A Business Intelligence Infrastructure Supporting Respiratory Health Analysis

R. Dinis, A. Ribeiro, Maribel Y. Santos
Centro Algoritmi, Universidade do Minho
Guimarães, Portugal
{pg15527, pg15287}@alunos.uminho.pt,
maribel@dsi.uminho.pt

Jorge Cruz*, A. Teles de Araújo**
*Faculdade de Medicina de Lisboa
**Fundação Portuguesa do Pulmão
Lisboa, Portugal
costacruzjorge@gmail.com, artur@telesdearaujo.com

***Abstract*—Business Intelligence Systems are being designed and implemented to support data analysis tasks that help organizations in the achievement of their goals. These tasks are accomplished using several technologies aimed to store and analyze data. This paper presents the particular case of the design and implementation of a business intelligence system to support health care specialists in the analysis and characterization of symptoms related with the chronic obstructive pulmonary disease. For this specific application domain, a data mart model is proposed, implemented and loaded, allowing the analysis of the available data using on-line analytical processing technology and a spatial data mining algorithm. The results obtained so far are promising, demonstrating the usefulness of the proposed business intelligence approach to characterize the key factors in the comprehension of the disease under analysis.**

Keywords-business intelligence; data mining; on-line analytical processing; chronic obstructive pulmonary disease

I. INTRODUCTION

The collection and storage of huge amounts of data is increasing as organizations need to analyze these data and identify useful patterns, trends or models that support the decision making process. Independently of the application domain, this is the reality of nowadays organizations. This paper addresses the particular reality of a non-profit organization, the Portuguese Lung Foundation (*Fundação Portuguesa do Pulmão* - FPP), that carry out several activities to collect, store and analyze data related with several diseases. The obtained results are used to characterize the actual reality and to set up campaigns aiming to improve the citizens' quality of life.

In particular, this paper is focused on the Chronic Obstructive Pulmonary Disease (COPD) for which Business Intelligence concepts and technologies are used to store and analyze the related data. The COPD is an airflow limitation that is not fully reversible and that affects up to one quarter of the adults with 40 or more years [1]. This disease is characterized by some of the following symptoms: chronic cough, sputum production and dyspnea. It can be confirmed in a clinical exam called spirometry, if the obtained values are 80% below of the Forced Expiratory Volume in 1 second (FEV1) and the ratio FEV1/FVC (Forced Vital Capacity) is lower than 0,7 [2]. The risk factors usually include: masculine gender, tobacco smoke, exposure to dusts and chemicals, air pollution, asthma, and genetic factors as a rare hereditary deficiency of α_1-antitrypsin [2]. This disease can be classified in four stages, according to the degree of severity. The first stage, or Mild COPD, is characterized by a FEV1 value above or equal to 80% and the presence, or not, of chronic cough and sputum production. COPD is not usually detected at this first stage. The second stage, or Moderated COPD, is characterized by a value of the FEV1 between 50% and 79%, shortness of breath during exertion, chronic cough and sputum production. The third stage, or Severe COPD, is characterized by a value of the FEV1 between 30% and 49%, greater shortness of breath, reduced exercise capacity and fatigue. The fourth stage, or Very Severe COPD, is characterized by a value of FEV1 below 30% and the presence of chronic respiratory failure [1] [2].

The analysis of the incidence of COPD and its geographical characterization is needed in order to provide health specialists with decision support indicators. To accomplish this goal, this paper presents the analysis of a data set made available by the FPP with data collected during 2007. The objective of this work is to design and implement a business intelligence system, analyze the data using On-Line Analytical Processing (OLAP) technology, and apply a spatial data mining algorithm for the identification of the spatial incidence and distribution of COPD. For the identified patterns, a characterization of them will be carried out in order to better understand, treat and prevent this disease. With the proposal of this business intelligence system, an integrated environment for the collection, storage and analysis of data is made available to the FPP. To the best of our knowledge, no similar system has been proposed and implemented. This business intelligence system includes a data mart model that enhances the data analysis tasks.

The advantage of applying business intelligence systems on real cases is associated to the delivery of innovative structures aimed to solve real world problems, giving a contribution to the theory of the area. This kind of approach has already been applied. We can find examples in the study of the toxic vigilance in France [3] and the study of sharing adverse drug event data [4].

