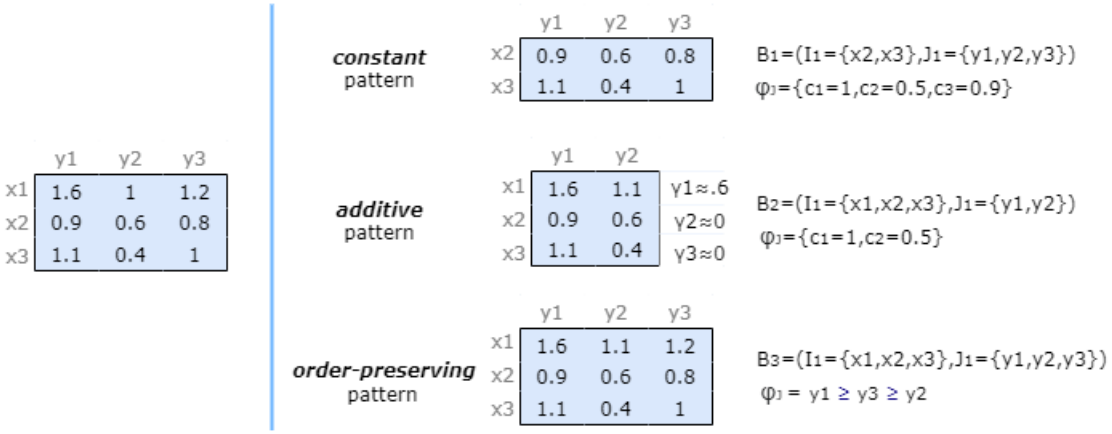


Figure 3.6: Biclustering with varying homogeneity criteria.

Due to the alarming increase in the rate of criminal activities in Nigeria, Ahmed and Salihu [1] performed a study using Geographic Information Systems (GIS) and a ST dataset of crime activity. Their focus was to find characteristics that help detecting, analysing and mapping of crime hotspots, along with other trends and patterns. Given a large number of existing labels for a crime, they grouped crime activity into four categories: offence against people, offence against property, offence against authority and offence against local act. GIS also allows overlaying other datasets (census demographics, locations of the police stations, dispatching to emergencies, banks, etc.) to better understand the underlying causes of crime and help law enforcement administrators to devise strategies. The identification of relations between certain types of crime and the locations where they occurred can lead to a proper allocation of resources, helping in their combat and prevention.

This study focused on three years, from 2008 to 2010, and the data used was collected by two police units, at Dala and Jakara Division. ArcGIS was the GIS software used for the analysis and results reveal. This software merges three categories of functions: database management, spatial analysis and visualization. They also explored the evolution of the four categories of crimes by region.

The results showed that the spatial patterns of crimes tended to be clustered outside the city wall, probably connected to the absence of police stations. Outside the city wall, there are police outposts without enough manpower and facilities/equipment for policing activities. These results help to identify places that need police stations. They stated that many organizations require an improved system of data manipulation and analysis that can link information to their geographic location. GIS was presented as an alternative that could aid in this area and improve the decision-making process.

This chapter surveyed relevant advances on spatiotemporal data modeling with applications in different domains, including emergency medical services, with particular incidence on clustering and pattern-based solutions. In spite of the advances, pattern-based solutions face notable challenges associated with pattern numerosity, coverage, statistical significance, and actionability. In this context, the target work will be primarily centered on the discovery of actionable clustering solutions, being the role of spatiotemporal patterns seen as complementary.

Part II

Data Exploration and Spatiotemporal Clustering

4

Data Exploration

Contents

As described in the introductory chapter, this thesis is developed in the ambit of project Data2Help. Previously, in this project, a multi-dimensional database was developed using the data provided by INEM.

As a starting point, we conducted an exploration of the multi-dimensional database in order to assess its organization and extract the relevant features related to an emergency call event (Section 4.1). Aiming to obtain knowledge about the behaviour of emergency calls at a national level, the second step taken was to analyse the last five years of data available, period 2015-1019 (Section 4.2). As the last step, we examined the flow of emergency calls during the week (aggregating emergencies by day of the week) and during the day (aggregating emergencies by hour) (Section 4.3).

4.1 Case Study

The data warehouse at disposal allowed the extraction of emergency call events with the following attributes: space (location of the occurrence), time (several timestamps recorded per occurrence), priority of emergency, and type of emergency.

For this project, the nation-wide analysis of emergency events between 2015–2019 is selected as our study case. In this period, a total of 6685099 emergency calls were registered, where 99.97% of them have a location associated (district and county). The same does not occur for the several timestamps associated with an emergency, as shown in Table 4.1. The lack of emergencies with all timestamps (only 4.99%), arrival of unit timestamp (only 5.68% have a value different of NULL) and unit on the way timestamp (only 13.11%), only gave the possibility to explore one time interval - the activation time. The activation time of an emergency is the difference between the annotation of the occurrence during the call and the unit dispatch (81.11% have both).

Interesting timestamps	Percentage with values
Ocurrence creation	100
Unit dispatch	81.11
Unit on the way	13.11
Arrival of the unit	5.68
Occurrence conclusion	99.99
All timestamps	4.99

Table 4.1: Timestamps in each emergency and the correspondent percentage that contain a value not NULL.

In terms of priority levels across the emergencies, 87.24% of emergencies are labeled with a priority level higher than 2, as presented in Figure 4.1. A priority level higher than 2 means that the occurrence needed an emergency vehicle allocated.

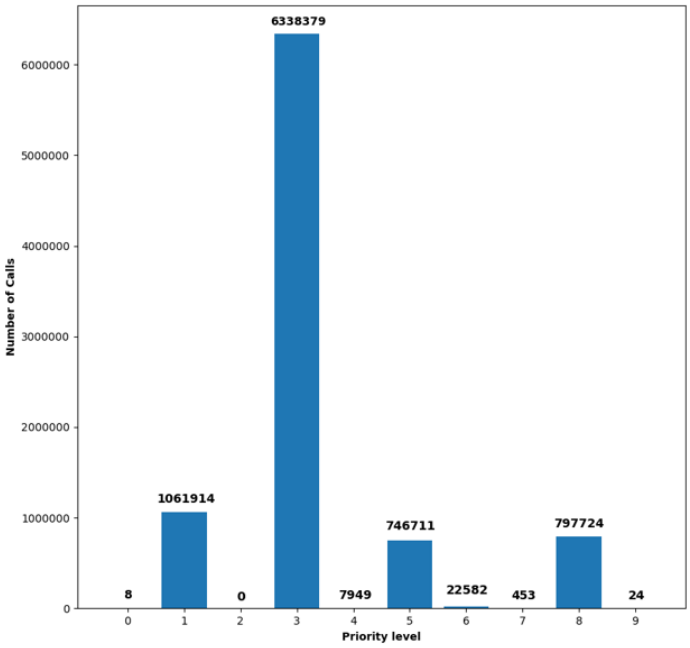


Figure 4.1: Bar chart containing emergency calls grouped by priority level.

The last crucial attribute extracted from the data warehouse was the type of emergency. INEM categorises this attribute with 41 different labels. In the scope of this project, those 41 types of emergencies were called the *initial pathologies*. The right side of Table 4.2 displays the initial pathologies and their prevalence among the case of study. Having 41 types of pathologies to inspect, represent and analyse in every step of the exploration and clustering lead to the creation of new pathologies, the *merged pathologies*. On the left side of Table 4.2 are presented the 13 merged pathologies, they were created by agglomerating the initial pathologies by similarity. While some agglomerations were more trivial to perform (e.g., merge types of pain into one category), others were not so clear (e.g., merge ”Intoxicação” with ”Diabetes”) but were executed in order to obtain a more reduced final number of pathologies.

Merged Pathology	Percentage	Initial Pathology	Percentage
OutrosProblemas	19.25	Outros Problemas	14.29
		Geral	2.95
		Pedido de Apoio Diferenciado	1.95
		Afogamento/Acidente Mergulho	0.04
		CODU MAR	0.00
		Ocorrências Complexas/NBQ	0.00
Trauma/Queimadura/Electrocussão/Hemorragia	17.16	Trauma	14.54
		Hemorragia	2.44
		Queimadura / Electrocussão	0.17
DorAbdominal/ProblemasUrinários/DorTorácica/DorCostas	13.44	Dor abdominal/Problemas Urinários	6.19
		Dor Torácica	4.79
		Dor nas costas	2.45
AlteraçãoEstadoConsciência	12.77	Alteração de Estado de Consciência	12.77
Dispneia/ParagemCardiorrespiratória/ObstruçãoViaAérea	11.71	Dispneia	10.05
		Paragem Cardiorrespiratória	1.39
		Obstrução Via Aérea	0.26
DéficeMotorSensitivo/Convulsões/Cefaleias	5.48	Défice Motor Sensitivo	2.48
		Convulsões	1.58
		Cefaleias	1.41
Lixo	4.89	Ocorrência Transferida do 112PT	3.06
		Não Ocorrencia	1.72
		Ocorrência Transferida do 112L	0.04
		Chamada Falsa	0.02
		Decisão não transporte	0.01
		Ocorrência Transferida do 112PT (Notificação)	0.00
		zTST	0.00
		Transporte Secundário	0.00
		Helitransporte	0.00
Intoxicação/Diabetes	3.80	Intoxicação	2.71
		Diabetes	1.08
AcidenteViação	3.43	Acidente Viação	3.43
Ginecologia/Gravidez/Parto/RecémNascidos/SAVP/CriançaDoente	3.24	Criança Doente	2.28
		Ginecologia/Gravidez	0.59
		Parto	0.37
		Recém Nascidos/SAVP	0.00
ProblemasPsiquiátricos/Suicídio/CAPIC	2.49	Problemas Psiquiátricos/Suicídio	2.32
		CAPIC	0.16
Agressão/Negligência/ViolênciaDoméstica/MausTratos	1.42	Agressão	1.26
		Negligência/Violência Doméstica/Maus Tratos	0.16
Alergias/Olhos/Ouvidos/Nariz/Garganta	0.86	Olhos/Ouvidos/Nariz/Garganta	0.44
		Alergias	0.42

Table 4.2: Merged pathologies and Initial pathologies percentages.

4.2 Medical emergency profiling (2015–2019)

This section shows the trail followed to acquire awareness on how the emergency domain of Portugal has evolved, with respect to the period 2015–2019. At first, we explored the total number of emergencies per day, week and month (4.2.1). After this, the data from all types of emergencies was splitted into the merged pathologies (Table 4.2) and we studied how the pathology emergencies vary during the distinct phases of the year (4.2.2).

4.2.1 National Global Emergencies

At start, we calculated the average number of emergencies per specific day (e.g. mean value of the five times 1st of January occurred in 2015-2019). The result of those values is presented as a heatmap in Figure 4.2. The conclusions taken from observing Figure 4.2 were that human mobility has an impact over the number of emergencies. More mobility correlates with a higher number of emergencies, for example, the 1st week of August (when more people go on vacations to the south of Portugal) and the week before and after Christmas (when people travel to be with their families). The days with fewer average emergencies take place on Portuguese holidays (25th of April, 1st of May and 10th of June), days typically where people is less stressed. The alarming value of this heatmap is for the 1st of January, being every year the day with more emergencies due to new year's celebrations.

