Spatiotemporal patterns of emergency prevalence and response in Portugal

Francisco Faria Barata
Instituto Superior Técnico
Lisboa, Portugal
franciscofariabarata@tecnico.ulisboa.pt

***Abstract*—The national-wide record of medical emergencies, monitored by Instituto Nacional de Emergência Médica (INEM), shows that every year the number of emergency occurrences continues increasing. 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 at all times and resources always available to be allocated, this can be viewed as an optimization problem. Therefore, having more knowledge about the domain and how emergencies behave over time and space, can lead to a better preparation and consequently a better response in time. Given the importance of the addressed problem this research can be helpful in saving more lives. The core purpose of this thesis is to answer the following question: "Is location correlated with the features from the medical emergency data?". With the desire to answer this question the best possible way, we implemented partitioning and hierarchical clustering algorithms, using data aggregated by location. By using these two clustering algorithms we were able to verify if the elements nearby in location were also been aggregated within the same cluster. The clustering results showed that medical emergencies have a correlation with the location, in terms of the number of emergencies, type of emergencies as well as unit dispatch time.**

***Index Terms*—Spatiotemporal Data Mining, Medical Emergencies, Time Series, Clustering, Patterns**

I. Introduction

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

In continental Portugal, emergency medical services are coordinated by 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 diagnose, 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). We wanted to "From where this growing is coming? It is general or only some pathologies?" It is necessary a good optimization of our resources in order to guarantee efficient responses to emergencies. We conducted a exploration over the data to figure out from where the growing was coming This optimization passes trough study more fine granularities and understand their specific behaviour Currently, the allocation of resources and the forecasting of emergencies is done by relying on rules based on the experience of senior staff at INEM. For the proper allocation of resources, an important reality that needs to be accounted is population movements (e.g. summer months due to tourism) in certain areas.

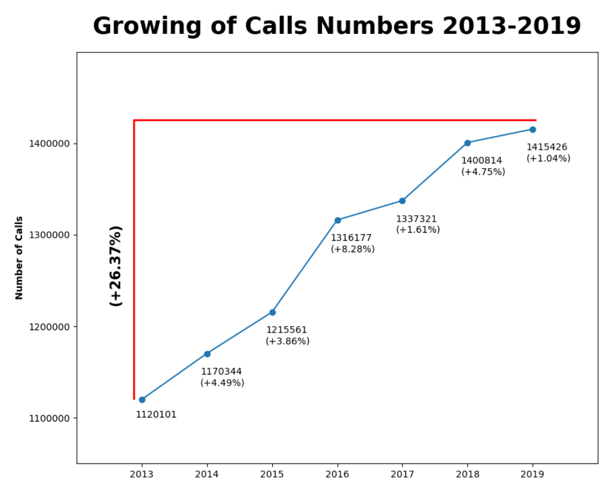

Fig. 1. Growing of emergency calls registered by INEM in the last years. The percentage below each point is the variation relatively to the year before.

Identify applicable funding agency here. If none, delete this.

A. Major Contributions

With this we contributed with:

1) conduct a comprehensive exploratory analysis of medical emergencies in continental Portugal over the past few years, with particular focus on: (1) the distribution of the prevalence of emergencies by type and severity, and (2) in the distribution of emergency response with a focus on service, dispatch and arrival times;
2) discover discriminative patterns of incidence and type of emergency by region;
3) incorporate meteorological and calendar information (events) in the discovery of patterns;
4) validate the discovered patterns according to their statistical significance, relevance and discriminative power;
5) develop associative forecasting models based on the patterns discovered;
6) extend the discovery of discriminative patterns to study the adequacy of emergency responses, focusing on the analysis of susceptibilities in the response.

II. Background

A. Time Series

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 [1].

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 addictive or multiplicative models, as explained by Brockwell et al. [3]. Classical approaches for time series analysis and forecasting generally rely on statistical principles, including *auto-regression*, *differencing* and *exponential smoothing operations* [6]. An example of a time series is presented in Figure 2, where is visible the upward trend in the number of emergency calls between 2013-2019.