This paper is organized as follows. Section II presents a brief overview on the available data and the transformations carried out to clean and put the data in the proper format for analysis. Section III describes the architecture of the proposed business intelligence system and the data mart

 ISBN: 978-1-61208-160-1

model used to store the data. Section IV presents the results obtained analyzing the data with OLAP technology. Section V describes the clustering approach and the SNN (Shared Nearest Neighbor) algorithm used to spatially characterize the incidence of COPD. Section VI concludes with some remarks and guidelines for future work.

II. AVAILABLE DATA

The data set available for this study was collected by the FPP in initiatives undertaken in Portugal in 2007. These initiatives are open to everyone who wants to participate. In them, the participants are asked to answer a questionnaire that integrates questions related to the symptoms and the risk factors of the COPD. The questionnaire also includes information about the geographical location where patients live (in a qualitative form), and their gender, age, height and weight. The result of the spirometry exam is also recorded.

The collected data, 1880 records, were made available in an Excel file. After an extensible analysis of the data, missing data fields and errors in data were identified.

Some of the tasks needed to clean (or to prepare) the data set include: i) the labeling of records without geographical localization (these records cannot be used in the spatial data mining task); ii) the replacing of *null* values on the height and weight with the mode of each one of these attributes; and, iii) the filling of the *do not know* stamp in all categorical attributes containing *null* values.

After the data cleaning process, a transformation phase took place. It was necessary to add the coordinates (x, y) of the geographical locations to each record. This will allow the use of the spatial data mining algorithm taking into consideration the geographical positions of the patients (the places where they live).

In the OLAP analysis presented afterwards, a subset of the available data is used. Only those records (275) related to patients with a FEV1 value lower than 80% are considered. This allows the characterization of the symptoms of individuals with COPD. For the spatial data mining task two different approaches are used. In the first, the whole data set is used, providing an overall characterization of the available data. After that, the subset with the 275 records, as in the OLAP analysis, is used.

For confidentiality reasons, no details about the available data are provided. Only aggregated results, resulting from OLAP analysis and the data mining process, are provided in Sections IV and V.

III. BUSINESS INTELLIGENCE SYSTEM

Business Intelligence systems are defined as analytical tools that aim the analysis of organizational data to provide further information to managers, improving the decision making process [8]. The Business Intelligence (BI) concept emerged as an evolution of the Decision Support Systems (DSS) [9]. This new concept replaced the data-oriented DSS because many approaches developed to assist the decision making process include OLAP and Data Mining technologies. OLAP arises to overcome some of the difficulties in the analysis of data stored in Operational Databases (ODB), which are exposed to continuous updates. The analytical process should access another database, specifically designed to support these analyses, the Data Warehouse (DW). It is on the data stored in this repository that OLAP and Data Mining technologies are usually applied in a BI context.

In this study, a BI system was designed and implemented to store and analyze the data collected by the FPP. The ETL (Extract, Transform and Load) process will take as input the data available in a spreadsheet and will run the appropriate mechanisms to clean, transform and load the data into the proposed data mart. This data mart will support the OLAP technology and the data mining algorithms.

To support forthcoming initiatives of the FPP, a web application dedicated to data collection tasks was implemented. The data collected through this web application is stored in an ODB enhancing the ETL process and improving the quality of the collected data.

The architecture of the BI system envisaged for the FPP is illustrated in Figure 1.

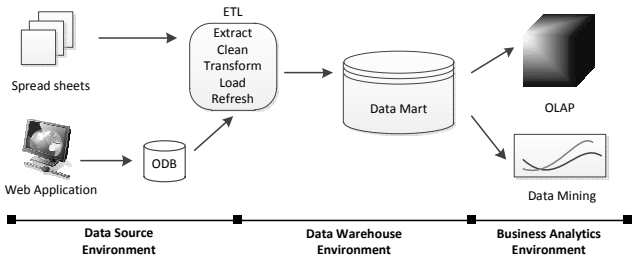

Figure 1. BI System Architecture

Related to implementation, the web application was developed using client/server pages (.ASP) and the programming languages HTML and Java Script. The ODB, the ETL process and the OLAP cubes were implemented using SQL Server 2008© technology. The Spatial Data Mining component makes use of a clustering algorithm implemented in Visual Basic© (more details can be found in Section V).

