NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

ANALYZING AND QUANTIFYING SECURITY DATA THROUGH A GEOSPATIAL VISUALIZATION FRAMEWORK

Empirical research regarding social and economic indicators in Portugal

Marcel Motta do Nascimento (M2016337)

Dissertation proposal presented as partial requirement for obtaining the Master's degree in Information Management with a Specialization in Business Intelligence and Knowledge Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

ANALYZING AND QUANTIFYING SECURITY DATA THROUGH A GEOSPATIAL VISUALIZATION FRAMEWORK

by

Marcel Motta do Nascimento (M2016337)

Dissertation proposal presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence and Knowledge Management

Advisor: *Miguel de Castro Neto*

ABSTRACT

This report has been built as a thesis proposal based on the work previously developed and submitted to the 18th Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI 2018) and later published in the books "Modelos Preditivos & Segurança Pública", by Teresa Rodrigues and Marco Painho and "Information Systems for Industry 4.0", by Isabel Ramos, Rui Quaresma, Paulo Silva and Tiago Oliveira. The original work was conceived in the scope of the SIM4SECURITY project; it proposed a set of techniques and predictive models for analyzing security by using geospatial data and quantifiable indicators. The current work not only aims to report on the Business Intelligence tools and techniques developed in the aforementioned project, but also, it should take us one step further on establishing a common framework for structuring, quantifying and visualizing data, with the purpose of analyzing social and economic indicators and their impact on crime rates in Portugal.

The framework developed herein could constitute as a strong decision support tool by helping security forces on optimizing resources allocation, tailoring more efficient security policies and finding patterns, challenges and threats in the social landscape for the next years to come.

KEYWORDS

Business Intelligence; Geographic Information System; Security; Data Modeling; SIM4SECURITY.

INDEX

LIST OF ABBREVIATIONS AND ACRONYMS

BI:	Business Intelligence
DW:	Data Warehouse
ETL:	Extract, Transform and Load
GIS:	Geographic Information System
KPI:	Key Performance Indicator
OLAP:	Online Analytical Processing
OLTP:	Online Transaction Processing

CHAPTER I: INTRODUCTION

Security stands for one of the oldest and most important social constructs in our society. It ensures the maintenance of order among individuals by the hands of the state through law enforcement and civil obedience. The establishment and prosperity of social structures for civil organization have always been an inherent part of human survival and development. Throughout history, these structures not only served to legitimize political leaderships but to safeguard fundamental rights, to enforce the maintenance of law and order to ensure the protection of its individuals and its own interests. To fulfill its purpose, military and police forces are a key component to the maintenance of these structures, and perhaps the oldest institutions of any communitarian structure.

In the same fashion, the foundations of modern society share similar core concepts as more rudimentary social structures; they serve as organization models for promoting security and protecting common interests. In fact, society, security and defense are terms that quite often intertwine and cannot be treated separately as part of the human condition for thriving.

In Max Weber's essay "Politics as a Vocation", the state is considered the sole source of the "right" to use violence, in a way to provide security to its citizens; this legitimate monopoly of violence manifests the state sovereignty and the reason itself for its existence. Therefore, the state cannot sustain itself if it's unable to provide security to the civil society, as the civil society cannot survive if it is unable to safeguard its individuals and their fundamental rights.

As those organizations evolved over time and adapted according to the political landscape, the role and importance of security forces followed suit, in face of a dramatic change in the underlying threats to society, the perimeter of social structures and the definition of security. This ever-changing scenario for the security forces called for a constant adjustment in public policies. That is, the application and development of new methods and technologies for tackling the current challenges for security became a critical demand from policy makers, leaders and other institutional forces, highlighting the relevance of extensive studies in this field. However, we need to make sure that new, efficient methods and technologies developed for this purpose cannot be used without transparent rules, liability, supervision and accountability (Rodrigues and Painho, 2018). In other words, the role of security forces should not jeopardize civil liberties guaranteed by the democratic rule of law. An unrestricted surveillance state willing to abridge individual freedom over security concerns are often associated with authoritarian and anti-democratic political structures, as explored by George Orwell and Aldous Huxley in modern literature.

This complex security paradigm forces the State to establish an appropriate institutional framework for internal security action and resources allocation (Teixeira, Lourenço & Piçarra, 2006), while the safety of the population becomes a central issue and demographics a strategic vector (Rodrigues, 2014).

