What Do the Data Say in 10 Years of Pneumonia Victims? A Geo-Spatial Data Analytics Perspective

Maribel Yasmina Santos[1(✉)], António Carvalheira Santos[2], and Artur Teles de Araújo[2]

[1] ALGORITMI Research Centre, University of Minho, Guimarães, Portugal
maribel@dsi.uminho.pt
[2] Portuguese Lung Foundation, Lisboa, Portugal
antonio.carvalheira@gmail.com, artur@telesdearaujo.com

Abstract. The need to integrate, store, process and analyse data is continuously growing as information technologies facilitate the collection of vast amounts of data. These data can be in different repositories, have different data formats and present data quality issues, requiring the adoption of appropriate strategies for data cleaning, integration and storage. After that, suitable data analytics and visualization mechanisms can be used for the analysis of the available data and for the identification of relevant knowledge that support the decision-making process. This paper presents a data analytics perspective over 10 years of pneumonia incidence in Portugal, pointing the evolution and characterization of the mortal victims of this disease. The available data about the individuals was complemented with statistical data of the country, in order to characterize the overall incidence of this disease, following a spatial analysis and visualization perspective that is supported by several analytical dashboards.

Keywords: Business intelligence · (Spatial) data warehouse · Data analytics · Pneumonia

1 Introduction

Business intelligence and analytics have become increasingly relevant over the past two decades, reflecting the magnitude and impact of data-related problems [1]. This is a field of knowledge that has been using data warehouses as data repositories, providing an integrated and homogeneous set of data used in analytical contexts to support the decision making process [2]. A data warehouse can then be analysed using different supporting technologies as on-line analytical processing [3] or data mining algorithms [4], among others.

When data includes spatial attributes, like locations, the data model of a data warehouse can include spatial dimensions or attributes, allowing the analysis of the available data under this spatial perspective. Data warehouses with spatial characteristics have also become a topic of growing interest in recent years [5], being their logical design based on the multidimensional model, providing support for the definition of spatial data

M.E. Renda et al. (Eds.): ITBAM 2016, LNCS 9832, pp. 3–21, 2016.
DOI: 10.1007/978-3-319-43949-5_1

dimensions and/or spatial measures. Dimensions represent the analysis axes, while measures are the variables being analysed against the different dimensions. The implementation of spatial On-Line Analytical Processing (OLAP) tools can be achieved through solutions that are OLAP dominant, Geographical Information Systems (GIS) dominant, or both in a mixed solution [6]. Those tools are powerful decision-making instruments as they allow users to explore and analyse data in user-friendly applications and to formulate *ah-doc* queries on these data.

This paper presents a data analytics perspective using the data available in a data warehouse, with spatial characteristics, integrating data related with the incidence of pneumonia in Portugal, from 2002 to 2011, integrating 369 160 records. Besides these data, with the characterization of the affected individuals and other related pathologies, it was possible to integrate statistical data collected in the last census exercise undertaken in Portugal in 2011 [7].

The work here presented shows how several dashboards with spatial data, implemented over the mentioned data warehouse, were used in a data-driven analytical approach for an interactive analysis of the data, highlighting valuable information to characterize the incidence of a disease that, for respiratory infections, is the leading cause of death and hospital admissions in Portugal [8], following a global trend, as stated by the World Health Organization, mentioning that the lower respiratory infections are among the 10 leading causes of death at a Mundial level [9].

This paper is organized as follows. Section 2 presents related work. Section 3 summarizes the adopted methodology. Section 4 describes the data available for analysis. Section 5 summarizes some of the main findings in understanding pneumonia fatalities. Section 6 concludes with some remarks about the described work and guidelines for future work.

2 Related Work

Several works in the literature show the analysis of data about respiratory diseases, and some of them about pneumonia, following data analysis strategies that try to point out tendencies, patterns or models that can be useful in the decision-making process. Some of these works use statistical approaches, or techniques usually used in business intelligence contexts like OLAP or data mining. Although with relevant contributions to the community, none of these works was able to integrate such vast volume of data, providing a comprehensive knowledge about the incidence of this disease and, more important, its fatalities. This is of upmost importance for decision-makers in the definition of adequate actions to fight this disease.

The work of [10] presents a descriptive analysis of data retrieved from the medical reports at the Tawau General Hospital in Malaysia, where patients filled a special form that required information such as the patient age, area of origin, parent's smoking background, parent's medical background (if known), patient medical background (if known), among other relevant information. The performed analyses identified the profile of the patients who were admitted to this hospital. The authors report that there are several factors that may have caused the pneumonia, such as family background, or genetic and environmental factors, alerting the government authorities and doctors for

the need of taking appropriate actions. In total, data from 102 patients were used in this study. As main results, the authors point that 86.27 % of the patients are from rural areas, underlining poor hygiene as an important factor in the origin of pneumonia in Malaysia.

With a higher number of studied individuals, the work of [11] reported that pneumonia is a disease most often fatal, which can be acquired by patients during their stay in intensive care units. Data from patients admitted to the intensive care unit at the Friedrich Schiller University Jena were collected and stored in a real-time database, totalizing 11 726 cases in two years. Based on these, the authors developed an early warning system for the onset of pneumonia that combines Alternating Decision Trees for supervised learning and Sequential Pattern Mining. The implemented detection system estimates a prognosis of pneumonia every 12 h for each patient. In case of a positive prognosis, an alert is generated. In this case, data mining algorithms, one of the data analysis techniques used by business intelligence systems, showed to be useful in the analysis of the collected data.

In [12], the authors show a study that allowed the development and validation of an ALI (Acute Lung Injury) prediction score in a population-based sample of patients at risk. For the prediction score, the authors used a logistic regression analysis. Patients at risk of acquiring an acute respiratory distress syndrome, the most severe form of ALI, were first identified in an electronic alert system that uses a Microsoft SQL-based database and a data mart for storing data about patients in an intensive care unit. A total of 876 records were analyzed, divided in 409 patients for the retrospective derivation cohort and 467 for the validation cohort.

More recently, [13] proposed the use of Disjunctive Normal Forms for predicting hospital and 90-day mortality from instance-based patient data, comprising demographic, genetic, and physiologic information in a cohort of patients admitted with severe acquired pneumonia. The authors developed two algorithms for learning Disjunctive Normal Forms, which make available a set of rules that map data to the outcome of interest. The authors show that Disjunctive Normal Forms achieve higher prediction performance quality when compared to a set of state-of-the-art machine learning models. Regarding data, patients with community-acquired pneumonia, a common cause of sepsis, were recruited as part of a study conducted in the United States (Western Pennsylvania, Connecticut, Michigan, and Tennessee) between November 2001−November 2003. Eligible subjects had 18 or more years old and had a clinical and radiologic diagnosis of pneumonia. Among the 2 320 patients enrolled, the authors restricted their analysis to 1 815 individuals admitted to the hospital.