The Data Mart model was defined with the decision process in mind. This data model integrates one vector of analysis, represented by the fact table **FactFPP**. The star schema is shown in Figure 2. The **FactFPP** fact table allows the storage of the relevant information collected with the questionnaire. This fact table is linked to a set of dimension tables allowing the analysis of the available data in different perspectives. If we look to the data mart model, we can see that this table is linked to the **Time**, **Location**, **Profession**, **Patient**, **Smoke Characterization**, **Allergy Characterization**, **Cough Characterization**, **Fatigue Characterization** and **Pulmonary Diseases Characterization** dimension tables, meaning that the **FVC**, **FEV1**, **FEF 25-75** (Forced Expiratory Flow 25%–75%), three values obtained during the spirometry exam that characterize COPD, and the **Severity Stage**, can be analyzed by when, where the COPD is verified, who (with the information of the associated individuals such as age, gender, weight, among other attributes) and how (with the several questions of the questionnaire grouped in five distinct dimensions). The fact **Patient** is an event counter used to quantify the

number of individuals with specific symptoms or characteristics. It should be noted that the dimension **Profession** and the attributes marked with * in Figure 2 (attributes present in other dimensions or even in the fact table) will not be used in this first analysis (both OLAP and data mining) because these data are not present in the available data set (they will be included in future analysis when the web application is used for data collection).

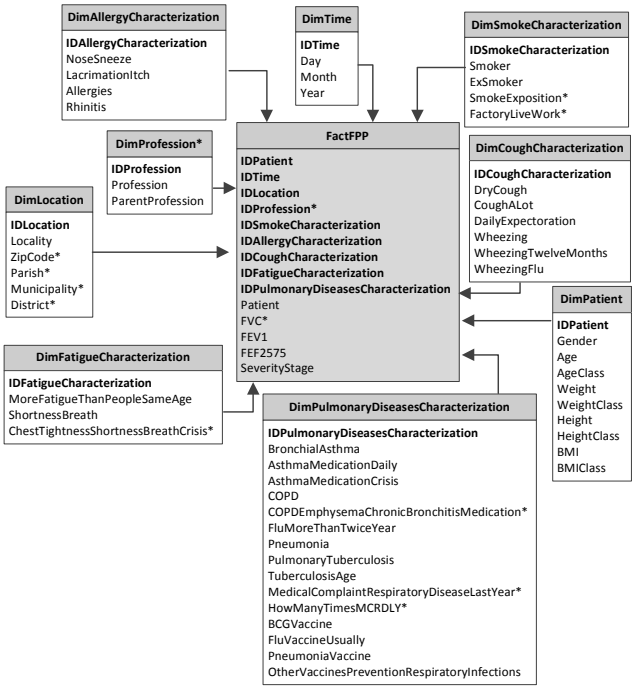

Figure 2. The Data Mart model

The presented data mart was designed in such a way that the star schema can evolve to a constellation schema as new fact tables are added. At this moment, two additional fact tables are envisaged: one for the study of pneumonia and the other for the study of lung cancer. As the constellation grows, more symptoms and data about the individuals can be related in the study of one or more diseases.

IV. DATA ANALYSIS WITH OLAP

After the presentation of the proposed BI system and the data model that stores all these data, this section presents the results obtained from the analysis of the available data using the OLAP technology. This technology is used to analyze the fact table along different perspectives.

The first analysis verifies the COPD severity stage of the individuals. Figure 3 shows that almost all the patients, 254 (92.7%), are at the 2nd COPD severity stage (Moderate). Only 18 patients are at the 3rd COPD severity stage (Severe) and 3 patients are at the 4th COPD severity stage (Very Severe).

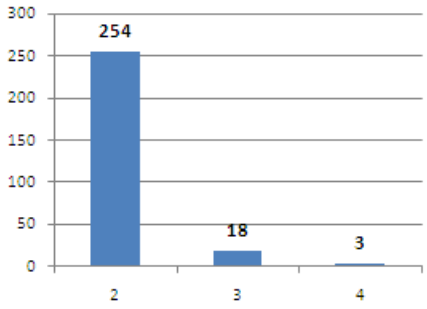

Figure 3. Severity Stages

After analyzing the COPD severity stages of the patients, the next analyses are focused on the answers given by the patients to the questions that allow their characterization. As already mentioned, all these individuals have a diagnosis of COPD after the spirometry exam. The characterization obtained to the group of questions related to **Smoke** can be seen in Figure 4. The results show that despite the tobacco is one of the most obvious risk factors for this disease, on the analyzed data, 168 patients (61.1%) with a diagnosis of COPD revealed that they never smoked.