Over the course of history, technology has been increasingly used by security forces but more recently technology itself also created a new, flourishing environment for crime networks and "cybercrimes", such as identity theft, money laundering, terrorism financing, drug trafficking, among many other felonies. The development process associated with the globalization, especially in the field of communications, created optimal conditions for the rise of new threats, organized crime networks and, ultimately, a redefinition of the concept of security. To put it into

perspective, in 2018 the annual costs related to cybercrime were estimated at € 600 billion, at around 1% of the global GDP (Lewis, 2018).

In 2003, in the context of the European Union, this new outlook was tackled by the adoption of a new model which consisted in the creation of a single, common architecture for managing security in a wider perspective extending to several strategic sectors (economy, food, health, environment, politics and community) (Rodrigues and Painho, 2018). The adoption of this strategy served not only to address a more efficient law enforcement mechanism, adapted to the current scenario, but also to promote a joint effort on distinguishing the causes and the effects of insecurity and creating policies focused on preemptive measures.

By looking at the European Union, Portugal can be considered an interesting case study: a modern democracy committed to the principles of freedom and security, one of the safest countries in the world according to the Global Peace Index (Institute for Economics and Peace, 2018), possessing a plural and intricate security system. Additionally, as demographic projections were prepared up to 2040 (Bravo, 2016), new demographic trends in the territory can be described as a triple ageing phenomenon: less youngsters, a progressively older working population group and a high increase of elderly population. The potential burden imposed to the social welfare system with less taxpayers and more pensioners leads local stakeholders to acknowledge the need for an increasingly strategic management, reduction of redundancies and convergence of resources and information among them.

Facing these challenges, the application of an Information Systems framework could serve as a mean of putting into evidence the current issues and limitations of public security policies, rationalize the allocation of resources, support the decision-making process and identify potential threats and risks to be addressed by policy makers. In this context, the SIM4SECURITY project was developed in order to build a technology-driven solution to assist the decision-making process in the public security. To do so, the project was split into five specific tasks:

1. Analysis and diagnosis of the current national situation: assessment of current demographic and social-economic indicators at a countrywide level;
2. Demographic forecast and scenario development: projections of the demographic composition and distribution until 2040, split by gender, age groups and region.
3. Development and implementation of a Geographic Information System (GIS) and design of a dynamic geoprocessing model;
4. Implementation of advanced spatial analysis methods (spatially dynamic clusters and modeling land cover change predictive model);
5. Modeling the distribution of security forces (number of officers and facilities location) according to the developed scenarios.

The findings obtained from these five tasks developed throughout the project were gathered and delivered into a visual, intuitive and interactive interface by using a Business Intelligence (BI) framework leveraged with Microsoft technologies, namely SQL Server and PowerBI. This solution will allow decision makers to assess its policies and resources, tailored according to local demands, for preventing and fighting crime.

CHAPTER II: DATA COLLECTION AND UNDERSTANDING

The required data providers can be split into two categories:

1. Non-sensitive data providers, such as the Instituto Nacional de Estatística (INE) and Direção-Geral da Polícia de Justiça (DGPJ);
2. Sensitive data providers, namely the Guarda Nacional da República (GNR).

For the development of the proposed data model, the following information was collected:

- Demographic and social economic data obtained through Statistics Portugal (INE);
- Crime data by municipality obtained through Direcção-Geral da Política de Justiça (DGPJ);
- Police stations, location and allocation of police officers provided by Guarda Nacional Republicana (GNR);
- Complimentary to external sources, demographic and crime projections computed by researchers of NOVA Information Management School (NOVA IMS).

After a filtering process, the indicators and attributes were selected, resulting in seven tables. Through these tables with transactional data, the indicators and attributes were extracted, transformed and loaded in analytical tables through an extraction, transformation and load (ETL) process that allowed the construction of a Data Warehouse (DW). The indicators and the description of their attributes are presented in the following tables (see Tables 1 to 7).

Attributes	**Description**
DICOFRE	Parish code
Actuação	Security force responsible for the parish
Comando	GNR *Comando* responsible for the parish
Destacamento	GNR *Destacamento* responsible for the parish
Posto	GNR *Posto territorial* responsible for the parish
PostoID	*Posto territorial* code
Efectivo	Number of police officers allocated to a *Posto territorial*
Longitude	Longitude of the *Posto territorial* location
Latitude	Latitude of the *Posto territorial* location

Table 1. Police officers table (Efectivos) (data privacy: private; source: GNR).

Attributes	**Description**
Município	Municipality designation
DICO	Municipallity code
Índice	Social economic index code
Valor	Social economic value

Table 2. Social economic table (socecon_data2011) (data privacy: public; source: INE).