3 Methodological Approach

The analysis of vast amounts of data with the aim of identifying useful patterns or insights can be achieved following an exploratory data analysis approach, which aims identifying relationships between different variables that seem interesting, checking if there is any evidence for or against a stating hypothesis [14]. In this process, it is very important looking for problems in the available data, as well as identifying complementary data that could add value to the data under analysis. In this sense, exploratory

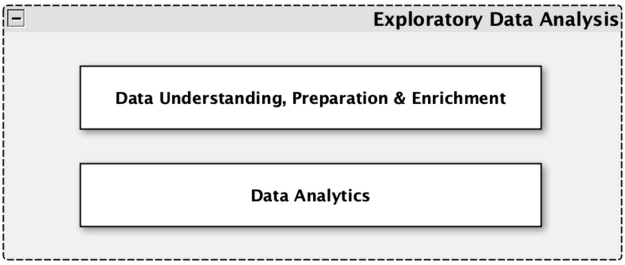


Fig. 1. Exploratory data analysis (different roles)

data analysis is useful in a preliminary analysis of the data, in order to understand, prepare and enrich it, and later, for the analysis itself in the data analytics approach, supporting the decision making process (Fig. 1).

Starting with the data understanding, preparation and enrichment, this allows the enhancement of a data set for data analysis purposes. In our previous work [7], it was possible to do an extensive analysis of the data, in order to get a deep knowledge about it, analyzing the available attributes, verifying all possible values, identifying data quality problems, enriching the data with external data sources, modeling the analytical repository for storing the data for analysis and, finally, implementing that repository. All these stages iteratively add value to the initial collected data, either cleaning the data (removing errors or problems) or completing it with additional sources (sometimes external to the organizations). For the concretization of such an analytical data repository, Fig. 2 summarizes the main followed steps, some of them possible through exploratory data analysis.

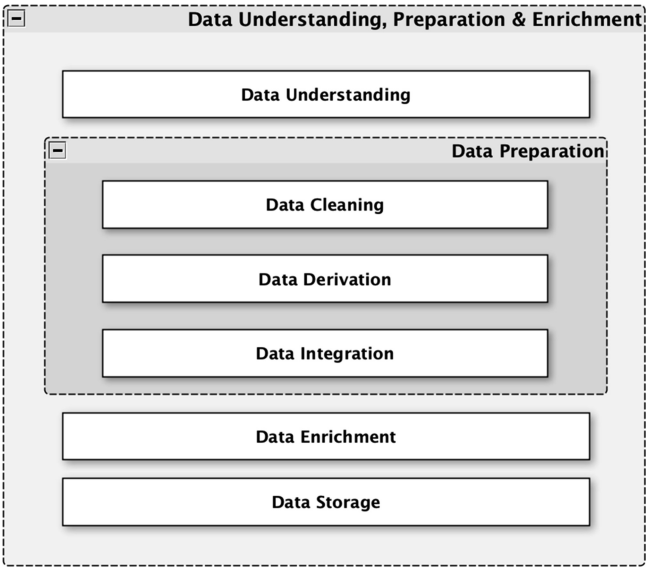


Fig. 2. Steps in the data understanding, preparation and enrichment

After the understanding, preparation and cleaning of the data, exploratory data analysis can be used for data analytics, making use of tables or specific charts or graphs to obtain useful insights on data. In this task, the user/researcher must do critical evaluations of the findings, identifying interesting paths for analysis and, also, those that do not worth pursuing, as data are not providing useful or enough evidence of results [14]. The overall goal is to show the data, summarizing the relevant evidences and identifying interesting patterns.

For data analytics with exploratory data analysis, this work makes use of analytical graphics (in this case with a geo-spatial focus), trying to make informative and useful data graphics [15, 16]. For Tufte [15], excellent graphics exemplify the deep fundamental principles of analytical design in action, mentioning 6 fundamental principles of the analytical design: 1. Show comparisons, contrasts, differences; 2. Causality, mechanism, structure, explanation; 3. Multivariate analysis; 4. Integration of evidence; 5. Documentation; and, 6. Content counts most of all (Fig. 3).

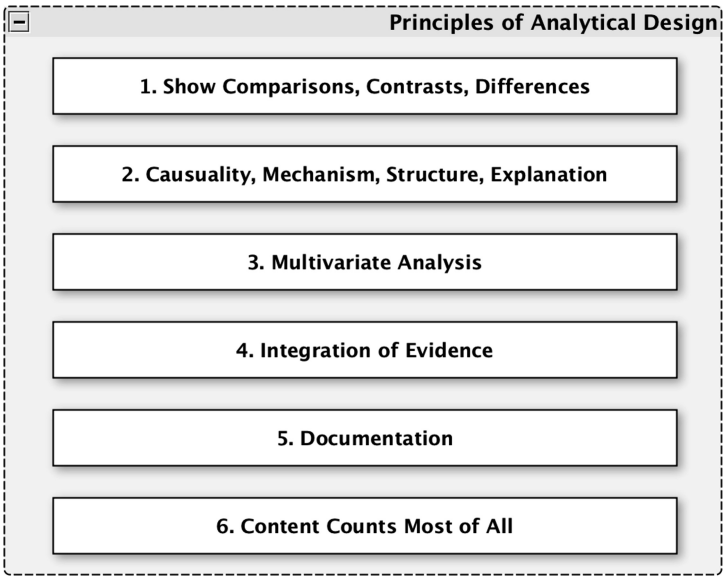


Fig. 3. Principles for analytical design (Source: [15])

Going through these principles, showing comparison is considered the basis of all scientific investigation, as showing evidence for a hypothesis is always relative to another competing hypothesis. Also, it is useful to show the causal framework when thinking about a question, meaning that data graphics could include information about possible causes, useful in suggesting hypotheses or refuting them. The most important is that this will raise new questions that can be followed up with new data analyses, which should be multivariate, as usually there are many attributes that can be measured or analyzed. Data graphics should attempt to show this information as much as

possible, rather than reducing things down to one or two features. In those data graphics, numbers, words, images and diagrams can be included to tell a story, making use of many modes of data presentation and integrating as much evidence as possible. When describing and documenting the evidences, data graphics must be properly documented with labels, scales and sources, telling a completely story by itself, avoiding the need for extra texts or descriptions for interpreting a plot. For presenting the results, the content includes a good question, the approach for addressing it and the information that is necessary for answering that question [14].

