Universidade de Lisboa
Faculdade de Ciências
Departamento de Engenharia Geográfica, Energia e Geofísica

Spatially: A Spatial Analysis Web GIS Prototype System

Alexandra Dias

Projeto
Mestrado em Engenharia Geográfica
2014

Universidade de Lisboa
Faculdade de Ciências
Departamento de Engenharia Geográfica, Energia e Geofísica

Spatially: A Spatial Analysis Web GIS Prototype System

Alexandra Dias

Projeto para obtenção de Grau de Mestre orientado por:
Prof.ª Doutora Cristina Catita (Faculdade de Ciências da Universidade de Lisboa) e
Prof. Doutor Michael Flaxman (Geodesign Technologies Inc.)
Mestrado em Engenharia Geográfica
2014

Abstract

Geographic information is increasingly more present in several areas of knowledge, resulting in the fact that around 80% (Zhang et al., 2010) of the databases include spatial information.

Processes of visualization and application of algorithms to spatial data are already common in the daily routine of any internet user.

The accentuated development of information and communication technologies have resulted in the appearance of increasingly more efficient technologies with the capacity of dealing with big data volumes with a geospatial component. Despite the effort over the last years in the development of Open source technologies related with this area, options that do not require the installation of software and that are accessible to a diverse range of publics are still lacking.

In Portugal, Census represent the biggest source of information about population, family, housing available having a defined spatial character and comprehending information with potential interest in a wide range of investigation themes.

As a result of the above, what is proposed in this project is the creation of a platform on the internet that allows the users to perform spatial analysis over their and/or a default data set from Census made available by the platform. The methods to be applied include visualization, spatial autocorrelation analysis and building spatial regression models over a defined dataset. Besides the evident functionality to achieve it is also pretended to gather an overview over the open source technologies available for the construction of a Web Gis (from the database to the geospatial data server and the display and analysis of the mentioned data) as well as their potentialities and limitations at the moment. Finally it is aimed to integrate three elements: an understanding of spatial analysis methods to be applied to geospatial data in areal sections, an exploration of the technologies available for the proposed goals mainly within the Open Source software area and the definition of system architectures according to the proposed objectives.

Key - words: Spatial Analysis, Web Architecture, Open Source, Census, Web-based GIS

Sumário

A informação geográfica está cada vez mais presente em diversas áreas de conhecimento, sendo que cerca de 80 % das bases das bases de dados inclui informação espacial. Processos de visualização de informação geográfica e aplicação de algoritmos com componente geográfica são já comuns no dia-a-dia de qualquer utilizador da internet.

O acentuado desenvolvimento das tecnologias de informação e comunicação tem resultado no aparecimento de tecnologias cada vez mais eficientes, e com capacidade de lidar com grandes volumes de dados com componente geoespacial. Apesar dos esforços dos últimos anos no desenvolvimento das tecnologias de código aberto e licença livre (*open source)* relacionados com esta área, são ainda escassas as opções que não impliquem instalação de *software* e que sejam acessíveis a diversos públicos.

Os Censos em Portugal representam a maior fonte de informação sobre população, família e habitação disponível, tendo esta um carácter espacial muito definido e compreendendo informação com potencial interesse para diversos tema de investigação.

Assim, o que se propõem neste projecto é a criação de um protótipo de plataforma na web que permita ao utilizador fazer análise espacial sobre os seus dados e/ou um conjunto de dados dos censos disponibilizado pela plataforma. Os métodos incluem visualização, análise de autocorrelação espacial e criação de modelos de regressão espacial com base nos dados mencionados. Além da evidente funcionalidade a atingir, pretende-se ainda fazer um levantamento das tecnologias (*software* e ferramentas) *Open Source* disponíveis para a construção de um Web Sig (desde a base de dados, ao sevidor de dados espaciais e disponibilização e análise dos mesmos), bem como das suas potencialidades e limitações actualmente. Finalmente, pretende-se fazer uma integração de três elementos: uma compreensão dos métodos estatísticos de análise espacial referidos, uma exploração das tecnologias (ao nível de software) *Open Source* disponíveis para os fins definidos e a definição de estruturas de arquitectura de acordo com a finalidade proposta.

Palvras-chave: Análise espacial, Arquitectura web, Open Source, Censos, WebSig

Resumo

A informação espacial tem cada vez mais expressão numa sociedade com uma crescente tendência para a tecnologia, sendo inclusivamente referido pela literatura que cerca de oitenta por cento da informação contida em bases de dados corporativas contém uma vertente espacial (Zhang et al., 2010).

A evolução das capacidades computacionais, e a sua passagem para um ambiente virtual tem-se revelado de extremo interesse para a indústria tecnológica com reflexo na utilização de ambientes *cloud* para o devido efeito em diversas aplicações na *internet* (Song et al., 2010; She et al,. 2012; Anselin et al., 2004). Associando a estes factores a recente tendência das comunidades tecnológicas para o *software* de código aberto e licença livre, obtém-se o ambiente perfeito de aprendizagem para um projeto de mestrado.

Este projeto nasce da escassez documentada pela bibliografia (Zhang et al., 2010; She et al., 2012) de ambientes *online* que permitam fazer análise espacial sobre os dados de forma intuitiva e com instrução de acompanhamento. Propondo-se neste contexto a realização de um protótipo de sistema de SIG na internet com incorporação de funcionalidades de análise espacial.

Os objectos definidos para este projeto compreendem:

- A compreensão e exploração de métodos e aplicações de análise espacial;
- O desenho da arquitectura de um protótipo de um sistema *Web SIG*;
- A exploração das opções software de código aberto disponíveis para a disponibilização de serviços SIG na internet;
- O desenho e o preenchimento de uma base de dados espacial;
- A produção de um protótipo de *website* de acordo com as conclusões dos tópicos anteriores;
- A exploração e implementação de funcionalidades de análise espacial num *website* com ferramentas interactivas;
- A implementação de um sistema de relatório de forma a produzir um documento PDF com os resultados dos métodos de análise espacial aplicados.

Com vista a permitir que esta plataforma tenha aplicabilidade não só entre a comunidade científica, mas também entre utilizadores comuns, disponibiliza-se, além da opção de utilização dos próprios dados através da transferência do ficheiro de dados

próprios, a opção de definir de entre um dos conjuntos de dados disponibilizados (dados referentes ao Censos 2011 disponibilizados pelo Instituto Nacional de Estatística (INE)).

O projeto está definido em três fases principais (a consultar na Figura 1 do referido documento) que consistem sucintamente em:

- Fundamentação teórica - compreendendo conceitos, aplicações e métodos tanto de análise espacial como de *Web* SIG;
- Implementação da plataforma – abarcando a aplicação dos conceitos definidos na fase anterior e incluindo grande parte do processo de implementação da plataforma (desde a colecção de dados até à implementação da base de dados e posteriormente construção da aplicação per se);
- Definição de módulos, articulação da estrutura – onde se integram todos os componentes integrantes da plataforma com a sua total funcionalidade. Esta fase inclui o desenvolvimento dos módulos de análise espacial, bem como de funcionalidades que estendem o protótipo base, tal como a funcionalidade de *login* e os formulários de interacção entre o servidor e o cliente.

A análise espacial é uma área com interesse em inúmeras aplicações amplamente referidas na bibliografia (Beale et al., 2010; Rey, 2007; Druck et al., 2004). Ciências Sociais, Biologia, Criminologia, Epidemiologia (...) são algumas das áreas que fazem extensivo uso da análise espacial, sem terem, contudo, formação específica para lidar com dados com uma componente espacial.

Assim, são apresentadas três etapas de análise espacial (a implementar no protótipo) que contemplam a visualização dos dados, a análise exploratória dos mesmo (englobando a análise de autocorrelação espacial) e a regressão espacial.

Estes três conjuntos de métodos estão encadeados num processo cíclico, e requerem uma interpretação informada. Um dos factores contemplados no capítulo 2 do documento está relacionado com a conceptualização do problema em causa, que é um dos maiores desafios do processo de análise espacial.

A definição da estrutura do protótipo é definida de acordo com (Mao 2005). Neste sentido definem-se os quatro componentes essenciais da arquitectura do protótipo do sistema de *Web* SIG : Base de Dados, Servidor, *Web Framework* e Linguagens de programação associadas (tanto no lado do servidor como do cliente).

Numa comparação superficial sobre o *software* de código aberto disponível, com uma forte base num estudo de Ballatore (Ballatore et al. 2011) que propõem uma comparação com base nos parâmetros definidos pelo utilizador. Os projetos *open source* considerados neste documento têm já como pré-requisito o facto de serem direccionados a dados com componentes geospaciais. Desta investigação resulta a definição do *software* e das tecnologias a utilizar na implementação do protótipo:

PostgreSQL (com extenção *PostGIS*), *GeoServer*, *Django* (extenção *GeoDjango*), *Python* – como linguagem de programação no lado do servidor, e *Javascript*, HTML, css como linguagens de *scripting* do lado do cliente.

A implementação do protótipo é feita com base em dados da Carta Administrativa Oficial Portuguesa (CAOP) do continente disponibilizada pelo INE, sendo os dados processados de acordo com os requisitos do sistema e agrupados em áreas de diferentes dimensões (freguesias, concelhos e distritos), com fim a serem inseridos na base de dados. As operações às quais a informação geográfica foi sujeita podem ser consultados de forma esquemática na página 46, sendo que o esquema final em Modelo de dados OMT-G de bases de dados geográfica na figura 20.

O *interface* do utilizador é também definido nesta segunda fase do processo onde se apresenta ferramentas para tal efeito. Este tópico, é no entanto considerado menos relevante no âmbito do projeto, e sendo consequentemente abordado de forma mais superficial.

A terceira fase do projeto compreende o desenvolvimento dos módulos em *Python*, integrando uma lista de bibliotecas da mesma linguagem, cuja estrutura e dependências é apresentada na figura 15. No capítulo 5 definem-se os módulos a construir e as funcionalidades a implementar nos mesmos, definindo-se ainda as variáveis de entrada e saída de cada função. Funcionalidades como o *login* e os formulários são também documentadas nesta etapa, que resulta na implementação do protótipo de sistema final com todas as funcionalidades propostas.

Os relatórios produzidos são finalmente apresentados, sendo os resultados organizados num ficheiro PDF com informações referentes aos dados e às análises efectuadas.

O projeto apresentado, apesar de contemplar algumas áreas de elevado interesse: tanto no meio académico como na indústria, teve bastantes limitações relacionadas maioritariamente com assuntos de carácter técnico. Dentre os quais se destacam as dificuldades de instalação de bibliotecas, os obstáculos na definição da arquitectura de forma a servir a informação *online* e os recursos em termos de servidor. Durante o

projeto, foram várias as abordagens relativamente a estes assuntos (especialmente no que toca à configuração do ambiente *Python* e o serviço dos dados), que foram sendo ultrapassados recorrendo a servidores externos.

De uma maneira geral, o projeto atinge os objectivos a que se propunha, tendo no entanto algumas limitações em termos de funcionalidades e uma generosa lista de trabalho proposto que no âmbito da disciplina não faria sentido desenvolver. As potencialidades computacionais deste tipo de projeto podem ser variadas, conforme os recursos disponíveis no servidor, sendo inclusivamente proposta a passagem do ambiente para um servidor com mais recursos, e mais flexível que permita uma interacção entre utilizadores numa comunidade de dados espaciais com capacidade de análise e investigação sobre dados de outros cientistas.

Em termos de *software opensource*, conclui-se que a sua utilização depende em grande parte da utilização pretendida, e que apesar da considerável documentação, problemas técnicos implicam necessariamente mais tempo do que num caso de *software* comercial em que há a possibilidade de apoio técnico. No entanto, e no caso deste projeto, a sua flexibilidade permite a execução de funcionalidades que dificilmente seriam implementadas com código fechado. Revelando-se particularmente interessante devido à flexibilidade de implementação e tendo, para o efeito requerido, sido sobejamente eficiente.

Palvras-chave: Análise espacial, Arquitectura web, Open Source, Censos, WebSig

Acknowledgements

No one is an Island, and we are all a sum of everything we choose to absorb from the people who pass us by.

With this declaration I must be thankful to my parents, the most important people in my life who made me the complete person I am today. I am also very thankful for my siblings: friends, accomplices and companions who are so very present in every aspect of my life and to whom I look up to for many reasons.
My grandparents, whose lessons and ideals will always accompany me as a great part of my character. To my friends and family, whose presence and support made this journey a lot easier, especially João for being so comprehensive and supportive.

I am especially thankful to Prof. Dr. Cristina Catita, for giving me the space to grow while providing support, orientating me in the right direction and for teaching me that most of the times the journey will teach us far more than the destination. To Prof. Dr. Michael Flaxman, for giving me such an amazing opportunity and for the orientation regarding technologies and to Eng. Rita Semedo, my internship coordinator, for being so flexible and comprehensive during the thesis process.

Furthermore, I dedicate this work to my grandfather, whose kindness, strong character and ethics will always be very present in my journey.

List of Figures

List of tables

Acronyms

GIS – Geographic Information Systems

SDI – Spatial Data Infrastructure

OS – Open Source

EDA – Exploratory Data Analysis

ESDA – Exploratory Spatial Data Analysis

ESRI – Environmental Systems Research Institute

MAUP – Modifiable Areal Unit Problem

OLS – Ordinary Least Squares

DBMS – Database Management Software

SQL – Structures Query Language

PK – Primary Key

OGC – Open Geospatial Consortium

HTML – Hypertext Markup Language

CSS – Cascading Style Sheet

GEOS – Geometry Engine Open Source

GDAL – Geospatial Data Abstraction Library

WAF – Web Application Framework

MVC – Model - View - Controller

URL – Uniform Resource Locator

CAOP – Carta Administrativa Oficial de Portugal (Portugal's Oficial Administrative Map)

EPSG – European Petroleum Survey Group

ETRS89 – European Terrestrial Reference System 1989

GRS – Geodetic Reference System

OGP - (International Association of) Oil & Gas Producers

IGP – Instituto Geográfico Português (Portuguese Geographical Institute

INE – Instituto Nacional de Estatítica (National Statistics Institute)

GIF- Graphics Interchange Format

JPEG – Joint Photographic Experts Group

IT – Information Technology

OSGEO4W – Open Source Geospatial For Windows

CI – Cyber Infrastructure

ICT – Information and Communication Technology

Index

"The application of GIS is limited only by the imagination of those who use it".

Jack Dangermo

1. Introduction

1.1. Research Objectives

Geographic Information Systems (GIS) have added many interesting options to the applications of Spatial Information. Data collection, storage, visualization, manipulation and analysis have been becoming less complex tasks throughout the years with the appearance of more developed and complex data infrastructures that propose themselves to accomplish the computational effort that all these processes involve.

This project has the objective of producing a simple tool for common users and scientists to analyze geospatial data, considering that the common user is not very familiar with this type of data nor with the spatial analysis functionalities.

The proposition consists in building a web infrastructure capable of receiving, analyzing and reporting on a specific input dataset or on a dataset collection made available by the platform, and according to a set of methods of spatial analysis defined by the user.

This kind of platform often receives the name of Spatial Data Infrastructure (SDI) and it is a data infrastructure that implements a framework of geographic data, metadata, users and tools that are interactively connected in order to use spatial data in an efficient and flexible way (Pascaul et al., 2012). According to Scholten et al. (2009), an SDI is a coordinated series of agreements on technology standards, institutional arrangements, and policies that enable the discovery and use of geospatial information by users and for purposes other than those it was created for.

More specifically this infrastructure aims at two specific users: Scientists who, being familiar with this type of data, are not very comfortable with the process of analyzing it, or the common user who will have a dataset available with statistical data of Portugal which can be explored according to the user's need.

In the first user's case, the input will be one of the most common spatial data types: shapefiles (.shp) (ESRI, 1998) and the user will be guided through a selection process in which there will be a theoretical aid on the selection of the analysis methods to perform.

The second user's case the platform will present itself as an intuitive tool, presenting the user with options regarding the data to analyze and the extent of the analysis, yet again with an assisted process of choice.

The methods proposed for this end comprehend data visualization, autocorrelation and spatial and non-spatial regression of data. The selection of spatial analysis methods to

be implemented was based on the spatial data to be analyzed by default (Census data (Instituto Nacional de Estatística, 2011)) which is distributed by administrative sections. Moreover, in an attempt to explore of the emerging Open Source philosophy and software, this study will be performed using mainly Open Source software, being one of the objectives to explore and present some useful tools towards the web GIS technologies within this area and contemplating a brief introduction to spatial data standards.

This study will not comprehend several technical issues related either to the database or to server configurations, it will however, focus on the Spatial Analysis to be performed, on the spatial data to be used and on facilitating and orienting the user's task when preforming such analysis.

The objectives of this project are listed below in a summarized way:

- Understanding and Exploring Spatial Analysis methods and applications ;
- Designing the architecture of a prototype Web GIS system;
- Exploring existent Open Source software options to provide GIS services online;
- Designing and populating a spatial database;
- Producing a prototype website according to the conclusions reached in the previous points;
- Exploring and implementing spatial analysis functionalities in the website with interactive tools;
- Implementing a report system to provide a report document (.pdf format).

1.2. Problem Statement

Desktop GIS now-a-days involves several spatial analysis components within a sophisticated and intuitive environment. Skilled professionals, environmental scientists and even the most curious common user can be attracted to this kind of software and make it useful in their own subject of study.

Spatial data has become a quite common data type and people are often drawn to websites or applications that include maps, maps' queries and several layers of connected information. There is no expertise required for a common user to deal with websites such as Google Maps, it is quite intuitive to discover that when a geographical place is searched, the map will display its location and even preform some spatial algorithms to allow the display of the shortest path for example. The common user

wants to have access to spatial information, even though the common knowledge makes it difficult to understand and integrate all the components that spatial data requires and that is one of the main motivations to develop this project.

Haining (2004) points out that the usefulness of operations on spatial data is dependent on how well reality that is represented on that data has been captured, so one of the goals of this project is to ensure that this fact is being comprehended by the reader, since the software available has little to no reference to this matter.

Census is a dataset with several applications, which can be applied to a wide range of study areas, it also possesses a temporal, spatial and institutional dimension that makes it very interesting within this project's scope.

In general this project has the following presumptions:

- Decision making implies huge difficulties that often result in expensive and long-lasting processes that are a consequence of the inexistence of adequate analysis tools.
- There is in general an insufficient capacity of integration and usage of existent spatial information.
- It is possible to develop an efficient platform which can be accessible and used by a diverse sample of users.

1.3. Methodology

The project was roughly divided in three main phases each one with some tasks and a resulting answer for a proposed question or outcome.

The referred information is presented below.

Phase 1: Analysis and Problem definition

- Literature Review for Spatial Analysis
- Literature Review for Web GIS
- Question: How can spatial Analysis be inserted in a Web GIS?