Attributes	Description
Índice	Social economic index code
Description	Social economic index description

Table 3. Social economic metadata table (socecon_meta) (data privacy: public; source: INE).

Attributes	Description
DICOFRE	Parish code
Nome	Parish designation
Pop_2011	Population by parish for 2011
Pop_2030	Estimated population by parish for 2030
Pop_2040	Estimated population by parish for 2040

Table 4. Population table (pop_summary) (data privacy: public/private; source: INE/NOVA IMS).

Attributes	Description
DICOFRE	Parish code
Classe	Land cover/land use (LCLU) class
Area_Km2	LCLU area in square kilometres
Area_ha	LCLU area in square hectares
Ano	LCLU reference year

Table 5. LCLU table (uso_solo) (data privacy: private; source: NOVA IMS).

Attributes	Description
Território	Territorial unit nomenclature
Distrito	District designation
Município	Municipality designation
Ano	Recorded year for the respective crime
Eventos	Number of occurrences regarding the respective crime
Índice	Intern nomenclature for the crime description

Table 6. Crime table (crime_hist) (data privacy: public; source: DGPJ).

Attributes	Description
Classe	Class regarding crime classification
SubClasse	Sub class regarding crime classification
Crime	Crime designation and respective code
Descrição	Crime designation
Índice	Intern nomenclature for the crime description

Table 7. Crime metadata table (crime_meta) (data privacy: public; source: DGPJ).

CHAPTER III: ANALYTICAL MODEL DESIGN AND IMPLEMENTATION

This chapter goes in depth on the design and deployment of a database normalization schema for the Data Warehouse (DW) architecture, establishing the facts and dimensions tables, implementing the ETL (Extract, Transform and Load) process for treating transactional data and integrating the proposed multi-dimensional, analytical data model into a dashboard.

Given that the data sources were not made available by direct access to servers, the input tables (typically, ".csv", ".xls", ".json" and ".xml" file types) required additional treatment during the Extraction stage of the ETL process. This initial process was performed using VBA subroutines running on top of Microsoft Excel.

After preprocessing the raw data into a tabular data frame, the latter could be imported into a relational database managed with Microsoft SQL Server by combining the capabilities of Integration Services (SSIS) and Management Studio (SSMS). Subsequently, the deployment of the Transformation stage (feature engineering, data aggregation, variable indexing) and the Loading stage (from a transactional data model into an analytical data model) was also made possible by using this same framework (see Figure 1).

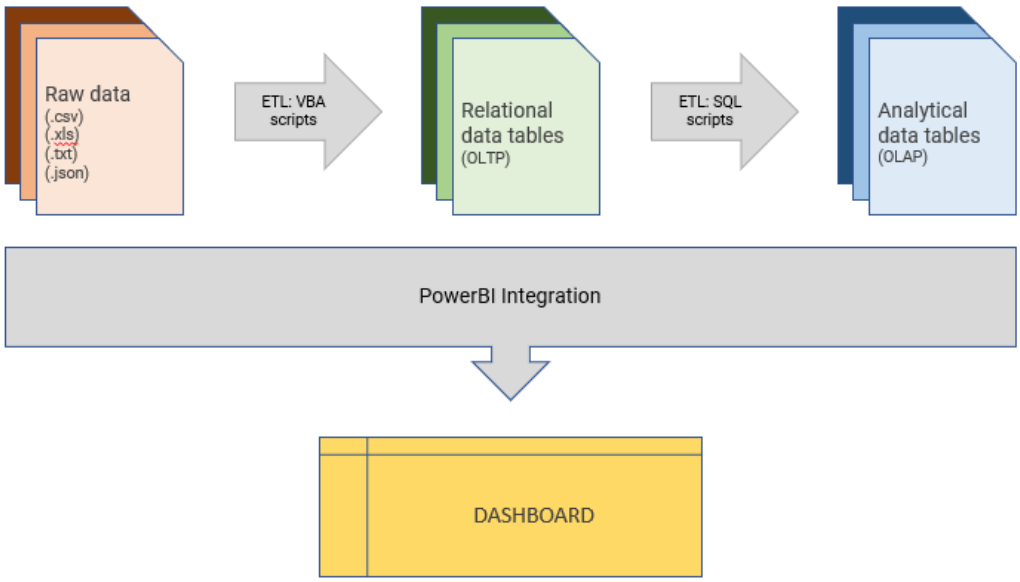

Figure 1. Overview of the ETL process.