All these principles of analytical design when included in data analytics through exploratory data analysis give support to the Data Analytics Cycle followed in this work, in which a question starts the cycle, being followed by data exploration. The analysis of results looks into the obtained findings in order to identify new questions or analytical paths for data analysis (Fig. 4).

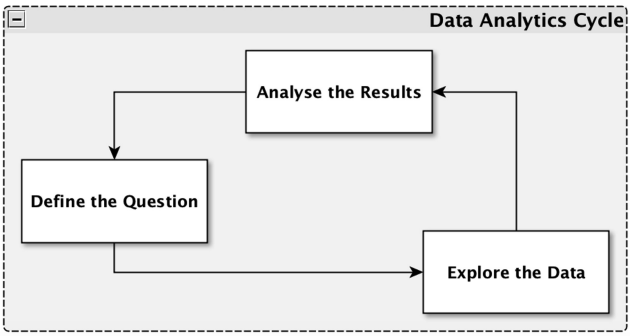


Fig. 4. Data analytics cycle

4 Overview of the Available Data

In this work, data from 10 years of incidence and victims of pneumonia were used, selected from a data warehouse that includes 369 169 records of individuals that had pneumonia, from 2002 to 2011, in continental Portugal. This extensive set of data was extracted from the HDGs database (*Homogeneous Diagnosis Groups*) of the Central Administration of Health Services - ACSS (*Administração Central dos Serviços de Saúde*). All the data, after an extensive work of extraction, transformation and loading, was stored in an analytical data repository now used for data analytics [7]. Besides the information of the individuals and their characteristics, this analytical repository also includes statistical data collected in the latest census exercise carried out in Portugal, in 2011 [17]. This will allow the verification of the most affected regions, regarding the number of mortal victims and the living population.

In our previous works [7, 18], the available data was analysed to characterize the disease and its evolution along the years. It was possible to verify that the consequences of the disease change depending on the age of the patients that are affected, on their

Table 1. Data attributes for analysis

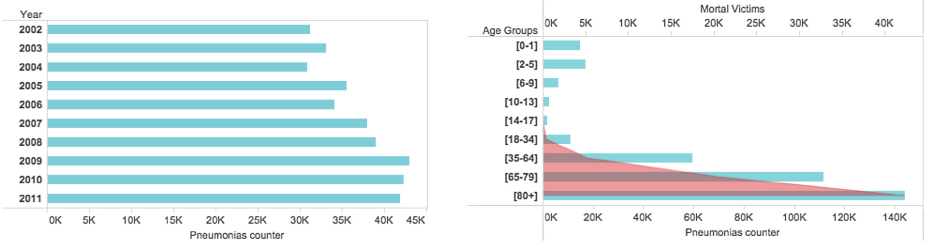
Attribute	Description	Type	Values
Admission days	Total number of days in a healthcare facility	Integer	Min: 0, Max: 1032, Median: 8, Standard deviation: 11.7
Admission days class	Classes for the number of days in a healthcare facility	Categorical	[0–3], [4–6], [7–10], [11–29], [30+]
Age	Age of the patient	Integer	Min: 0, Max: 111, Median: 76, Standard deviation: 26.9
Age groups	Classes for the age of the patient	Categorical	[0–1], [2–5], [6–9], [10–13], [14–17], [18–34], [35–64], [65–79], [80+]
District	District of the patient	Categorical	18 Districts (Continental Portugal): Aveiro, Braga, Porto, Lisboa, Coimbra,…
Gender	Gender of the patient	Categorical	F (Female), M (Male)
Longitude	Longitude coordinate	Numeric	Min: –9.462, Max: –6.210
Latitude	Latitude coordinate	Numeric	Min: 37.000, Max: 42.140
Mortal victim	Flag that states if the patient was, or not, a mortal victim	Binary	0: Not a mortal victim 1: Mortal victim
Municipality	Municipality of the patient	Categorical	279 Municipalities of Continental Portugal
Number of residents	Number of residents in a given parish	Integer	Min: 31, Max: 66 250, Median: 820, Standard Deviation: 5 083
Parish	Parish of the patient	Categorical	3445 Parishes of Continental Portugal
Pneumonias counter	Event-tracking measure to summarize data	Integer	1
Readmissions number	Number of readmissions in a healthcare facility	Integer	Min: 0, Max: 13, Median: 0, Standard deviation: 0.63
Year	Year of the admission/visit to the healthcare facility	Integer	[2002–2011]

physical condition, as well as other pathologies that may affect the course of the disease. These studies have shown that the number of cases of pneumonia has increased 33.9 % in the decade under analysis and that the number of fatalities increased at a higher rate, reaching 65.3 % from 2002 to 2011 [7]. Moreover, it was possible to verify that a significant number of patients that died, as consequence of this disease, had a very short admission in the hospital, in terms of staying there for treatment. Regarding related pathologies, some patients with pneumonia also presented other diseases like the chronic pulmonary disease, the chronic cardiac disease, the chronic renal disease,

the chronic pancreatic disease, the chronic hepatic disease, and the diabetes mellitus disease [18].

Having this preliminary knowledge about the incidence of the disease, this paper follows a data-driven analytics approach for a deepest analysis of a subset of the available data, trying to understand the course of the disease, in terms of fatalities, focusing in its geo-spatial incidence and in the identification of the more affected regions, considering several dimensions of analysis. With regard to location, it is important to mention that due to privacy concerns, the location where the patients' live/lived is associated with the centroids of the corresponding parishes and not to a specific street, for instance. To allow the proper visualization of the available information on a map, the centroids' coordinates were shacked in order to slightly distribute them in a map, around the corresponding parishes, showing the number of patients in each location. For the study presented in this paper, the relevant data attributes for analysis are summarized in Table 1, presenting the attribute name, description, type, and its possible values.

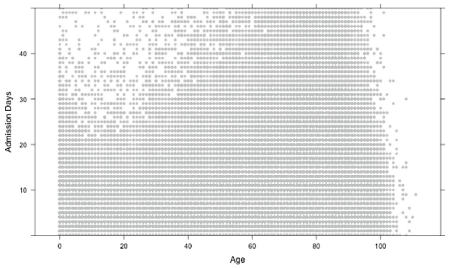
Before proceeding with the data analytics approach, let us briefly explore the available data in order to provide some background knowledge about the phenomena under analysis. Figure 5 shows two distribution graphs with the number of cases of pneumonia by year (Fig. 5(a)), and the number of cases by age (Fig. 5(b)). In the first case, it is possible to verify the increase that the disease has presented along these ten years. In the second, the incidence of cases increases substantially after the sixties, reaching the highest value in patients in the eighties. Also, as shown in the red area of Fig. 5(b), the number of mortal victims increases with age. Regarding the classes for the age, this is the first time that these specific ranges are used and the aim is to provide a deeper insight in these several groups.