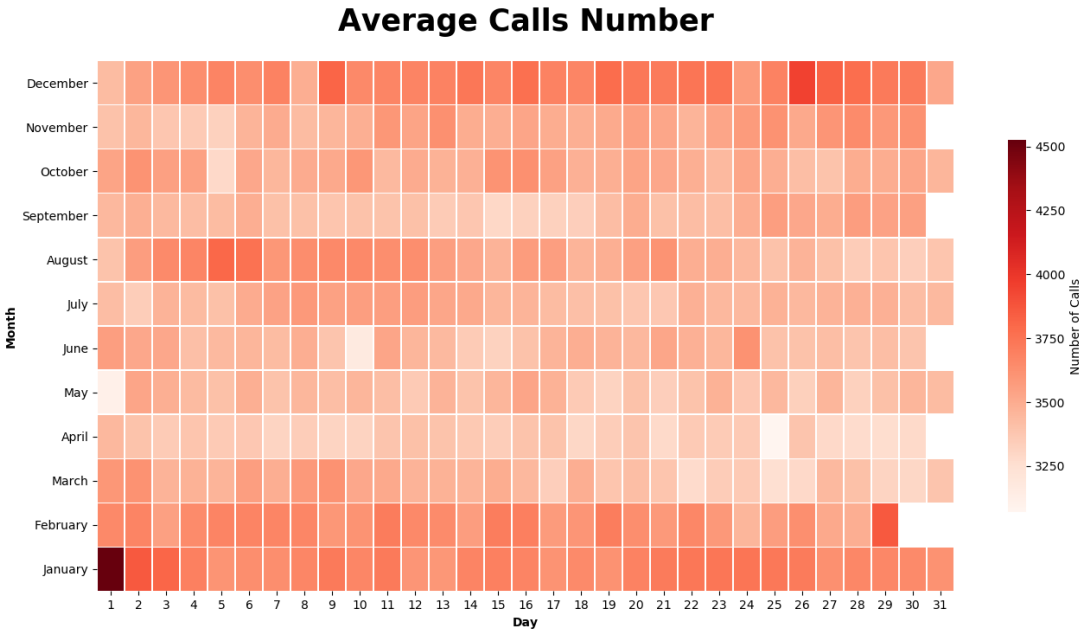
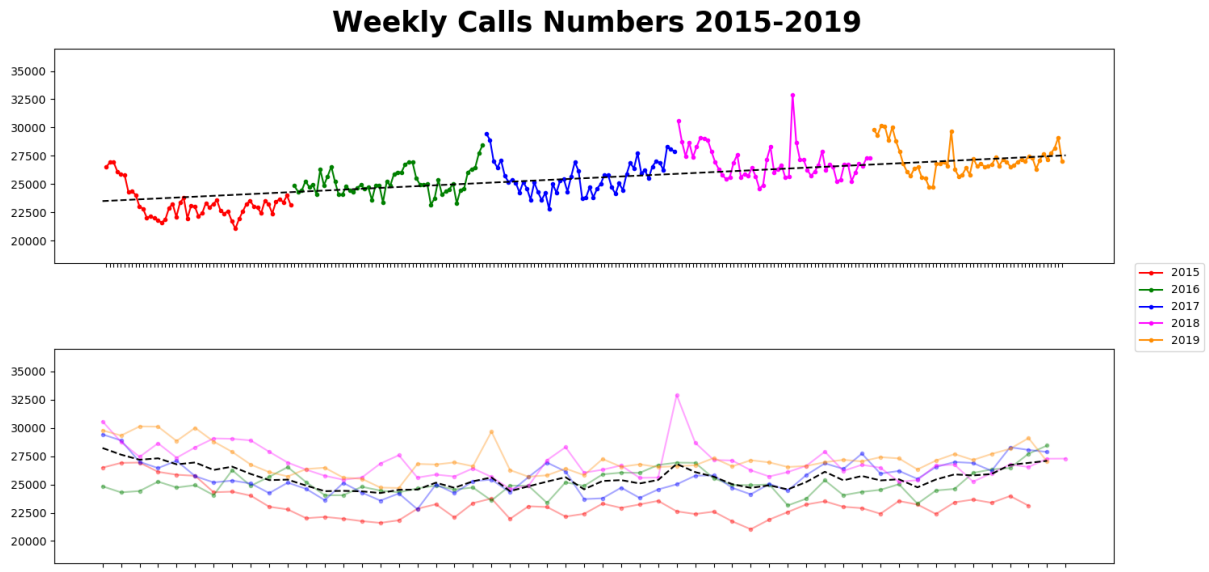
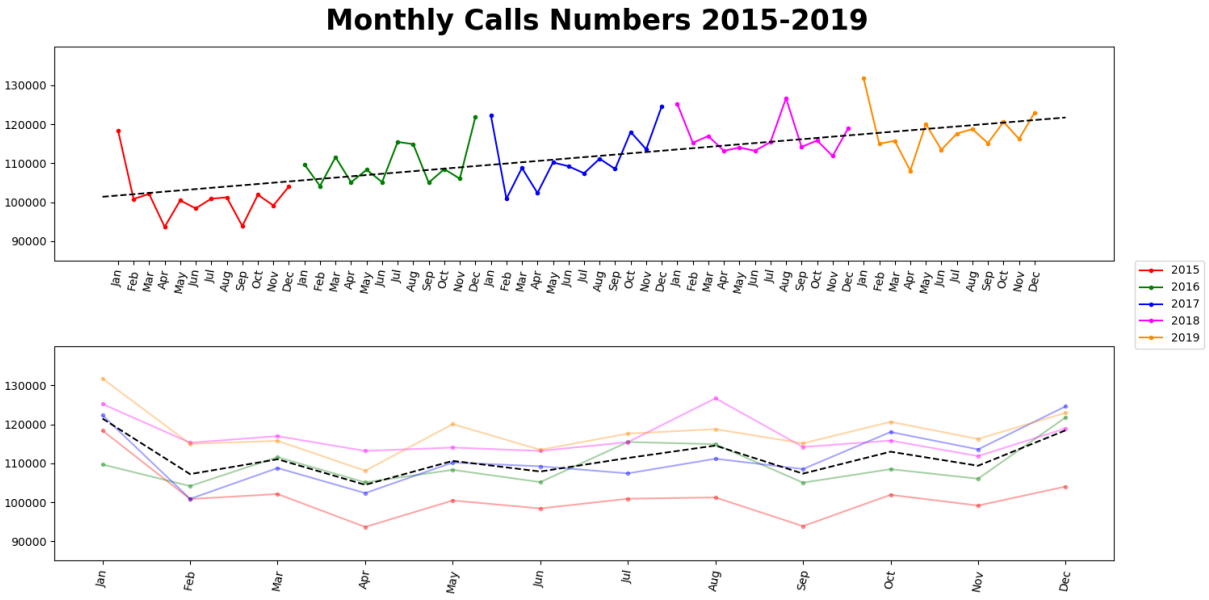


Figure 4.2: Average number of calls in each day (2015-2019).

To analyse the evolution of the emergency number of calls over time, the data was agglomerated by week (Figure 4.3a) and by month (Figure 4.3b). The values were labelled by year, having each year a colour associated. On the top part of both figures, we computed a linear regression (dashed line) that expresses the clear upward trend of the emergency calls number in this period of time. On the



(a) Values of emergency calls agglomerated by week 2015-2019.



(b) Values of emergency calls agglomerated by month 2015-2019.

Figure 4.3: Evolution of the total number of emergency calls between 2015-2019.

bottom part of the figures, it was overlaid the values for the 5 years and calculated the mean value for each position (dashed line). Inspecting the bottom part of Figure 4.3b, the conclusion taken was that, although the values for average emergencies per month are similar, January, August and December tend to have more emergencies.

4.2.2 National Merged Pathologies Emergencies

At this phase, we conducted a study of the merged pathologies with two objectives in mind: (1) find if the merged pathologies have any sort of seasonal behaviour and (2) discover which pathologies have been increasing over the period 2015-2019, since was known (Figure 1.1) that the total number of emergency has been growing every year.

The visualization that lead to a better understating of the pathologies domain was the computation of the collection of heatmaps presented in Figure 4.4. Each heatmap corresponds to a merged pathology and each cell contains the average amount of occurrences registered for that month of that year. Figure 4.4 reveals the following conclusions:

1. 'OutrosProblemas', 'DorAbdominal/ProblemasUrinários/DorTorácica/DorCostas' and 'Ginecologia/-Gravidez/Parto/RecémNascidos/SAVP/CriançaDoente' emergency numbers have been increasing in the last years.

2. 'Trauma/Queimadura/Electrocussão/Hemorragia' and 'AcidenteViação' present more emergencies in the second half of the year. The 'AcidenteViação' number of emergencies also has been growing due to the fact that nowadays the amount of vehicles per capita is higher.

3. 'Intoxicação/Diabetes' have been decreasing over this period of time, related to more tracking and prevention.

4. 'AlteraçãoEstadoConsciência', 'Dispneia/ParagemCardiorrespiratória/ObstruçãoViaAérea' and 'DéficeMotorSensitivo/Convulsões/Cefaleias' present their critic time during the winter.

5. 'ProblemasPsiquiátricos/Suicídio/CAPIC', 'Agressão/Negligência/ViolênciaDoméstica/MausTratos' and 'Alergias/Olhos/Ouvidos/Nariz/Garganta' have more occurrence during the summer.

The analysis was extended to model the merged pathologies (Figure 4.5). In this figure, the left side presents the values for each year in a continuous way (showing the evolution over time) and at the right the same data but overlapped to help the visualization of seasonal patterns. Upon the analysis of the left part, we conclude that the growth of emergency calls over time comes from the rising number of cases of the following pathologies: 'OutrosProblemas', 'Trauma/Queimadura/Electrocussão/Hemorragia', 'DorAbdominal/ProblemasUrinários/DorTorácica/DorCostas', 'AlteraçãoEstado

Consciência', 'Ginecologia/Gravidez/Parto/RecémNascidos/SAVP/CriançaDoente', 'ProbPsiquiátricos/Suicídio/CAPIC' and 'AcidenteViação'. Inspecting the right part of Figure 4.5, it was identifiable the pathologies that had an identical pattern every year ('AlteraçãoEstadoConsciência', 'Dispneia/ Paragem-Cardiorrespiratória/ObstruçãoViaAérea', 'Intoxicação/Diabetes', 'AcidenteViação', 'Ginecologia/ Gravidez/-Parto/RecémNascidos/SAVP/CriançaDoente', 'ProblemasPsiquiátricos/Suicídio/CAPIC', 'Agressão/ Negligência/ViolênciaDoméstica/MausTratos' and 'Alergias/Olhos/Ouvidos/Nariz/Garganta').

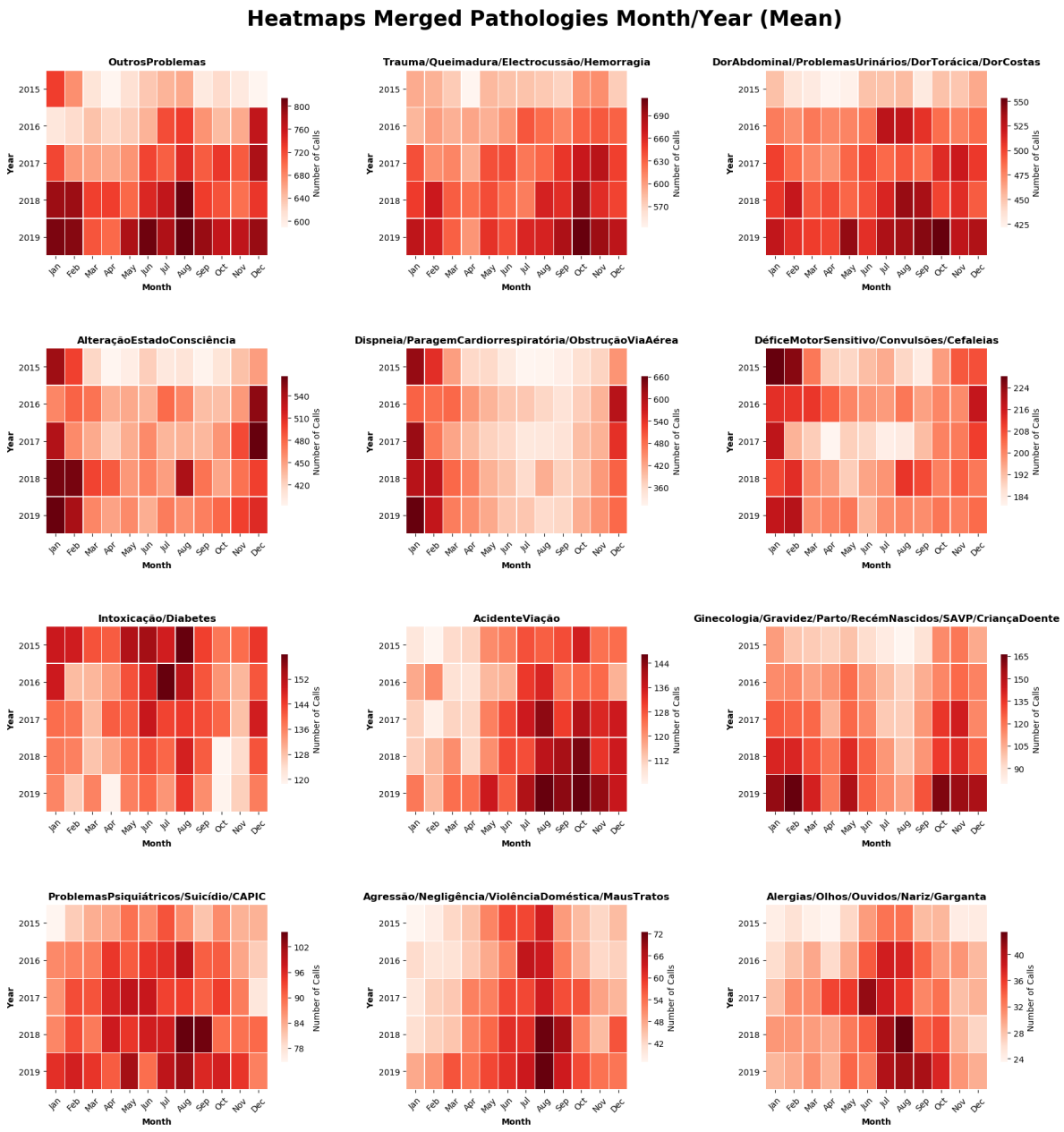


Figure 4.4: Heatmap collection for the merged pathologies (each cell contains the average amount of occurrences registered for that month of that year).