Phase 2: Design and Implementation

- Census Data : Collection, Exploring, Processing
- Database : Design and Implementation
- Technology exploration and research
- Web GIS design

- Result: Web GIS website setup.

Phase 3: Spatial Analysis Implementation

- Structuring and exposing spatial analysis methods
- Building Spatial Analysis Modules
- Implementing a report capability
- Result: Web GIS with Spatial Analysis capacity, login and input option.

The figure 1 below structures the defined phases of the project in a schematized way.

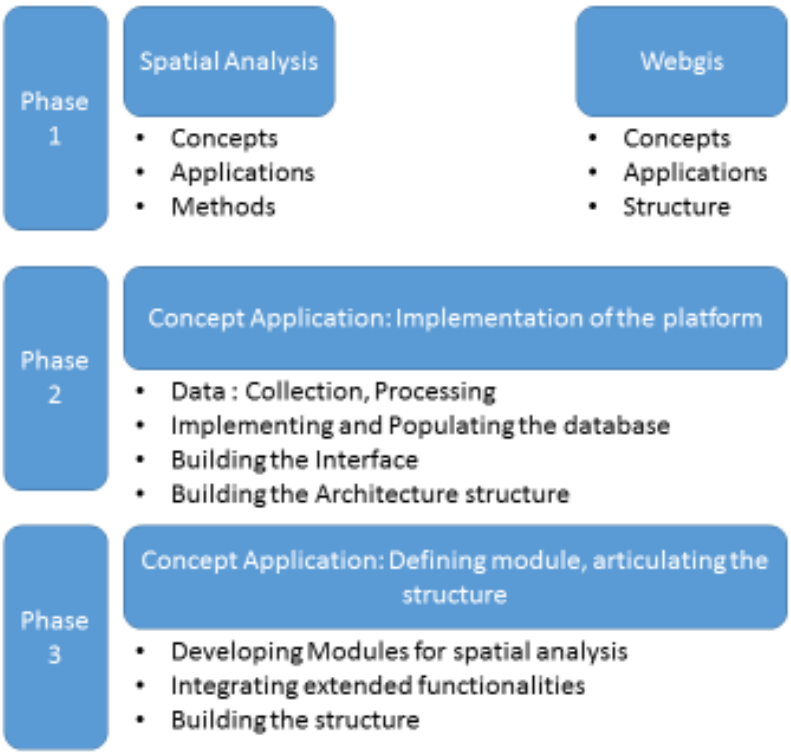


Figure 1 : Schematized summary of the phases of the present project.

1.4. Research Significance

In a fast paced technological era, several solutions start to appear to suppress the needs of the scientific community. The grown interest in Spatial analysis is recognized by the bibliography (Haining, 2004) as a reflection of the existence of well formulated questions or hypothesis in which location has a significant role, aligned with the sufficient quality of the spatial data and the appropriate statistical methodology. Spatial data is becoming increasingly more useful as scientists start to discover its potential,

leading to an increasing request for tools which can efficiently manage and analyze this type of data. It has also been estimated that about eighty percent of all data stored in corporate databases has a spatial component (Zhang et al., 2010). She et al. (2012) refers that despite the research done to put spatial data analysis functions online, there is still a gap between the spatial data and the analysis procedures.

Even though this type of software is to be used by several types of users (most of them with a limited knowledge of geospatial information or GIS), an urgent need for feasible approaches is starting to emerge. Despite the awareness of the software industry for this expressive reality, there is little option available online with a significant simplicity and practical results that can fully satisfy the demands of a scientist without a technological background need. As new challenges for geographic information systems and spatial statistics from different disciplines arise with new technologies of data service and demanding need in spatial data sharing and advanced analyses (Zhang et al., 2010).The promotion of spatial intelligence is pointed by the mentioned bibliography as essential due to high benefits in high level decision making.

Furthermore, as the cloud environment starts to develop, and the server-computing technologies start to become capable of dealing with large datasets it becomes increasingly more interesting to allow this virtual environment to store and process these data that are often of considerable dimensions. According to (Kwakkel et al., 2012) the internet itself, with the onset of the semantic web, is increasingly becoming a distributed repository of diverse information, including information which is relevant for regional studies of science, technology and innovation. The same author refers the need to look for open source solutions, and loose frameworks from which a complete solution for analysis and visualization can be developed.

As a result of the mentioned problem, and with the full awareness that a modern Geographical Engineer is a hybrid between an information technology professional and a knowledgeable individual with a comprehensive perspective over geospatial data, geospatial analysis and interpretation, it started to make sense to apply all this skills into a complex project that may be suitable for overcoming the mentioned limitations.

1.5. Thesis Structure

The presented document will be structured in six main chapters.

The present chapter has the objective of determining the project to be executed and document the motivation behind it and the scientific contribution it can represent. Objectives for the project, expectations and intentions are reported within this chapter, along with the resumed structure of the document.

In the second chapter most of the Spatial Analysis related theory to support this project will be presented. Since Spatial Analysis will be the main task performed by the project's outcome it becomes necessary to explore both the basic concepts as well as some data related topics and spatial analysis methods to be applied during the implementation of the system.

Web-Based GIS will be explored throughout the third chapter with an overview of the existent technology, possible architectures and state of the art. In this chapter other projects' architecture will be taken into consideration with the objective of defining the most suitable approach towards this subject.

The Prototype and System Implementation will be the theme for the fourth chapter where the most practical part of this project will be explored. The workflow of the project will be presented in the beginning of the chapter and followed through the whole implementation with a more considerate explanation for all the steps from the data collection to the prototype's implementation. Obstacles and unexpected situations will also be included in this chapter that culminates in the basic implementation of the prototype.

The Project's outcome will be presented and discussed in the sixth chapter. The achieved result will be explored and analyzed. Modules' implementation will also be considered in this part of the project as well as the more thorough explanation for each application to be built for this project.

In the last part of this document some conclusions will be drawn regarding this projects' outcome. Some options for further development within this project's theme will also be exposed as well as an overview of the technologies used and methods implemented.

Figure 2, presented below, summarizes the topics covered in each chapter.

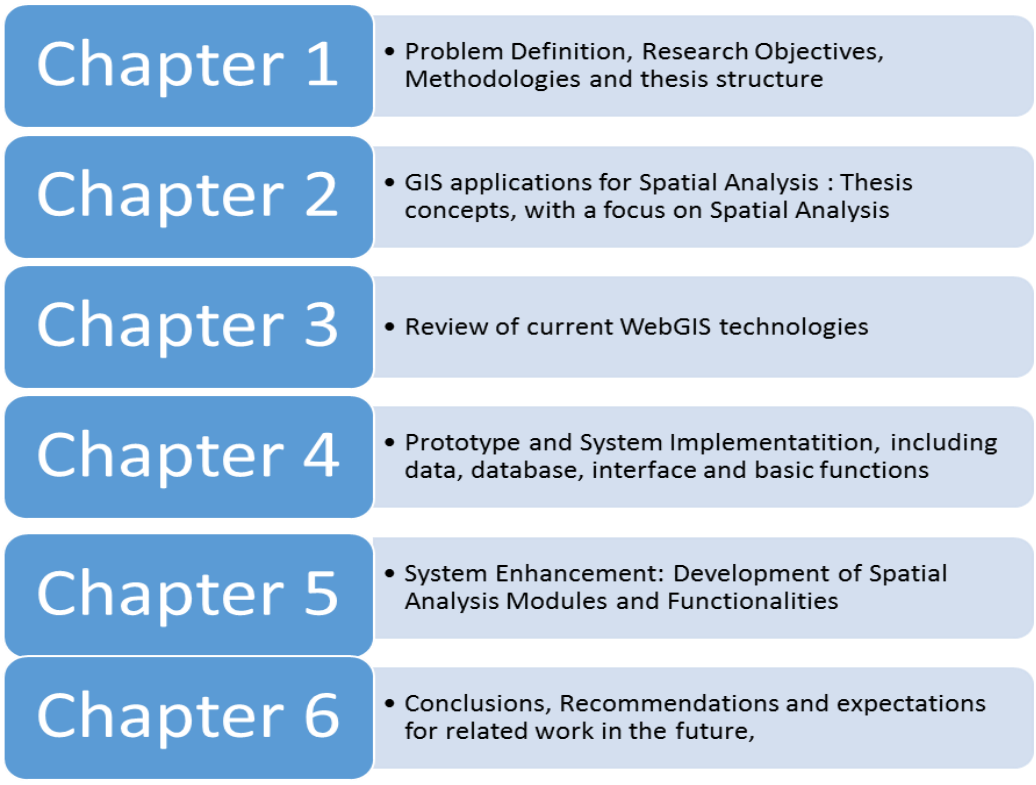


Figure 2 : Summary of the topics covered in each chapter of the present document.

2. GIS applications for Spatial Analysis

2.1. Thesis Concepts

2.1.1. Spatial Data

Geographic information concerns objects or phenomena with a specific location in space. Geographical data comprehends geometrical aspects (positions and dimensions), attribute data and temporal data (moment in time in which the data is valid). This logical division is intended to provide an answer to the questions Where? What? And When? Respectively. As can be consulted in figure 3 (Kraak & Ormeling, 2010).

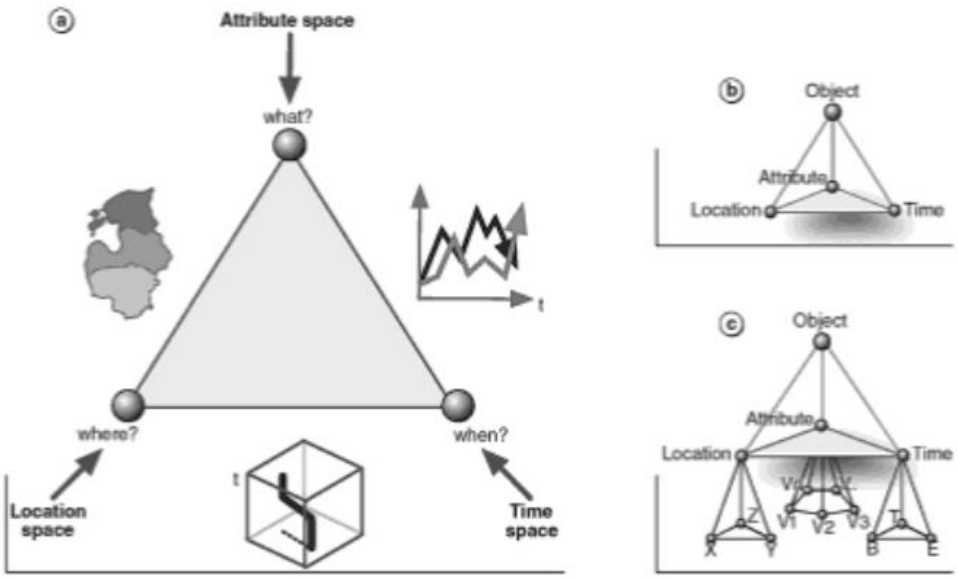


Figure 3 : The characteristics of geospatial data. a. Location, attribute and time, related with the elementary questions Where, What and When, b. the object view, c. detailed characteristics of data components (Source : (Kraak & Ormeling, 2010)*, p. 4)*

Geospatial information has several characteristics that can be further explored in bibliography (Fischer & Getis, 2010; Kraak & Ormeling, 2010; Haining, 2004; Kwakkel et al., 2012), nonetheless, they are summarized in the next paragraph.

Its scale can either be local (as for example analyzing the presence of a species in a specific region) or global (when evaluating sea level rise).

Time is also an important characteristic of geospatial data, since it can either be used to analyze historical data (millions of years to decades), present data (current distribution of an element) or to predict scenarios based on the historical data.

In order to preform data analysis accurately the data must be comparable and compatible. Some of the questions that must be considered when preforming analysis operations comprehend the date when the data was collected, in which way it was

collected and for what purpose. These questions make it possible to assure that the utility, reliability and accuracy of the data is compatible with the purpose it was collected for. Different purposes may require different answers to the above referred questions. Data nature defines data according to four different categories: point like objects, linear objects, areal objects and volumetric objects. Measurement units and whether the change is gradual or not (continuous or discrete phenomena's) will also influence spatial data analysis.

2.1.2. Spatial analysis

'Spatial Analysis' is a term that dates back to the 1950s and has a historical evolution that can be consulted in (Berry & Marble, 1968) (p 1- 9).

Even though the most significant advances in geospatial analysis were achieved in with the appearance of GIS, its principles are based on quantitative and statistic geography. Methods of spatial analysis are robust and capable of operating over a range of spatial and temporal scales.

Goodchild & Haining (2004) define it as the 'collection of techniques and models that explicitly use the spatial referencing associated with each data value or object that is specified within the system under study', its methods require assumptions about data, describing the spatial relationships or spatial interactions between cases.

Spatial analysis relies on the idea that there is a similarity to nearby attribute values in geographic space, this property was referred by Tobler (1970) as the 'First Law of Geography' and is mentioned by (Druck et al., 2004) as spatial dependency.

Cartography, Mathematical modeling and the development and application of statistical techniques are mentioned in the bibliography (Haining, 2004) as the three main elements of spatial analysis, underlining that it is of great importance that the spatial data is adequate to the question to be solved.

2.1.3. Conceptualization and data proprieties

Any process made upon spatial data requires a conceptualization of the real world, in this context it is necessary to identify the proprieties that are relevant to the application (Haining, 2004). It is also necessary to define the adequate presentation to the dataset in question: Level of spatial aggregation and geometric class (point, lines, areas or surfaces).

The measurement process is to be considered as the attributes and spatial measurements have to be as precise as possible and defined according to their application. More on scales can be read in the bibliography (Smith et al., 2013).

To resume, "Modeling geographic reality means the process of capturing the complexity of the real world in a finite representation so that digital storage is possible" (Haining, 2004).

Points, lines, areas and surfaces are the classes of digital objects for representing geographic phenomena, some examples of this are presented in the figure 4 below. Depending on the application of the data it is common to use spatial aggregation to analyze a phenomenon, this leads to a change in the mentioned class.

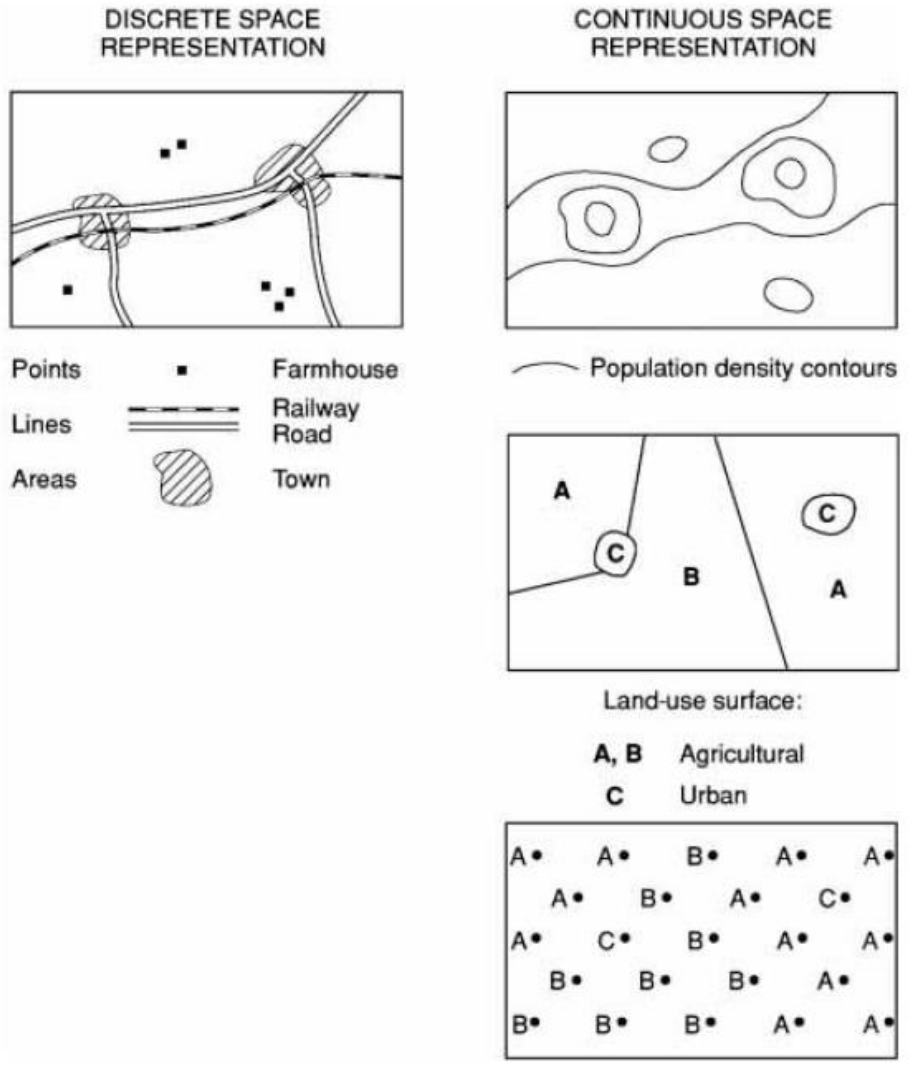


Figure 4 : Discrete and continuous space representations. (Source:(Haining 2004)*, p. 45)*

Attribute characteristics can refer to the spatial object themselves or to entities that are associated with or attached to the spatial objects but not directly dependent on them. Conceptualization comes as an essential part of spatial analysis as it refers to the definition and meaning of the attribute. Representation, on the other hand refers to how an attribute is operationalized into variable for the purpose of acquiring and storing data on the attribute and to enable analysis.

Analysis is undertaken on data collected with respect to one or more variables that measure attributes associated with geographic reality that is typically represented into the form of spatial objects. Conceptualization and representation issues of attributes are specific to each particular application.

Haining (2004) presents figure 5 as the schema that defines this process. According to this schema, the real world is represented through to the selected attributes, time and space which constitute the model. Model quality may be assessed in terms of the precision of a representation, its clarity, its completeness, its consistency and resolution. The data matrix (the data structure obtained in order to represent the real world through the mentioned model) will therefore have uncertainty related both to the data model and the data quality.

Buttenfield & Beard (1994) suggest the use of the term accuracy to reflect the correspondence between a representation or conceptualization and what the analyst wishes to measure.