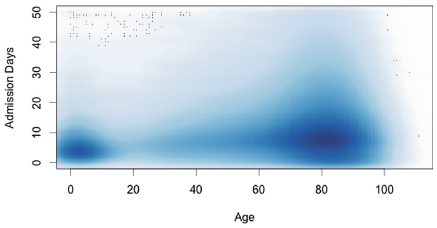
a) Distribution of Cases by Year b) Distribution of Cases by Age

Fig. 5. Number of cases by year and age (Color figure online)

Patients with pneumonia can have shorter or longer stays in the healthcare facilities for treatment. In many cases, severe conditions require longer stays or, in some cases, very short stays are verified when the patients died because it was too late for treatment, for instance. As we can see in Fig. 6(a), very long stays, superior to 30 days, are mainly associated to individuals with more than forty years old, while shorter stays can be verified in all ages. This is better seen in the graph of Fig. 6(b), which depicts a smoothed colour density representation of a scatterplot, obtained through a kernel density estimate [19].



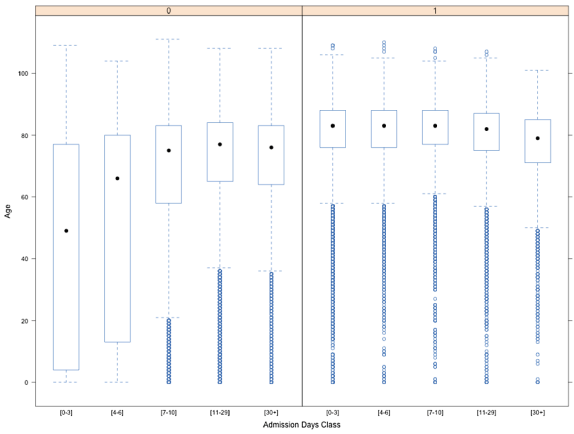
a) Plot of Age and Admission Days



b) Smooth Graph

Fig. 6. Analysis of ages and number of admission days

When we look into the relation between the age of the patients, the classes that were created for the number of days in the hospital, and if the patient is, or not, a mortal victim, the pattern previously mentioned emerges even stronger. For those that died as consequence of the disease, flag mortal victim equal to 1 in the right part of Fig. 7(a), the patients had an average age of approximately eighty years old, being this value very homogeneous for all the classes of admission days. In the case of patients that were not mortal victims, flag mortal victim equal to 0 in the left part of Fig. 7(a), stays in the healthcare facilities tend to be longer as age increases. The information obtained from Fig. 7(b) is very relevant as shows that, for a significant number of mortal victims, shorter stays in the hospital were verified, meaning that for many of these patients it was too late for treatment. Given the spatial component of the used analytical data model, it is now possible to characterize where theses patients lived and the regions that are more affected by this disease.



a) Ages and Admission Days

Age Groups	Admission Days (Group) [0-3]	[4-6]	[7-10]	[11-29]	[30+]
[0-1]	17	7	9	15	6
[2-5]	16	11	2	9	
[6-9]	15	3	5	9	3
[10-13]	4	3	8	3	1
[14-17]	25	3	13	5	7
[18-34]	179	60	51	106	50
[35-64]	1 915	846	637	1 259	587
[65-79]	6 948	3 612	2 979	5 203	1 763
[80+]	15 557	8 237	6 672	9 684	2 188

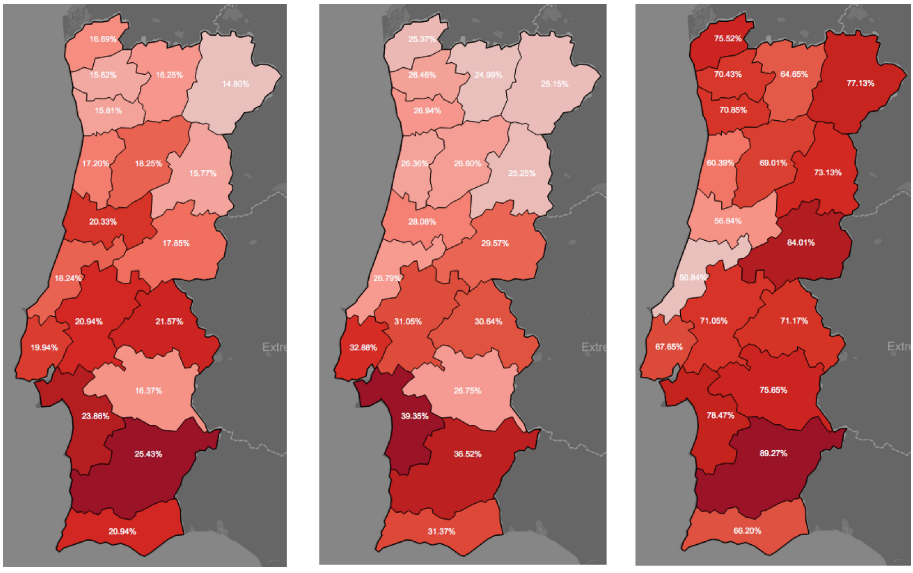
b) Ages and Admission Days for Mortal Victims

Fig. 7. Relation between ages and number of days in the healthcare facility

Before proceeding with the data analytics study, and for a technological characterization of the used tools, it is worth mentioning that all the dashboards presented in the following section were implemented using Tableau [20], while the graphs presented in this section were implemented using Tableau or R [19].