As shown in Figure 2, the Data Warehouse design was conceived in the Third Normal Form (3NF) configured in a snowflake schema, composed by:

a. Four facts tables: "Fact_Demografia" (Demography), "Fact_Crime" (Crime), "Fact_Território" (Territory) and "Fact_Segurança" (Security);
b. And seven dimension tables: "Dim_Ano" (Year), "Dim_SocEcon" (Socioeconomic), "Dim_Crime" (Crime), "Dim_UsoSolo" (Land Use), "Dim_Postos" (Stations), "Dim_Mun" (Municipality) and "Dim_Freg" (Parish).

In the dimensions tables the attributes and metadata were stored for each component of the data warehouse, while in the facts tables the metrics and indicators were stored, aggregated by type of transaction.

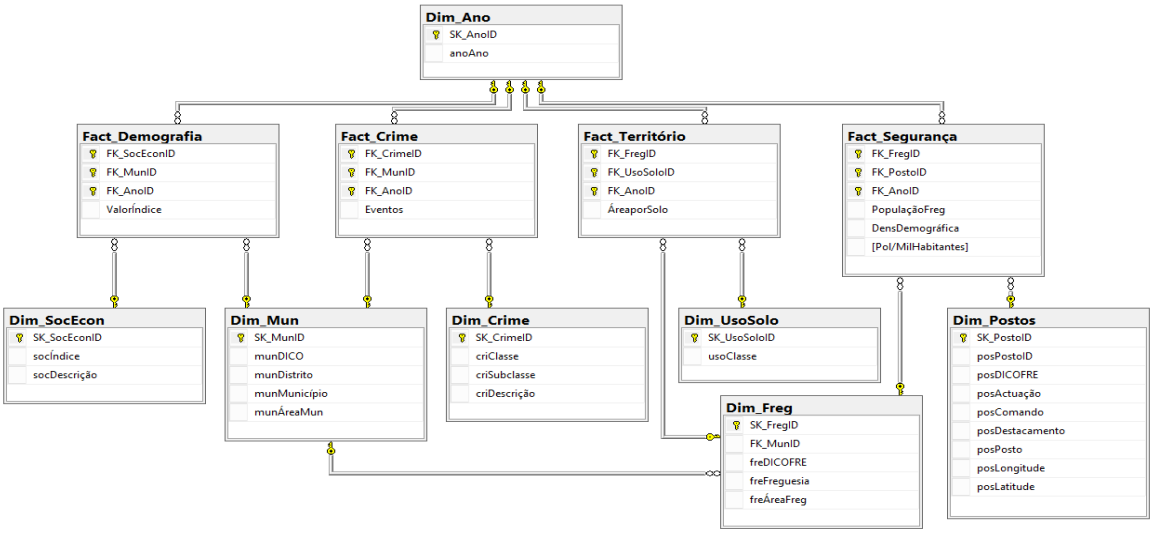

Figure 2. Data Warehouse design: Dimensions and Facts.

Having the DW structured and populated with the desired data model, the integration to Microsoft PowerBI was made by using a TCP/IP connection to the local SQL Server host. Once connected, several dashboards and visuals were created by leveraging ArcGIS, R and DAX capabilities, in order to highlight the most relevant findings gathered by the SIM4SECURITY research group (see Figure 3).