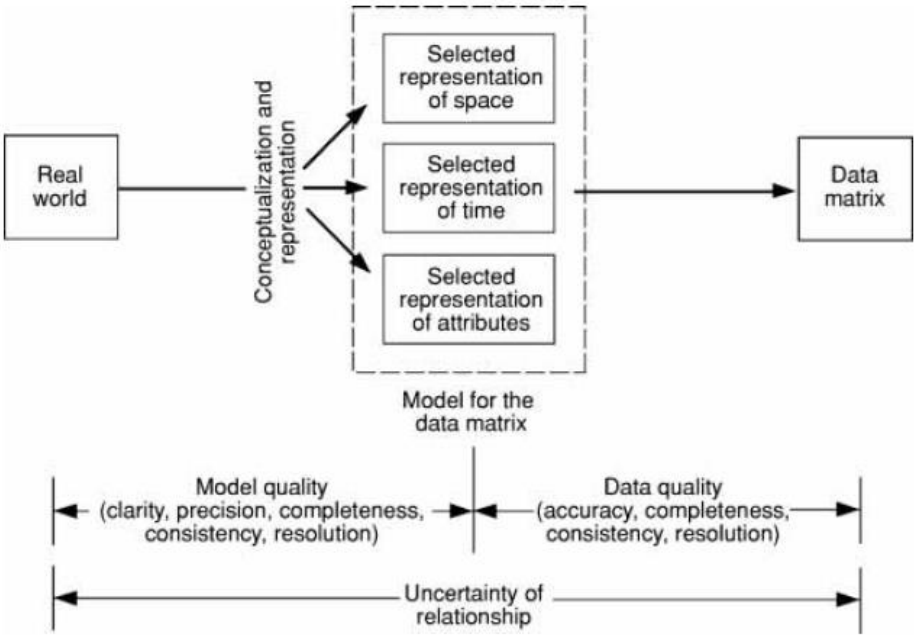


Figure 5 : Conceptualization and Representation, the relationship between the Real World and the Data Matrix. (Source: Haining, 2004)

Proprieties of spatial data (Druck et al., 2004):

- Spatial dependence is defined by the similarity of values for the same attribute measured at locations near to one another, which is bigger than amongst values separated by larger distance.
- If the similarity is constant through the whole area, the dependency structure is defined as stationary, if it varies across the map, it is non-stationary. In the absence of these situations, the structure of dependency is heterogeneous.
- The structure of a spatial dependency can display some differences across the axes (north/south, east/west). If the dependency structure is similar along both axes it is called an isotropic dependency structure, if it is not it is called non-isotropic or anisotropic.

2.1.4. Data Quality

The definition on data quality is defined by the bibliography as the performance of the data set given the specification of the model, as it depends both on the objective of the appliance of the dataset and the conceptual model to which the data was integrated. According to Salgé (1995), any assessment of data quality from the users' or producers' perspective relies on determining how closely data values represent reality given the chosen model for representing that reality.

Even though from an Engineering perspective there's usually a strong need for precise numbers and accurate approaches to this issue, this approach is far more reasonable when considering to the web environment and to situations where data precision may be far more flexible than for engineering purposes. Regardless the fact that there are several applications in which precision has to be rigorously defined, for this specific project the definition above will suffice, or even be more adequate.

There are several factors to consider when preforming spatial analysis over a dataset, these factors are summarized in table 1 according to the characteristic of data quality concerned and the phase of spatial analysis affected. Even though this topic will not be thoroughly explored within this project, it is of great importance to know the implications and approaches to different degrees of data quality which can be consulted in the literature (Haining, 2004).

Table 1 – Dimensions of data quality in relation to stages of spatial analysis. (Source: (Haining, 2004) *p. 178)*

	Accuracy	Resolution	Consistency	Completeness
Data collection and preparation of final database	Concerns about the presence of gross errors	Creating a common spatial framework for data collected on different frameworks	Incompatible data values (e.g. disease cases reported in areas without population)	Presence of missing data; need to interpolate or predict
Form and conduct of statistical analysis	Choice of error model. Need for robust and resistant statistical methods of analysis	Differences in variable precision across spatial units. Sensitivity of results to different methods of areal interpolation		Modelling in the presence of missing data.
Interpretation of results		Ecological bias. Forming invalid individual level hypotheses from aggregate analyses		Concerns about spatial variation in undercounting. Model misspecification due to the effects of missing variables

2.1.5. Areal Data Problems

Census data and other statistical data are often gathered in areal units for confidentiality purposes or statistical reasons. These areal units are usually delimited by closed polygons inside which it is presupposed to be homogeneous (Druck et al., 2004).

Aggregated data may be used to infer individual-level relationships. The lack of reliable individual-level data is often the cause of using areas to aggregate data.

For the Census case, for example, the target may be of areal level if analyzing the data for administrative sections, regarding the municipality management or the economic factors that are dependent on municipalities' investment on specific areas. For the Census data case, spatially defined groups may present some degree of homogeneity since as pointed out by Holt et al. (2010), 'individuals who live in the same area are exposed to common influences and as a result exhibit similarities individuals with similar characteristics choose to live in the same area'. However, aggregation bias is always present, given the rare existence, or inexistence of purely homogenous areal aggregates.

The modifiable areal units problem (MAUP) is explained by Holt et al. (1996, p. 181) as 'If a statistic is calculated for two different sets of areal units which cover the same population, or sample, a difference will usually be observed even though the same basic

data have been used in both analyses. This difference is cited as evidence of the modifiable areal units' problem'.

For this specific cases, neither the scale of the analysis (choice of number of spatial units) nor their particular configuration (the selected partitioning on zoning given the scale of analysis are fundamental and could, therefor be modified and thus the term 'modifiable'. The effects of it are addressed by Tobler (1989) and several approaches towards this problem are defined by the bibliography (Haining, 2004, pp.160-173).

In this project the aggregation level goes from Civil Parishes to Districts, having three levels of aggregation that can be used according to the user's object of study. The zoning of the statistical data is aggregated based on administrative boundaries, which are independently managed and have different premises for important areas such as education, family, housing and health.

2.2. Spatial Analysis Methods

Rosenberg (2011), defines four main categories according to the application of the spatial analysis methods. The categories presented bellow will be further explored in the next pages, where special attention will be given to modules to be implemented in the present project.

- Selection : Evolves all the processes from database navigation to display of simple cloropeth maps;
- Manipulation: comprehends all functions that create spatial data, it may be map algebra, geoprocessing, augmenting the capacity of analysis and correlation;
- Exploratory analysis: allows the description and visualization of spatial data by describing patterns of spatial association, suggesting the existence of spatial association, spatial instabilities and identifying atypical observations.
- Confirmatory analysis: includes several models of estimation and validation procedures.

There is a wide range of spatial analysis methods that can be applied to spatial data, depending on the type of data considered. In this particular project, the main data type is areal, as a result of it, the methods presented in this chapter will be limited to those applied for areal data analysis

2.2.1. Data Visualization

Graphic display of data aids on detecting data properties. Its visualization can enlighten the viewer (who may or may not be familiar with the data) about specific characteristics of the dataset, being part of a process of understanding and gaining insight into the data.
Buja et al. (1960, p.80) define the approaches to data visualization into rendering and manipulation. According to the author, the decision as to what to show in a plot and in particular in deciding what type of plot to construct is what rendering refers to.
Techniques for displaying distributions such as histograms, boxplot, Q-Q and rankit plots and time series (plots) are inserted within univariate data. As for multivariate data, scatterplots, traces and glyphs are mentioned techniques. How to operate on individual plots and how to organize multiple plots in order to explore the data is, according to the mentioned author, what defines manipulation. Tasks included in this approach include finding gestalt (identifying patterns, shapes and other proprieties in the data set), posing queries and making comparisons.

Graphical methods for vizualizing data

- Histogram

Histograms are often applied when there are a large number of observations. It consists of a graphical method for displaying the shape of a distribution. A frequency table is used to organize the occurrences o each valued and further along displayed in a bar chart (Lane, 2007).

- Box Plot

Box plot or whisker diagram displays the distribution of data based on five characteristics of the dataset: minimum, first quartile, median, third quartile and maximum. The interquartile range (IQR) is represented by the length of the box that is delimited by the first and third quartile. The median is represented by a segment inside the rectangle and the maximum and minimum values by the 'whiskers'.
A value is considered an outlier if it is 3xIQR above the third quartile or below the first and suspected outlier if it is 1,5xIQR above the third quartile or below the first (Lane 2007).
Table 2 resumes the graphical methods used for data visualization.

Table 2 – Summary of graphical methods for data visualization. (Source: (Haining,2004), *p. 194)*

	Univariate distributions	Bivariate distributions	Multivariate distributions
Categorical	Barcharts	Trellis plots; mosaic plots; glyphs (e.g. rays and trees)	
Quantitative	Boxplots; histograms; dotplots; quantile plots; rankit plots; residual plots; level and spread plots; time series plots	Scatterplots; Q-Q plots; level and spread plots; mean difference plots	Matrix and trellis scatterplots; parallel co-ordinate plots; projection pursuit; 3-D scatterplots; coplots; glyphs (e.g. Chernoff faces); grand tour

2.2.2. Weights

Weights take a crucial role in several areas of spatial analysis. The spatial matrix expresses, in general terms, the potential for interaction between observations at each pair i,j of locations for a spatial data set composed of n. The structure of these weights can be specified in various ways, and this structure is defined by the spatial weights matrix.

Conceptually, spatial weights define the diagonal (w_{ii}) of a n x n matrix to zero, while the other elements of the matrix (w_{ij}) reflect the potential of interaction.

There are three main types of weights, according to the elements taken into consideration to the definition of the weight values (Rey, 2013).

1. Contiguity weight matrices reflect the neighbors and weights attributes. There are three different criteria, depending on the definition of neighborhood:

- Rook, which takes as neighbors any pair of cells that share an edge;
- Queen, which includes the vertices of the lattice to define contiguities;
- Bishop, which designates pairs of cells as neighbors if they share only a vertex.

2. Distance based weights are generated by taking into consideration the distance between observations. This methods often considers a flat surface which implies that the data should be projected in advanced.

- K-nearest neighbor weights, considers a number (k) of nearest neighbors:

- Distance band weights, relies on distance bans or thresholds to define the neighbor set for each spatial unit as those other units falling within a threshold distance of the focal unit;
- Bandwidth determines the distance threshold, the form of the kernel function defines the distance decay in the derived continuous weights.

3. Kernel Weights combines the distance based thresholds together with continuously valued weights.

2.2.3. Exploratory spatial data analysis

Exploratory data analysis (EDA) is defined by She et al. (2012) as a collection of techniques for summarizing data properties (descriptive statistics) and detecting patterns in data, identifying unusual or interesting features in data, detecting data errors, distinguishing accidental from important features in the data set, formulating hypothesis from data .

Furthermore, examining model results, proving whether model assumptions are met and whether there are influential data effects in model fits are also referred as applications of EDA techniques. This set of techniques quantitative summaries of the data that may have a visual representation such as graphs, charts and figures (She et al., 2012).

Spatial Data has its own set of techniques – ESDA, which includes summarizing spatial properties of the data, detecting spatial patterns in data, formulating hypotheses which refer to the geographical distribution of the data, identify cases or subsets of cases that are unusual given their location on the map (Anselin, 2009).The main difference between EDA and ESDA is the spatial component as ESDA extends EDA by adding methods to address special queries that arise as a consequence of the spatial referencing of the data. As a result of it, the map becomes a crucial element in the analysis of the data or examining the model results.

Spatial Autocorrelation

The computational expression of the overall tendency for similar values to be found near together on a map or pertains to the non-random pattern of attribute values over a set of spatial units is called spatial autocorrelation. It can either be positive, meaning that similar values have a tendency to be found together, or negative which reflects a value dissimilarity in space. In both of the referred situations of autocorrelation, the

observed pattern is different from what would be expected under a random process operating in space.

There are two different perspective from which autocorrelation can be analyzed (Rey, 2013):

- Global autocorrelation involves the study of the entire map pattern and presents the question of whether the pattern displays clustering or not.
- Local autocorrelation aims to explore within the global pattern to identify clusters or so called hot spots that may be either driving the overall clustered pattern or that reflect heterogeneities that depart from global pattern.

Spatial Autocorrelation indicators are cross product statistics that derive from the expression (Druck et al., 2004):

$$\Gamma = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\xi_{ij} \qquad \textbf{(1.)}$$

Where w_{ij} is a spatial weight that reflects the spatial relationship between spatial units i and j and ξ_{ij} is a measurement of correlation between variables.

Global autocorrelation.

Gamma Index of Spatial Autocorrelation.

The principle behind a general cross-product statistic to measuring spatial autocorrelation is applied in the Gamma Index of spatial autocorrelation. In this method, two similarity matrices for n objects are accessed in order to define if they measure the same type of similarity.

Gamma index is defined by the expression $\Gamma = \sum_i \sum_j a_{ij} b_{ij}$ (Hubert et al., 1981), consisting of the sum over all cross-products of matching elements (i, j) in the two matrices.

Fundamentally, the first similarity matrix will store a measure of attribute similarity such as cross product, squared difference or absolute difference while the second matrix will be a measure of locational similarity – a spatial weight matrix. Meaning that formally the Gamma Index will be represented by:

$$\Gamma = \sum_i \sum_j a_{ij} w_{ij} \qquad \textbf{(2.)}$$

Where w_{ij} represents the elements of the weights matrix and a_{ij} are corresponding measures of attribute similarity.

Moran's I

The global spatial autocorrelation can be measured using one of the oldest indicators of spatial autocorrelation (Moran, 1948). This indicator is often applied for continuous variables and compares the value of the variable at any location with the value at all other locations. The attribute y measured over n spatial units and is given by Moran's I as :

$$I = n/s_0 \sum_i \sum_j z_i w_{i,j} z_j / \sum_i z_i z_i \qquad \textbf{(3.)}$$

Where $w_{i,j}$a spatial weight, n is the number of areas that form the study region. $z_i = y_i - \bar{y}$ where y_i is the value that the attribute takes in area i (analogous for the j area) and $\bar{y}$ is the mean value of the attribute in the study region and $S_o = \sum_i \sum_j w_{i,j}$.

Moran's I can reach values between -1.0 and 1.0, being the amount of autocorrelation defined by the module of the coefficient. It is inexistent when I equals zero and positive or negative according to the index's signal. (Druck et al., 2004)

Geary's C

In Geary's C (Geary, 1954) case the interaction is reflected by the deviations in intensities of each observation location with one another. This indicator is similar to Moran's I. And is represented by:

$$C = ((n-1)/2S_0) \sum_i \sum_j w_{i,j} (y_i - y_j)^2 / \sum_i {z_i}^2 \qquad \textbf{(4.)}$$

Where $w_{i,j}$a spatial weight, n is is the number of areas that form the study region. $z_i = y_i - \bar{y}$ where y_i is the value that the attribute takes in area i and $\bar{y}$ is the mean value of the attribute in the study region and $S_o = \sum_i \sum_j w_{i,j}$.

The index's result can take values from 0 to 2. Values below 1 indicate negative autocorrelation whereas values equal to 1 indicate no correlation.

Getis and Ord's G

Similarly to the previous statistics, Getis and Ord's G (1992), is a global measure of spatial association. Representing a multiplicative measure of spatial association of values that fall within a critical distance of each other. However, this index takes values from 0 to 1 and can only be interpreted in comparison with the expected index, which is calculated considering a random distribution.

$$G = \sum_i \sum_j w_{i,j}(d) y_i y_j / \sum_i y_i y_j \qquad \textbf{(5.)}$$

Where d is a threshold distance used to define a spatial weight and y_i is the value that the attribute takes in area i (analogous for the j area).

LOCAL Autocorrelation

Local Indicators of Spatial Autocorrelation (LISAs) for Moran's I and Getis and Ord's G can be applied to determine clustering.

Local Moran's I (LISA)

Local Indicators for Spatial Association (Anselin, 1995) are applied to identify local association between and observation and its neighbors up to a specified distance from the said observation. Helps on determining the nature of local distribution.

$$I_i = \sum_j z_i w_{i,j} z_j / \sum_i z_i z_i \qquad \textbf{(6.)}$$

Where $w_{i,j}$ is a spatial weight, $z_i = y_i - \bar{y}$ where y_i is the value that the attribute takes in area i (analogous for the j area) and $\bar{y}$ is the mean value of the attribute in the study region.

Local G and G*

Getis and Ord can be formalized in two forms: G_i *and* G_i^*(Ord & Getis, 1995). Comparing local averages to global averages. While Gi* statistic includes the value of the point in its calculation, GI excludes this value, considering the value of its nearest neighbors (within d) against the global average (which also does not include the value of the point itself). It takes values closer to one if there's a cluster and a small value if there's a disperse pattern.

$$G_i(d) = \frac{\sum_j w_{i,j}(d) y_j - W_i \bar{y}(i)}{s(i)\{[(n-1)S_{1i} - W_i^2]/(n-2)\}^{(1/2)}}, j \neq i \qquad \textbf{(7.)}$$

$$G_i^*(d) = \frac{\sum_j w_{i,j}(d) y_j - W_i^* \bar{y}}{s\left\{\frac{[(nS_{1i}^*) - (W_i^*)^2]}{n-1}\right\}^{\left(\frac{1}{2}\right)}}, j = i \qquad \textbf{(8.)}$$

Where

$W_i = \sum_{i \neq j} w_{i,j}(d),\ \bar{y}(i) = \frac{\sum_j y_j}{(n-1)},\ s^2(i) = \frac{\sum_j y_j^2}{(n-1)} - [\bar{y}(i)]^2,\ W_i^* = W_i + wi, i,\ S_{1i} = \sum_j w_{i,j}^2\ (j \neq i),\ and\ S_{1i}{}^* = \sum_j w_{i,j}^2\ (\forall j),\ \bar{y}\ and\ s^2$

Considering that $w_{i,j}$ is a spatial weight, n is the number of areas that form the study region, y_j is the value that the attribute takes in area j and $\bar{y}$ is the mean value of the attribute in the study region.

Testing indexes' significance

The Autocorrelation coefficients previously presented need to be tested for statistical significance. This procedure can be performed under two different model assumptions (Smith et al., 2013):

The classical statistical assumption of normality, assuming that the observed value of the coefficient is resultant of a set $\{z_i\}$ of independent and identically distributed values from a Normal distribution.

In order to perform this tests, the null hypothesis, has to be identified. In this case, for the pattern analysis it is assumed that there is Complete Spatial Randomness (CSR) either of the features themselves or of the values associated with those features.

Z-scores and p-values are therefore calculated to determine whether the null hypothesis must be rejected or not, in this case, rejecting the null hypothesis reveals a statistically significant spatial pattern. The confidence level defines the amount of risk the user is willing to accept for making a false rejection of the null hypothesis (ESRI, 2013). They are summarized in table 3:

Table 3 - Normal distibution , z-score, p-value and confidence level (Source:(ESRI, 2013)*).*

z-score (Standard Deviations)	p-value (Probability)	Confidence level
< -1.65 or > +1.65	< 0.10	90%
< -1.96 or > +1.96	< 0.05	95%
< -2.58 or > +2.58	< 0.01	99%

The second model assumes that the observed pattern of the set $\{z_i\}$ of values is considered just one realization amongst all the possible random permutations of the observed values across all zones. A permutation approach is taken in order to get inference for this statistic, taking the randomization null hypothesis as the basis for statistical significance testing (Smith et al., 2013).