5 Geo-Spatial Characterization of Pneumonia Victims

Given the context of the previous section, the number of fatalities, its increase all over the years, and the fact that this disease seriously affects particular groups of people, this section provides a geo-spatial characterization of these victims, trying to understand this phenomena, knowledge that is essential for the appropriate definition of actions to fight it. As shown in Fig. 8(a), with the overall percentage of victims attending to the number of cases, the *Beja* district stands out with an average of 25.43 % of victims. In general, the South and the interior part of the country are more affected by this disease. If we restrict the data to those individuals with 80 or more years old (Fig. 8(b)), the difference between North and South is even more noticeable, but now with the district of *Setúbal* being more affected, with an average fatality rate of 39.35 %. If we continue filtering data to consider now those victims with 80 or more years old and with very short stays in the hospital ([0–3]), we can see that the percentage increases in all cases, with an overall percentage of victims that is very high, reaching almost 90 % in districts like *Beja* (89.27 %) or *Guarda* (84.01 %).



a) Overall Percentage b) Ages: [80+] c) Ages: [80+], Stays: [0-3]

Fig. 8. Percentage of mortal victims

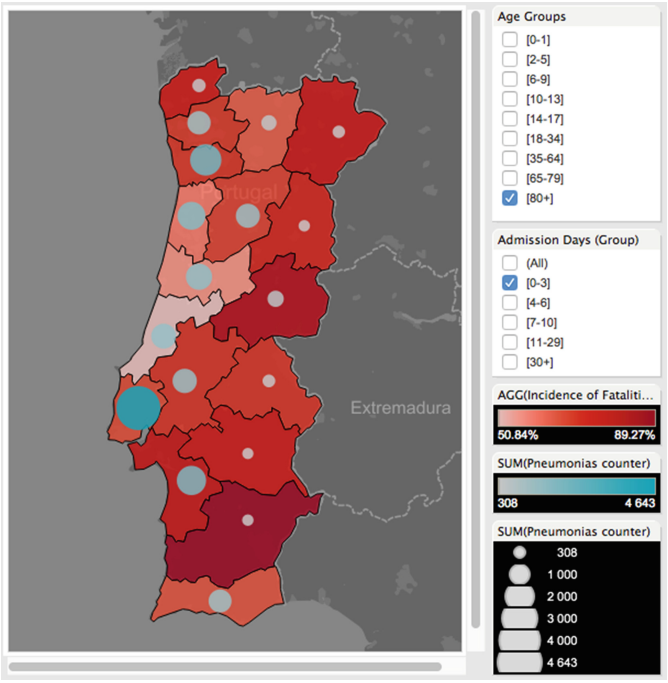


Fig. 9. Number of cases and percentage of victims ([80+], [0–3])

It is also important to stress that this behaviour is not only associated to these individuals, 80 or more years, as for the age class of [65–79], although with a smaller incidence, *Beja* presents, for example, a percentage of victims of 70.67 %. This is even more relevant if we consider that, for these regions, usually few cases of pneumonia are verified, although it seems that more severe. Considering the age class of 80 or more years old, the more affected one, Fig. 9 shows a dashboard applying a filter to this age class ([80+]), and to the shorter stays ([0–3]), and, as can be seen, more cases of pneumonia are verified in the metropolitan areas of *Lisboa* and *Porto*, but with a percentage of victims of 67.65 % for 4 643/3 141 cases of pneumonia/victims and 70.81 % for 2 364/1 675 cases of pneumonia/victims, respectively, contrasting with *Beja* and its 89.27 % for 317/283 cases of pneumonia/victims.

Looking to the particular case of *Beja*, it is now needed to drill-down and see what is the scenario inside the district. For that, the analysis of the several municipalities and parishes is useful, obtaining a higher detail in the geo-spatial characterization.

Figure 10 depicts the indicators under analysis for the municipality of *Beja* and an interesting pattern emerges. Six of the municipalities present 100 % of victims ([80+] for ages and [0–3] for stays) and all are located in the interior of the district. In this figure, the percentage of incidence of victims ranges from 73.33 % to 100 %, while the number of cases by municipality ranges from 2 to 75.

The analysis of this percentage, district by district, allowed the verification that different districts present different geo-spatial incidences, either with higher mortality to

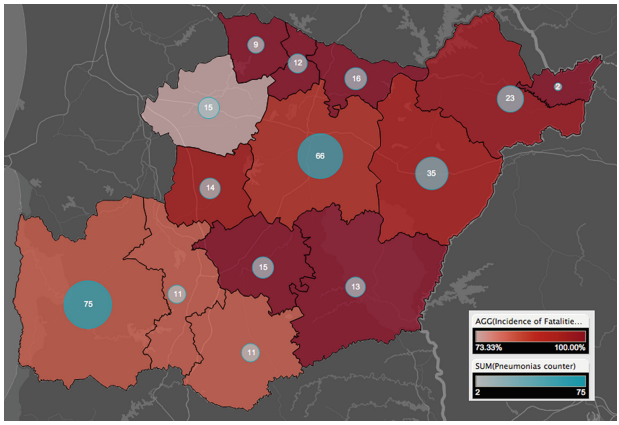
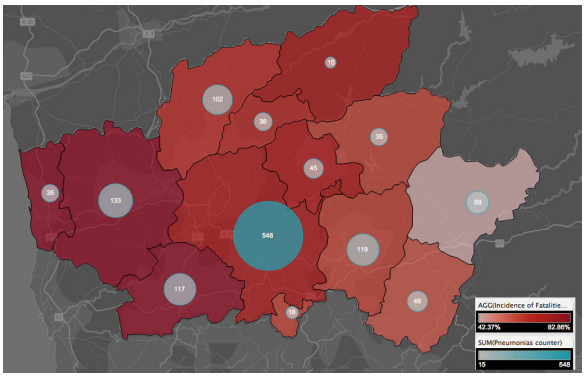


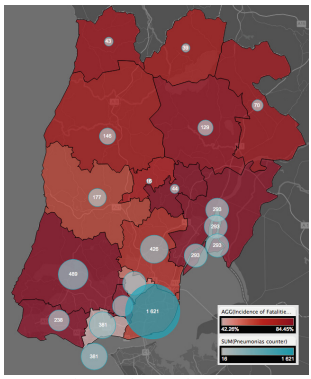
Fig. 10. Number of cases and percentage of victims for *Beja* ([80+], [0–3])

the interior of the country, like *Beja* (Fig. 10), to the littoral, like *Braga* (Fig. 11(a)), or with an undifferentiated pattern, like *Lisboa* (Fig. 11(b)).

Having all regions individuals with 80 or more years old, it is now important to verify why the percentage of victims is so different from one district to another. Figure 12(a) presents a map of *Beja* with a red circle marking each victim in the age class of [80+]. The colour of the circle is indexed to the age of the victim. As darker the circle, as older the victim, ranging ages from 80 to 101 years old. In this case, it seems that the municipalities with higher rates of mortality are the ones with eldest people, although no strong correlation was found between these two metrics. Figure 12(b) presents the values of the median and average for age in each municipality of *Beja* and the average value for the percentage of mortality. As can be seen, the difference between genders is relevant, being male in general affected sooner that female. This trend was verified in all the 18 districts of continental Portugal.