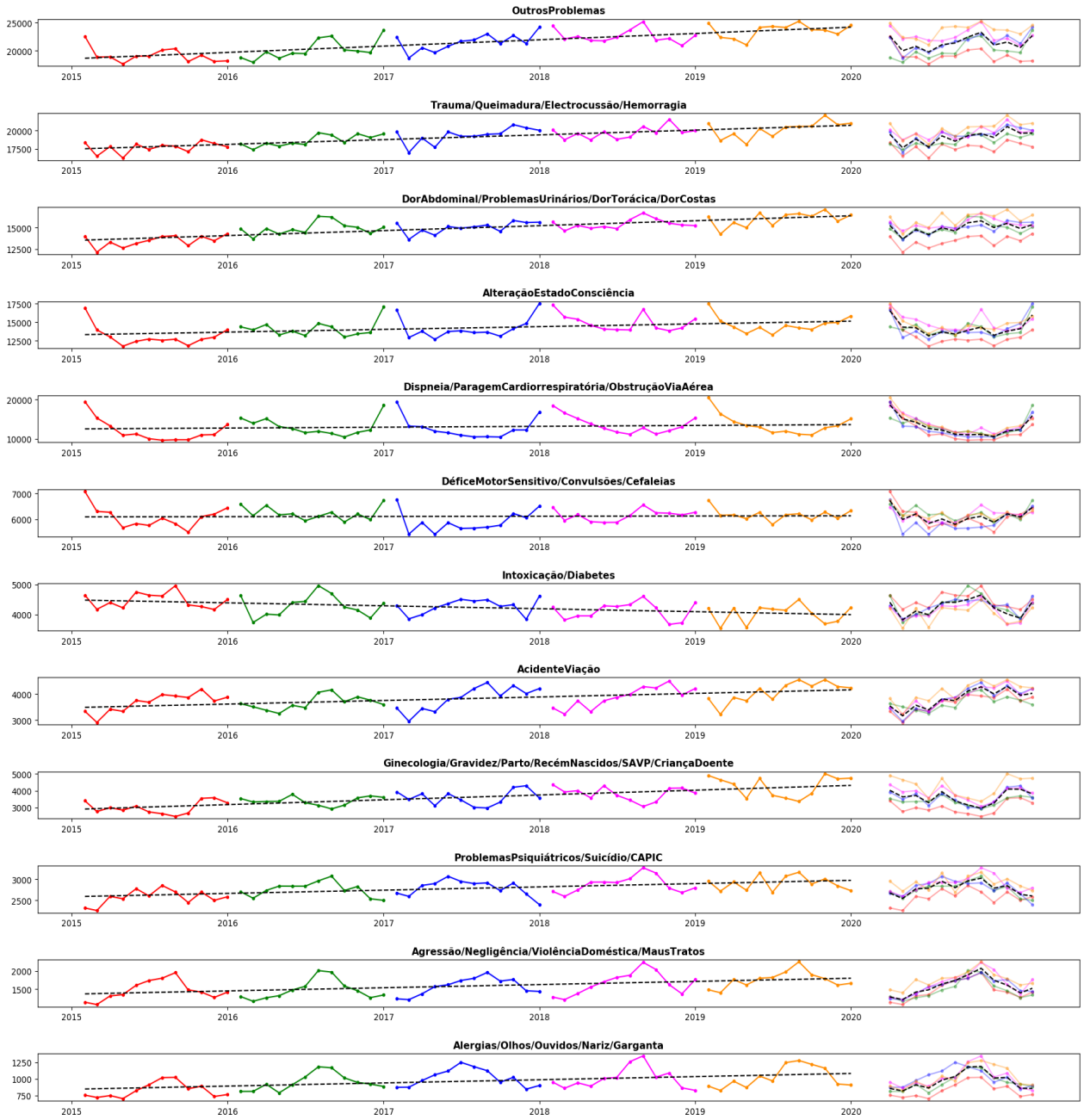


Figure 4.5: Values of emergency calls for the merged pathologies agglomerated by month 2015-2019.

4.3 Weekday and hour impact on emergencies

As a last exploration study, we researched how emergency prevalence per emergency type varied along the time of the and day of the week.

Considering Figure 4.6a, a major observation is that the total number of emergency calls does not have a strong association with the day of the week (only that Mondays have a slightly higher amount of emergencies). On the other hand, a completely different scenario occurs for the period of the day. The day was splitted into 4 hours intervals (0h-4h, 4h-8h, 8h-12h, 12h-16h, 16h-20h and 20h-0h) and the emergencies were agglomerated according with these intervals. In Figure 4.6b, can be analysed how the number of emergencies have a relation with the time intervals. The time intervals with more emergencies are 4h-8h and 8h-12h. Time periods with fewer emergencies (0h-4h and 20h-0h) present a high number of outliers (red circles) that are from specific days of the week (Fridays and Saturdays when people tend to go sleep later than usual).

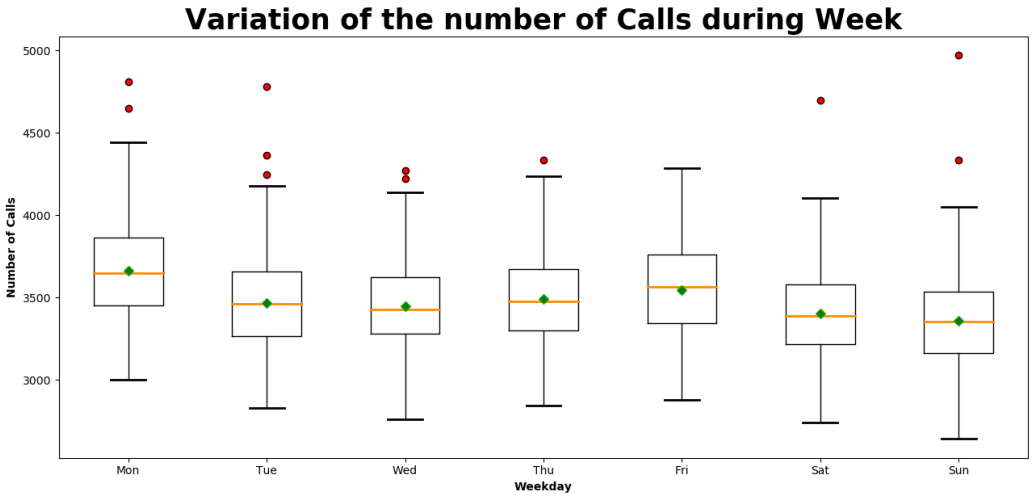
Complementarily, to conclude the exploration of the data, we want to assess how the emergencies per group of pathologies are distributed along a day and a week. To this end, we computed the collection of heatmaps present in Figure 4.7. After the examination of the heatmaps the reached conclusion was that this aggregation showed very clear patterns:

1. 'OutrosProblemas', 'Trauma/Queimadura/Electrocussão/Hemorragia', 'DorAbdominal/Problemas Urinários/DorTorácica/DorCostas' and 'Ginecologia/Gravidez/Parto/RecémNascidos/SAVP/Criança Doente' have more occurrences between 4h-8h and 8h-12h and fewer occurrences at weekends;

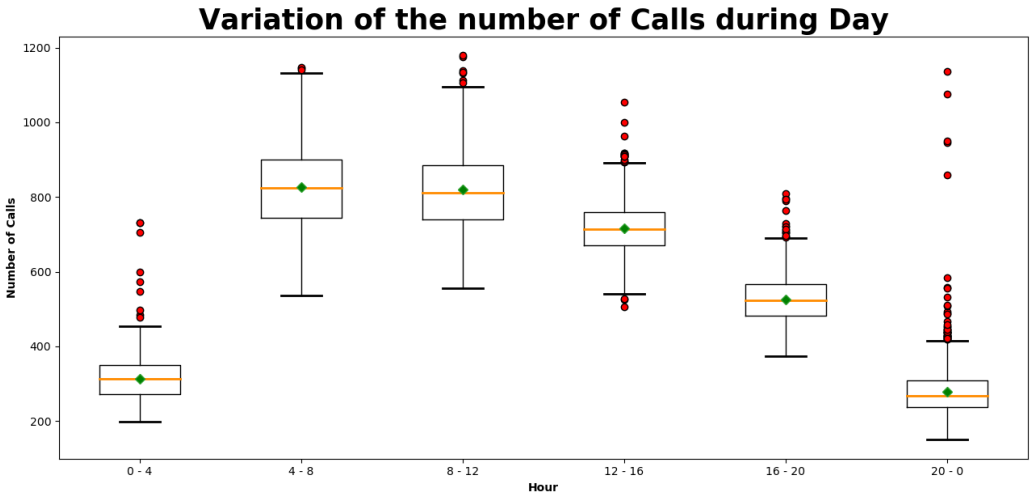
2. 'AcidenteViação' shows the highest value Friday at 12h-16h, when people's mobility increase due to the beginning of the weekend;

3. 'Intoxicação/Diabetes' and 'Agressão/Negligência/ViolênciaDoméstica/MausTratos' values of emergency prevalence start increasing after 12h, presenting the highest value at Sundays 20h-0h;

4. Analysing all heatmaps a conclusion was taken, Monday 4h-8h is the worst period of the week in terms of emergencies. Five different patolhogies ('OutrosProblemas', 'Trauma/Queimadura/Electrocussão/Hemorragia', 'DorAbdominal/ProblemasUrinários/DorTorácica/DorCostas', 'Dispneia/ParagemCardiorrespiratória/ObstruçãoViaAérea' and 'DéficeMotorSensitivo/Convulsões/Cefaleias') have this cell as a hotspot (cell with higher value).



(a) Variation of the number of calls during the week.



(b) Variation of the number of calls during the day.

Figure 4.6: Boxplots showing how medical emergency prevalence varies according to the day of the week (4.6a) and the period of the day (4.6b).