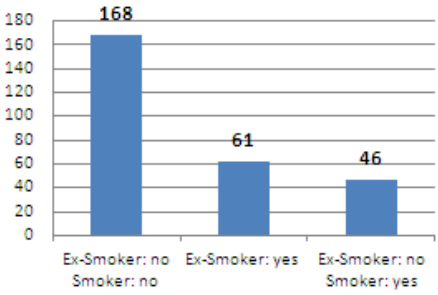

Figure 4. Smoke Characterization

Concerning the characterization of the group of questions **Fatigue**, the results are presented in Figure 5. Although it is clear that most of the patients, 187 (68.0%), feel more fatigue than people who have the same age (**MFTPSA**) and/or have shortness of breath (**SB**), and 118 of them (42.9%) even feel both symptoms, there are 88 individuals (32.0%) who have COPD that do not present any of these two symptoms.

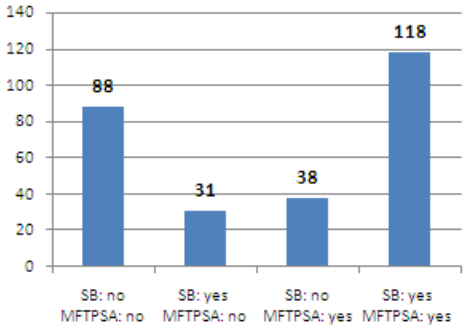

Figure 5. Fatigue Characterization

Analyzing the answers obtained to the group of questions **Cough** (Figure 6), and more precisely the two symptoms of COPD present in this group of questions, cough (**Dry Cough**) and expectoration (**Daily Expectoration – DE**), we can verify that 171 (62.2%) individuals with COPD have dry cough and daily expectoration. These two symptoms seem to be related. Indeed, when the individuals do not have dry cough, only 17.3% of them have daily expectoration. However, when the individuals have dry cough, the percentage of them who also have daily expectoration increases considerably (reaching 42.7%).

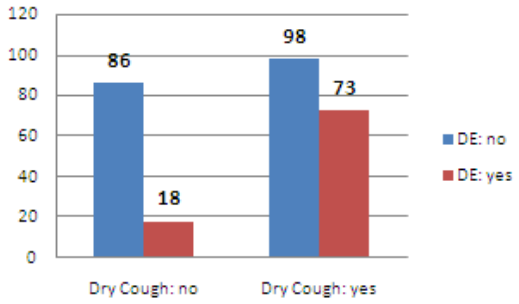

Figure 6. Cough Characterization

Concerning the **Pulmonary Diseases** group (TABLE I), we analyzed the three pulmonary diseases characterized in this group of questions, which could have any relation with COPD. These diseases are the following: bronchial asthma (a COPD risk factor), pneumonia and pulmonary tuberculosis (**Pulm Tuber**).

Analyzing the cube at TABLE I, none of the individuals suffered from all these three diseases (there are cases of **dk** – *do not know* – that are not considered in this observation). Another observation that can be pointed from TABLE I is that only a low percentage of these patients who have COPD suffered from at least one of these three diseases. Indeed, 65 (23.6%) individuals have or had bronchial asthma, 57 (20.7%) individuals suffer or suffered from pneumonia and 17 (6.2%) from pulmonary tuberculosis. As bronchial asthma is a risk factor of COPD, it could be expected that the percentage of incidence of this disease was significantly higher in patients who have COPD.

TABLE I. PULMONARY DISEASES CHARACTERIZATION

		Bronchial Asthma		
		no	yes	Grand Total
Pneumonia	Pulm Tuber	Fact FPP Count	Fact FPP Count	Fact FPP Count
⊟ dk	dk	3	2	5
	no	5	2	7
	Total	8	4	12
⊟ no	dk	2	2	4
	no	155	35	190
	yes	9	3	12
	Total	166	40	206
⊟ yes	dk	6	1	7
	no	25	20	45
	yes	5		5
	Total	36	21	57
Grand Total		210	65	275

Regarding the answers obtained to the group of questions **Allergy**, the results are presented in Figure 7. We analyzed the incidence of three characteristics from this group of questions: the eyes lacrimation and itch (**LI**), runny nose and sneeze (**NS**), and rhinitis. The results showed that 202 (73.5%) individuals with COPD also have runny nose and sneeze when they are not with flu. This fact can reveal a link between these symptoms and COPD. However, there is almost the same number of patients, 201 (73.1%), who claims that they do not suffer from rhinitis. Another fact observed in this group of questions was that from the 73 patients with COPD who do not have runny nose and sneeze, 10 (13.7%) have eyes lacrimation and itch. From the 202 patients with COPD who have runny nose and sneeze, 128 (63.4%) have eyes lacrimation and itch.