Zhao and Bhowmick [8] 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 1.

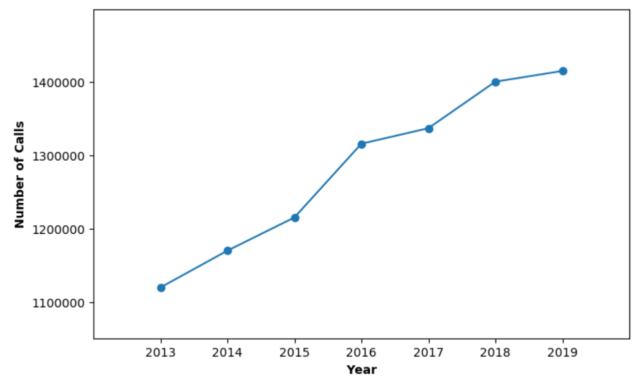

Fig. 2. Evolution of the emergency occurrences number in Portugal (continental territory) recorded by INEM between 2013-2019.

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 DTW and Fréchet distance are able to capture such forms of similarity among time series, as demonstrated by Atluri et al. [2]. 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 **??**, 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.

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

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.

B. Georeferenced Time Series

Time series describing evolving behaviour of one or more features recorded at fixed locations and uniform intervals are referred as *georeferenced* [7]. Neves et al. [5] formulates that GTS can be represented as a tuple $(\phi, \mathbf{x})$, where ϕ is a pair (*latitude*, *longitude*) describing the location in which the series $\mathbf{x}$ 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 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.

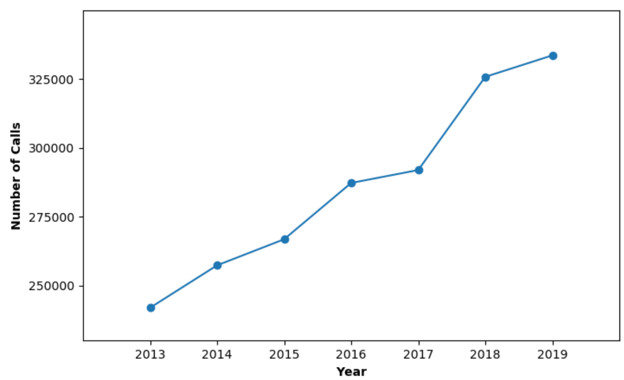

Fig. 3. Evolution of the emergency calls in Lisbon between 2013-2019.

C. Clustering Time Series

In the presence 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 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.

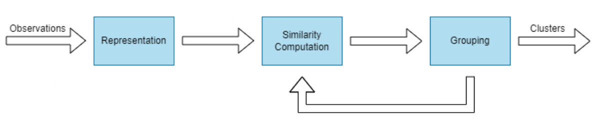

Fig. 4. Overview of the clustering process

As shown by Giorgino [4], 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 1.

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 [2] (e.g. districts that are connected with the same patterns).

III. Data Exploration

A. National Merged Pathologies Emergencies

At this phase it was 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.

The visualization that lead to a better understating of the pathologies domain was the computation of the collection of heatmaps presented in Figure 5. Each heatmap corresponds to a merged pathology and each cell contains the average amount of occurrences registered for that month of that year. Figure 5 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.