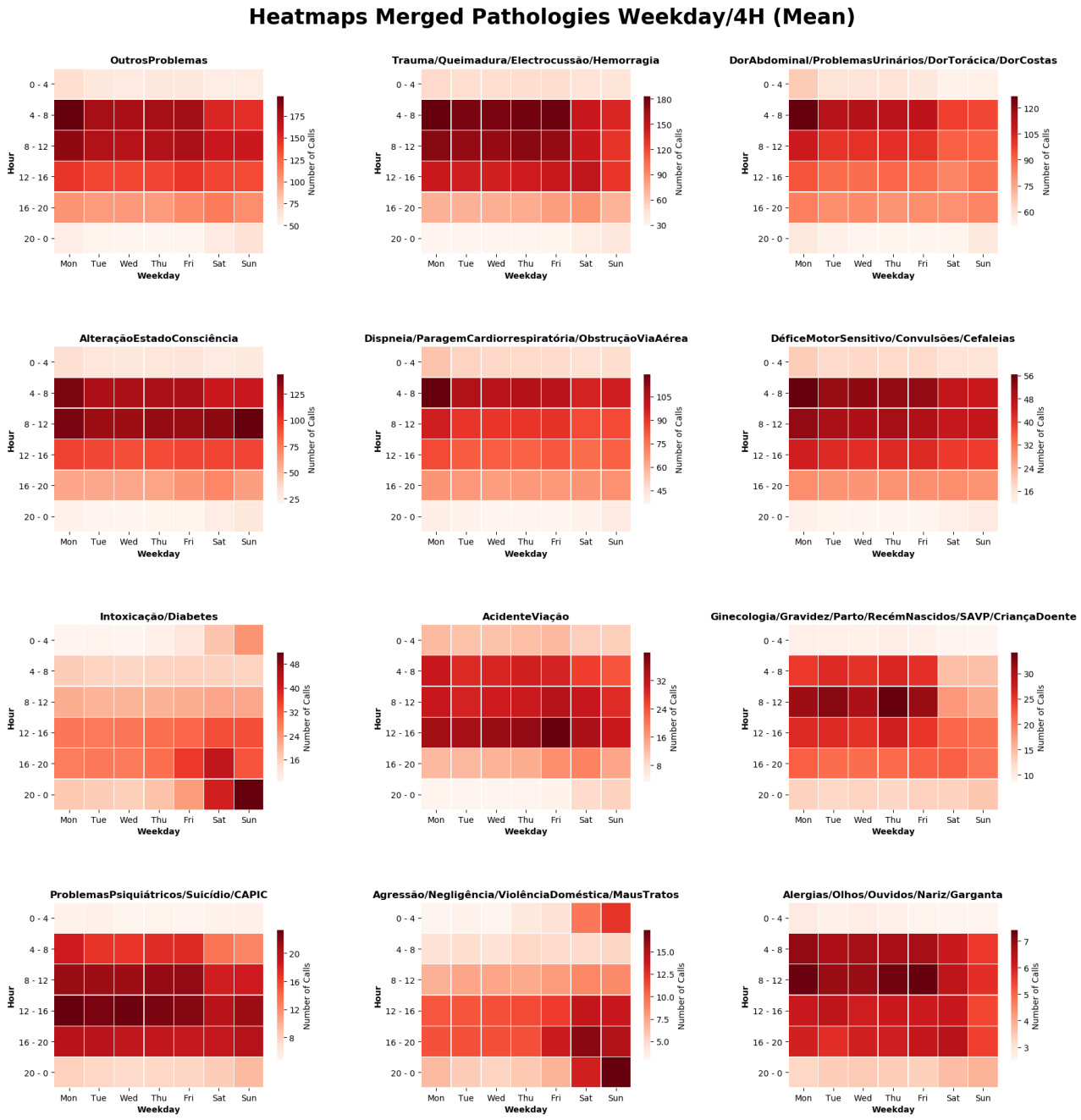


Figure 4.7: Heatmap collection for the grouped pathologies (each cell contains the average amount of occurrences registered for that time interval and that day of the week).

5

Clustering Solution

Contents

With the focus on answering the question that gave origin to this thesis: "Is there an underlying regional structure according to prevalence of medical emergencies and quality of response?", we apply spatiotemporal clustering methods over the occurrences by location. For this purpose, two different spatial granularities were selected: Portugal divided by counties (278) and by districts (18). In terms of temporal granularities two were explored: weekly and monthly. Our focus was to find synergies among different regions (forming clusters with similar behaviour) and correlate the clusters formed with their geographical location. This leads to the opportunity to allocate human resources, medical material and funds in a more decentralized way, resulting in a better response to emergencies.

This Chapter presents the steps taken in the implementation of our solution for spatiotemporal clustering. We first conducted a preprocessing step to prepare the data and form the time series for the clustering stage (Section 5.1). In our search for regions with similarities, we implemented two types of clustering algorithms (Partitioning and Hierarchical), as explained in detail in Section 5.2. In the end, we developed a geographic representation of Portugal to properly analyse the clusters formed by the learnt clustering models (Section 5.3).

5.1 Preprocessing and GTS formation

To reduce the computation time in the clustering stage, we process the raw event data to produce the targeted time series *a priori*. We formed four types of time series:

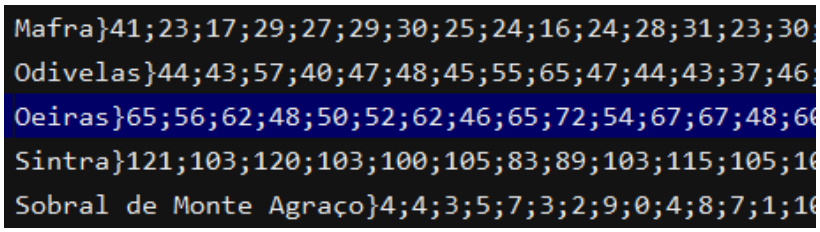
1. **Global** - univariate time series, where each position is an integer value corresponding to the number of occurrences, as shown in Figure 5.1a.

2. **Activation Time** - univariate time series, where each position is a float value corresponding to the average activation time.

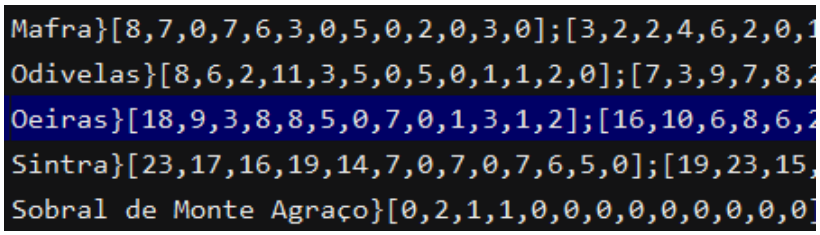
3. **Pathologies** - multivariate time series, where each position is a list with 13 elements, as shown in Figure 5.1b. Each element in this list is an integer value that corresponds to the number of occurrences for a specific group of pathologies. The order of the pathologies within the list is shown in accordance with their prevalence, represented in Table 4.2.

4. **Priority** - univariate time series, where each position is a integer value corresponding to a severity level.

For each type of time series, a folder was created that contained 18 text files (one for each district). Each text file stores the values for the counties that belong to that district (one county by line). Figure 5.1 presents Lisbon district file example for the global series (5.1a) and for the pathologies series (5.1b)



(a) Lisbon district text file for the global series.



(b) Lisbon district text file for the pathology-specific multivariate series.

Figure 5.1: Text files created for Lisbon district.

5.2 Clustering Implementation

In our implementation, we tested two types of clustering algorithms: a partitioning algorithm (*k-means*) and a hierarchical algorithm (*agglomerative*). We also develop an extension for the hierarchical algorithm, where a spatial restriction was imposed, detailed in 5.2.2

After starting experimenting our cluster implementation, a conclusion was instantaneously reached. Due to the big variation of density population in Portugal's territory, the clusters were not forming with basis on the shape of the time series, instead, the primary factor was the dimension of the number of occurrences. We solved this problem by normalizing the time series with the respective population values for each location, using values available from Censos 2011.

5.2.1 Partitioning clustering

Partitioning algorithms represent each cluster by a prototype. For *k-means*, a prototype is an object representing the center of mass of the cluster, in the context of time series this object is called barycenter. For this type of clustering, we used *tslearn* package to model the clusters and internal validation methods, SSE(5.1) and Silhouette(5.2), to measure the compactness and the separation of the clusters.

$$SSE = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathbf{C}_k} dist(\mathbf{x}_i, \mathbf{c}_k)^2, \tag{5.1}$$

where K is the number of clusters, $\mathbf{x}_i$ is the element and $\mathbf{c}_k$ is the barycenter of the cluster $\mathbf{C}_k$.

$$Silhouette = \frac{b - a}{max(a, b)}, \tag{5.2}$$

where a is the average intra-cluster distance (average distance between each point within a cluster) and b is the average inter-cluster distance (average distance between all clusters). The value of the silhouette ranges between $[-1, 1]$, where a high value (1) indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

5.2.2 Hierarchical clustering

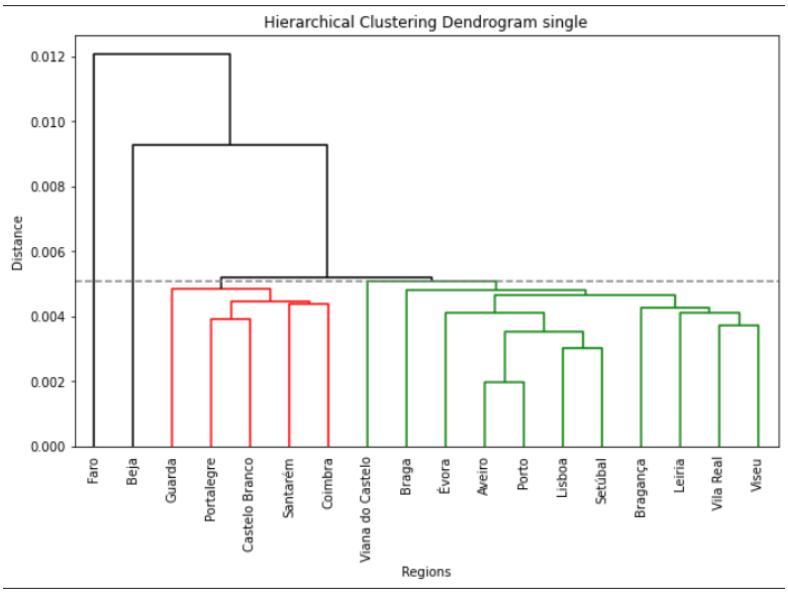


Figure 5.2: Example of a dendogram from our project.

Hierarchical algorithms group objects in a tree-like structure, called *dendrogram* (Figure 5.2). We used an agglomerative hierarchical algorithm from *scipy* package in our implementation, that start by placing every object in different groups and iteratively joining the two most similar groups until a threshold, defined by us, was reached (horizontal dashed line in Figure 5.2). To measure similarity between groups three linkage criteria were used:

1. **single link** - where the distance between the groups is the distance of the two closest objects.
2. **complete link** - where the distance between the groups is the distance between the two furthest objects.
3. **average link** - where the distance between the groups is the average distance between every object of every cluster.

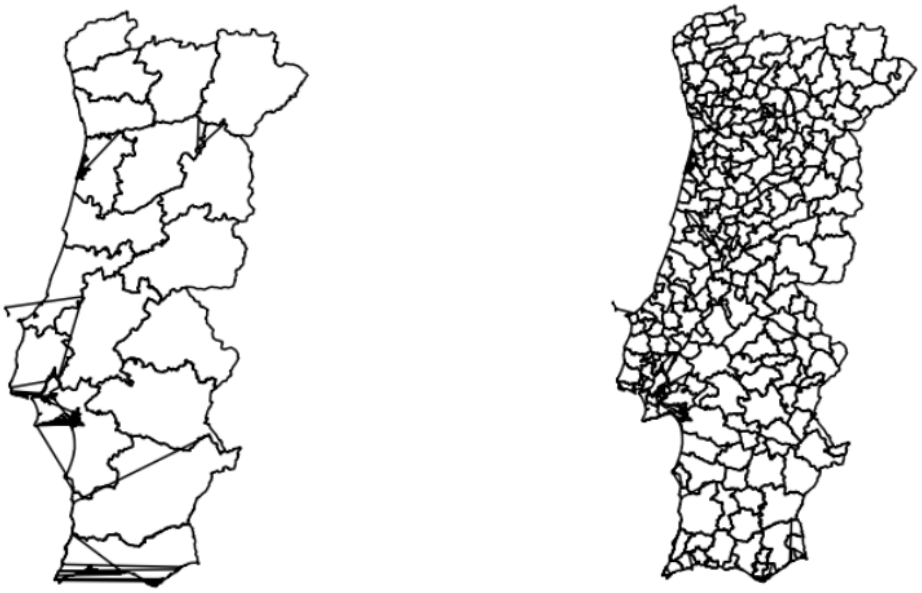
As an extension to this algorithm, we created a constraint to potency the modeling of clusters where the members are nearby. We created a method, *areRegionsConnected*, that returns a boolean value,

meaning if the two locations are close in the geographic context. For the district granularity, we considered ”connected” the districts that share boundaries. In the county granularity, to be considered ”connected” a county needed to belong to the same district or belong to an adjacent district. With this method built, when computing the values for the DTW matrices:

1. **if the regions are ”connected”:** the real distance value between the time series is stored.
2. **else:** a high-value distance is stored, forcing the regions to have more difficulty clustering.

5.3 Clusters Visualization

In order to have a geographic visualization of the clusters formed, with both district (Figure 5.3a) and county (5.3b) granularity, we develop a map to be filled with the same colors in which the time series were printed.



(a) District granularity.

(b) County granularity.

Figure 5.3: Portuguese continental territory divided by district (5.3a) and by county (5.3b).

Complementarily, the barycenter series per cluster is visualized against the time series grouped within that cluster using line charts.

6

Results

Contents

This chapter gathers the majer discoveries found during our exploration of spatiotemporal patterns in the target emergency study case (Section 4.1). By dividing the emergency occurrences by location (district/county), we explored their relationships with the aim on finding if exists any correlation in time and space. The number of occurrences per capita shows the lowest values in the north region of Portugal and starts increasing when moving south Figure (6.8b).