2.2.4. Spatial Regression

Regression analysis allows the process of modeling, examining and exploring spatial relationships facilitating the understanding of the factors that caused observed spatial patterns. These models can be used to predict outcomes based on the used independent variables. A Spatial Regression's goal is to explain or predict a dependent variable (y), recurring to explanatory variable(s) (x) that is (are) believed to have influence on the dependent variable, as it is illustrated by figure 6.

Such explanation is given by coefficients (β) which values are computed by the regression tool and that reflect both the relationship and strength of each explanatory variable to the dependent variable (Scott, 2009). There is also a part of the dependent variable which is not explained by the model (may be either under or over predicted) which has de name of residuals (ε).

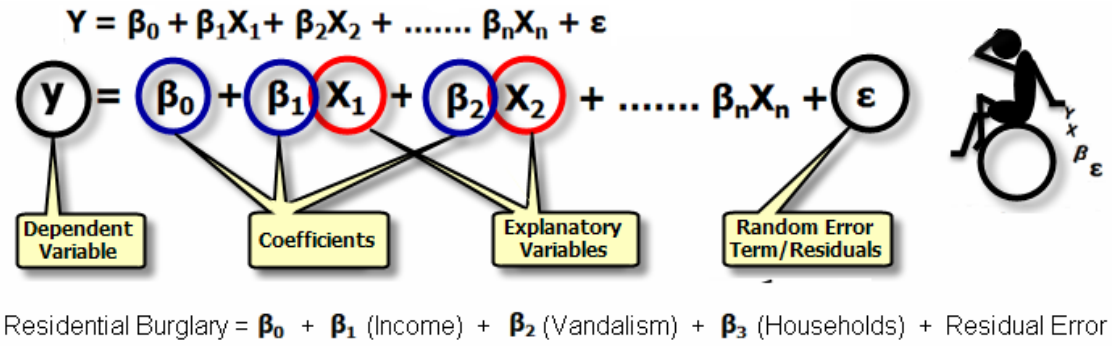


Figure 6 : Spatial Regression Model Equation Explanation with example. (Source: (Scott, 2009))

In order to define a successful spatial regression model, the variables that have a contribution on the dependent variable have to be thoughtfully chosen.
A (linear) relationship may be of the form (Druck et al., 2004):

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + \varepsilon_i \quad or \quad y = X\beta + \varepsilon \qquad \textbf{(9.)}$$

Where β is a column vector of p parameters to be determined and x is a row of independent variables with a 1 in the first column.
Taking y ={y_i} as a set of n observations or measurements of the response variable, with corresponding recording of matching values for the set of independent variables, then a series of linear equations such as the above can be formulated (Druck et al., 2004). Nevertheless, since the number of observations is usually greater than the number of coefficients, the best fit solution can be a possible approach to this situation. The best fit in this case will be the solution for vector β that minimizes the difference between the fitted model and observed values at these data points.

- Ordinary Least Squares (OLS)

Least squares is the term applied to the procedure of minimizing the sum of the squared differences and is often applied. In this case ε is a vector of errors that in conceptual terms is assumed to represent the effects of unobserved variables and measurement errors in the observations. The expected value of this error E(ε)=0 and the variance E(εε)=σ^2 I is constant. Where I is the identity matrix.
In this context, the set of n equations is typically solved for the vector β that minimizes the sum of the squares error terms, $\varepsilon\varepsilon^T$, hence the name Ordinary Least Squares or OLS (Smith et al., 2013).
The solution for the coefficients using this approach is obtained from the matrix expression(Smith et al., 2013):

$$\hat{\beta} = (X^T X)^{-1} X^T y \qquad \textbf{(10.)}$$

And

$$var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \qquad \textbf{(11.)}$$

The variance for such models is usually estimated from the residuals of the fitted model using the expression:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p} \qquad \textbf{(12.)}$$

In the spatial context, the objective is to model the variation in some spatially distributed dependent variable, from a set of independent variables. Conceptually, the form of the chosen model should be as simple as possible, both in terms of the expression employed and the number of independent variables included. Moreover, the correlation between different independent variables should be as low as possible and the proportion of the variation in the dependent variable(s), y, explained by the independent variable (X) should be as high as possible. When there is a high correlation among some or all the independent variables (x), (usually reflected by a correlation coefficient above 0.8) the model is almost certain to have redundant information and may be described as being over-specific (Haining, 2004).

The coefficient signal (+/-) of each explanatory variable indicates if the relationship is either positive or negative, as illustrated by figure 7.

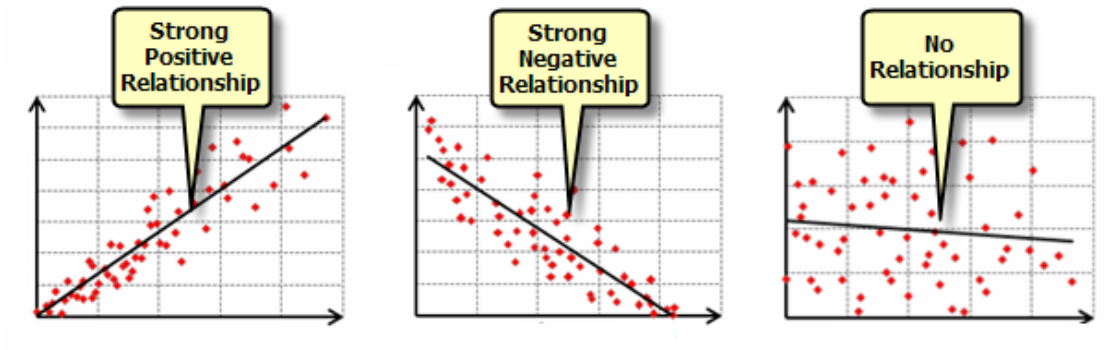


Figure 7 : Illustration of spatial relationship according to the explanatory variables signal. (Source: (Scott, 2009))

Multi-colinearity is said to exist if the there is a strong relationship between selected independent variables and is also broadly linear. To reduce multi-colinearity there are several techniques, such as: applying a so-called centering transform (by deducting the mean of the relevant independent variable from each measured values), increasing the sample size (used specially for small samples); removing the most inter-correlated variable or combining inter-correlated variables into a new single composite variable.

Heteroskedasticity is said to exist if the spread of errors is not constant. When fitting a model using OLS it is generally assumed that the errors (residuals for sample points) are identical and independently distributed.

In this case the estimated variance under OLS will be biased, resulting in non-reliable specification for confidence intervals and standard statistical significance tests (e.g. F tests)

Model performance can be assessed using several statistics mentioned in standard statistical bibliography (Scott, 2009).

- Spatial autoregressive modeling

Pure spatial autoregressive model consists of a spatially lagged version of the dependent variable,y (Druck et al., 2004):

$$y = \rho W y + \varepsilon \qquad \textbf{(13.)}$$

Despite its similarity to a standard linear regression model, the first term is constructed by a predefined n by n spatial weighting matrix, W, applied to the observed variable y. together with a spatial auto regression parameter, p, which typically has to be estimated. According to Anselin (2008, p257), spatial lag models are 'a formal representation of the equilibrium outcome of processes of social and spatial interaction'.

A spatial lag model can reflect some kind of interaction effect by expressing the notion that the value of a variable at a given location is related to the values of the same variable measured at nearby locations.

Adjusted R^2 (a modification of the statistic mentioned above), considers the complexity of the model in terms of the number of variables that are specified.

The dynamics of longitudinal spatial data or observations on fixed areal units over multiple time periods can be analyzed by several exploratory approaches.

- Statistical Significance

Squared coefficient of (multiple) Correlation or coefficient of determination (R2).

This statistic records the proportion of variation in the data that is explained by the models, as it is the function of the squared residuals with standardization being achieved using the sum of squared deviations of observations from the overall means:

$$R^2 = 1 - \frac{\sum \varepsilon_i^2}{\sum (y_i - \bar{y})^2}, R^2 \epsilon\ [0,1] \qquad \textbf{(14.)}$$

Where ε_i represents the residual for the observation, y_i is the value for that specific observation and $\bar{y}$ is the mean value for the observations.

If determined under appropriate conditions, this coefficient will take values close to 1 if highly significant. However, it is frequent not to be able to determine the statistical

significance of the measure due to distribution conditions not being met, as a result of it the value should be taken as an indicator of goodness of fit.

In the context of spatial analysis, the conditions that should be met if preforming inferential analysis are often not met. The conditions are mentioned below:

1. The set $\{y_i\}$ is comprised of independent (uncorrelated) observable random variables;
2. The set $\{x_i\}$ is comprised of independently and identically distributed unobservable random variables with mean 0 and constant variance,σ^2 , where σ^2 is not a function of β or x. (heteroskedasticity);
3. The set $\{x_i\}$ is Normally distributed;
4. The model applied is appropriate, complete and global. This assumption includes assuming that the independent variables, x, are themselves uncorrelated and the parameter β are global constants.

- Results:

If there's no structure that reflects an exaggerated amount of clustered miss prediction, it is said to be random noise. If there is a tendency of miss prediction in a clustered area, the model is probably missing one or more key explanatory variable. The process of determined the model (the variables included) is often an iterative process.

R2 – Is defined by the bibliography (Scott, 2009) as the percentage of explaining this model has of the dependent variable.

In order to properly analyze a spatial regression there are at least six main points to consider on the report:

1. Coefficients have the expected sign – positive signals indicate positive autocorrelation and negative signals indicate the opposite;
2. No redundancy among explanatory variables – defined by the variance inflation factor that must be below 7.5;
3. Coefficients are statistically significant;
4. Jarque Bera is not statistically significant. This coefficient measures weather the residuals from a regression model have a normal distribution. Random noise is normal in a properly defined model. It reveals a random spatial pattern. (otherwise the model is biased, meaning that there is probably one or more variables missing);

5. Model performance:
    a. Adjusted R2 should have values between 0 and 1, usually taking values above 0.5;
    b. AIC should be the lowest value possible (if in doubt among different variables, consider the lowest value of AIc);
6. Model residuals should be free from spatial autocorrelation (it is frequent to run a spatial autocorrelation test on the residuals).

2.3. Application of spatial analysis

In the bibliography there are several examples of applications for spatial analysis. These applications are mainly related to areas of observational sciences such as environmental criminology, geographical and environmental (spatial) epidemiology, regional economics and the new economic geography, urban studies, environmental science, policy area (decision making), etc. (Beale et al., 2010; Rey, 2007; Druck et al., 2004). The continuous interest in this area can be verified by the wide range of dates present in the bibliography (Moran, 1948; Anselin, 1995; Fischer & Getis, 2010). The availability of good-quality data, the emergence of well- formulated hypotheses that can be expressed in mathematical terms, the availability of appropriate mathematical and statistical tools and techniques and the availability of technology for facilitating analysis are the main factors presented by (Goodchild & Haining, 1998) that explain this continuous and still growing interest in this area that has been experiencing an expansion in its use connected to the awareness of the importance of location in theorizing disciplinary approaches to describing spatial events.

2.4. Chapter Summary

Throughout the present chapter several concepts related with both spatial data and spatial analysis were presented. The importance of data quality and conceptualization took an important role in this chapter since as John Tukey often remarked, "better an approximate answer to the right question than an exact answer to the wrong question" (Tukey, 1962). The presented quote is very suitable for this topic since the conceptualization is the tool that allows the representation of the real world, and if the conceptualization is not faithful to the reality to be represented, the question will be wrong from the beginning leading to an untruthful answer.

Furthermore, and since this system aims at giving some support on the analysis process, all the methods to be implemented were presented and explored in an accessible way,

contemplating results and inputs in order to guide the user through the process. The summary for the implemented project can be consulted in the figure 8 below, within a schematized format to facilitate the general understanding of the processes' defined.

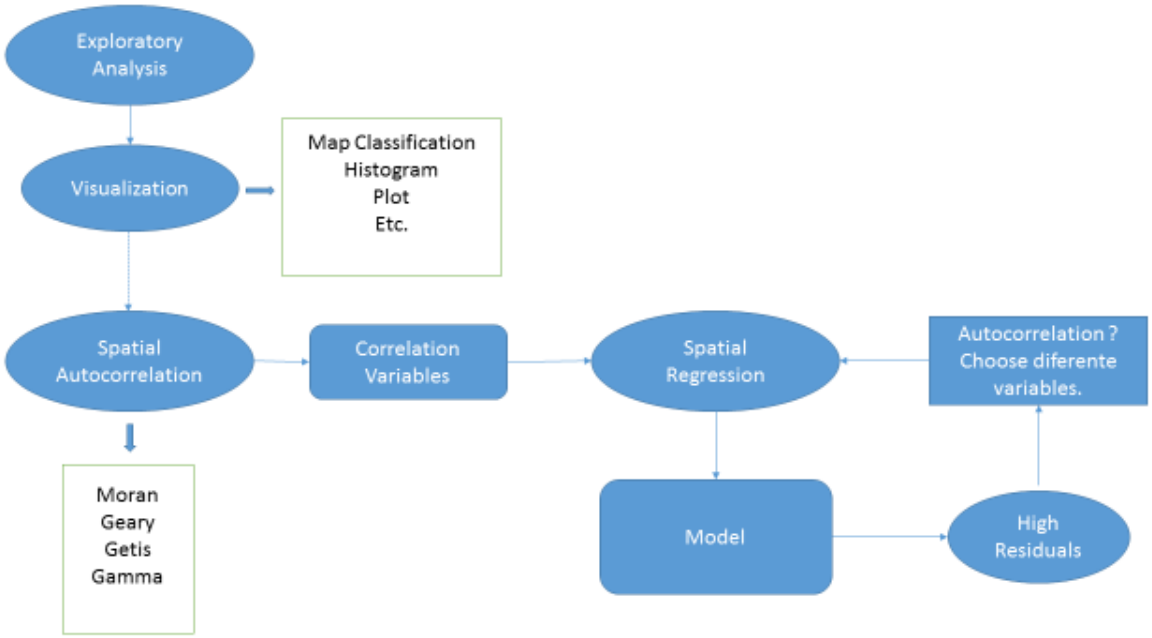


Figure 8 : Spatial Analysis functionalities to be implemented. Schematized structure.

3.Web-Based GIS

3.1. Web-Based GIS Applications

There is a wide range of geostatistical software available (a brief summary may be consulted in (Druck et al., 2004)), and the availability of so called Web GIS is already a growing tendency, having its most common emphasis in map delivery, cartographic presentation and geographic information. However, combining both components – Web GIS with spatial analysis is an area with little exploration, with very limited bibliography.

Anselin et al. (2004) present a solution using Geotools open source mapping toolkit, which comprehends a collection of Java classes. The mentioned project, is considered an initial framework, having some defined situations to be further explored such as limitations in performance issues, the download time for the Java applet, the limited amount of functionality, the fact that Java is not the most appropriate language for intensive numerical operations any web platform serving GIS data has to have at least four core components.

Other approaches are for example spatial weights creation online (Anselin et al., 2004), which is still not significant to this particular case. The mentioned author proposes an integrated web-based environment incorporating open source software packages to provide geoprocessing services. Even though it is in fact an interesting paper and a similar proposition to the one this project has to offer, the scope of the tool is no longer public, requiring an expensive membership fee. Some interesting perspectives over the technology used and implementation are documented in the same article.

A very interesting approach is also proposed by Schrader-patton et al. (2010), where the author presents a web service to allow online land management. Several interesting concepts are considered in the document, even though all the project was developed within a commercial software environment. Other interesting work in this context is a thesis work by Cabral (2001) that explores spatial data management for storing purposes and the online exploration of spatial data within a web GIS environment.

Probably the most similar work to this project is presented by Lu et al. (2013), whom propose a website for spatial analysis purposes with very interesting functionalities that comprehend steps from visualization of spatial data to spatial autocorrelation. The website is also very intuitive, lacking only in accessibility for common users since the explanation is somehow limited. The same author revises other spatial software,

referring that from the existent software, either there is a strong spatial analysis component, but the presence is limited to desktop software or GIS online rarely include further spatial analysis functionalities. Azavea (Azavea, 2012), GIScloud (Cloud, 2012) and SKE (SKE, 2012) are mentioned as examples of websites, regardless, their spatial analysis functionalities are referred to as limited.

The presented studies all have interesting perspectives that where considered for this project's research. All of them consider a similar WebGIS structure composed by four main components. These components are presented below and schematized in figure 9 (Amrita, 2012). There is a wide range of possibilities when building a Web GIS platform including open source software and proprietary (license required software). Given this projects' scope, all the options made and presented from now on will be limited to open source software. In the addition to the main structure presented below, it is common to include a Web framework as an integration component.

Database

- Store and manipulate spatial and non-spatial data.

GIS Server

- Provides visualization, spatial data analysis, mapping, and spatial data management services.
- Supports complex workflow activities, including versioning.

Web application server

- HTTP server: The Web server that processes the HTTP requests.
- Application server: Contains the Web application and supports client-side APIs (such as JavaScript) and server side logic (such as servlets, Enterprise JavaBeans (EJBs)) to invoke GIS server tasks.
- Database connection: Java Database Connectivity (JDBC) or Open Database Connectivity (ODBC) API to connect to the database.

User interface

- Web browsers: Increasingly popular choice for interaction with the GIS.
- Desktop software: Used for complex spatial data manipulation and visualization tasks with direct connection to the GIS server.

- Mobile devices: Support one-way and two-way data replication tasks.

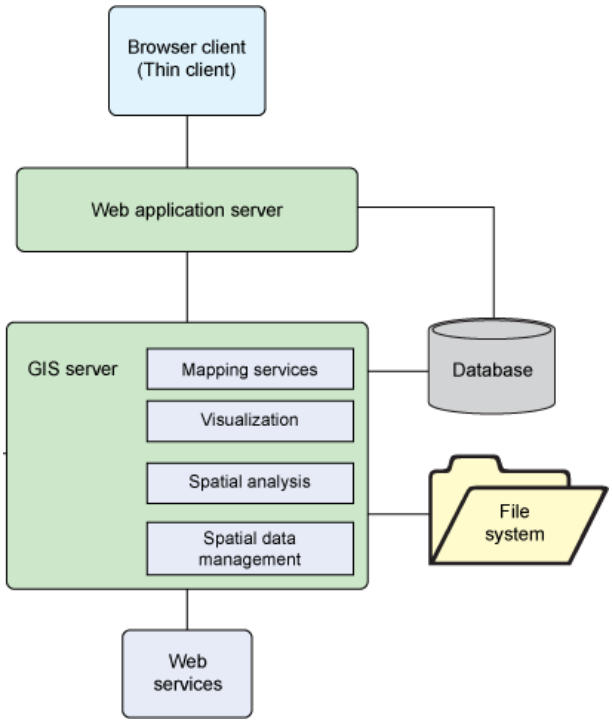


Figure 9 : Schema of a Web GIS application (Source: (Amrita, 2012)).