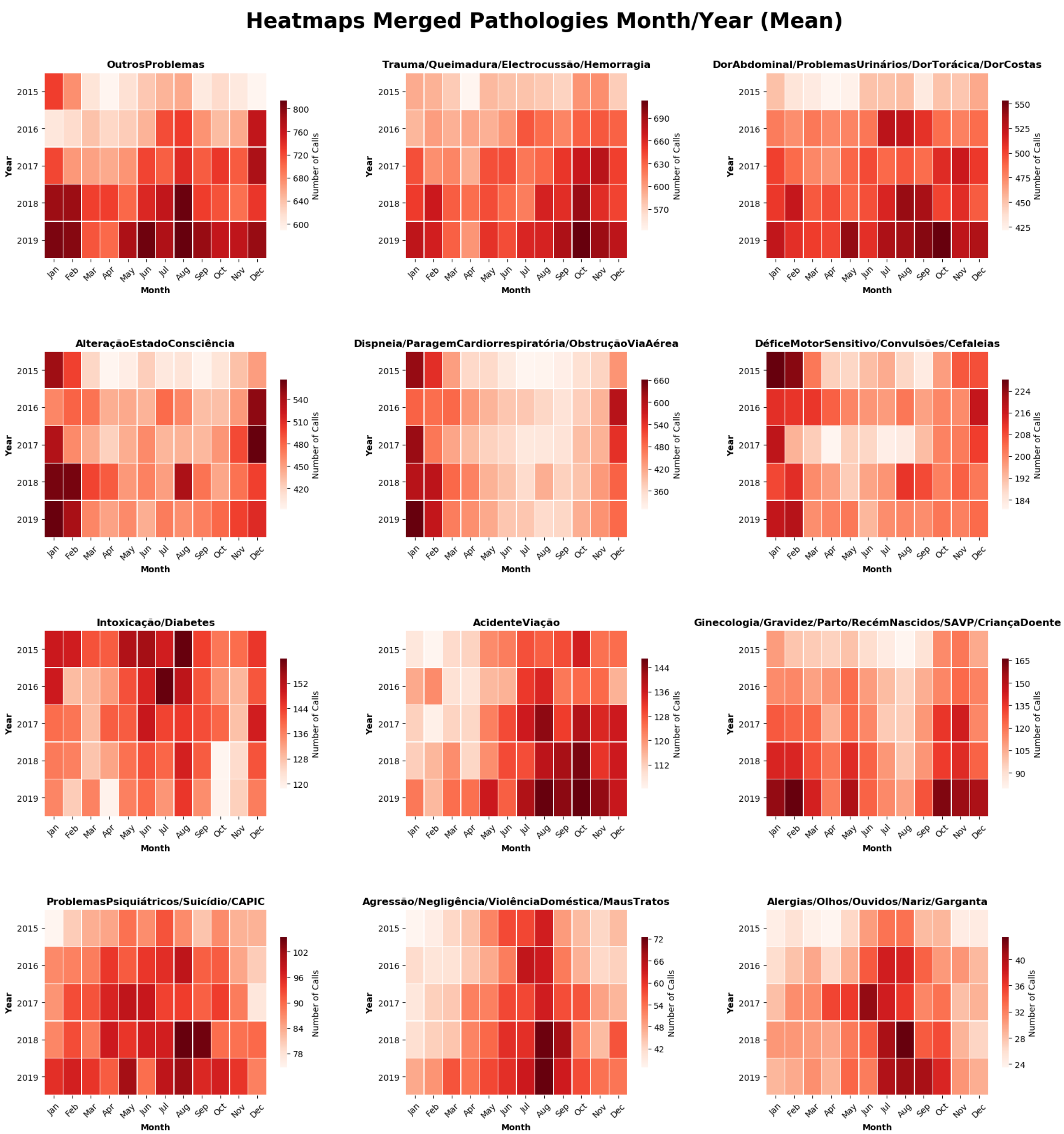

Fig. 5. Heatmap collection for the merged pathologies (each cell contains the average amount of occurrences registered for that month of that year)

IV. Clustering solution

With the focus on answering the question that gave origin to this thesis: "Have close locations in space a similar behaviour of number/type of emergencies?", the best solution discovered was to perform clustering methods over the occurrences agglomerated 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 IV-A). In our search for regions with similarities, we implemented two types of clustering algorithms (Partitioning and Hierarchical), as explained in detail in Section IV-B. In the end, we developed a geographic representation of Portugal to properly analyse the clusters formed by our models.