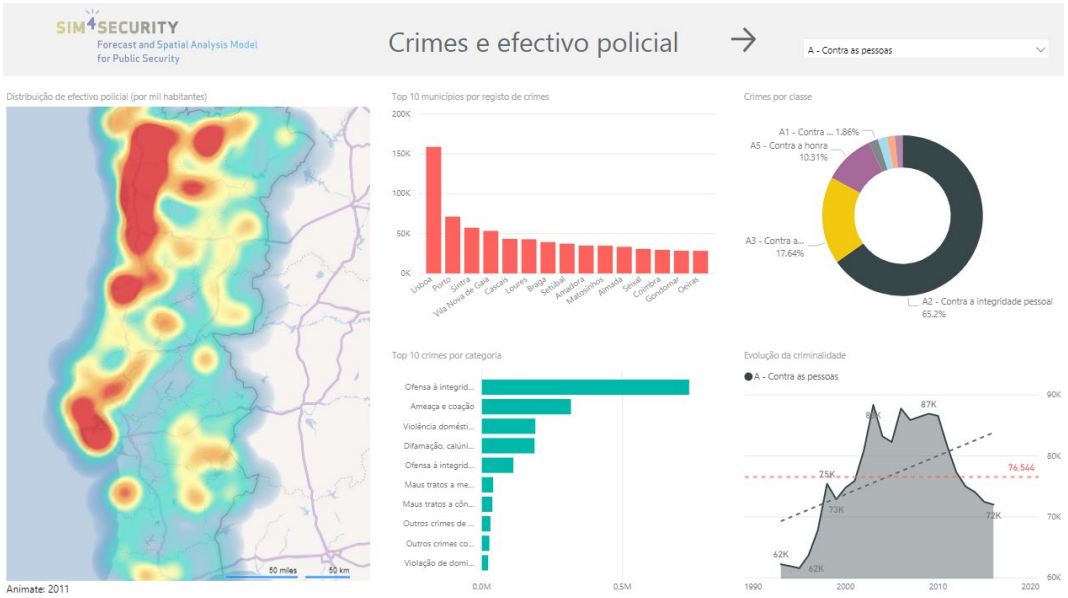

Figure 3. PowerBI Dashboard implementation: crime and security coverage.

CHAPTER IV: DATA ANALYSIS AND FURTHER MODELLING

The following steps and considerations were made throughout the data processing and modelling process:

4.1 Extracting index values from a multidimensional model

The index and their corresponding values were structured for an OLAP model. That means, for optimizing analytical tasks, each individual index was assigned to a unique identifier within the attribute "socDescrição" as part of the dimension "Dim_Mun". In order to concurrently process and display these indices and their corresponding values, a matrix transformation was performed by using the "Matrix" visual in PowerBI and DAX expressions;

4.2 Evaluate the relevance of each index

Each index was analyzed as how they contributed to the occurrence of crime events by location and time range. Finding the most relevant contributors to our target variable is a necessary step for selecting the optimal number of features according to their discriminative power in our analytical model. The following assessments were initially proposed:

- Pearson's correlation coefficient and correlation matrix: estimate the linear association between two variables. In this case, an independent and the target variable;
- Granger's causality: discriminatory power of an independent variable towards the target variable, given that the two variables are a time series. However, given that several variables did not present a consistent time series structure throughout the dataset, this analysis was ruled out;
- Distribution test: evaluate the distribution for the most relevant variables under study.

The aforementioned measures gathered in the dashboard were created using R and DAX scripts in PowerBI.

As shown in Figure 4, population size presents an outstanding positive correlation with reported crimes (Pearson's correlation coefficient = 0.96). Thus, as population rates increase, crime rates follow suit. In addition to that, both variables present a log-normal distribution, positive skewness and kurtosis, having over 75% of the sample below the mean of the sample (see Figure 5).

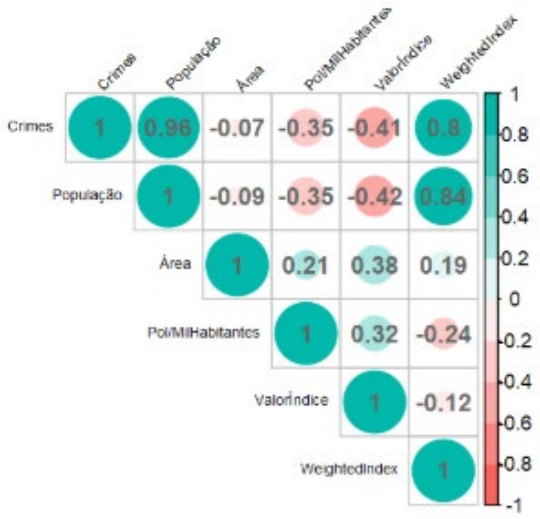

Figure 4. Correlation matrix: Correlation among main indicators and indices.