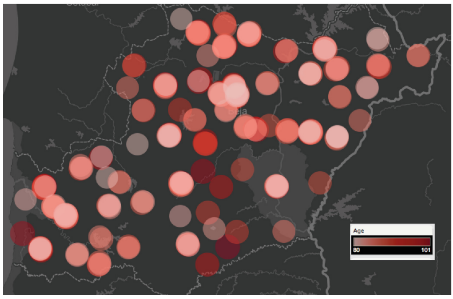


a) *Braga* District



b) *Lisboa* District

Fig. 11. Number of cases and percentage of victims for other districts ([80+], [0–3])



a) Location of the Victims

	Gender					
	F			M		
Municipality	Median Age	Avg. Age	Incidence of Fatalities..	Median Age	Avg. Age	Incidence of Fatalities..
Aljustrel	88.00	87.67	100.00%	83.00	83.63	87.50%
Almodôvar	83.50	84.67	83.33%	89.00	86.60	80.00%
Alvito	91.00	90.33	100.00%	86.00	85.83	100.00%
Barrancos	85.50	85.50	100.00%			
Beja	90.00	89.03	82.76%	87.00	86.46	91.89%
Castro Verde	87.00	88.88	100.00%	87.00	87.00	100.00%
Cuba	85.50	86.63	100.00%	85.50	87.00	100.00%
Ferreira do Alentejo	87.00	87.33	77.78%	86.00	88.33	66.67%
Mértola	87.00	88.00	100.00%	86.00	87.00	100.00%
Moura	85.00	86.57	92.86%	84.00	85.00	100.00%
Odemira	86.00	88.34	72.41%	84.50	86.07	89.13%
Ourique	85.00	85.00	80.00%	84.00	84.67	83.33%
Serpa	86.00	87.00	81.25%	85.00	85.05	100.00%
Vidigueira	86.00	85.67	100.00%	86.00	87.57	100.00%
Grand Total	86.00	87.59	86.00%	85.00	86.07	92.22%

b) Age and Percentage of Mortality

Fig. 12. Spatial distribution of the victims in *Beja* ([80+], [0–3]) (Color figure online)

In general, and taking as an example the three districts more detailed until now, we can look into the number of readmissions each patient had (Fig. 13). Considering all patients, all ages and limiting the analysis to the shorter stays ([0–3] days), in general *Beja* presents fewer readmissions for each patient and, as already seen, higher mortality, a phenomenon that is, for this district, also verified in younger patients. In the case of no readmission, *Braga* and *Lisboa* present a crescent trend pattern related with age, which is associated to the number of pneumonia cases. In the case of 1 or more readmissions, they are mostly verified after the sixties for *Beja*, after the forties for *Braga*, and after the twenties for *Lisboa*. Figure 13 limits the visualization to a maximum of two readmissions, although in some cases more readmissions were verified. In this figure, colours are associated with the defined age groups.

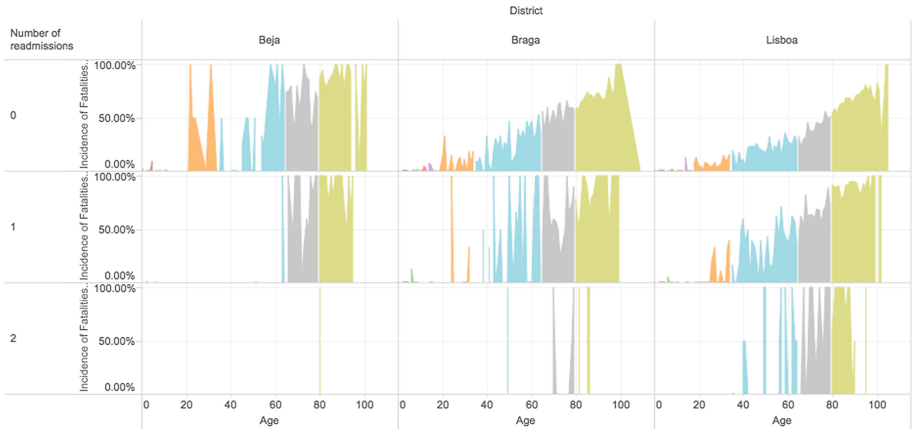


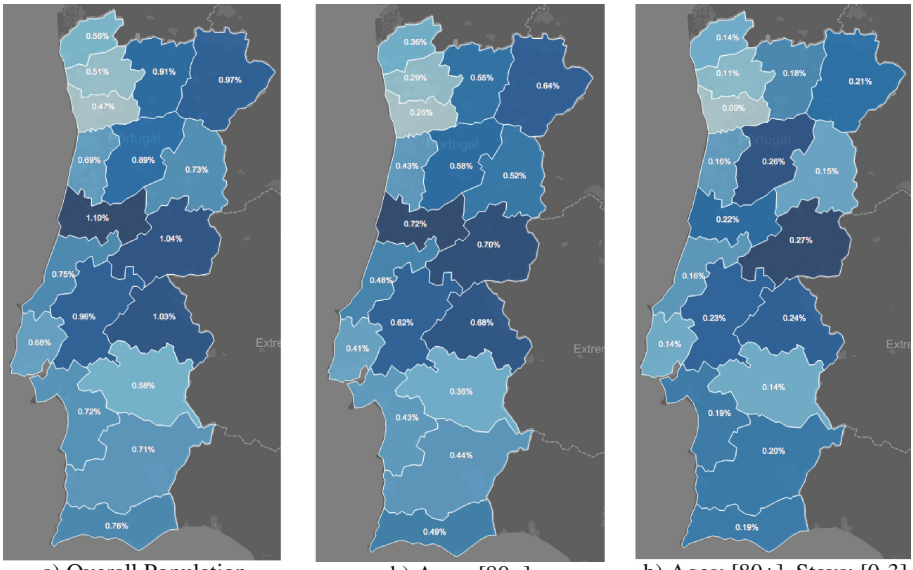
Fig. 13. Number of readmissions for shorter stays ([0–3] days)

Fig. 14. Overall percentage of mortality for shorter stays ([0–3] days)

In an overall characterization of the several districts, the other 15 of continental Portugal, Fig. 14 shows that some interesting patterns emerge with districts that have higher rates of mortality in youngest people, like *Aveiro*, *Faro*, *Portalegre*, *Santarém*, *Vila Real* or *Viseu* around the twenties, and *Évora*, *Portalegre* or *Viseu* around the forties, just to mention some cases. It is interesting to see that some districts present several similarities, while others show almost no cases in younger people like *Évora*.

It is now important to look into the other data available in the analytical data repository, like the statistical information, to understand if the high incidence of mortality in some regions cloud be influenced or explained by other factors.