A. Preprocessing and GTS formation

To reduce the computation time in the clustering stage, we performed the time series generation *a priori*. We formed three 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 6.
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 7. Each element in this list is an integer value that corresponds to the number of occurrences for a specific merged pathology. The order of the pathologies within the list follows their prevalence, represented in Table **??**.

For each type of time series was created a folder that contained 18 text files (one for each district). Inside each text file were stored the values for the counties that belong to that district (one county by line). Figure 8 presents Lisbon district file example for the global series (6) and for the pathologies series (7)

B. 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 IV-B2

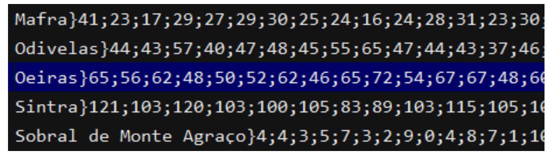
```
Mafra}41;23;17;29;27;29;30;25;24;16;24;28;31;23;30
Odivelas}44;43;57;40;47;48;45;55;65;47;44;43;37;46
Oeiras}65;56;62;48;50;52;62;46;65;72;54;67;67;48;6
Sintra}121;103;120;103;100;105;83;89;103;115;105;1
Sobral de Monte Agraço}4;4;3;5;7;3;2;9;0;4;8;7;1;1
```

Fig. 6. Lisbon district text file for global series

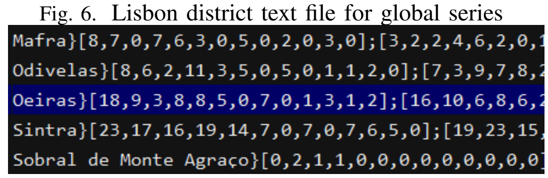
```
Mafra}[8,7,0,7,6,3,0,5,0,2,0,3,0];[3,2,2,4,6,2,0,
Odivelas}[8,6,2,11,3,5,0,5,0,1,1,2,0];[7,3,9,7,8,
Oeiras}[18,9,3,8,8,5,0,7,0,1,3,1,2];[16,10,6,8,6,
Sintra}[23,17,16,19,14,7,0,7,0,7,6,5,0];[19,23,15,
Sobral de Monte Agraço}[0,2,1,1,0,0,0,0,0,0,0,0,0]
```

Fig. 7. Lisbon district text file for pathologies series

Fig. 8. Text files created for Lisbon district

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.

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(2) and Silhouette(3), to measure the compactness and the separation of the clusters.

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

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

$$Silhouette = \frac{b - a}{max(a, b)}, \tag{3}$$

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.

2) Hierarchical clustering: Hierarchical algorithms group objects in a tree-like structure, called *dendrogram* (Figure 9). 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 9). To measure similarity between groups three methods were used:

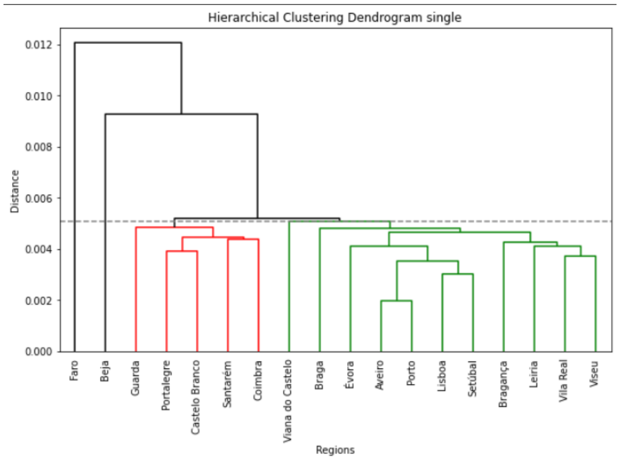

Fig. 9. Example of a dendogram from our project

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 were "connected":** we stored the real distance value between the time series.
2) **else:** we stored a high-value distance, forcing the regions to have more difficulty clustering.