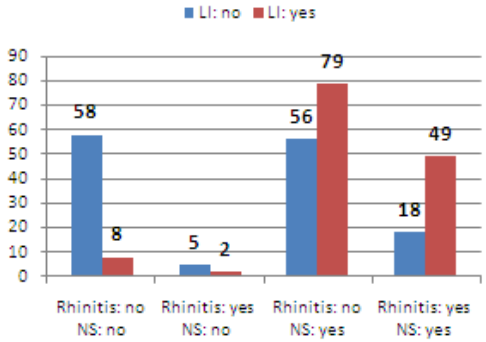

Figure 7. Allergy Characterization

This last data analysis with OLAP has as objective to show the importance of the initiatives and campaigns made in favor of this respiratory disease and the need to participate in them. TABLE II shows the obtained results. Patients were subdivided by **Gender** and by **Age Class**. The **COPD** attribute indicates if the patient already knew that he/she suffers from COPD before the spirometry exam. We can say, based on these data, that 229 patients (83.3%) with COPD did not know that they had this disease before participating in these initiatives promoted by the FPP. This cube also shows age as a risk factor for COPD. Indeed, 204 individuals (74.2%) with more than 41 years old have this disease. It was also noted that only 7 (2.5%) individuals, between 0 and 17 years old, have COPD.

TABLE II. IDENTIFICATION OF COPD AT THE FPP'S INITIAVIVES

		COPD		
		no	yes	Grand Total
Gender	Age Class	Fact FPP Count	Fact FPP Count	Fact FPP Count
⊟ f	[0-17]	3		3
	[18-40]	37	2	39
	[41-64]	54	8	62
	65+	44	9	53
	Total	138	19	157
⊟ m	[0-17]	4		4
	[18-40]	22	3	25
	[41-64]	29	4	33
	65+	36	20	56
	Total	91	27	118
Grand Total		229	46	275

 ISBN: 978-1-61208-160-1

V. DATA ANALYSIS WITH SPATIAL DATA MINING

This section presents the analysis of the available data using a data mining algorithm, given particular attention to the spatial component of the data. Firstly, it will be presented the data mining algorithm that will be used and then the obtained results.

A. The Shared Nearest Neighbor (SNN) Algorithm

The SNN algorithm was first presented by Levent Ertöz, Michael Steinbach and Vipin Kumar in 2003 on the "SIAM International Conference on Data Mining". This algorithm can find clusters with different sizes, shapes and densities, and automatically identifies the number of clusters in the data set [10]. Usually, this algorithm has three input parameters, *k* – the size of the list of neighbors, *Eps* – the density threshold, and the *MinPts* – the threshold value used to classify the core points [10] (Figure 8).

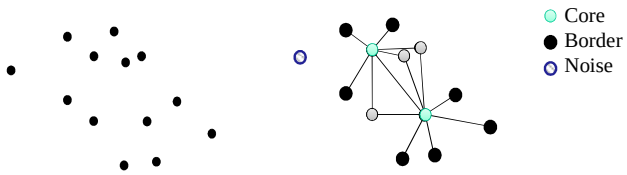

Figure 8. SNN: an example

The implementation used in this paper was coded from scratch in Visual Basic. It requires only an input parameter: *k*. *Eps* and *MinPts* are calculated based on *k* as follows: *Eps = 3k/10* and *MinPts = 7k/10* [11] and includes the following steps [11]:

1. Identify the "*k*" nearest neighbors for each point (using a distance function to calculate similarity);
2. Calculate the SNN similarity between pairs of points as the number of nearest neighbors that the two points share;
3. Calculate the SNN density of each point;
4. Detect the core points;
5. Form the clusters from the core points;
6. Identify the noise points;
7. Assign the remainder points to the cluster that contains the most similar core point.