3.2. State of the art of Web-Based GIS architecture

There are several architecture options to be considered in this project. Geospatial technology has been developing significantly as spatial data becomes more popular. Environmental subjects, Transportation planning and Biology studies often appear in the related areas and there are some interesting approaches to the problems to be considered (Rangel et al., 2010; Woodall & Graham, 2004; Carneiro & Santos, 2003).

Even though the architecture chosen varies from study to study, PostGIS (PostgreSQL, 2012) is the most common component in the articles consulted, so the database option is quite often PostgreSQL(2014) with a PostGIS extension.

As for the remaining components of the architecture, there are some interesting options being presented and discussed over the next few paragraphs.

D'Amore et al. (2012) built a spatial data infrastructure (SDI) for atmospheric pollution monitoring and modeling. Its functionalities include storing, mining and visualizing information with the main goal of evaluating the impact of atmospheric pollution ecosystems and human health. The software used in the mentioned project was PostGIS for data storage, Geoserver to export services, GeoNetwork for metadata generation and

management and Javascript libraries embedded in OpenLayers were used to display geographic web services. This made the SDI a pluggable system, built trough components plugged together which requires some effort to integrate the different components. In this case, there was an information and communication technology (ICT) pluggable framework called GeoInt, which purpose was to decrease the SDI component's complexity for end users (by supporting data input from different sources and data management).

Wang & Zhang (2010) also used a similar infrastructure, when creating a platform to protect the power distribution system from lightning related damage and faults, some of the main functionalities include capture, storage, management, visualization and analysis of a Lighting Database. In the capturing part Tomcat was introduced, but the platform structure was very similar to the previous mentioned. The interesting conclusion on this prototype was that the lack of advanced special processing capabilities and static functions revealed itself a big obstacle.

Other interesting projects include the implementation of a land monitoring portal to combine multiple sources and multiple types of data (Lee, 2009). The Open Source comparison made has revealed itself to be quite interesting and some analytic views of the architecture of a Geo-Portal were coherently exposed.

On all of the projects consulted, the authors use PostgreSQL as a database, having been related that its reliability and convenience due to the spatial extension (PostGIS) making it more accessible. Geoserver (2012) is often the server utilized; the reasons on this subject are not revealed on most projects, but the interactive tool that allows the data management without any complicated code associated is probably the most interesting feature for this choice. As for the front end, most of the projects presented so far have introduced Javascript embedded in OpenLayers (2014) library as their main tool.

The answer to analytic functionalities appears in some articles associated with R or Grass software but Cagnacci & Urbano (2008) mention that the lack of advanced spatial processing capabilities and static function was an obstacle while developing the project. It is important to refer that neither R nor Grass were mention throughout the article.

Two other interesting projects on very different areas were also consulted to explore the possibility of introducing a Web Framework. The first one (Oussalah et al., 2013), had the objective of handling a considerate amount of data with location information and suitable for geo-location analysis. There were three main domains defined: User interface, specific representation of the data and logic that includes functional

algorithms handling information exchange between database and user interface. Even though the project's aim was to study people's behavior on social networks based on their location and semantics, it presented a complex architecture perspective which includes the possibility of input and managing data from three different domains, which can certainly be a useful feature.

Weigel (Weigel et al., 2010), also presented a similar architecture allowing input, storage, harmonization and output in an application to track devices off the road in urban areas.

Both these two last documents introduced Django web Framework (and GeoDjango extension) to receive and integrate spatial information data on a background database, which is extremely interesting considering this project's objective.

3.3. Overview of Web-Based GIS Technologies

3.3.1. Database

As mentioned by Gillenson (2011) even though the technology has evolved and the storage capability has been significantly improved, a database is, as it was in the beginning of times the vehicles needed to store and utilize the data that is important in our environment. Data can be a very powerful tool, and this fact was the main reason why database management system software (DBMS) and the 'database environment' appeared. The basic schema of a DBMS is displayed in figure 10.

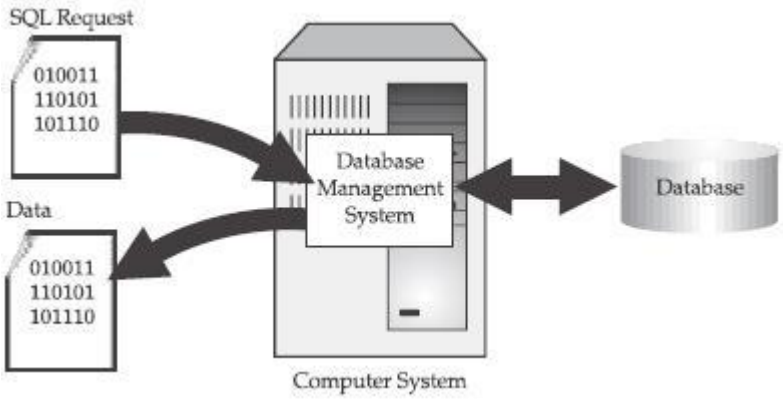


Figure 10 : Basic DBMS schema (Source: (Gillenson, 2011)).

Encouraging data sharing and the control of data redundancy as well as improvements in data accuracy were some of the advantages of their creation. The main consequences of these features are the vast storage capacity, acceptable access and response for database queries. Data security, data privacy and back up recovery are also pointed improvements.

A comparison between the two of the most mature OpenSource projects for spatial DBMS is presented in figure 11, regarding a different list of parameters that were considered the most influent factors in the user's decisions according to Ballatore et al. (2011).

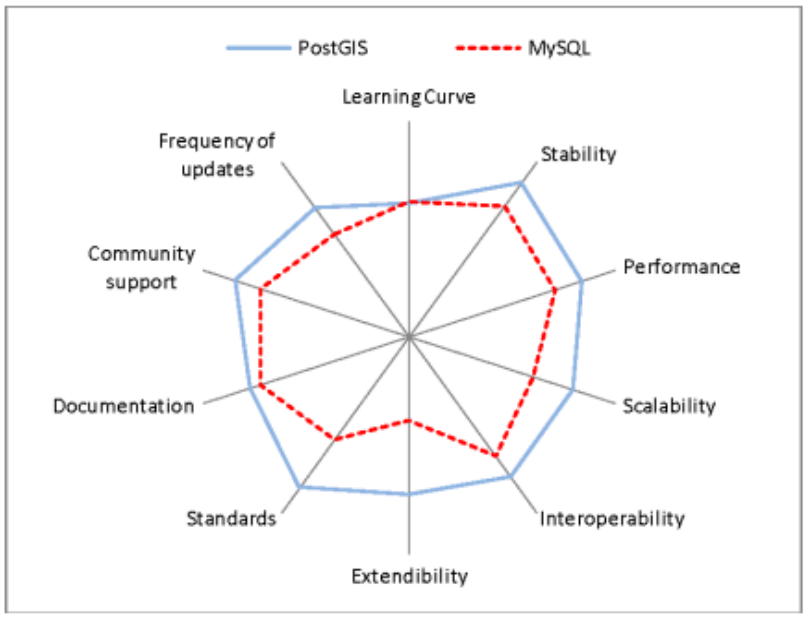


Figure 11 : Spatial database management systems (Source: (Ballatore et al., 2011) p.17)

- PostgreSQL – with a PostGIS extension

PostgreSQL (2014) is an open source object-relational database system. It is one of the most mature open source database projects, having more than 15 years of active development. Its strong reputation for reliability, data integrity and correctness were supported by its proven architecture. Some of its most useful features include supporting foreign keys, joins, views, triggers and stored procedures in multiple languages.

It is very important that besides the reliability and integrity (that are unarguably needed), the capacity of including spatial data in the database is also taken into consideration. PostGIS (PostgreSQL, 2012) extends PostgreSQL's data type support, adding geographic objects to the database. Some of PostGIS' most interesting features in this project's scope include automatic geometry columns inserted in tables, topology and index based nearest-neighbor searching (high performance).

3.3.2. Server

The server is the connector between the database and the client's page. In the GIS case, the server has the crucial function of rendering geospatial information contained in the database to images, so it can be visualized on a map or as an image on the web page.
Its applicability comprehends a wide range of purposes, from querying spatial DBMS to projection support including integration with other geographic libraries.
Figure 12 classifies both servers according Ballatore et al. (2011) In which a considerable sample of users were questioned about several aspects of the mentioned software.

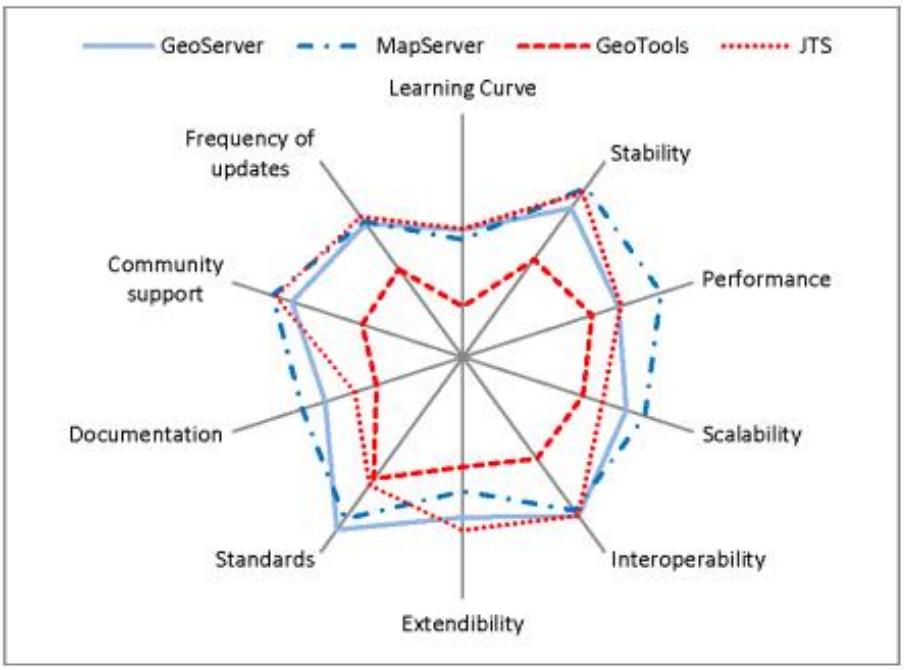


Figure 12 : Comparison between sever web mapping servers and spatial libraries (Source: (Ballatore et al., 2011), p. 16)

3.3.3. Hypertext Markup Language (HTML)

According to Meloni (2011) what happens in the internet is that users are allowed access to the web content. This content (text, images or other multimedia content) are rendered by browsers which are given certain instructions found in individual files.
The mentioned files contain text marked up, or surrounded by HTLM codes that can describe the exact way in which the information will be displayed. A web presence requires files that contain text to display or codes that can be interpreted by the server in order to send a graphic along to the user's web browser. Such content must be planned, design and integrated with all the pieces to be included in the web presence.

Browsers are able to organize the web content components and manage those parts according to the HTML commands in the file.
In order to publish web content, it is required to have a web server. So a web hosting provider will be necessary to make any web content public to other users.

3.3.4. Cascading Style Sheets (CSS)

Cascading Style Sheets (CSS) are used to define the display of the web content present in the HTML. This document is supposed to specify fonts, colors, spacing and other characteristics that will define the aspects of a website. It is define by Meloni (2011) as a grouping of formatting instructions that controls the appearance of several HTML pages at once.

3.3.5. Javascript

Javascript is a tool that allows a variety of visual and interactive features in addition to useful content: graphic, sounds, animation, and video. JavaScript commands can be inserted in the HTML documents.
Whereas HTML is a simple markup language it is not able to respond to the users, make decisions or automates repetitive tasks, a programming language or scripting language is able to do all this. Scripting languages are usually simple, and they make available interactive tasks that require more sophisticated languages.
Figure 13 illustrates the interaction between html, CSS and JavaScript.

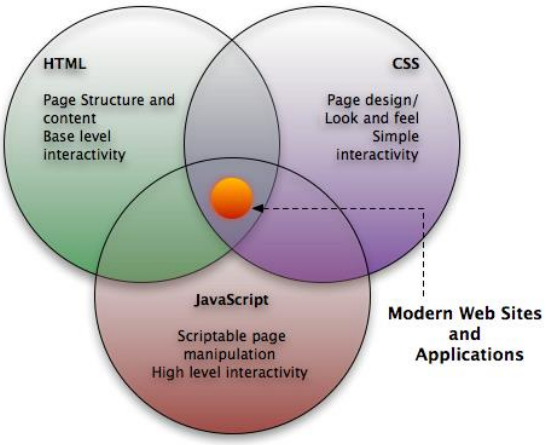


Figure 13 : Interaction between JavaScript, CSS and HTML in a modern Web Site or Application (Source: (MASS MEDIA GROUP LTD., 2011))

JavaScript libraries considered have all the same goal: displaying and managing spatial data on the client side. Therefore a comparative graphic is displayed on figure 14 evaluating the mention libraries in a set of comprehensive parameters. This graphic, from Ballatore (2011), will be useful to define the architecture of this project.

Javascript Libraries - Client Side:

- Openlayers
- ExtJS
- MooTools
- Prototype

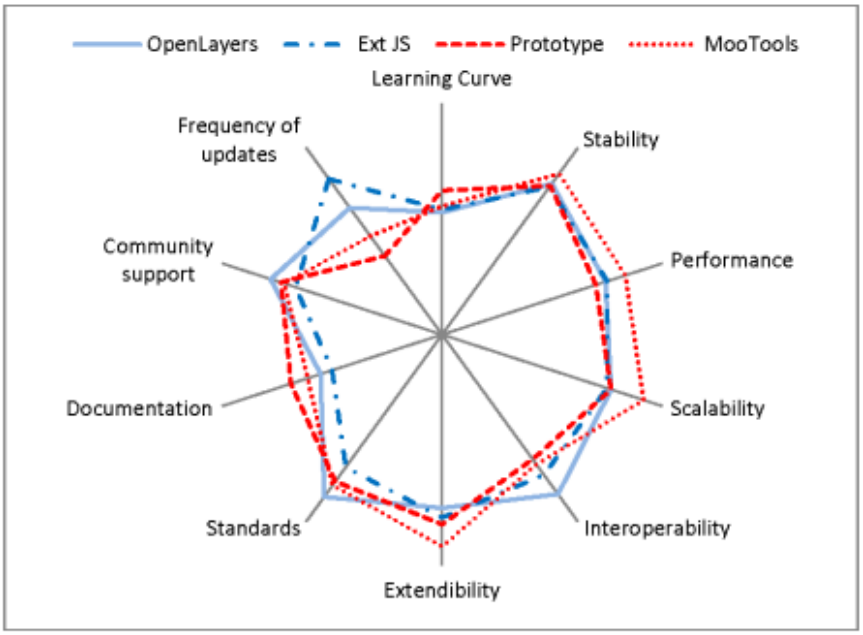


Figure 14 : Comparison of Javascript libraries and mapping services for map display.(Source: (Ballatore et al., 2011), p. 15)

3.3.6. Python Libraries

Python libraries are very useful tools that involve a rather challenging installation process, requiring several pre-requisites and often leading to the installation of other libraries. An interesting way to avoid this process is to use python development environment which provides all the required libraries and hosting services (pythonanywhere.com for example). There are several libraries that deal with geospatial analysis online, being the most important ones considering the scope of this project GDAL(GDAL, 2014), Matplotlib(John et al., 2012), Pysal(Pysal, 2014) and ReportLab(Reportlab, 2014) A dependency schema of the used libraries is presented in figure 15 as to illustrate the procedure for this particular project.

Each core Library is presented in Green rectangles, having dependencies in blue ellipses connected with each of the mentioned libraries.

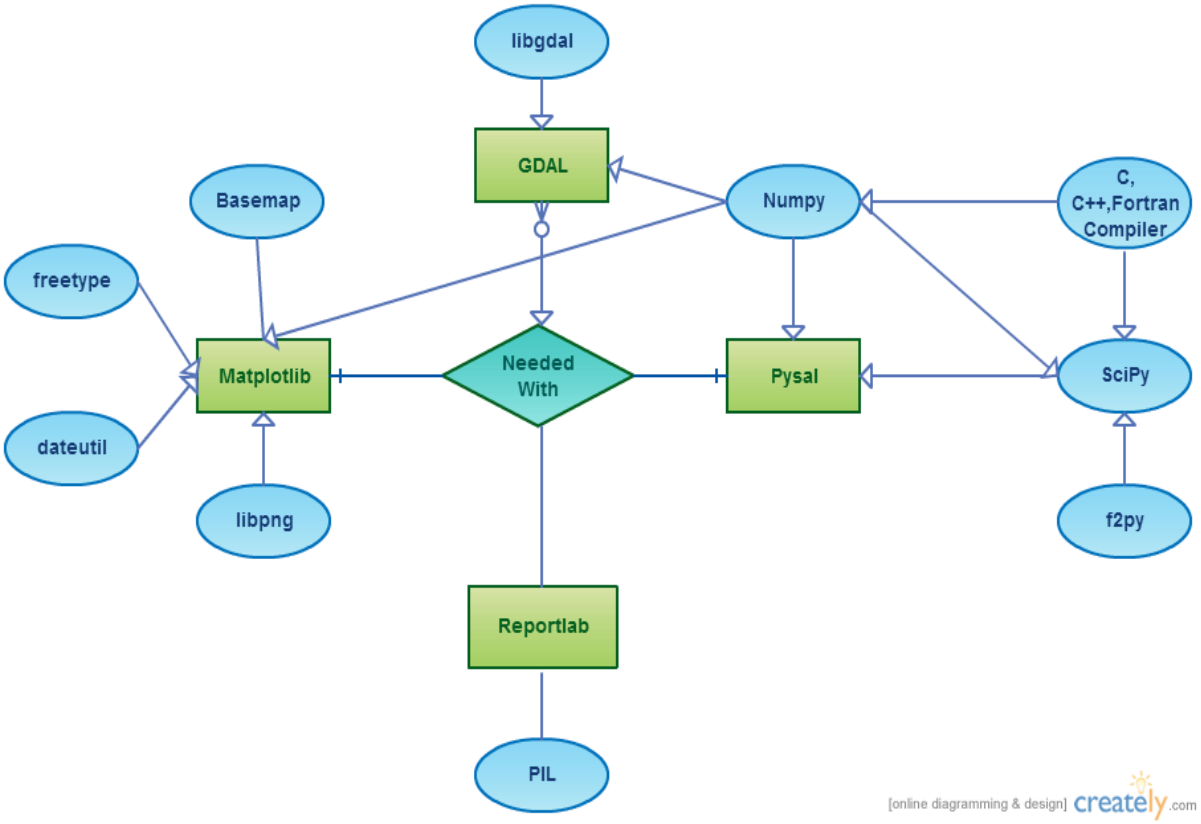


Figure 15 : Illustrating schema of the Python libraries to be applied and required dependencies.