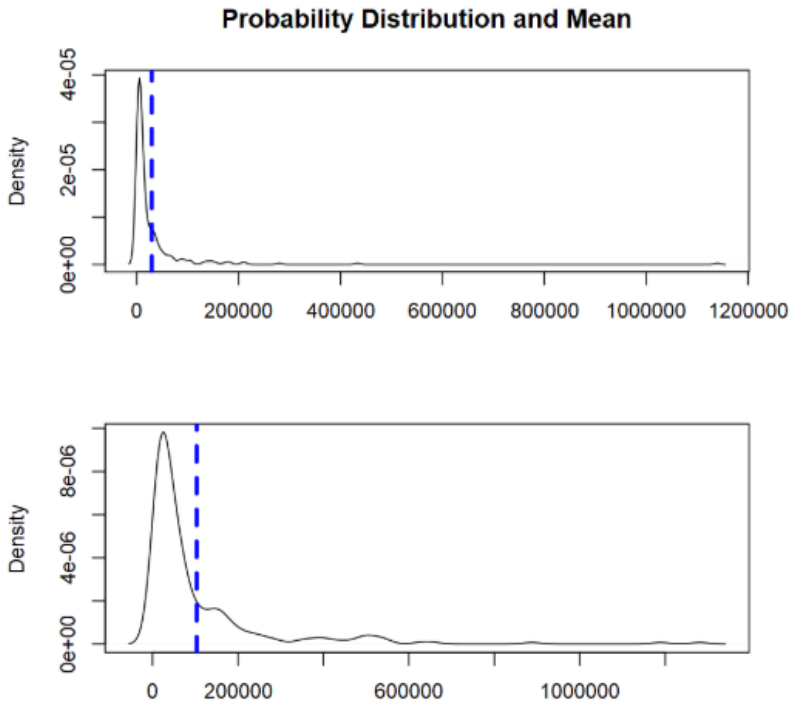

Figure 5. Probability Density plot: Crimes (top) and Population (bottom).

Therefore, it is proposed to use this measure as a primary weight for normalizing the indices obtained throughout this project. In that sense, we will have a better grasp of the occurrence of crime in each municipality, the overall security level and the effectiveness of allocated resources and policies.

However, two potential issues were observed and considered when handling the "Population" measurement:

1. After normalizing the indices by population, they acquire a similar distribution and generate a set of new correlations which once were not relevant or meaningful;
2. This variable takes into consideration permanent residents only. Areas subject to high, seasonal influx of tourists and temporary residents could present an abnormal behavior, that is, an elevated crime rate in contrast to the total population.

Following population, the most correlated socioeconomic indicators were: "Daily purchasing power per capita" (0.68), "Percentage of population with a university degree completed as education level" (0.62), "Averaged annual amount of retirement pensions" (0.55), "Average monthly earnings" (0.50), "Percentage of population without formal schooling" (-0.43) and "Percentage of population with elementary school completed as education level" (-0.45).

4.3 Define and implement KPIs

The main challenge on setting up KPIs is to infer quality measures at the municipality level as it was not feasible to define a reference measurement for performance which could be objectively quantifiable. In other words, the type of analysis performed throughout this project aims to represent relative security levels (e.g. municipality "A" has a more efficient and/or sufficient security system than municipality "B"). Conversely, stating that a certain municipality is safe in absolute terms or establishing a "gold standard" for evaluating performance could be a misleading assumption. Therefore, four different approaches were used for assessment:

1. Long Run Average and Simple Linear Regression: crime data was gathered to estimate the historical average of crime events and along with a simple linear regression, using the crime records over time as the single regressor, it was possible to find patterns in the security landscape and predict the occurrence of crimes in the coming years. However, it will be further discussed why these methods are not sufficiently reliable for predicting time-series events;

$$Long\ Run\ Average_j = \frac{\sum_{i=1993}^{k} crime_{ij}}{(k-i)+1}$$

2. Crime rates normalized by population: the outstanding correlation observed between population and crime occurrence called for an additional normalization step. By doing so, it was possible to assess the contribution of other intrinsic factors, such as socioeconomic indicators, on the occurrence of crime throughout the country;

$$Crime\ per\ Thousand\ Inhabitants_{ij} = \frac{crime_{ij}}{pop_{ij}} * 1000$$

3. Improvement over Downturn year: each municipality was assessed individually using its own historical data as reference. A Downturn period was defined for each municipality which consisted in the year with the highest amount of crimes in record. Then, the Downturn period was compared to the crimes reported for the latest year in our dataset (i.e. 2016) in order to assess the degree of improvement or degradation in comparison to the Downturn year (see Figure 6);

$$Crime\ Downturn\ Ratio_j = \frac{c_{kj} - max_{i=1993}^{k}(crime_{ij})}{max_{i=1993}^{k}(crime_{ij})}$$