V. Results

This Chapter exposes the discoveries found during our journey exploring spatiotemporal patterns using the data from our case of study. Dividing the occurrences data by location (district/county) in order to explore their relations between each other, always with the aim on finding if exists any correlation in space, lead to results that justify a different way of distribution of our national emergency resources. The main conclusion being observed was how the number of occurrences per capita presents the lowest values in the north region of Portugal and how this value starts increasing with moving south. 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) it was also computed the average values for Portugal (dashed line represented in clusters figures) to be visible which clusters are above and below the national average level.

The results and information obtained are organized by spatial granularity, district granularity in Section V-A and county granularity in Section V-B.

A. District Granularity

When clustering the global series with a district spatial granularity, using Spatial clustering with average linkage, it revealed a correlation between the district geographic location and the number of occurrences per capita, as verified in Figure 12. The districts from the north of Portugal have a smaller 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. 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 was using Spatial clustering with average linkage (Figure 15). 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 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.

B. County Granularity

For the global series, the best set of clusters was returned by the hierarchical clustering with average linkage. Figure 18 shows the results, we reached the following conclusions: the more density counties have lowers emergencies per capita (Porto and Lisboa counties belong to the red cluster) and counties with more occurrence 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. 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, justifying the spike observed; (2) Crato have a festival in August that match emerge of occurrences, and (3) Alcoutim was noticed by RTP as the more unpopulated and aged county of Portugal, what clarified this singularity.

With consideration to the activation time series, the algorithm that performed better outcome was the hierarchical with average linkage.

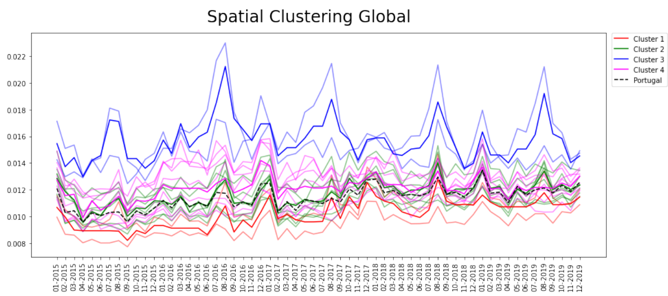

Fig. 10. Time series representation with cluster label

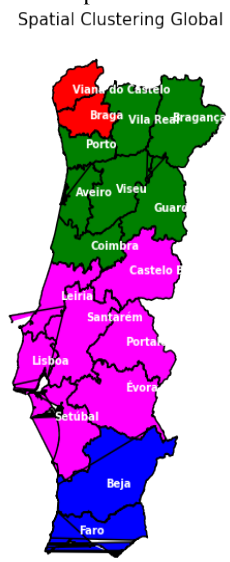

Fig. 11. Map divided by formed clusters

Fig. 12. Cluster results for global series with district and monthly granularities, using spatial algorithm with average linkage

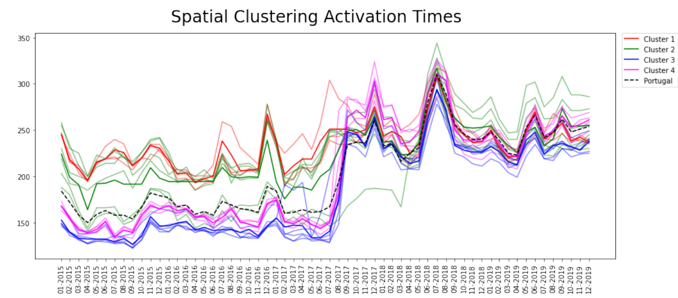

Fig. 13. Time series representation with cluster label

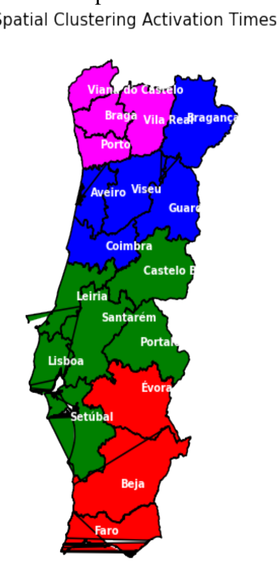

Fig. 14. Map divided by formed clusters

Fig. 15. Cluster results for activation time series with district and monthly granularities, using spatial algorithm with average linkage

VI. Conclusion

This thesis, within the context of project Data2Help, 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 give advantages on how we can adjust the available resources, 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 patterns of 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.