In the work presented in this paper were used two different distance functions: one for all the available data and another for records with FEV1<80%. This way it is possible to mine the available data and find spatial distributions taking into consideration the whole data set, patients with or without COPD, and also those records that are exclusively associated with patients that suffer from COPD. For this last data set, it is possible to group patients not only by their geographical proximity but also by their incidence of COPD. The equations to the distance functions are the following (adapted from [6]):

$$DistFunction(p_1, p_2) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (1)$$

$$DistFunction(p_1, p_2) = w_1 * \left(\frac{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}}{mDist} \right) + w_2 * \left(\frac{|s_1 - s_2|}{mFEV} \right) \quad (2)$$

In equation 2, w_1 and w_2 represent the weights assigned to the position and to the FEV1 parameters. Through these weights it is possible to verify the impact of the position and the FEV1 value on the clustering result. Also, in equation 2, *mDist* and *mFEV* are the maximum difference values between any two points (for the distance and for the FEV1, respectively). They are used to normalize the distance function value.

In order to identify the appropriate values for w_1 and w_2, some simulations were made with the data set. The obtained results are presented in Figure 9.

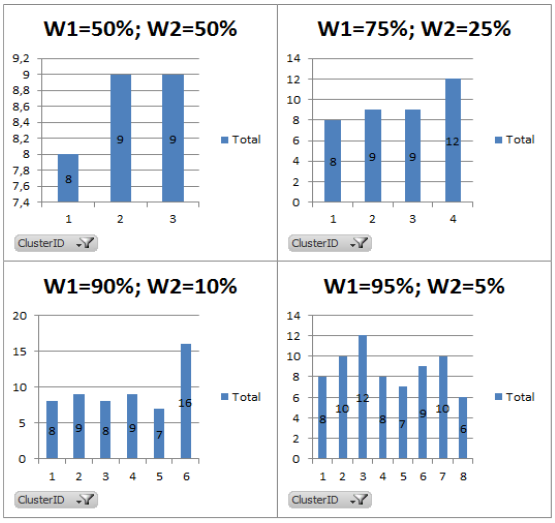

Figure 9. W_1 and W_2 analysis

After analyzing the obtained results, the weights w_1=95% and w_2=5% seem to be more appropriate as they allow the identification of more clusters, with different shapes and sizes. The other weights classify a large number of records as noise. These weights already showed to be appropriate in other contexts [6].

B. Obtained Results

As mentioned in Section II, we are using two different approaches. In the first, all the records available in the data set are used. After several simulations to identify a proper value for *k*, we decided to adopt a *k* value of 15, which returned 41 clusters, with the spatial distribution shown in Figure 10. In this figure each cluster is represented by a different color. As the coordinates were added at the Parish level, and not at a lower level of detail, a small agitation of the points coordinates was carried out in order to avoid points overlapping in the clusters visualization.

From the presented results it is possible to verify that the majority of the points are associated to the huge metropolitan areas of Oporto and Lisbon, despite the big clusters (in area) spread at the interior, central coast and south of the country. These results can have two main explanations: the fact that these two regions are the most populated in Portugal or a non-balanced data set with a higher incidence of individuals from these two metropolitan areas. This last case was the verified in the data set under analysis.

 ISBN: 978-1-61208-160-1

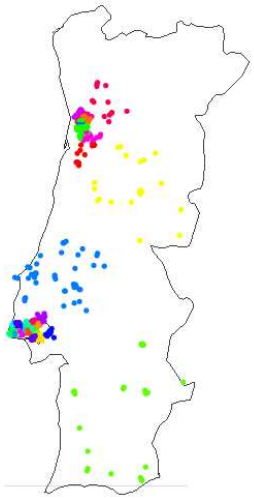

Figure 10. Clusters for the first approach to the data analysis

Associated with the presented results, we proceed verifying the risk factors associated to the patients of this first approach. One of the risk factors is easy to see just by looking at the graphical distribution of the points, the air pollution, because the majority of the clusters are located in the more populated regions of the country and where the incidence of factories, cars and other pollutants is higher. To analyze the other risk factors, the answers to the questionnaire were verified (Figure 11). Analyzing the risk factors we verify that 44% of the patients that are grouped in the several clusters are or were smokers, 45% are male, 17% are older than 65 years old, and 15% suffers from asthma.

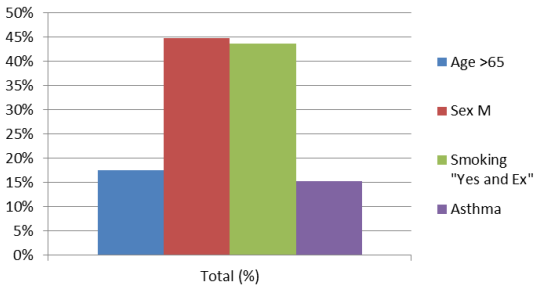

Figure 11. % total of risk factors (first set of results)

The other approach followed only uses the records that have a FEV1 value below 80%. In this second approach the equation of the distance function used is the equation 2 and after some simulations the best value found for *k* is 7. The obtained results are presented in Figure 12 where colors are associated to clusters (each cluster has a different color).