3.3.7. Web framework

Commonly called web application framework (WAF), a web framework is a software framework which function is to support the development of dynamic websites, web applications, web services and web resources. Its main advantage is that it provides a structure so that web development can be performed simultaneously on different knowledge areas. Usually database access, templateing frameworks and session management are some of the sections in which the web framework allows simultaneous work. Code reuse is also an interesting advantage of this type of software, since its applications can be often inserted in distinct web pages. The framework architecture may vary depending on the software utilized. For this particular case the Model- View – Controller architecture will be considered and therefore presented a little further.

MVC (Model-view-controller) architectural approach separates the data model from the business rules and from the user interface. Some of its most interesting advantages

include modularizing code, promoting code reuse and allowing multiple interfaces to be applied.

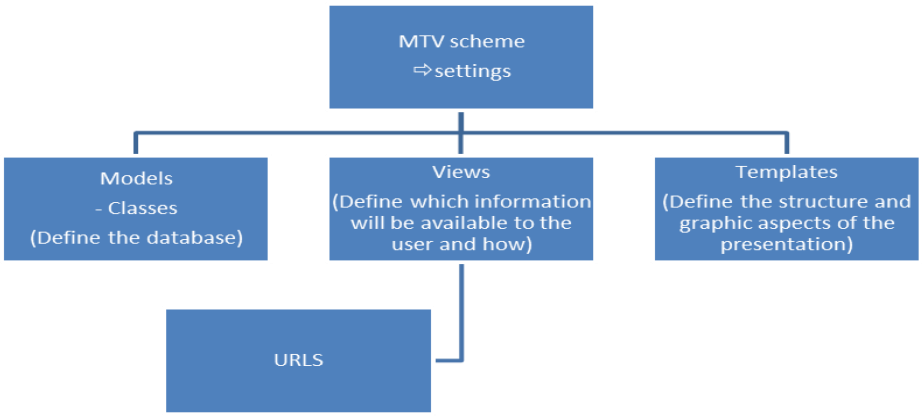


Figure 16 : MTV Schema: necessary files and core structure of a MTV model.

- Django

Django (Foundation 2014) web framework is one of the most popular ones and it is written in Python. It is structured in an MTV pattern, having Model, Template and Views as the main components. The Model represents the data structures, where classes are the relations and instances are the attributes of the relation. The template is an instrument allows the user to define the visual aspect of their web content using HTML, and the View is a function that renders from a template. The concept is fairly structured and can represent a valuable and simplifying tool for publishing web content and limiting the extent to which HTML needs to be sprinkled throughout the Python source.
It has been adapted to spatial data with GeoDjango (Django, 2014). This new extension includes tools to deal with a spatial database and with spatial data, incorporating several spatial libraries in python. Figure 17 represents a summarized schema of all the components that GeoDjango comprehends and the languages that enable the communication between them.

Figure 17 : GeoDjango Structure, including basic libraries, available databases, implemented standards and displaying format (Source:(Springmeyer, 2009)).

3.4. Chapter Summary

This chapter's objective was to introduce basic concepts that support the technologies and theory behind this project. While the first part explained basic concepts and tools that are required for a full understanding of this subject, the architecture part was used to define the best software combination to perform the task in hands.

According to the database comparison, the most suitable database to choose will be PostgreSQL with the PostGIS extension. This database is referred by numerous studies as a reliable and efficient Open Source Database.

Even though MapServer achieves overall better scores in the study by Ballatore (2011) Previously referred in this chapter, Standards and Interoperability were considered the most important feature of a server, leading to the preference o GeoServer for this specific purpose. This decision is strongly related with the fact that handling the data within the server implies the use of python spatial libraries which are built to deal with OGC standards, making the process of dealing with the data more flexible.

The client side library chosen was OpenLayers with the same arguments than Geoserver: Standards and Interoperability have a great importance in projects in which the data will have to be passed through different layers of software. Given the nature of the project, these are the most reasonable choices.

Besides the basic components it was also considered necessary to insert a web framework to provide a structure to the project. Django was web framework chosen especially because it supports geospatial data with the Geodjango extension but also due to its maturity and the fact that it is written in python. As suggested by (Westra, 2010), there are many spatial libraries written in python, thus facilitating the manipulation of the data.

4. Prototype and System Implementation

4.1. System requirement Analysis

4.1.1. Prototype implementation

Mao (2005) defines the following steps of the system implementation: data collection, data pre-processing, database design and populating, system architecture design, user interface design and functionality development. The same steps will be followed in this project to implement the prototype following the workflow presented in figure 18.

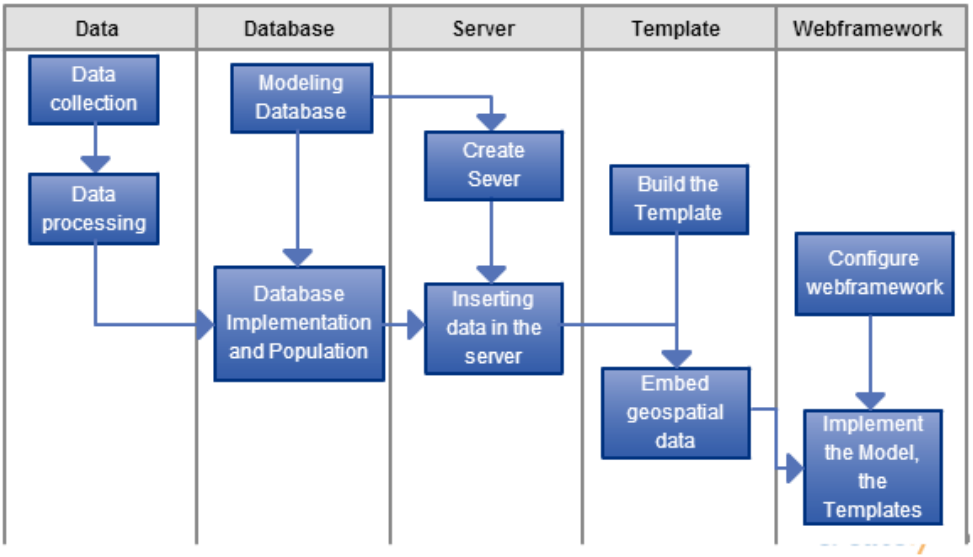


Figure 18 : Workflow of the prototype WebGIS system..

4.1.2. Data Collection

Data collection may represent a complex process, depending on the source. Metadata is often insufficient and data related problems (already mentioned in the 2nd chapter) must not be overlooked. As a result of it, the data source must be thoroughly chosen and the dataset have to be conveniently explored. Table 6 (attachments) presents several trustworthy data sources in Portugal.

Spatial data is becoming increasingly more accessible and the above table provides some examples of feasible data sources for data for the Portuguese territory.

Census data from INE (Instituto Nacional de Estatística) (INE, 2011) comprehends both spatial data files (in shape file format and xml) and a list of statistical data that is stored in a separated csv file, including a set of 122 statistical variables. This data has a spatial resolution that goes up to 'Freguesias' – the smallest administrative section in Portugal for the CAOP case and down to statistical subsections in the Census case. Statistical

sections (or subsections) represent the smallest partition of the Portuguese territory for statistical data (which is not particularly interesting for this project's scope).

4.1.3. Spatial data processing

- Coordinate System

Since all the data collected for this project was in the same coordinate system there was no need for transformations. The coordinate system used is presented below along with the correspondent EPSG.

5. Unit: meter
6. Geodetic CRS: ETRS89
7. Datum: European Terrestrial Reference System 1989
8. Ellipsoid: GRS 1980 (a = 6378137m , f=1/298.257222101)
9. Prime meridian: Greenwich
10. Data source: OGP
11. Information source: Instituto Geográfico Português (IGP).
12. Revision date: 2007-08-15

- Dataset Division

As presented in the previous section 4.2.2. the amount of data related to this topic represents a heavy dataset to process in practical time for a server side process. This situation leads to a division of the dataset.

Even though it is logical that for the database this option represents redundancy in storing (which is technically an undesirable situation), due to server capabilities and the spatial nature of the data it was defined that the division of the dataset was the more adequate way to deal with this question. Therefore, the dataset was divided in three areal units, according to Portuguese administrative boundaries:

- Civil Parishes ('Freguesias') : Total of 2882
- Municipalities ('Munícipios') : Total of 308
- Districts ('Distritos') : Total of 18

It is important to refer, that as was mentioned in the 2nd chapter, conceptualization is a crucial part of the spatial analysis. Therefore data has to be organized and managed in order to provide a model that is as faithful to the real world as possible, and

coordinating datasets to complement information may require some operations, as it was necessary in this case.

Furthermore, in order to facilitate user's comprehension of the data, a separation of the variables according to six defined theme groups was implemented:

- Buildings
- Demography
- Education
- Employment
- Families
- Housing

The variable list associated to each of the categories defined can be consulted in the attachments section.

- Operations over the dataset

The schema bellow summarizes the operations that the dataset was submitted to for storing and management purposes. It is frequently necessary to proceed to this type of operations in order to organize the data according to our conceptual model. Coordinating datasets may also require operations over the data as it was necessary in this specific case, nonetheless, the operations are different for each specific case. The processing operations applied to the dataset are presented in figure 19.

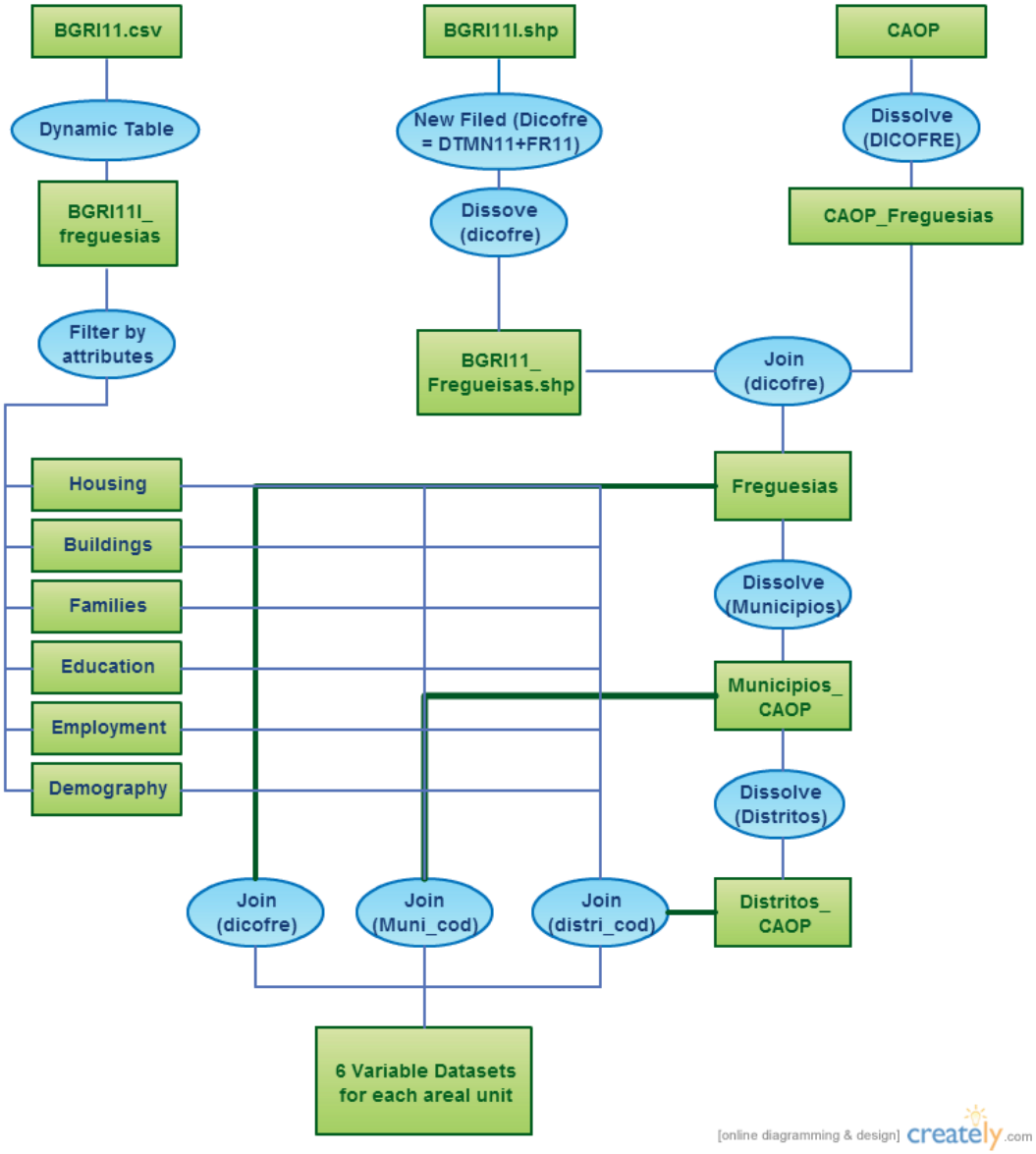


Figure 19 : Data processing diagram, including all the datasets used and geoprocessing operations applied.

Dataset particularities:

- CAOP's attribute TAA (attributes can be consulted in attachment), has two possible values: 'Área Principal' and 'Área Secundária), meaning that Civil Parishes are divided in main and secondary areas (subareas of the Civil Parish). This information is not relevant for this specific case, so a dissolve by the attribute 'Dicofre' (Civil Parishes unique identifier) was performed in order to reduce the amount of data (from 4414 records to 2882).
- BGRI11 is provided by INE with a division of files (already mentioned), this involves understanding the compatible files and joining the information through

this field. For practical reasons, the variables file was processed with the aid of dynamic tables, being subsequently joined to the spatial dataset.

- The Considered CAOP dated back to 2012, since the BGRI11 data dates back to 2011 and there was a recent reorganization of the administrative areas in Portugal that eliminated 2063 Civil Parishes, creating 895 new ones, which would lead to a mismatch of the data.
- Dicofre (a unique identifier of CAOP) is composed by three components:
  DI : disctrict identifier (‘Distrito’)
  CO : municipality identifier (‘Concelho’)
  FRE : civil parish identifier (‘Freguesia’)

4.1.4. Spatial Database design

The databased was modeled according to OMT-G (Borges et al. 2001).
The structure of the defined database is presented in OMT-G notation in figure 20.

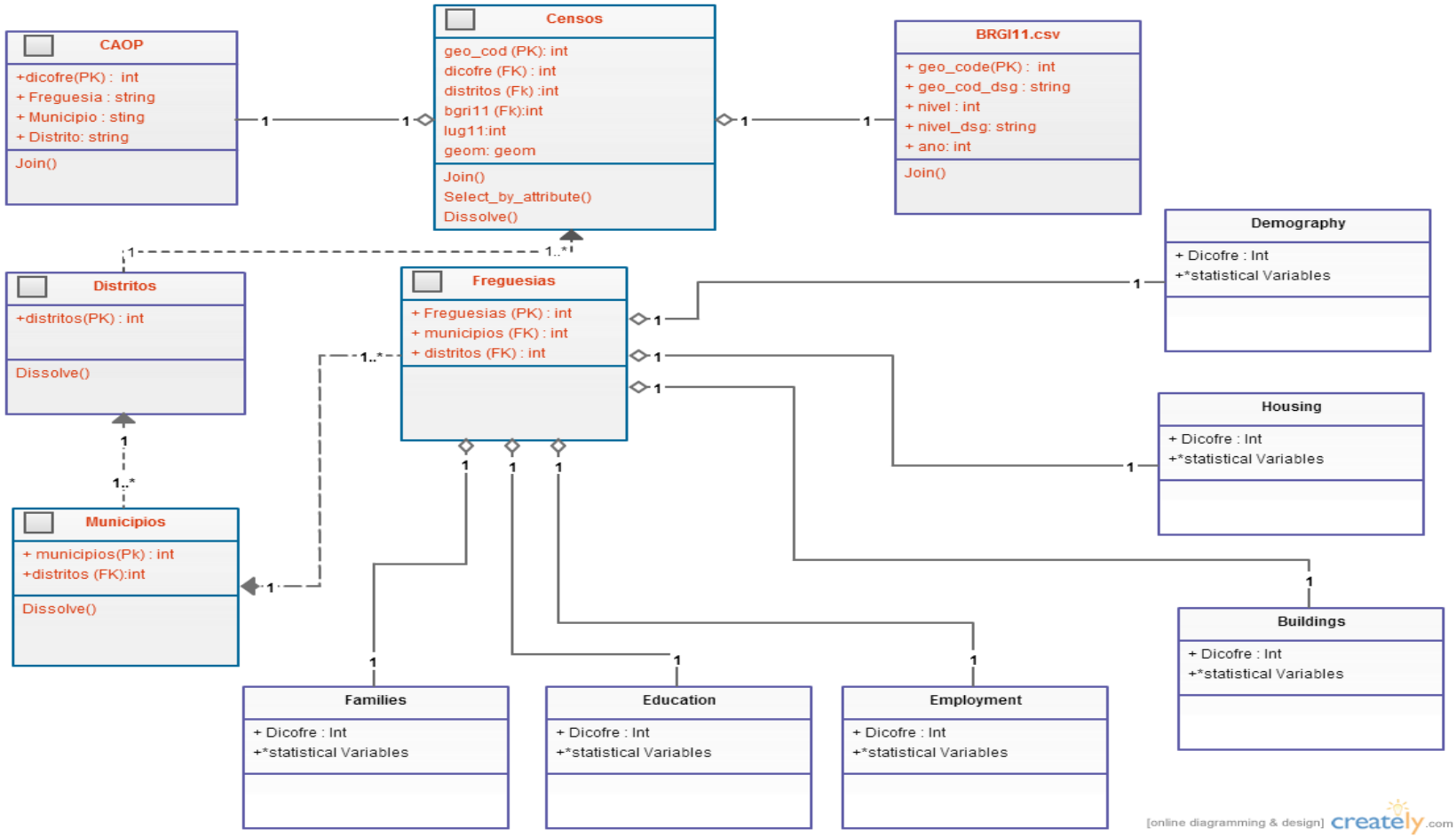


Figure 20 : Classes and transformations Diagram

(OMT-G Geographical Database Data Model)

4.1.5. Website Setup

As mentioned in previous chapters, a web framework was used in order to build the website. At this point the website's structure as well as a brief explanation of each file is presented. The files organized according to MTV model (defined in chapter 3) and will be further explored as most of their functionalities are to be contemplated within the 5th chapter.

The Website structure includes a PostgreSQL database with a PostGIS extension, a server to provide geographical information (Geoserver), and libraries to deal with geospatial data, the Web Framework chosen is Geodjango, requiring Python and Django.