The answer for our question came with the clustering results for the GTS. 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, with basis on the cluster needs. There is another possibility to improve the response time with basis on this clusters formation is: find clusters with opposed behaviour in pathologies and with them find periods of temporary allocation of means.

The data exploration of the national data revealed several interesting trends and patterns for the pathologies and total number of emergencies. The most alarming number is the continuous increase in the number of emergencies, making projects like this relevant for the continuous research and discovery of new forms to improve the response time, saving more lives.

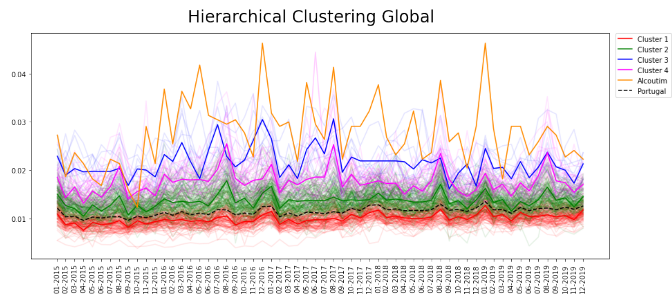

Fig. 16. Time series representation with cluster label

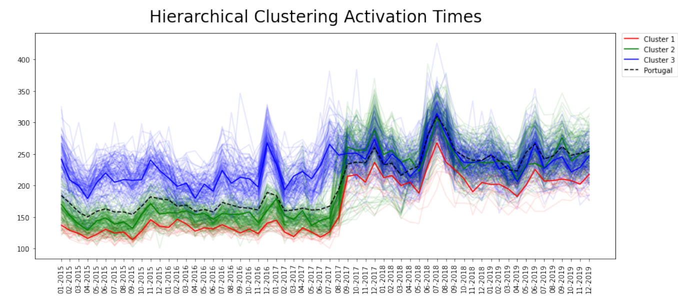

Fig. 19. a

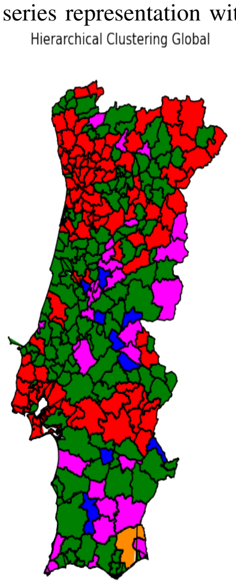

Fig. 17. Map divided by formed clusters

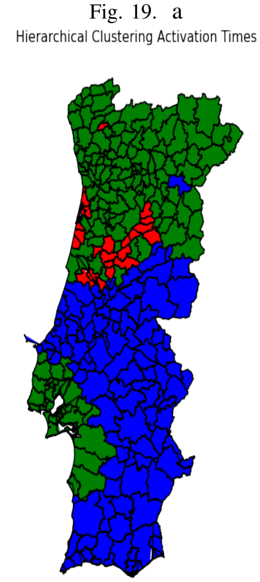

Fig. 20. b

Fig. 18. Cluster results for global series with county and monthly granularities, using hierarchical algorithm with average linkage

Fig. 21. Clusters results for activation time series with county and monthly granularities, using hierarchical algorithm with average linkage

References

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

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

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

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

[5] F. Neves, A. Finamore, and R. Henriques. 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)*, 2020.

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

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

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