Comparing Figure 12 with Figure 10 it is possible to verify that the clusters in Figure 12 are more concentrated in one region which is the region of Lisbon and surroundings. The other cluster appears at North, grouping individuals from several districts. Obviously, more data is needed to obtain a deeper understating on the spatial distribution of COPD.

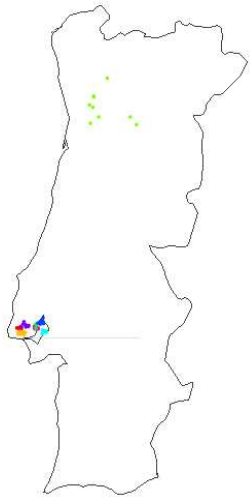

Figure 12. Clusters for the second approach to the data analysis

In a more detailed analysis to this second set of results is possible to see that, as in the first set of results, the more prominent risk factors are being male, smoking or advanced age (Figure 13), with exception of: cluster 2 where the asthma has more incidence than smoking; and, cluster 4 with asthma having the same incidence as male, smoking and advanced age.

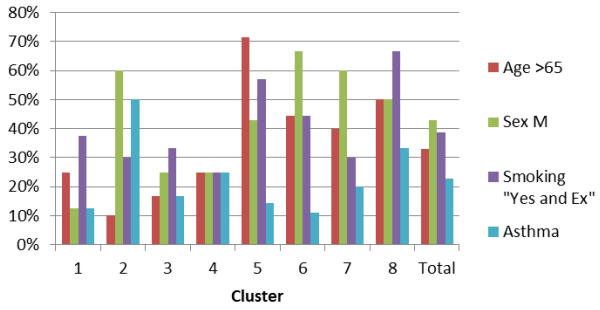

Figure 13. % Risk factors per cluster and total

Looking at the symptoms (Figure 14), the two symptoms that appear with more incidence are the fatigue and the shortness of breath, exception made to the first cluster where the chronic cough is the symptom that appear in first place. This incidence of fatigue and shortness of breath has huge impact on the quality of life of the patients. The fact that these patients are from the more populated regions in Portugal, and consequently with more pollution, and nearly 40% of those are smokers or ex-smokers, could help to explain why these two symptoms have the higher percentages. Since the answers to the questionnaire do not tell us if patients have chronic respiratory failure and the values of FEV1 are all above 50%, the stage of COPD in which the patients are is the stage 2 or moderate COPD (TABLE III). All the patients with the FEV1 below 50% were considered as noise by the clustering algorithm because they are only a few records, and the distribution of the points in terms of geographic locations and the FEV1 values do not

 ISBN: 978-1-61208-160-1

have sufficient similarities to group them in one or more clusters.

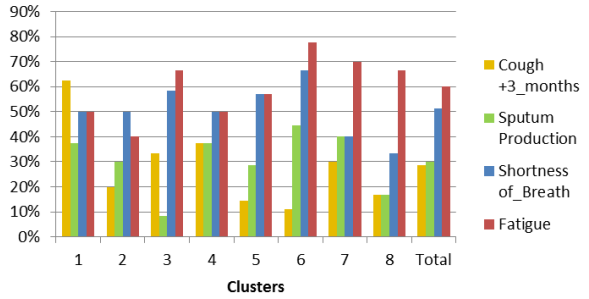

Figure 14. % Symptoms per cluster and total

TABLE III. COPD STAGES PER CLUSTER

Stages COPD	**Cluster**							
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***
Stage 2	8	10	12	8	7	9	10	6
Stage 3	0	0	0	0	0	0	0	0
Stage 4	0	0	0	0	0	0	0	0
Total	**8**	**10**	**12**	**8**	**7**	**9**	**10**	**6**

VI. CONCLUSION

This paper presented a business intelligence system for the analysis of data collected by the Portuguese Lung Foundation (FPP – *Fundação Portuguesa do Pulmão)*. The objectives set to this work were to design and implement a business intelligence system, analyze the data using OLAP, and apply a spatial data mining algorithm for the identification of the spatial incidence and distribution of COPD (Chronic Obstructive Pulmonary Disease). The implemented system integrated a data mart for the storage of data and OLAP and spatial data mining technologies for data analysis.