Taking the statistical information, data related with the latest census in Portugal (made in 2011) was selected. In this case, Fig. 15(a) shows the spatial distribution of the incidence of mortal victims considering the overall population of each district. In this case, three districts have percentages of incidence superior to 1 %, namely *Coimbra*, *Castelo Branco* and *Portalegre*, with 1.10 %, 1.04 % and 1.03 %, respectively. Other districts present values very close to 1 %. In the case of mortal victims with 80 or more years old, Fig. 15(b), the three districts already pointed out continue to have the higher values, now with 0.72 %, 0.70 % and 0.68 %, respectively. Only when the available information is filtered, considering the shorter stays in the hospital, Fig. 15(c), *Castelo Branco* presents the highest percentage of victims attending to the



a) Overall Population b) Ages: [80+] b) Ages: [80+], Stays: [0-3]

Fig. 15. Percentage of mortal victims regarding the overall population

population of that district, namely 0.27 %, being followed by *Viseu* and *Portalegre*, with 0.26 % and 0.24 %, respectively.

In Fig. 15(a), one district called our attention due to its dissimilarity with the others also located at the littoral of the country. *Coimbra* presents the highest percentage of mortal victims attending to the global population in that district. Along the years under analysis, and as already mentioned, the overall increase in terms of the number of mortal victims was more than 65 %, and *Coimbra*, as can be see in Fig. 16, follows this average trend, with 66.15 %, having an overall incidence of mortality of 19.52 %. *Beja* is again in the spot not only because this district has the highest overall incidence of mortality, 35.78 %, but also because the variation of the number of victims was, from 2002 to 2011, of 143.94 %. *Castelo Branco* presents the highest variation with 167.68 % being followed by *Leiria* with 147.01 %.

Giving this context of overall variation of the incidence of mortality, it is now relevant to verify the evolution of the number of mortal victims along the years (Fig. 17). In Fig. 17(a), the variation of mortal victims considering the different age groups and the several years shows that, in younger patients, the variation is usually higher although few cases are verified. In these cases, some outliers show variations occasionally higher than 100 %, either positive or negative (those cases were filtered from the image for the sake of clarity). The variation of cases for the several years tends to avoid huge variations with age, being the number of victims a more constant number considering the number of pneumonia cases. However, along the years, the number of victims has increased considerably in the age class of 80 or more years old, as can be seen in Fig. 17(b). In global terms, the incidence of victims is around 30 % for this age class, less than 20 % for [65–79], less than 10 % for [25–64], and so on.

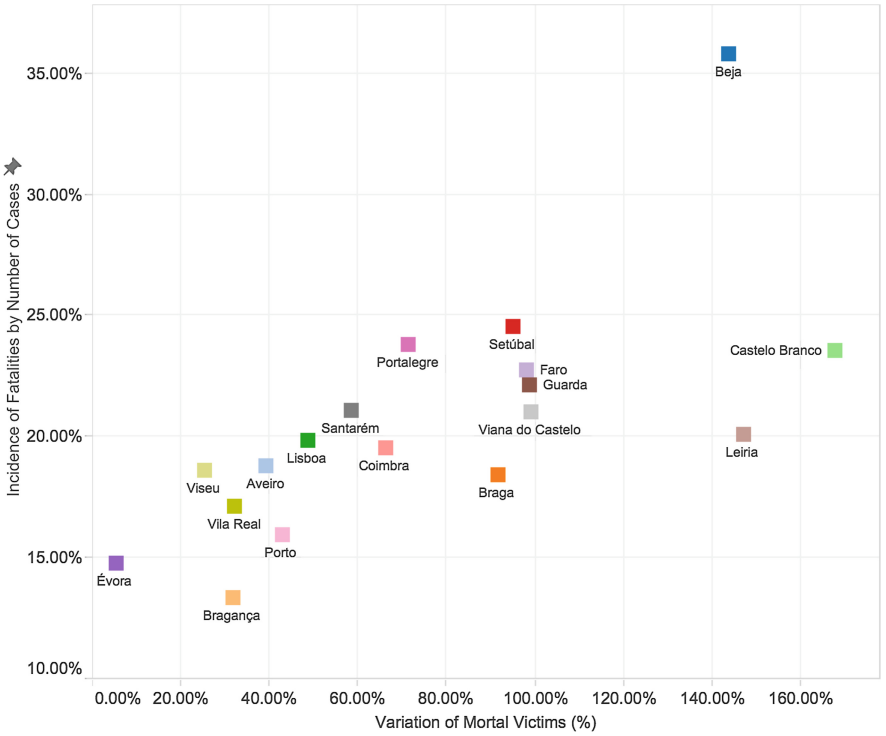
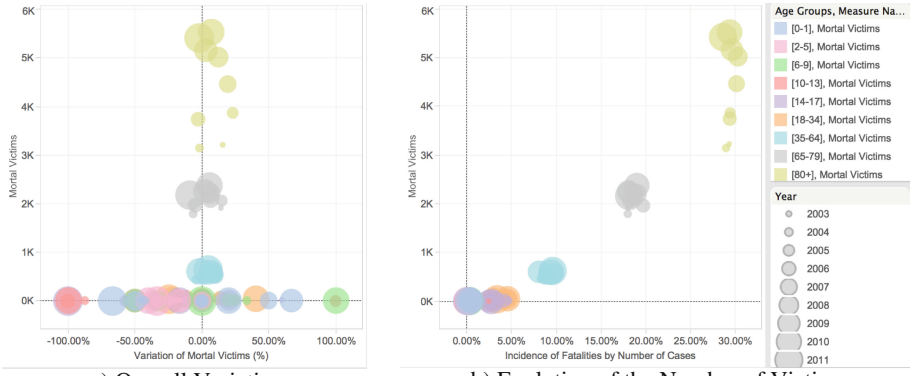


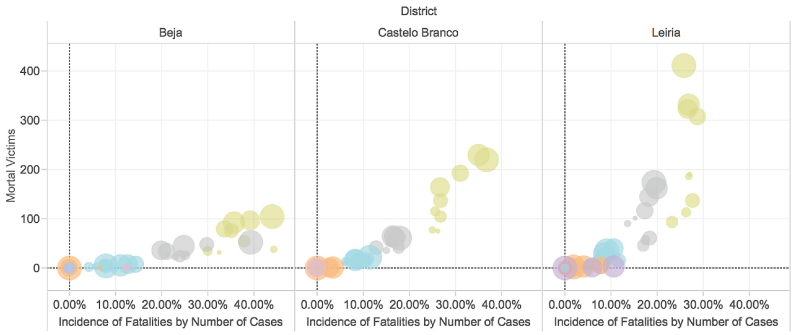
Fig. 16. Variation of the number of victims from 2002 to 2011