We also performed an analysis of the locations that presented a difficulty clustering (anomalies) because of their unique behaviour.

For each type of time series (global, activation time, and pathologies), we assessed their behavior against average values for Portugal (dashed line represented in clusters figures) in order to easily identify which clusters are above and below the national average levels of prevalence and response efficiency.

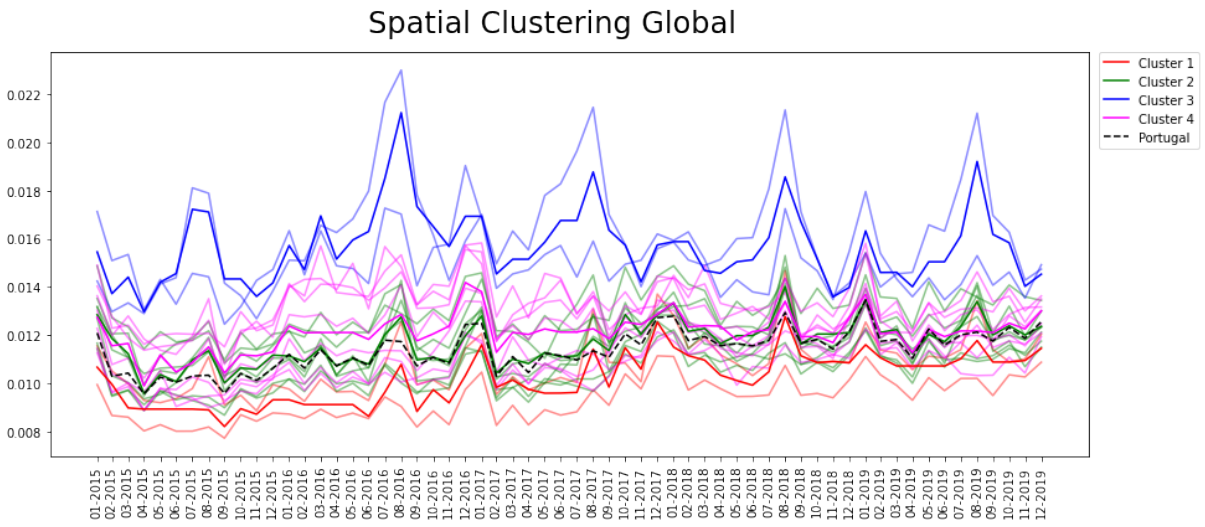
The results and information obtained are organized by spatial granularity: district granularity in Section 6.1 and county granularity in Section 6.2.

6.1 District Granularity

When clustering the global series with a district spatial granularity, under Spatial clustering with average linkage, a correlation between the district geographic location and the number of occurrences per capita is observed, as verified in Figure 6.1. The districts from the north of Portugal have a lower number of emergency calls per capita and those values increase with the more in the south a district is geographically located. The districts that presented an atypical flow of emergencies when compared with the others were: Faro, Beja and Braga (Figure 6.2). Faro and Beja stood out due to the seasonal behaviour, where in the summer months the number of emergencies escalate. Braga was considered an anomaly for a different reason, during the entire period have the lowest number of occurrences per capita.

For the activation time series, the solution that presented the better set of clusters recurred to Spatial clustering with average linkage (Figure 6.4). The evolution over time of the activation time series shows two completely distinct behaviours: pre and after August of 2017. Before, the northern districts had better activation times, but after it looks like all clusters have the same pattern. This led us with the idea that maybe before Centro de Orientação de Doentes Urgentes (CODU) operated at some type of regional level and later changed to a centralized center. But what is truly observable is that the average activation time increased. The study of how this time interval may affect the time of rescue would have been also interesting, however, the number of missing values in the timestamps of the occurrence made it impractical.

Figure 6.5 shows the presence of three anomalous regions detected in the clustering for the pathologies series. The enormous variation of people in the hot months in Faro makes the number of emergencies blow up.



(a) Time series representation with cluster label.

(b) Map divided by formed clusters.

Figure 6.1: Cluster results for global series with district and monthly granularities, using spatial algorithm with average linkage.

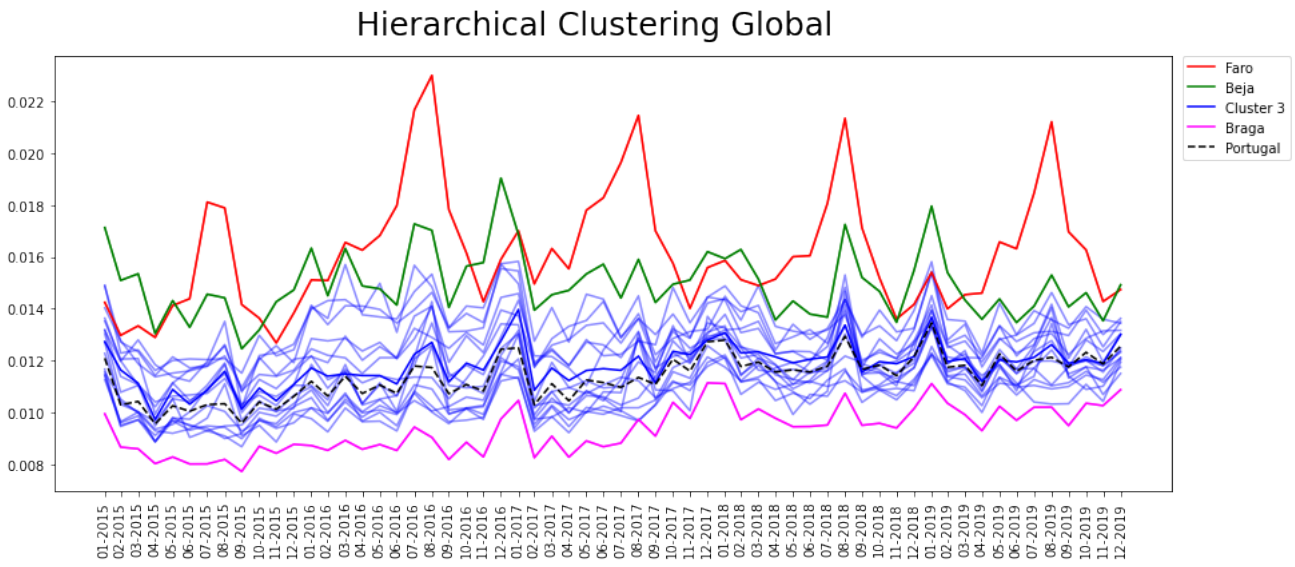


Figure 6.2: Faro, Beja and Braga districts highlighted as anomalies with global series.

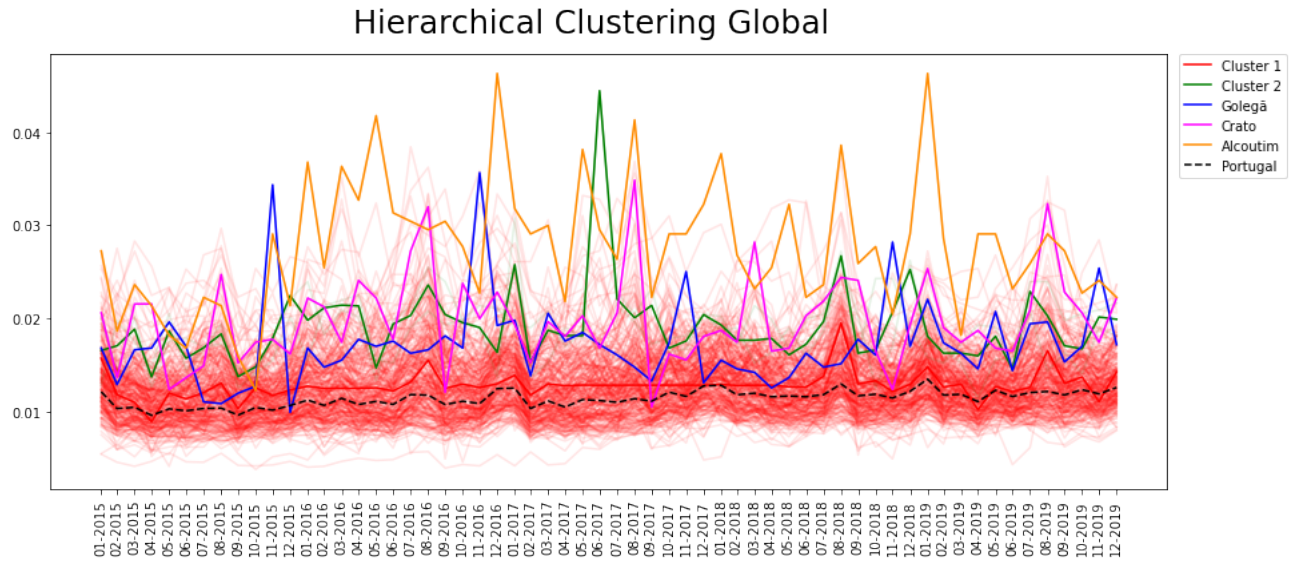
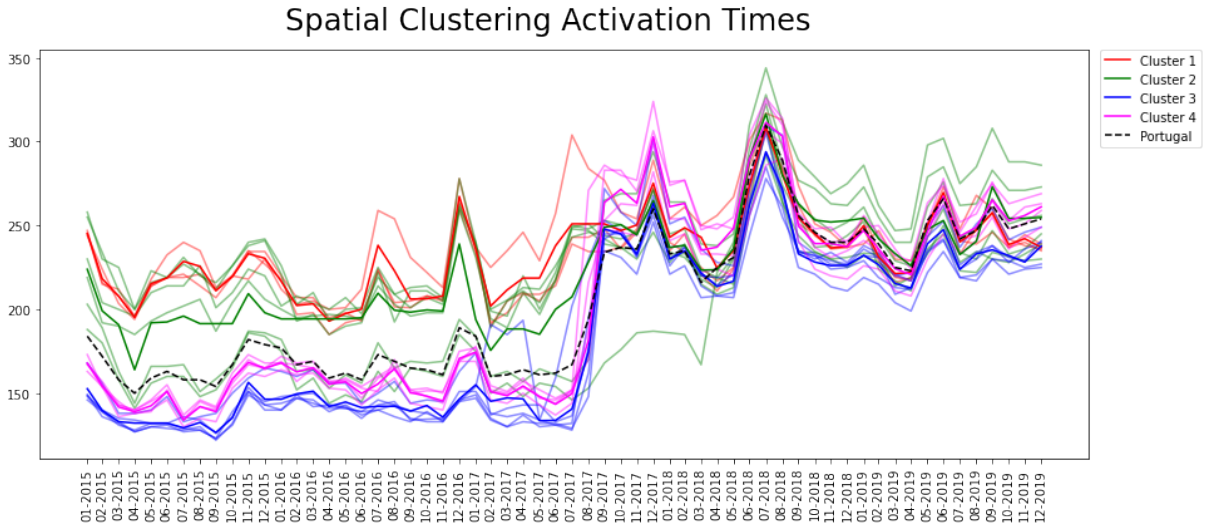


Figure 6.3: Alcoutim, Crato e Golegã counties highlighted as anomalies with global series.



(a) Time series representation with cluster label.

(b) Map divided by formed clusters.

Figure 6.4: Cluster results for activation time series with district and monthly granularities, using spatial algorithm with average linkage.