Data from 1880 patients were available from the FPP´s data set. Each patient answered a questionnaire and made a clinical exam called spirometry. The several analyses in terms of OLAP reinforced essentially the idea that this disease is very difficult to diagnose without the spirometry exam. That happens because despite we have confirmed some risk factors related to COPD and identified some patterns of this disease, there seems to be a kind of contradiction with their symptoms. This happens, for instance, with the characterization of the responses obtained in the question groups **Fatigue** and **Cough**. Many patients do not have any symptoms from these two groups of questions that are typical of this respiratory disease. Therefore, the difficulty in diagnosing the disease also reinforce the conclusion that a good prevention of COPD, making a spirometric exam periodically, is essential for people who fall within the risk factors of this disease.

In terms of spatial analysis we have two main conclusions. First, when comparing the data between the two set of results, the distribution of the clusters is similar and the percentages of the risk factors are also similar. The second conclusion is that the main symptoms that the patients mention to have are shortness of breath and fatigue, and all the patients grouped in the clusters are in the stage 2 of COPD.

The proposed BI system showed to be useful in the analysis of the available data, allowing the characterization of the key factors that are associated to COPD.

Related to future work, it is envisaged to perform the analysis with more data, to automate the ETL process between the operational database of the web application and the data mart, to incorporate a temporal variable tagging each patient in order to be possible a spatio-temporal analysis, and to perform analysis with other spatial cluster algorithms besides the SNN used in this paper.

REFERENCES

[1] GOLD, Global Strategy for Diagnosis, Management, and Prevention of COPD, Tecnical Report, Global Initiative for Chronic Obstructive Lung Disease, 2010.

[2] R. A. Pauwels, A. S. Buist, P. M. A. Calverley, C. R. Jenkins, and S. S. Hurd, “Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease,” American Journal of Respiratory and Critical Care Medicine, vol. 163, pp. 1256-1276, 2001.

[3] G. Guyodo, I. Blanc, J. L. Boulben, Lefevre B, F. De Bels, and R. Garnier, “The French Toxic Exposure Surveillance System: Adaptation of a Business Intelligence System for Toxicovigilance,” Clinical Toxicology, vol. 48, no. 3, 2010.

[4] M. Horvath, H. Cozart, A. Ahmad, M. K. Langman, and J. Ferranti, “Sharing Adverse Drug Event Data Using Business Intelligence Technology,” Journal of Patient Safety, vol. 5, no. 1, pp. 35-41, 2009.

[5] S. Brooker, S. Clarke, J. K. Njagi, S. Polack, B. Mugo, B. Estambale, E. Muchiri, P. Magnussen, and J. Cox, “Spatial Clustering of Malaria and Associated Risk Factors During an Epidemic in a Highland Area of Western Kenya,” Tropical Medicine & International Health, vol. 9, no. 7, pp. 757-766, 2004.

[6] A. Moreira, M. Y. Santos, M. Wachowicz, and D. Orellana, "The Impact of Data Quality in the Context of Pedestrian Movement Analysis," In M. Painho, M. Y. Santos and H. Pundt (Eds.), Geospatial Thinking, Lecture Notes in Geoinformation and Cartography: Springer, 2010.

[7] G. R. Jordan, N. Loveridge, K. L. Bell, J. Power, N. Rushton, and J. Reeve, “Spatial clustering of remodeling osteons in the femoral neck cortex: a cause of weakness in hip fracture?,” BONE, vol. 26, no. 3, pp. 305-313, 2000.

[8] W. F. Cody, J. T. Kreulen, V. Krishna, and W. S. Spangle, “The integration of business intelligence and knowledge management,” IBM Systems Journal vol. 41, pp. 697-713, 2002.

[9] S. Alter, Information Systems: A Management Perspective: Addison Wesley Longman, 1999.

[10] L. Ertöz, M. Steinbach, and V. Kumar, "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data", Proceedings of the Second SIAM International Conference on Data Mining, San Francisco, pp. 47-58, 2003.

[11] A. Moreira, M. Y. Santos, and S. Carneiro, "Density-based clustering algorithms – DBSCAN and SNN (Implementation Documentation)," http://ubicomp.algoritmi.uminho.pt/local/, Date of access: February, 2011.

 ISBN: 978-1-61208-160-1