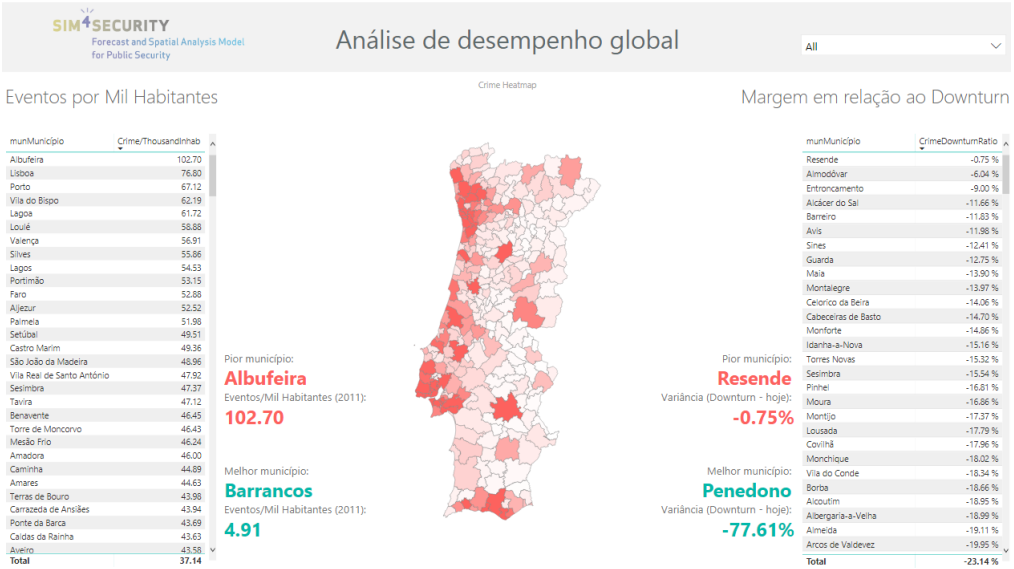

Figure 6. Global Analysis: Crime history variance and population-weighted crime rates.

4. Ranking system according to the socioeconomic indices of each municipality: given the diverse nature of the data comprised by the socioeconomic indices (several combinations of ratios, scales and data ranges), these measures were converted into ranks according to their corresponding value for each municipality, ordered from the highest to the lowest index values. This transformation process would allow a more simplistic and efficient approach of measuring relative performance, avoiding standardization while keeping interpretability (see figure 7).

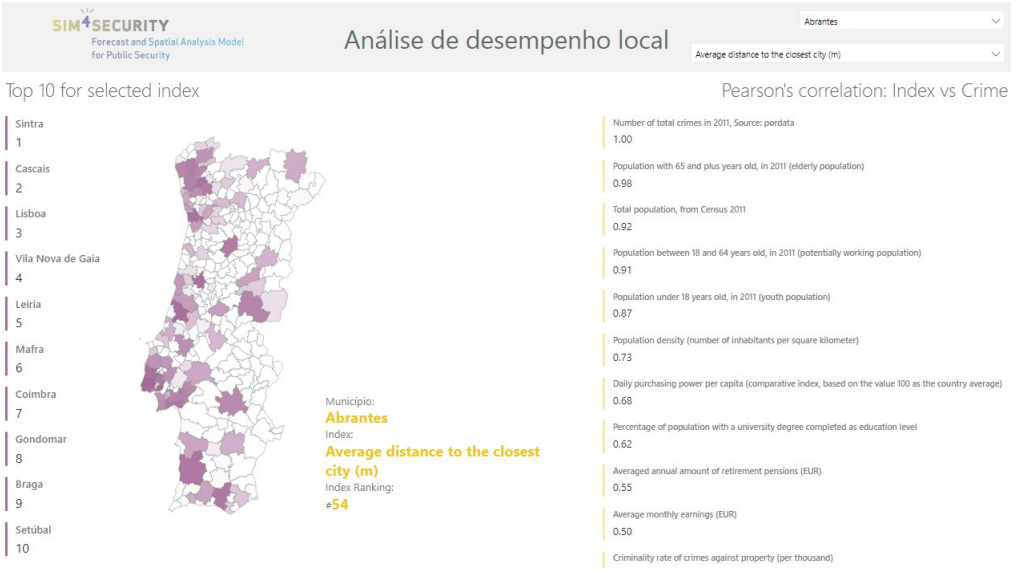

Figure 7. Local Analysis: Ranking system for socioeconomic indices

The performance indicators resulting from the four approaches allow us to look into the observed data points from the municipality, district and country perspectives. For their implementation,

the measurements available in the DW were manipulated and aggregated by using DAX expressions, a native feature on PowerBI for manipulating dynamic arrays and aggregating data.

CHAPTER V: CONCLUSION

The distribution optimization of police forces along Continental Portugal territory was the main goal to be achieved in SIM4SECURITY project. To achieve this goal a GIS model was elaborated along with the development of future demographic and crime scenarios. The amount of data that was collected from different sources represented a challenge in which for being able to have insights for decision making purposes, this data had to be subject to an ETL process. A data warehouse was then elaborated, where the data collected was stored. This data warehouse is very important in various aspects namely to provide interconnection between the various dimensions and metrics of the variables in study and to feed the dashboards developed in the framework of the SIM4SECURITY project.