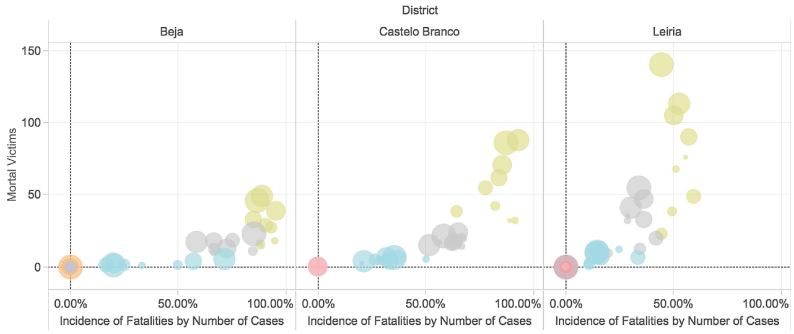
a) Overall Variation b) Evolution of the Number of Victims

Fig. 17. Variation of the number of mortal victims along the years

For the three districts with the highest variations in the decade under analysis, *Beja*, *Castelo Branco* and *Leiria*, Fig. 18(a) shows how these districts behaved along the years. In the case of *Leiria*, this district presents an increase in terms of the age classes [65–79] and [80+] that is very impressive. Although with a significant increase in the



a) For All Stays



b) For Shorter Stays ([0-3])

Fig. 18. Relevant variations of the number of mortal victims along the years

number of victims, the percentage of mortality is lower than the verified in *Beja* or *Castelo Branco*, even when the number of days of admissions, to consider the shorter stays, is filtered (Fig. 18(b)). Each one of these districts present a characteristic trend in the evolution of the disease and its consequences. Even in a small country like Portugal, the differences between districts is so high that justify a deepest analysis and the identification of the potentiating factors.

Given the presented analyses, it is possible to see that data, when properly stored in an analytical repository, can be analysed in an interactive way, combining different perspectives and applying different filters to data. The goal is to gain a deeper understanding of the phenomena under analysis, in order to support the decision making process. In this case, the purpose was to spatially characterize a disease that provokes so many deaths. The knowledge obtained should allow decision makers to define appropriate measures to fight this disease.

6 Conclusions and Future Work

This paper presented an overall geo-spatial characterization of pneumonia incidence in continental Portugal, taking into consideration, mostly, the mortal victims caused by this disease. Data from 10 years counting 369 160 records, available in an analytical

repository, were analysed in specific dashboards that take into consideration the spatial component of the data, mainly the location of residence of the patients, indexed to the corresponding parishes. All implemented dashboards make available maps, graphs or tables that allow user interaction for data selection or filtering, facilitating data exploration and supporting the identification of relevant patterns or trends in data.

As future work, it is envisaged the refreshing of the data warehouse in order to add data from the recent years, allowing the analysis until 2015, for instance. Moreover, it is envisaged the upgrade of the data model, in order to consider other vectors of analysis, like environmental data, crucial to verify how climacteric conditions or pollution affect the course of this disease.

Acknowledgement. This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – *Fundação para a Ciência e Tecnologia* within the Project Scope: UID/CEC/00319/2013.

References

1. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q. **36**, 1165–1188 (2012)
2. Kimball, R., Ross, M.: The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. John Wiley & Sons Inc., Indianapolis (2013)
3. Abelló, A., Darmont, J., Etcheverry, L., Golfarelli, M., Mazón, J.-N., Naumann, F., Pedersen, T., Rizzi, S.B., Trujillo, J., Vassiliadis, P., Vossen, G.: Fusion cubes: towards self-service business intelligence. Int. J. Data Warehouse. Mining **9**, 66–88 (2013)
4. Han, J., Kamber, M., Pei, J.: Data Mining: Concept and Techniques. Morgan Kaufmann Publishers, San Francisco (2012)
5. Viswanathan, G., Schneider, M.: On the requirements for user-centric spatial data warehousing and SOLAP. In: Xu, J., Yu, G., Zhou, S., Unland, R. (eds.) DASFAA Workshops 2011. LNCS, vol. 6637, pp. 144–155. Springer, Heidelberg (2011)
6. Rivest, S., Bédard, Y., Proulx, M.-J., Nadeau, M., Hubert, F., Pastor, J.: SOLAP technology: merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. ISPRS J. Photogrammetry Remote Sensing **60**, 17–33 (2005)
7. Santos, M.Y., Leite, V., Carvalheira, A., de Araújo, A.T., Cruz, J.: Characterization of pneumonia incidence supported by a business intelligence system. In: Ortuño, F., Rojas, I. (eds.) IWBBIO 2015, Part I. LNCS, vol. 9043, pp. 30–41. Springer, Heidelberg (2015)
8. Eurostat: Respiratory diseases statistics, June 2016. http://ec.europa.eu/eurostat/statistics-explained/index.php/Respiratory_diseases_statistics
9. WHO: World Health Organization. "The top 10 causes of death." 27 May 2015 (2015). http://who.int/mediacentre/factsheets/fs310/en/
10. Sufahani, S.F., Razali, S.N.A.M., Mormin, M.F., Khamis, A.: An analysis of the prevalence of pneumonia for children under 12 year old in Tawau general hospital, Malaysia. In: Proceedings of the International Seminar on the Application of Science & Mathematics, Kuala Lumpur (2011)
11. Oroszi, F., Ruhland, J.: An early warning system for hospital acquired pneumonia. In: Proceedings of the 18th European Conference on Information Systems (2010)

12. Trillo-Alvarez, C., Cartin-Ceba, R., Kor, D.J., Kojicic, M., Kashyap, R., Thakur, S., Thakur, L., Herasevich, V., Malinchoc, M., Gajic, O.: Acute lung injury prediction score: derivation and validation in a population-based sample. Eur. Respir. J. **37**, 604–609 (2011)
13. Wu, C., Rosenfeld, R., Clermont, G.: Using data-driven rules to predict mortality in severe community acquired pneumonia. PLoS ONE **9**, e89053 (2014)
14. Peng, R.: Exploratory data analysis with R (2015). http://Lulu.com
15. Tufte, E.R.: Beautiful Evidence, 1st edn. Graphics Press, Cheshire (2006)
16. Tufte, E.R., Graves-Morris, P.R.: The Visual Display of Quantitative Information. Graphics press, Cheshire (1983)
17. INE: Portugal census (2011). http://censos.ine.pt
18. Santos, M.Y., Carvalheira, A., de Araujo, A.T.: A data-driven analytics approach in the study of pneumonia's fatalities. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), 36678 2015, pp. 1–10. IEEE (2015)
19. R-project: the R project for statistical computing (2016). https://www.r-project.org
20. Tableau (2016). http://www.tableau.com