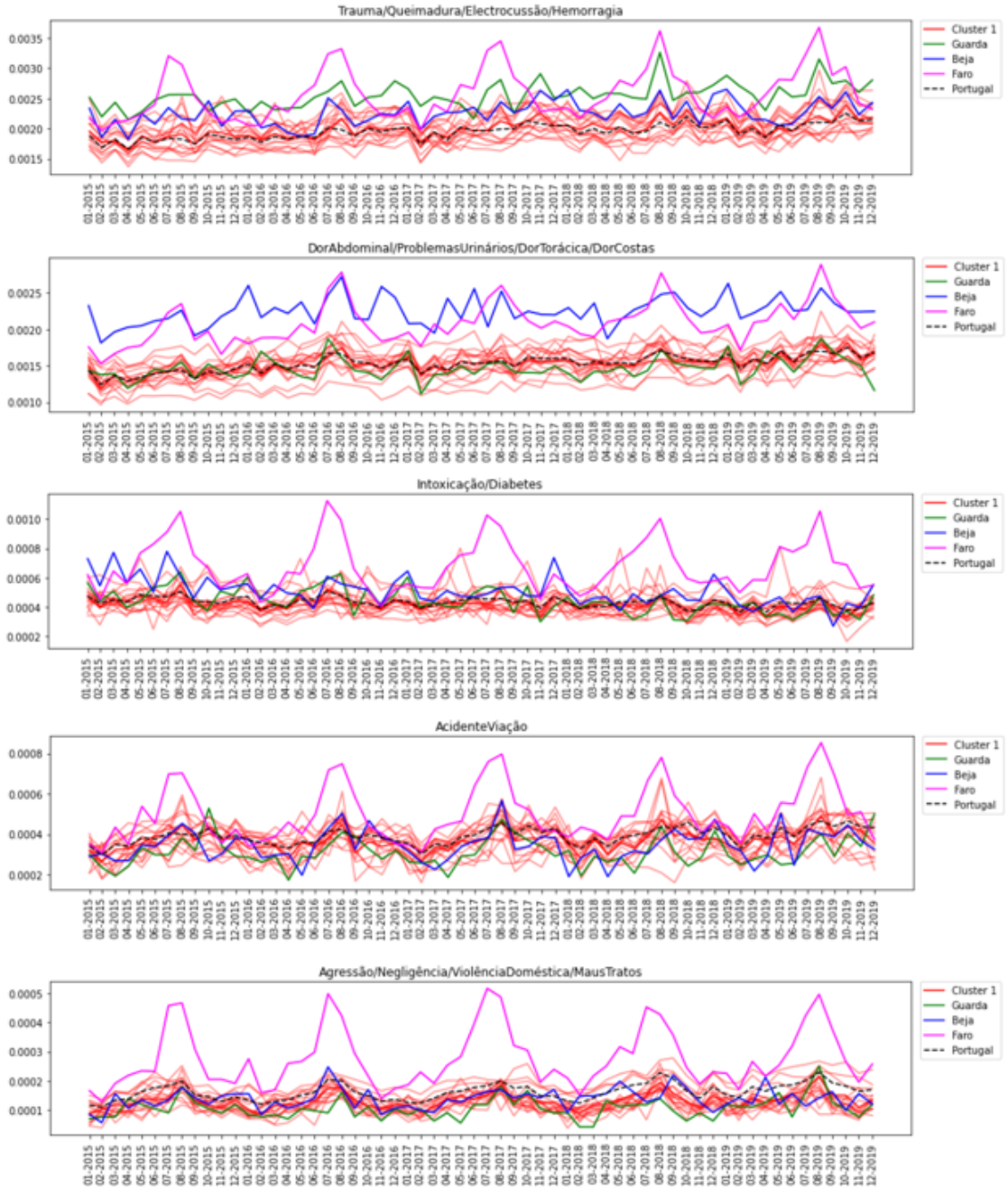


Figure 6.5: Faro, Beja and Guarda highlighted as anomalies in clustering with pathologies series.

6.2 County Granularity

For the global series, the best set of clusters was returned by the hierarchical clustering with average linkage. According to Figure 6.7, we reached the following conclusions: counties with higher citizen density have lower emergencies per capita (Porto and Lisboa counties belong to the red cluster) and counties with more occurrences per capita are more in the interior (pink and blue clusters).

For global series, the counties with the rarest conduct were: Golegã, Crato and Alcoutim (Figure 6.6). After searching for explanations for these types of abnormal situations we found: (1) Golegã hosts 'Feira da Golegã' always in the first week of August (https://feiradagolega.com), justifying the spike observed; (2) Crato has a festival in August (https://cm-crato.pt/cartaz-do-festival-do-crato-2022/) that match emerge of occurrences, and (3) Alcoutim was noticed by RTP as the more unpopulated and aged county of Portugal (https://www.rtp.pt/noticias/pais/linha-da-frente-alcoutim-o-concelho-mais-despovoado-e-envelhecido-do-pais$_v1383430$), what clarified this singularity.

With consideration to the activation time series, the algorithm that performed better outcome was the hierarchical with average linkage. The result is shown in Figure 6.8.

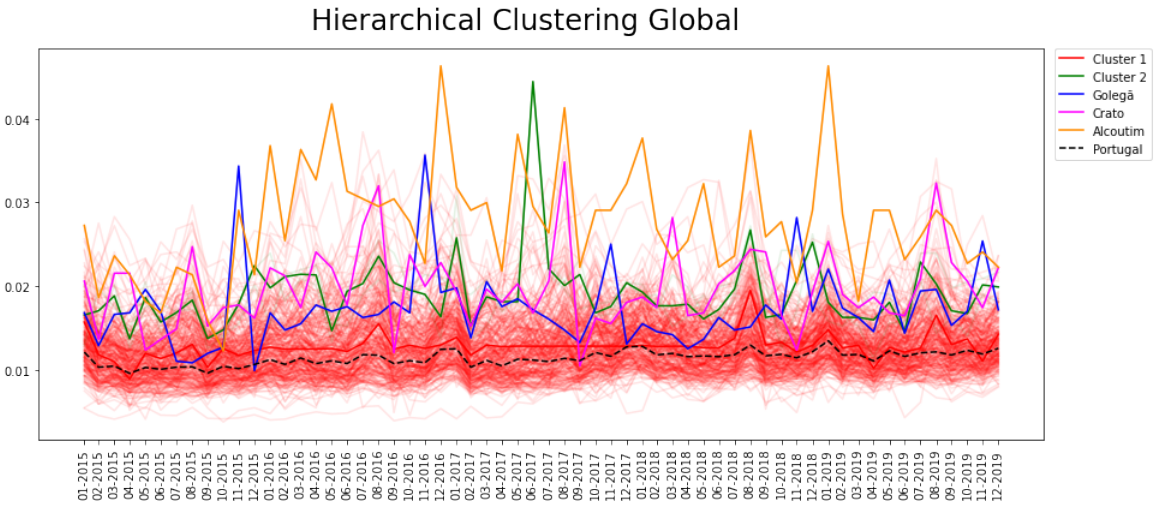
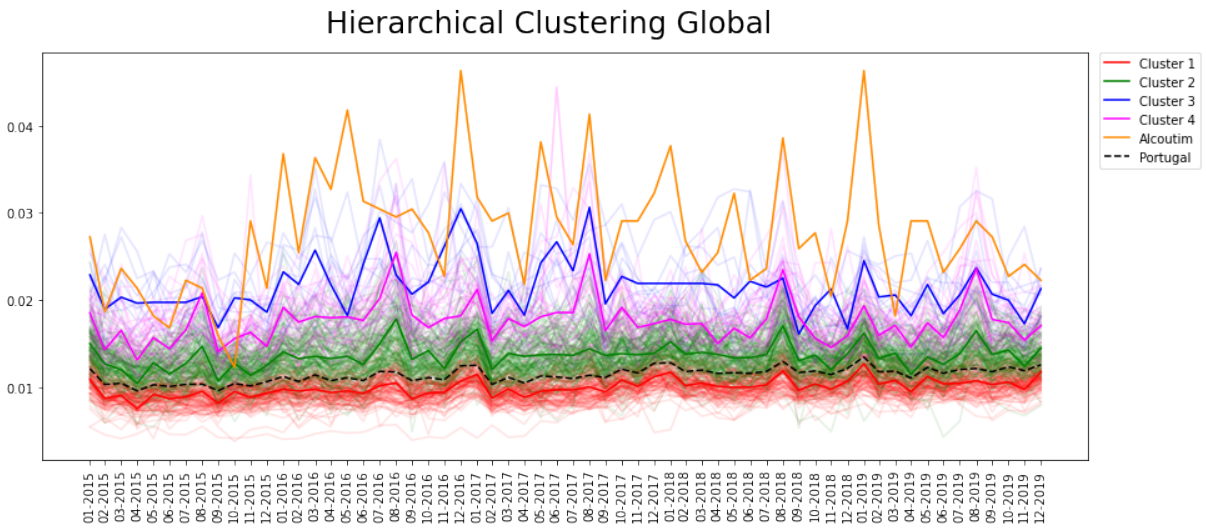
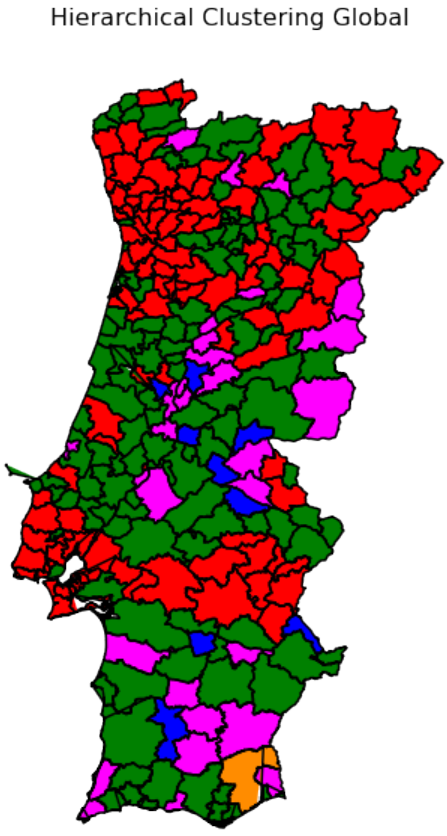


Figure 6.6: Alcoutim, Crato e Golegã counties highlighted as anomalies with global series.

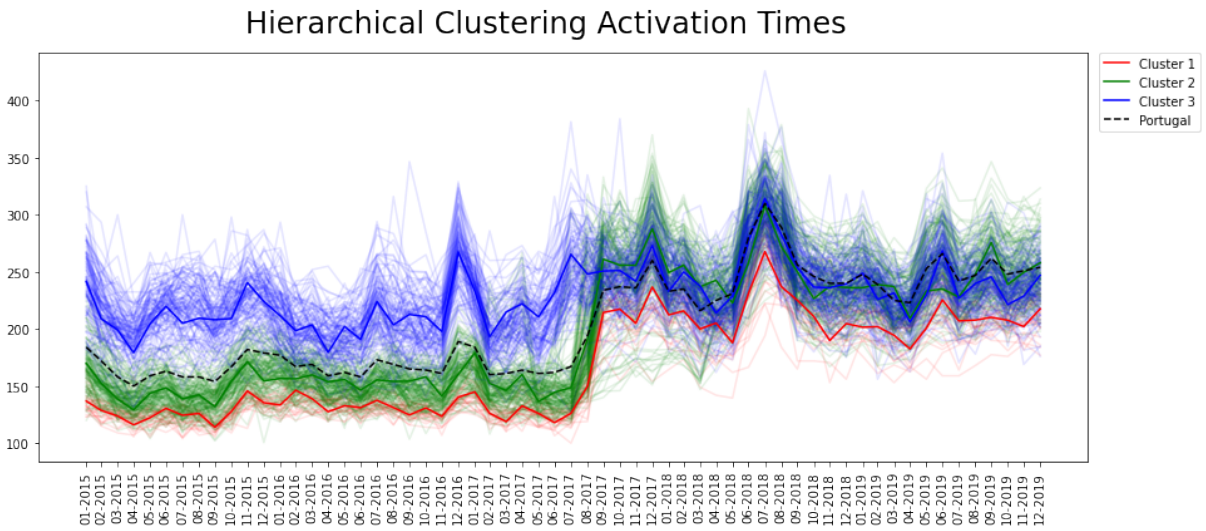


(a) Time series representation with cluster label.

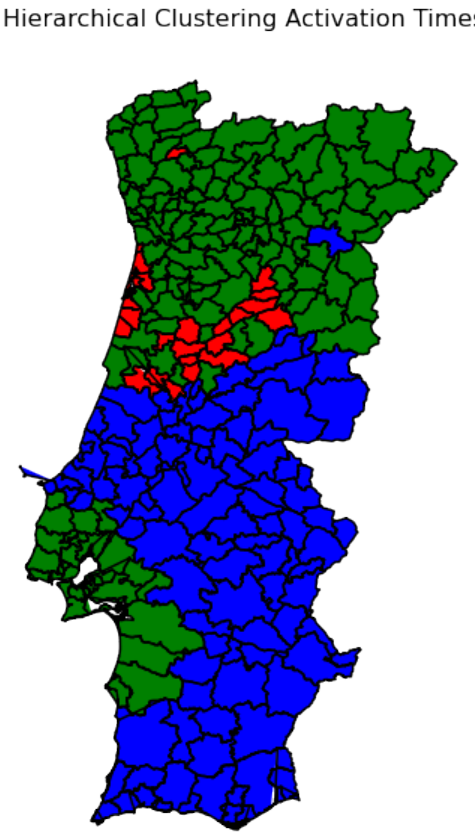


(b) Map divided by formed clusters.