The resulting dashboards allowed a better understanding of the relations between crime and demographic variables and showed that they are a very useful tool that could aid in the decision-making process to optimize the allocation of security forces along the Continental Portugal territory.

The key takeaway points observed by these reports are:

- Crimes and population are in constant decline over the past years. Population is expected to shrink 8.5% by 2040, while crime is down by over 23% since 2008 (downturn year), close to reaching an all-time low (considering the historical data collected for the past 25 years). These findings reinforce the high correlation shared between these two variables.
- GNR's security forces are mostly concentrated in the countryside, especially in northern Portugal (measured by one police officer per thousand inhabitants). Given that over 95% of the country's land usage refers to agriculture and forests and the urbanization rate is relatively low, this phenomenon could be caused by the higher pace to which non-urban areas are being "depopulated".
- Urban areas are more likely to have higher crime rates. They also present higher purchasing power per capita and education levels. Meanwhile, unemployment and school dropout rates have shown no correlation to the crime phenomena.
- Over 57% of all crime records in Portugal refer to crimes against property, also being shown as the top contributor throughout the years when compared to other crime categories.
- When analyzing the occurrence of crimes weighted by the population for each municipality, it was observed that areas initially with a low contribution to the overall crime records at country level were ranked to the top of the list (worst performers). It's assumed that areas subject to seasonal influx of tourists and/or temporary residents, which are not reported as part of the local population, yield weighted crime rates much higher when compared to non-weighted crime rates. Nonetheless, this phenomenon could be interesting when tailoring specific policies for improving the security coverage in these areas.

Despite the interesting insights provided by the dashboards, the data gathered to elaborate them presented several challenges which remained unsolved, namely:

- The crime data is disaggregated only until municipality level, while population data is disaggregated to parish level;

- The data regarding security forces is limited to the GNR, that have their main actuation areas in rural zones;
- The DGPJ database is not consistent by region and has several missing values regarding certain crimes/years, limiting further data modelling and time-series analysis (as identified in chapter 3, while implementing Granger's causality).

Overcoming these limitations would certainly improve the insights obtained from the dashboards. At the current state, the analytical capabilities of the dashboard are a good starting point but are limited to a more general/broad analysis for decision-makers and still lack in proposed actions and/or policies regarding crime and its underlying causes. With the latest AI features released for PowerBI over the recent months, implementing predictive modeling could prove to be a convenient and powerful decision support tool for improving the depth of the framework herein developed.

For future developments it would be interesting to have a finer granularity for crime records at parish level from both national Portuguese police forces (i.e. PSP and GNR) and a complete history of the security resources allocated for each region security at parish level from the aforementioned stakeholders. Furthermore, additional data regarding tourists and temporary residents could also provide a deeper understanding of crime events in particular areas of interest in the country. With this additional data it would be possible to perform a deeper analysis at local level and, ultimately, improve the decision support process while deploying an efficient location-allocation of security forces.

BIBLIOGRAPHY

Adam, F., Pomerol, J. C. (2008). *Developing Practical Decision Support Tools Using Dashboards of Information*, in Handbook on Decision Support Systems 2, Springer, Berlin, Heidelberg.

Bravo, J. (2016). *Projecções de População Residente a Nível Concelhio – Metodologia.* Universidade Nova de Lisboa, Information Management School, Novembro 2016.

Few, S. (2006). *Information Dashboard Design: The Effective Visual Communication of Data.* O'Reilly.

Global Peace Index 2018. Retrieved from Institute for Economics and Peace (IEP). (http://economicsandpeace.org/).

Lewis, J. (2018). *Economic Impact of Cybercrime – No Slowing Down.* Retrieved from https://www.mcafee.com/enterprise/en-us/assets/reports/restricted/rp-economic-impact-cybercrime.pdf.

Paine, K. D. (2004). *Using Dashboard Techniques to Track Communication.* Strat Comm Manage, 8, 5, 30-33.

Rodrigues, T. (2014). *Population dynamics. Demography matters*, in Globalization and International Security. An overview, NOVA Publishers, New York, 57-74.

Rodrigues, T., & Painho, M. (2018) Modelos Preditivos e Segurança Pública. Fronteira do Caos Editores Lda, 17.

Teixeira, N. S., Lourenço, N., & Piçarra, N. (2006). *Estudo Para a Reforma do Modelo de Organização do Sistema de Segurança Interna. Relatório preliminar,* IPRI.