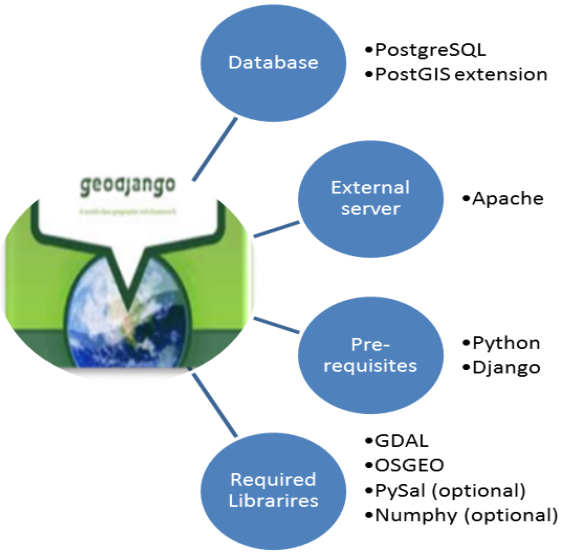


Figure 21 : Geodjango structure: Main core components and Required Libraries.

4.1.6. User Interface Development

The User Interface defines all the application interaction. Four main categories were defined towards this goal. The categories are presented in the table below along with a brief description of the page and the technologies applied.

Table 4 – User Interface pages, descriptions and technologies applied.

Page	Description	Technologies
Home	the introductory page of the project	Html, css, JavaScript Openlayers : data from geoserver
Spatially Instructions	Some basic information about the data and the methods to be applied	Html, css, JavaScript
Spatially– Spatial Analysis	The form through which the user defines the analysis to be performed	Html, css, JavaScript Forms
About	A summary description of the project and the default dataset	Html, css, JavaScript

HTML along with css and JavaScript were used to build the user interface. The user interface's appearance is presented in figures 22 to 25.

Internet provides a wide variety of html builders online which allow the users to easily construct their own website. In this case dotemplate.com (Ruiz, 2007) was used to produce the basic html code. Even though it represents a valuable help as a starting point, it is advisable to have some html knowledge in order to alter the code. In the styling case (css), it is generated along with the defined html, saving a considerable amount of time.

Pages:

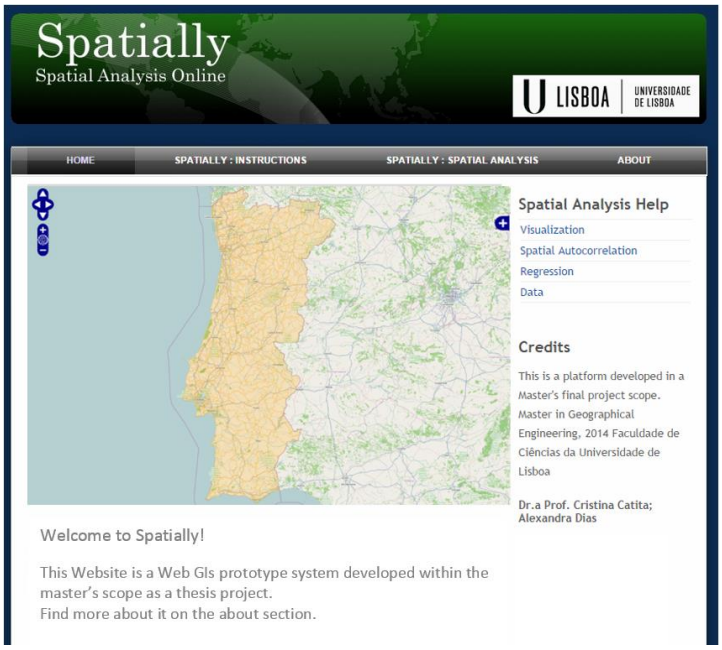


Figure 22 : Spatially: Web GIS Prototype System - Home Page

- **Home**

Home is the first page the user will have access to when entering Spatially. This page presents links to the other website pages. It also displays a map of Portugal and some brief information about the page.

Spatially: Instructions

The Instructions page is intended to provide the user with a guideline for using the methods of spatial analysis made available by the webpage. In this section, the user may consult a summary of the spatial analysis techniques applied, along with some bibliography to point the user to some additional information on the subject. The text can be consulted in attachment, section 3.

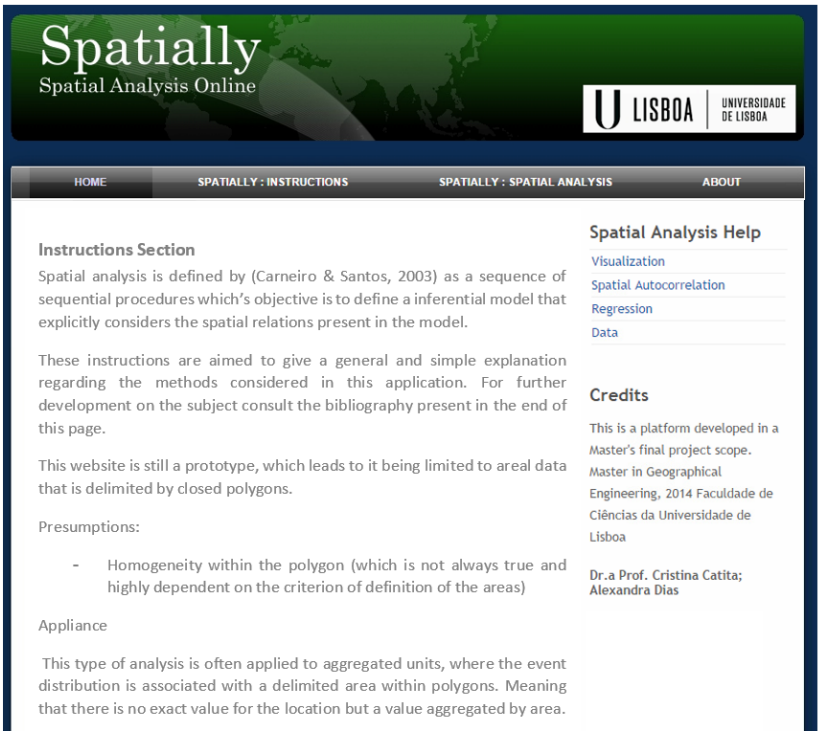


Figure 23 : Spatially: Web GIS Prototype System - Instructions Page

- **About Spatially**

As the title suggests, this page provides extra information regarding Spatially project. In this page, the project is presented along with some objectives and the description of the context within which the prototype was created.

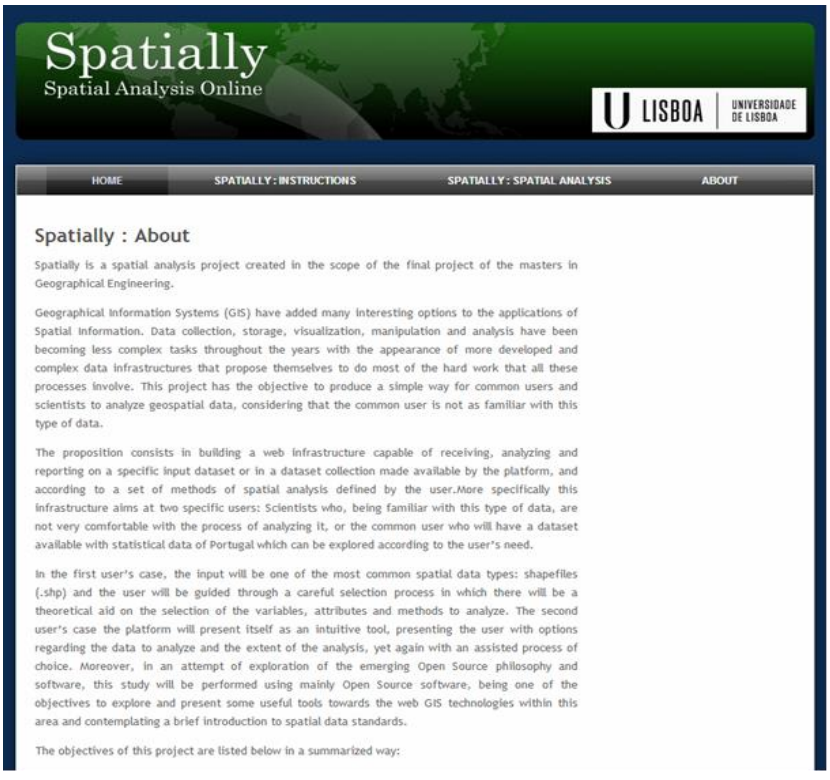


Figure 24 : Spatially: Web GIS Prototype System: About Page

Figure 25 : Spatially Web GIS Prototype System: Spatial Analysis Page.

- **Spatially : Spatial Analysis**

Spatial Analysis page provides a form for the user to fill with the parameters that define the spatial analysis methods to be applied and the dataset that they will be applied to. When submitted, this form will initiate the download of the pdf file that constitutes the report from the specified analysis. The report is presented in detail in the fifth chapter of the present document.

4.1.7. Website structure

- **Inner Structure**

As mentioned before, Django is based on an MVC model. This means that every page presented includes at least three core components that are dependent but separately managed. Django's main unit is called Project. It is created automatically using the command line and with it all the files that are necessary for the project. The project is the equivalent to a webpage since it includes all the parts that will make the connection between separate applications and the final client. The file structure created with a project is presented below:

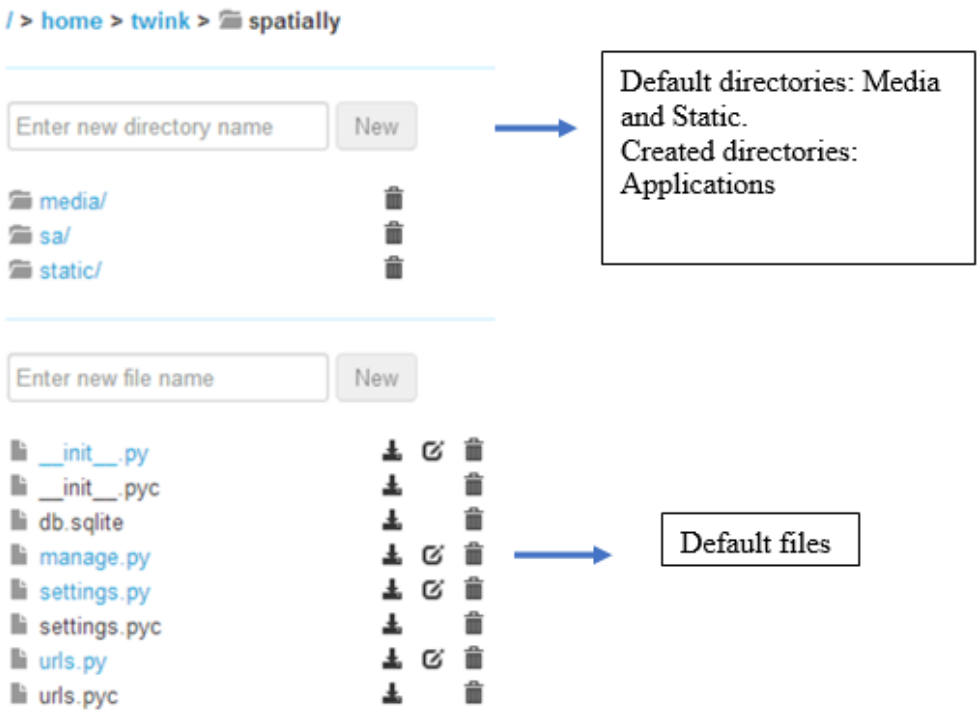


Figure 26 : Django Project File Structure - Spatially Example

Django supports different applications that are 'requested' using the urls file. This process guarantees that the structure is modular and independent and that the features in it are somehow reusable. The structure of an application is presented in figure 27.

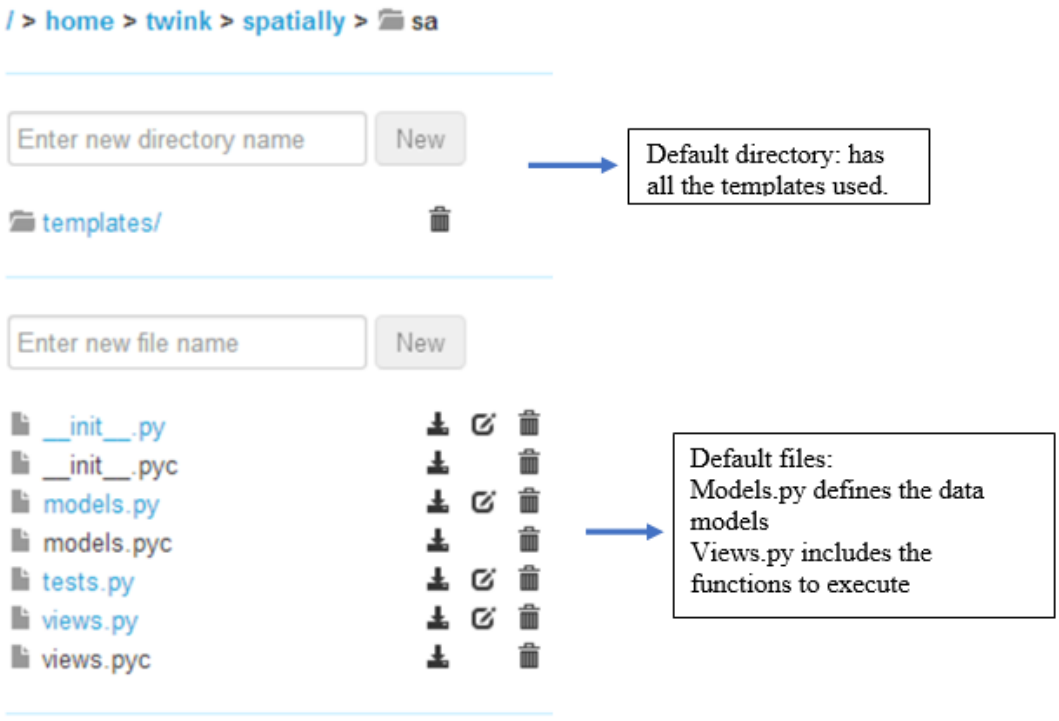


Figure 27 : Django Application File Structure - Spatially: Sa (spatial analysis).

- General Structure

Figure 28 displays the geral structure of the website prototype. All the pages implemented are schematized in the mentioned figure illustrating the relationship between the built pages. Furthermore, the Login functionality, the Forms and the implemented Modules will be discussed over the next chapter.

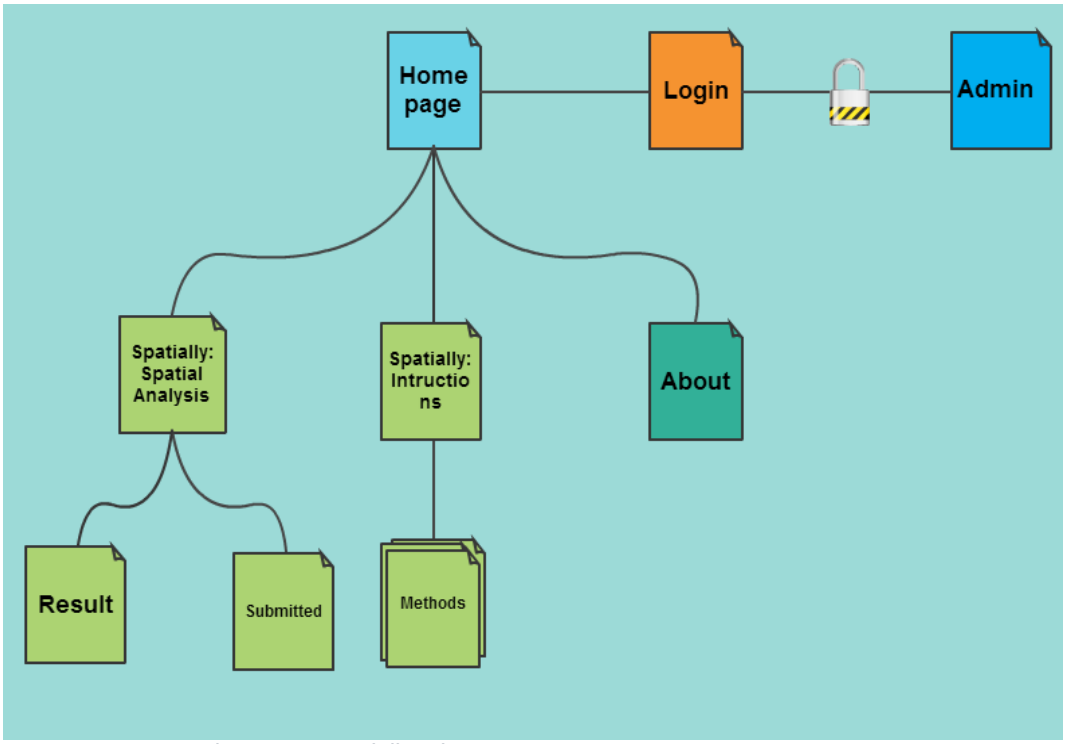


Figure 28 : Spatially: The Prototype System Structure.

4.2. Implementation obstacles

The most significant part of the problems encountered during this project are related with the software or its architecture.

Database installation and the external server have presented no problems in this project, the documentation (PostgreSQL, 2014; GeoServer, 2012) has all the proper procedures, nevertheless it may be dependent on the system used.

Pre-requisites and required libraries on the other hand may present themselves as a complex challenge. This part of the project was probably the most extensive part. Setting up the whole environment takes time, persistence and patience and is often facilitated by the help of experts in specific areas. Looking for help online, reaching out for help from specific problem or networking are all approaches to consider while setting up this type of structure.

Considering the problems in installing all the required dependencies, the approach towards the architecture has suffered some changes along the process.

Even though the first considerations were towards a local installation of the web framework while the server side would support the server and the database, as the problems arose it became increasingly more rational to include all the components on the server side.

This approach is considerably more efficient for publishing data, but not very practical for development. Most of the code was tested locally and transferred afterwards to the server.

In this specific case the first approach was to use the database from the server and use python on Django locally. This option revealed itself a lot more complex than predicted due to conflicts within the system, so the architecture was planned in order to have all the components installed on the server.

Import errors of libraries and prerequisites are frequent in this part of the project, resulting in a time consuming phase. For this purpose precompiled packages (such as OSGEO4W (OSGeo, 2014) if working on a windows environment) are recommended. Other possible solution is to host some of the components on a pre-configured external server, special attention has to be paid to security and access restrictions that may lead to loss of flexibility in the communication between the system components (user interface, web framework, server and database).

It is also important to consider that spatial data requires specific extensions and therefore there is no complete framework that allows its storage, visualization and analysis without using external tools to grant these capabilities.

In this specific case the final solution includes an external server to provide the database service and Geoserver as well, and Pythonanywhere.com was used to host Django and all the python libraries in order to avoid installation obstacles.