Figure 6.7: Cluster results for global series with county and monthly granularities, using hierarchical algorithm with average linkage.



(a) Time series representation with cluster label.



(b) Map divided by formed clusters.

Figure 6.8: Clusters results for activation time series with county and monthly granularities, using hierarchical algorithm with average linkage.

Part III

Conclusions and Future Work

7

Concluding Remarks

Contents

7.1 Discussion

This thesis, established within the context of the Data2Help project, aimed to study if nearby regions present similar behaviour in medical emergency features (number of occurrences per capita, time to activate a rescue vehicle, pathology seasonality, etc.). Uncovering these types of synergies are valuable to multiple ends, including the support to resource allocation, human and material, in order to have the minimal response time achievable. Looking at emergencies in a more decentralized way gives the chance to properly understand and model emergency patterns at specific locations, allocating the needed support *a priori* and also recognizing periods where resources are not being used, being accessible to be shifted to a nearby location in need.

To answer the targeted research question, clustering solutions were produced for the GTS data. Both district and county granularities confirm that, in fact, regions closer in space tend to have more look-alike behaviour. This can lead to the formation of regional clusters, where local decisions can be optimized on the basis of the specific cluster needs. There is another possibility to improve the response time based on the formed clusters: explore synergies by identifying clusters with opposed behaviour with regards to specific pathologies and, using them, find periods for the punctual allocation/exchange of means.

The exploration of the nation-wide emergency data revealed notable trends and patterns for the different pathologies and severities. A most alarming number is the continuous increase in the number of emergencies, making alike projects relevant for the continuous research and discovery of new forms to improve emergency responses, saving more lives.

Future Work

We highlight four major lines of future work. First, the translation of the collected empirical evidence in practical initiatives near INEM. Second, assessing response time association with the number and type of available vehicles at the different stations. Third, extending the proposed clustering solutions with similarity criteria based on inverse correlation to identify synergistic regions, i.e., candidate regions for the periodic exchange of resources. Fourth, the complementary discovery of spatiotemporal patterns from raw emergency data to augment the already acquired knowledge.

Bibliography

[1] Ahmed, M. and Salihu, R. (2013). Spatiotemporal pattern of crime using geographic information system (gis) approach in dala lga of kano state, nigeria. *American Journal of Engineering Research (AJER)*, 2(3):51–58.

[2] Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367.

[3] Antonelli, D. and Bruno, G. (2015). Application of process mining and semantic structuring towards a lean healthcare network. In *Working Conference on Virtual Enterprises*, pages 497–508. Springer.

[4] Atluri, G., Karpatne, A., and Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, 51(4):1–41.

[5] Aydin, B., Kempton, D., Akkineni, V., Gopavaram, S. R., Pillai, K. G., and Angryk, R. (2014). Spatiotemporal indexing techniques for efficiently mining spatiotemporal co-occurrence patterns. In *2014 IEEE international conference on big data (Big Data)*, pages 1–10. IEEE.

[6] Brockwell, P. J., Davis, R. A., and Fienberg, S. E. (1991). *Time series: theory and methods: theory and methods*. Springer Science & Business Media.

[7] Chen, M.-S., Han, J., and Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, 8(6):866–883.

[8] Dong, G. and Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52.

[9] Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31:1–24.

[10] Goethals, B. (2003). Survey on frequent pattern mining. *Univ. of Helsinki*, 19:840–852.

[11] Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[12] Jasso, H., Hodgkiss, W., Baru, C., Fountain, T., Reich, D., and Warner, K. (2007). Spatiotemporal characteristics of 9-1-1 emergency call hotspots. In *Proceedings of the National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation (NGDM'07)*, pages 10–12. Baltimore Maryland.

[13] Lamine, E., Fontanili, F., Di Mascolo, M., and Pingaud, H. (2015). Improving the management of an emergency call service by combining process mining and discrete event simulation approaches. In *Working Conference on Virtual Enterprises*, pages 535–546. Springer.

[14] Li, J., Fu, A. W.-c., He, H., Chen, J., Jin, H., McAullay, D., Williams, G., Sparks, R., and Kelman, C. (2005). Mining risk patterns in medical data. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 770–775.

[15] Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2):284–289.

[16] Neves, F., Finamore, A., and Henriques, R. (2020a). Efficient discovery of emerging patterns in heterogeneous spatiotemporal data from mobile sensors. *17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous)*.

[17] Neves, F., Finamore, A., Madeira, S., and Henriques, R. (2020b). Mining actionable traffic patterns of road mobility using biclustering.

[18] Selvam, A. and Thivakaran, T. (2016). Mining patterns from 9-1-1 calls dataset. *International Journal of Applied Information Systems (IJAIS)*.

[19] Shekhar, S., Jiang, Z., Ali, R. Y., Eftelioglu, E., Tang, X., Gunturi, V., and Zhou, X. (2015). Spatiotemporal data mining: a computational perspective. *ISPRS International Journal of Geo-Information*, 4(4):2306–2338.

[20] Wei, W. W. (2006). Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*. Oxford Library of Psychology.

[21] Wong, A. K. C. and Wang, Y. (2003). Pattern discovery: a data driven approach to decision support. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 33(1):114–124.

[22] Wu, X., Zurita-Milla, R., Izquierdo Verdiguier, E., and Kraak, M.-J. (2018). Triclustering georeferenced time series for analyzing patterns of intra-annual variability in temperature. *Annals of the American Association of Geographers*, 108(1):71–87.

[23] Zhao, Q. and Bhowmick, S. S. (2003). Sequential pattern mining: A survey. *ITechnical Report CAIS Nayang Technological University Singapore*, 1(26):135.

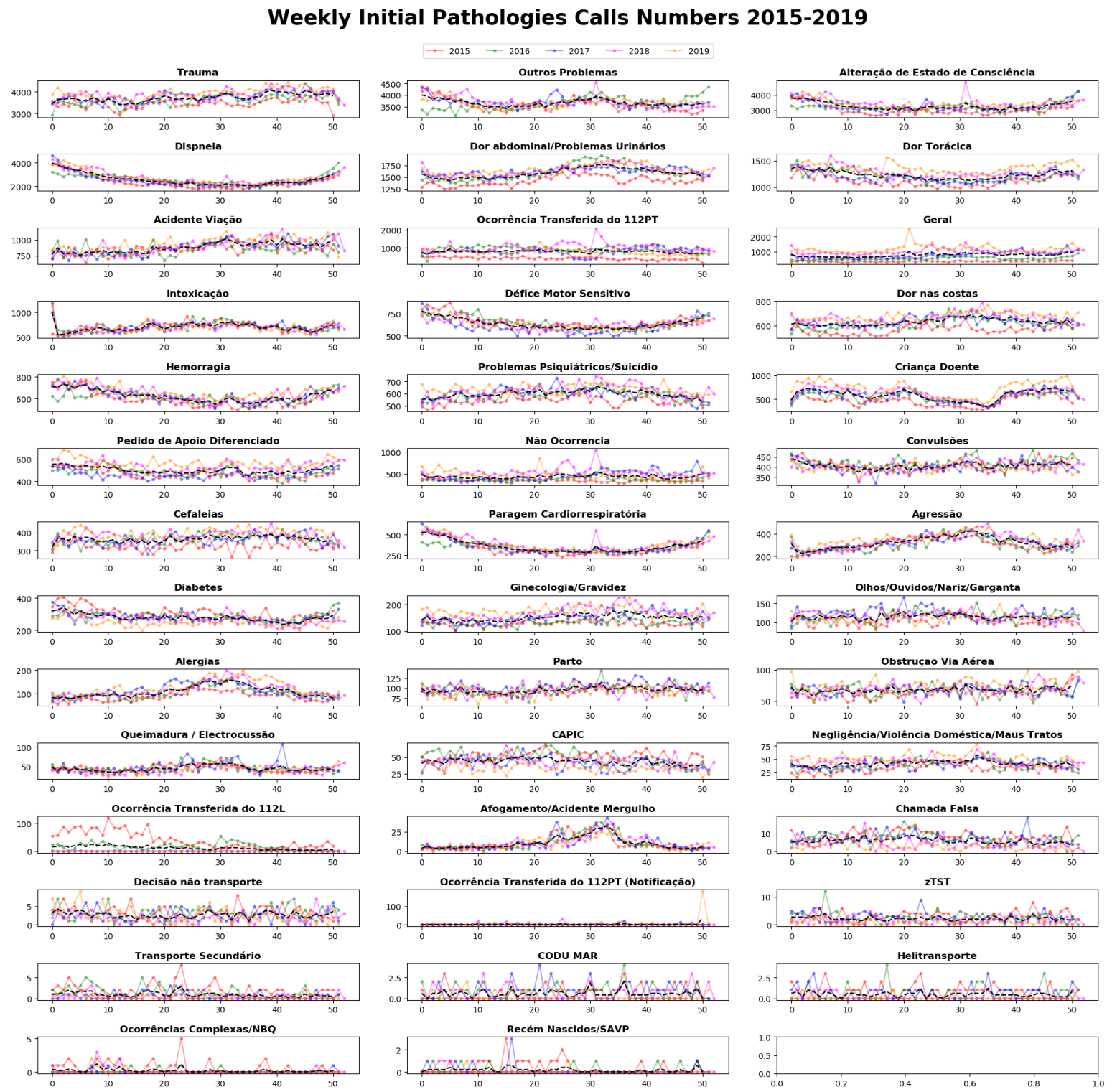


Figure 1: Initial Patholigies weekly.

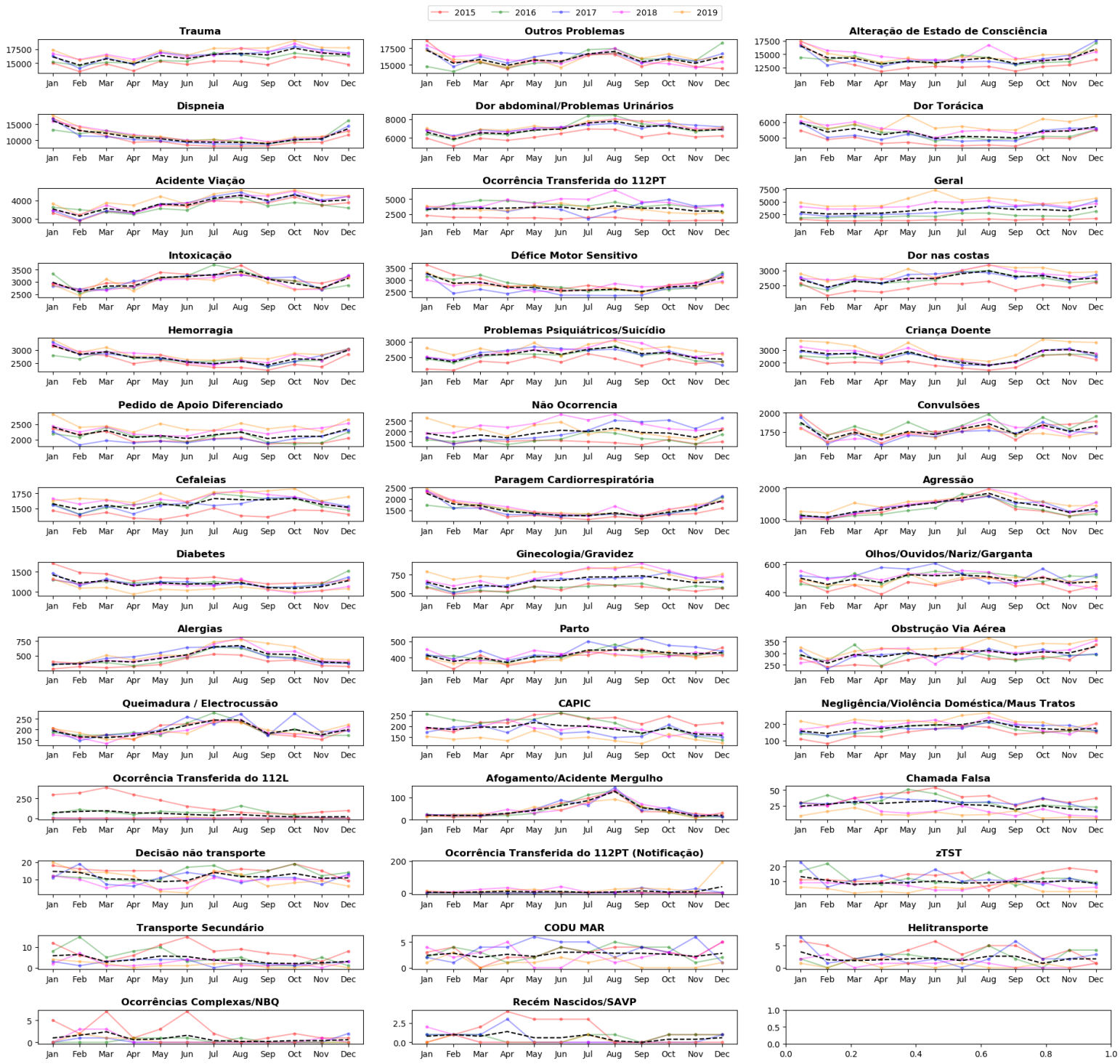


Figure 2: Initial Pathologies monthly.

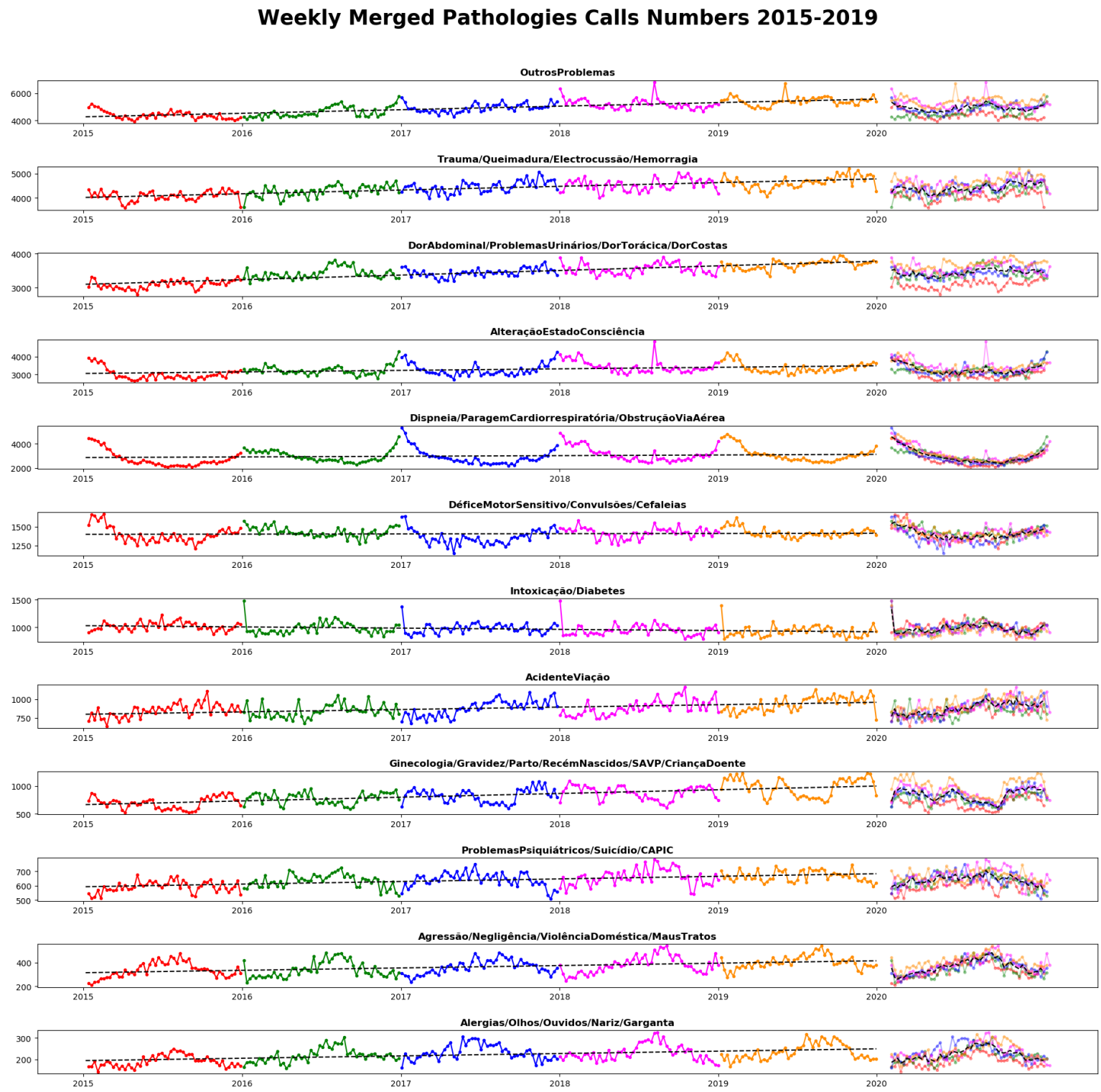


Figure 3: Merged Pathologies weekly.

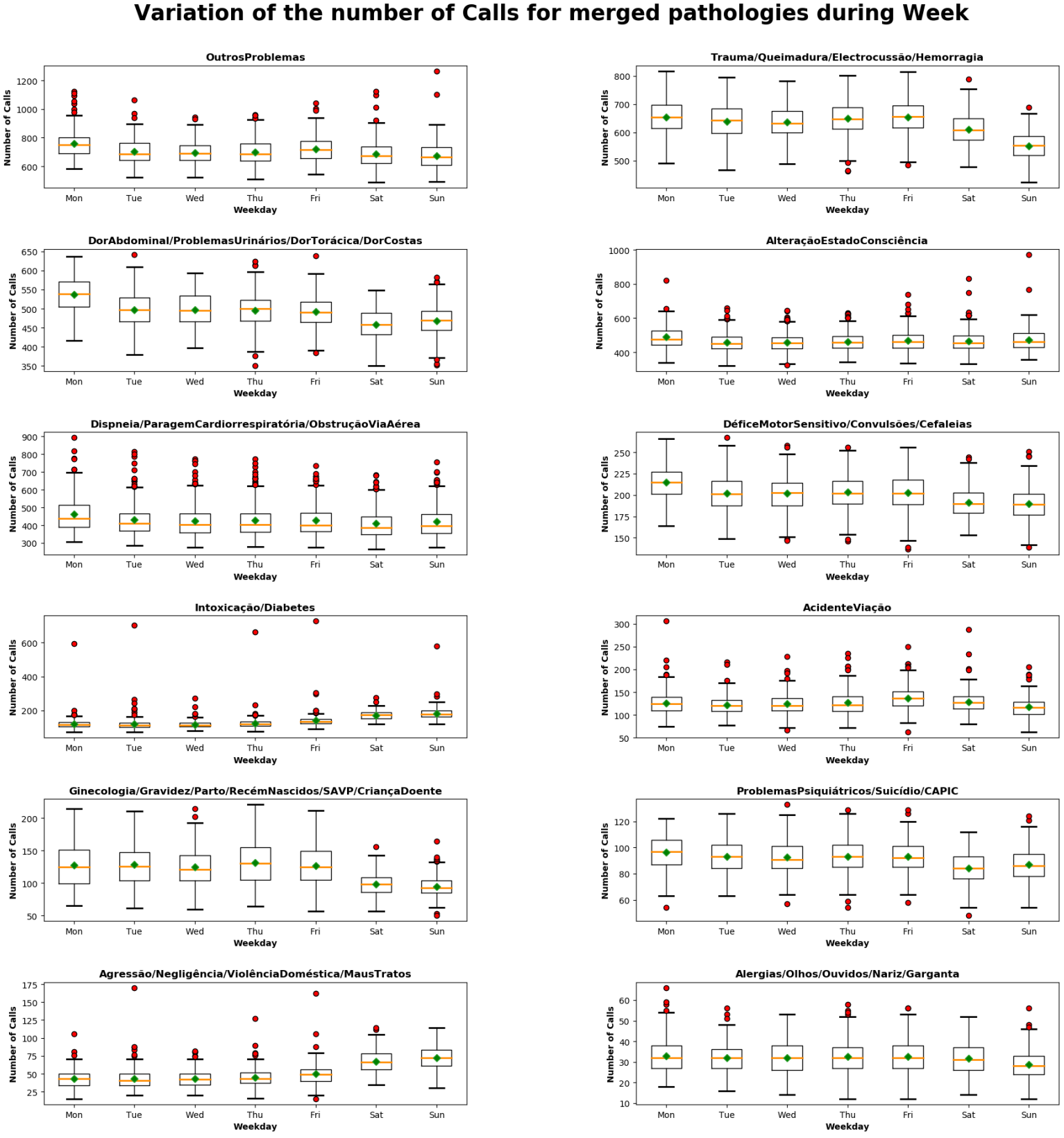


Figure 4: Boxplots Merged Pathologies weekday.

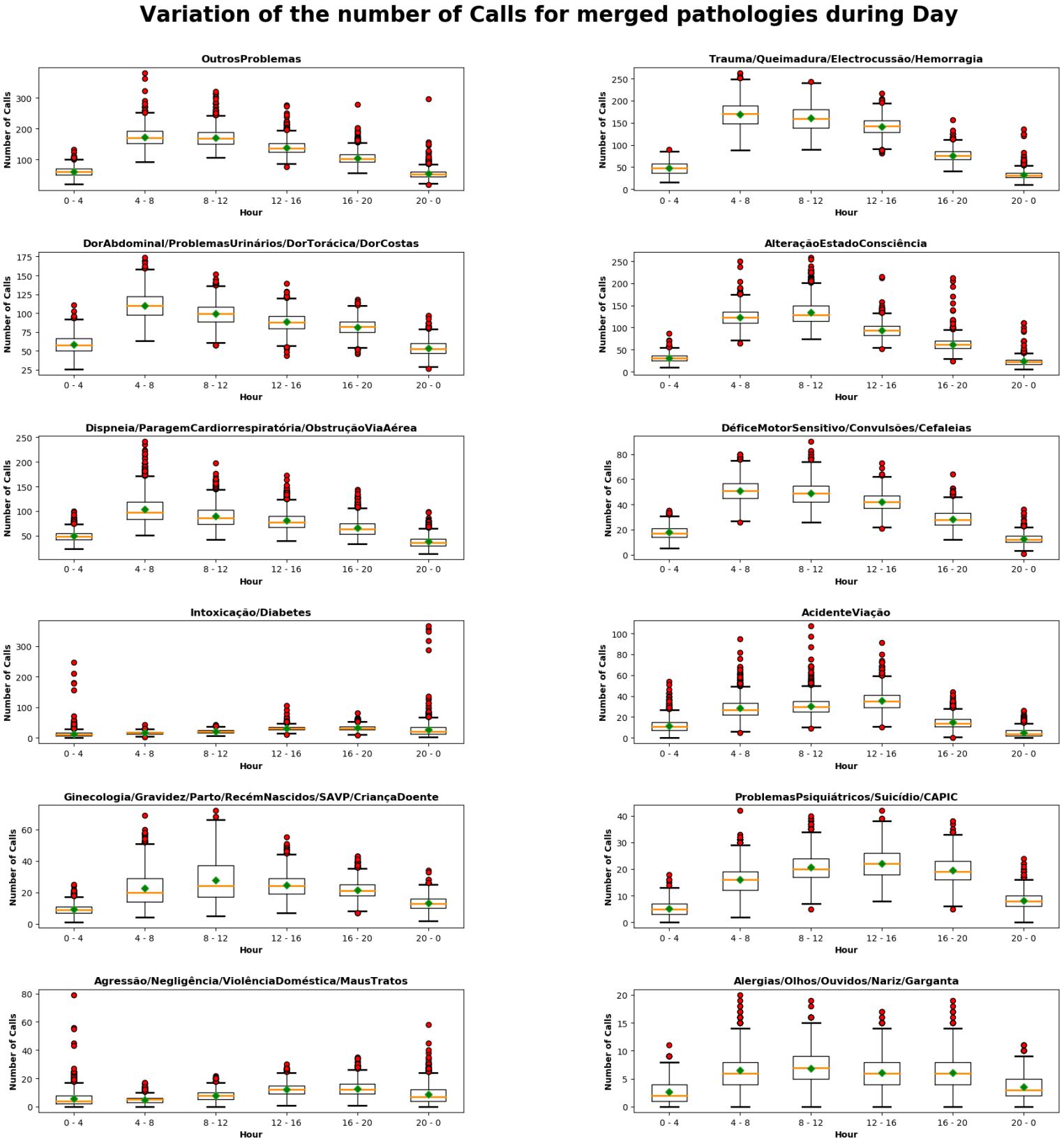


Figure 5: Boxplots Merged Pathologies period of the day.