4.3. Chapter Summary

This chapter comprehended the main steps of the implementation of the system prototype. Several questions arose during the process of implementing the system prototype, mostly related with its conceptual definition and architecture implementation.

Conclusions regarding this chapter point to an attentive exploration of the dataset before advancing to any database implementation.

Furthermore it is important to mention that installation errors are unavoidable and it may be necessary to shift the approach towards the architecture depending on the situation.

Technology is also an evolving field, which may lead to the appearance of different and more suitable tools for the task to be accomplished. It is therefore advisable to be alert for new possible approaches. Some examples of this are HTML builders and form builders that represent a useful tool for this end.

5. System Enhancement and Spatial Analysis Implementation

5.1. Prototype System Enhancement

5.1.1. Login

The login feature is provided by Django as one of the core functionalities the basic template is made available by the web framework and is activated when the application is set up in manage.py file. The database is also updated with the fields defined for the admin interface and with the creation of new users.

For development purposes a super user is created when the application is activated. This user is applied for developing and testing purposes while the website is still not fully functional (as it was in this case during the development process). The geospatial extension – GeoDjango - does not change the structure of the application neither the web framework, but it does implement modules that facilitate the usage of data with a spatial components. One of the examples is the database handling module that includes a PostGIS option. This option must be defined on the settings.py file and allows Django to communicate directly to a PostgreSQL database with a PostGIS extension, with the advantage of being able to deal with data with a spatial component present in the database. This characteristic is also extensible to the administrator interface as can be consulted in figure 29, implementing an open layers map element that allows the visualization and editing of the data present in the defined database.

The defined template is presented bellow along with the admin component.

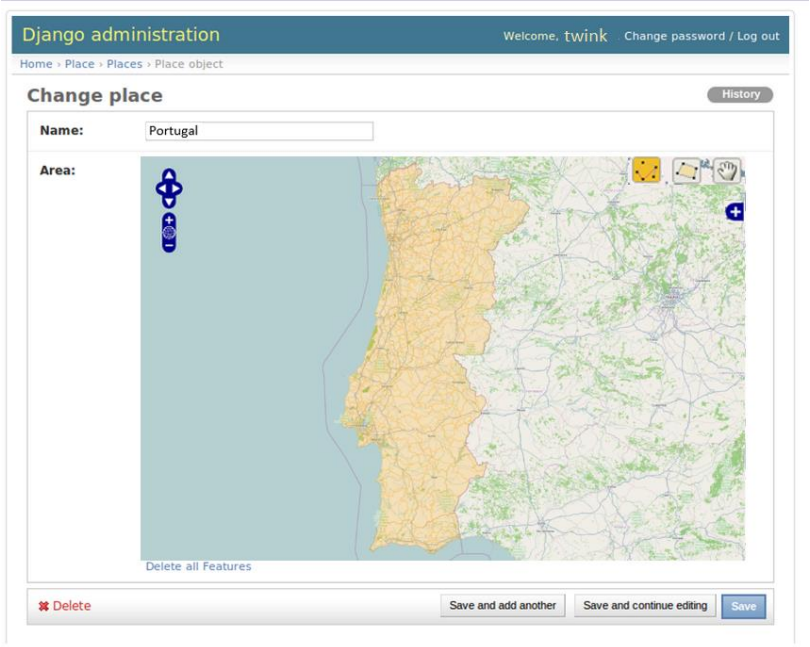


Figure 29 : Spatially: Web GIS Prototype System: Django Admin Web Page

5.1.2. Forms

Forms represent the communication between the user interface and the server. They integrate the variables to be considered and define the user interface in order to make the requirements understandable. Forms were necessary in this case for the user to define both the dataset to be used and the spatial analysis methods to execute. As for most of the templateing options, an external application that aids in the construction of forms was used (PEP, 2011) this app allows the user to define the fields to be inserted and the design aspects of it and provides a zip file with the components necessary to its implementation (HTML, css, images and JavaScript libraries). The resulting form is presented in figure 25 (chapter 4.2.7. on the Spatially: Spatial Analysis section).

5.2. Spatial Analysis Modules

5.2.1. Tools

As mentioned in section 4.2.8., the main python libraries used in this project were: OSGEO (OSGeo, 2014), GDAL (GDAL, 2014), OGR , Numpy (Developers, 2013), Pysal(Pysal, 2014), MatplotLib(John et al., 2012) and ReportLab(Reportlab, 2014).
From the above libraries the first three are used mainly to deal with spatial datasets, projections, shapefile interpretation and most of the operations related with spatial data management. Numpy is used for most of the spatial analysis operations internally, as it deals with array operations. Both Matplotlib and ReportLab are used for reporting and displaying operations, leaving Pysal the most important part of the process – Spatial Analysis.
All the Spatial Analysis modules comprehend all the mentioned libraries, though the most specific part of the process is executed by Pysal.

5.2.2. Implemented Modules

The functionalities implemented in the website were mainly developed in python and were structured according to the schema presented in figure 30. The schema illustrates the implicit relationships and the methodology followed to produce the report. Fundamentally, element.py

is the first code to be executed, calling all the functions specified by the user in the form. There are some hierarchical relationships according to the required inputs to the specified methods.

The output of elements is an array of elements (text, numerical results or images) to be printed in report.py into a pdf report to be presented to the user. In order to demonstrate the implemented functionalities one of the datasets made available by the website was used.

This dataset is composed by rates of several demographic characteristics. These rates were obtained by dividing the characteristic to evaluate by the number of people present in each municipality, in an attempt to create a more homogeneous areal unit.

As an example of the applied methods, a dependent variable was selected: rates of families without unemployed people (number of families without unemployed people divided by the number of families). All the visualization and exploratory methods were applied to the dependent variable for this example, while regression methods required additional independent variables. The presented results: both graphics and indexes were calculating using the implemented Python modules, and Getis and Ord Analysis are not implemented due to the existence of null values which cannot exist in this model.

The chosen independent variables (variables that are suggested to explain the dependent variable) were:

- Rate of retired individuals (number of retired individuals divided by the number or present individuals);
- Rate of individuals that work in the primary sector (number of individuals who work in the primary sector of activities divided by the number or present individuals);
- Rate of individuals with superior degree (number of individuals with superior degree divided by the number or present individuals).

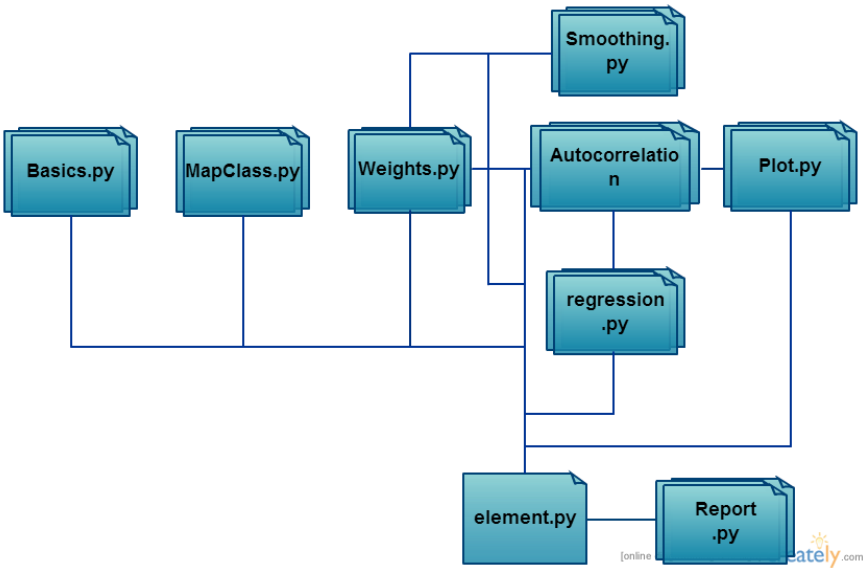
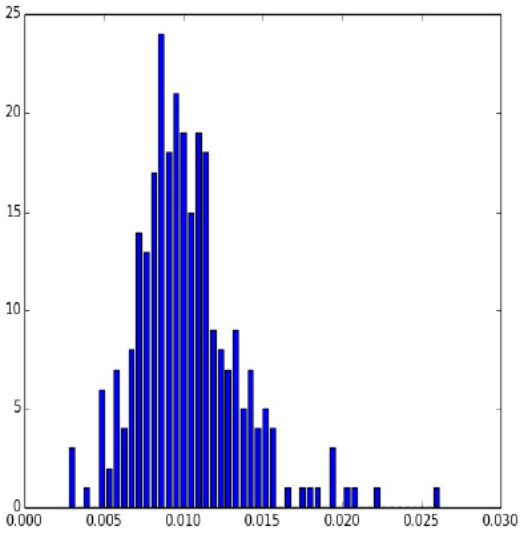


Figure 30 : Spatially Module Structure

- Visualization

Visualization is the first part of spatial analysis representing the first approach to a dataset. It is useful for having a global idea of the dataset characteristics. Histogram and box plot were already presented over the 2nd chapter, as for map classifiers, their programming definition in Pysal is mentioned bellow along with the methods available in Pysal.



- Histogram

The Histogram displays the amount of observations present in the dataset for a certain value. For this specific case, it can be concluded that there are more rates between 5% and 10% of the population. Nonetheless, the values take a range that goes from around 3% up to 27%.

Figure 31 : Histogram: Rate of unemployed people looking for their first jobs.

- Map Classifiers

For an array y of n values, a map classifier places each value y_i into one of k mutually exclusive and exhaustive classes. Each classifier defines the classes based on different criteria, but in all cases the following hold for the classifiers in PySAL:

$$C_j^l < y_i \leq C_j^u \ \forall i \in C_j \qquad \textbf{(15.)}$$

Where C_j denotes class j which has lower bound C_j^l and upper bound C_j^u.

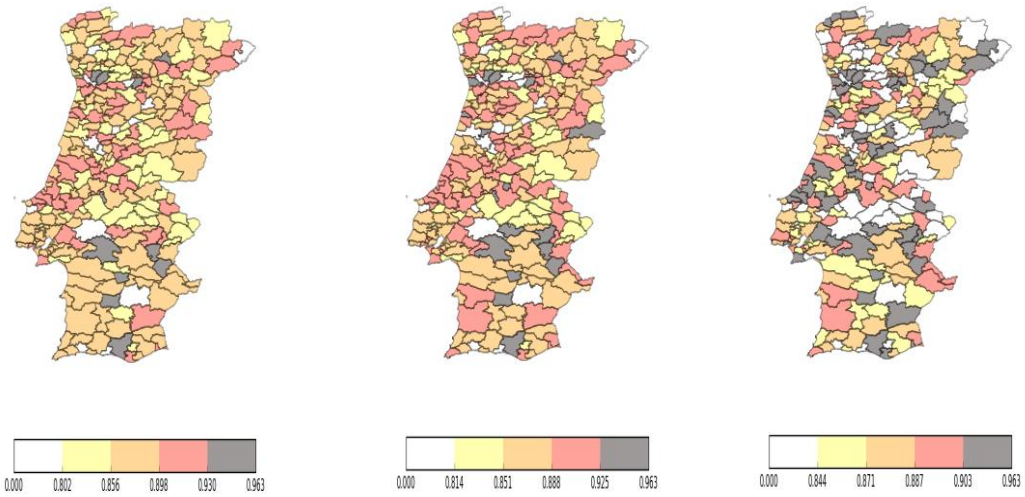


Figure 32 : Map Classification: Quantiles, Equal Intervals and Fisher Jenks Methods (correspondently) – projected in WGS84.

- Weights

Spatial weights matrix expresses the potential for interaction between observations at each pair i,j of locations. It is calculated by the modules based on contiguity criteria, distance criteria or according to kernel weights.

- Exploratory Spatial Analysis – Spatial Autocorrelation

- **Moran**

Moran's index indicates a positive global spatial correlation, indicating that there is a tendency for similar values to be found together (either positive or negative). Figure 33 below shows the empirical distribution of Moran's Index in relation to the expected value (vertical

red bar), while figure 34 presents the distribution of the permutation values in comparison with a normal distribution (in red). The values obtained by performing the permutations (999 by default) is approximately coincident to the normal distribution line presented.

Global Autocorrelation :Moran Index, Variable: taxas_fam

Moran's I: 0.608529731716

Moran's I expected value: -0.00361010830325

Average value of I from permutations: -0.00378044653297

Variance of I from permutations: 0.00138358604814

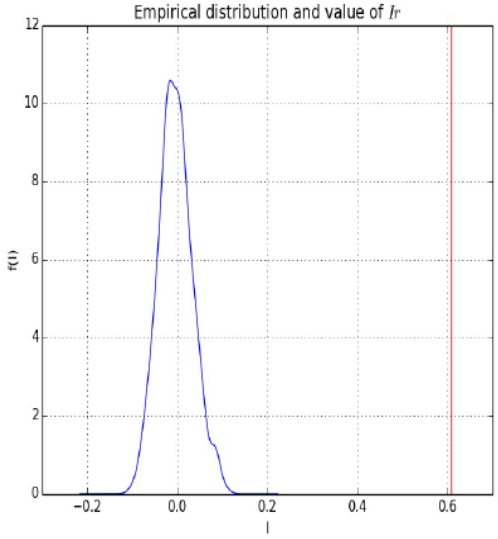


Figure 33 : Moran's I empirical distribution (expected distribution in blue and actual value in red).

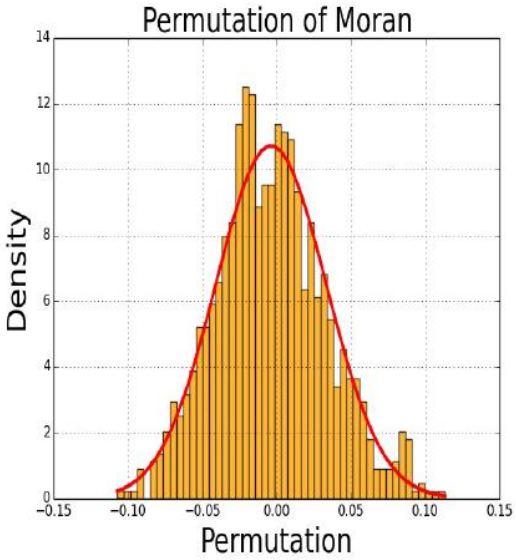
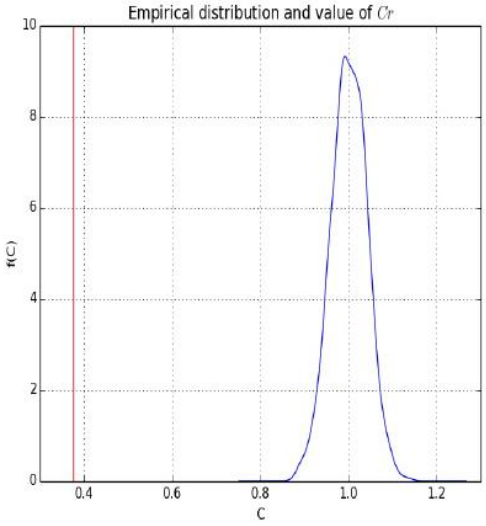


Figure 34 : Permutations of Moran's I index in comparison with the normal distribution line.

- **Geary**

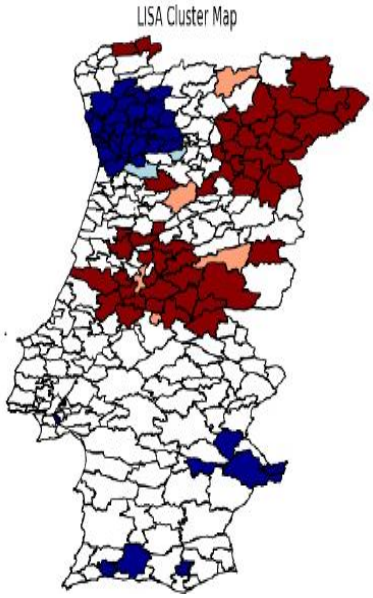


As expected for the result obtained for the Moran's I index, Geary's C takes a value close to zero, revealing a positive spatial autocorrelation. The empirical distribution of the index value obtained and the expected value for inexistent spatial autocorrelation is displayed in figure 35.

Figure 35 : Empirical distribution and value of Geary's C index (The expected distribution in blue and actual value in red).

Global Autocorrelation :Geary Index, Variable: taxas_fam

Geary's C: 0.376137654345
Geary's C expected value: 1.0
Average value of C from permutations: 1.00087846635
Variance of C from permutations: 0.00163355154272



- **Local Moran**

Local Moran index (also known as LISA) calculates spatial autocorrelation locally. The map shown in figure 36 indicates a high cluster tendency in the Center and Northeast regions of the Portuguese Continental Territory and a disperse tendency for the Northwest Territory. Besides these obvious two regions and a smaller disperse region in the south of the country, the country doesn't show clustered regions, meaning that in the other regions of the territory, the values of the neighboring municipalities is not similar between them.

Figure 36 : Lisa Cluster Map for rates of families without unemployed people.- projected in WGS84.

- Regression:

Regression is intended to capture spatial dependence by applying regression methods. It can either be Non spatial or spatial and further information on the subject can be found on chapter 2 (2.2.7.). Both the dependent and independent variables chosen for this purpose are presented in the beginning of this topic.

Ordinary Least Squares:

The Ordinary Least Squares output is presented in Figure 37. The R value obtained indicates that the independent variables (rate of retired people, rate of people with a higher education and rate of people working for the primary sector) explain around 45% of the variation of the dependent variable (rate of families without unemployed people). And the F test reveals some significance of the coefficients overall. Multicolinearity condition number is 13 which is

considerably below 30 (which is ideally the limit that this number should take, being below 100 still makes this number feasible (Scott, 2009)). Looking to the Hesteroskadasticty section of the report, there is probably a problem. Given the statistical significance of the Lagrange Multiplier results, the more adequate model for this case will probably be the error model, but both of them will be performed in this document's scope.

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var              Number of Observations:         278
Mean dependent var  :      0.8741              Number of Variables   :           4
S.D. dependent var  :      0.0350              Degrees of Freedom    :         274

R-squared           :    0.451485
Adjusted R-squared  :      0.4455
Sum squared residual:       0.186              F-statistic           :     75.1769
Sigma-square        :       0.001             Prob(F-statistic)     :    1.66e-35
S.E. of regression  :       0.026             Log likelihood        :     621.582
Sigma-square ML     :       0.001             Akaike info criterion :   -1235.163
S.E of regression ML:      0.0259             Schwarz criterion     :   -1220.653

------------------------------------------------------------------------------------
    |      Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
    |      CONSTANT       0.7654200       0.0093472      81.8874477       0.0000000
    |       taxas_1o      0.1077903       0.0790519       1.3635397       0.1738317
    |    taxas_enSu       0.1473941       0.0500595       2.9443754       0.0035136
    |     taxas_ref       0.3436237       0.0251556      13.6599352       0.0000000
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER        13.820521
```

```
TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2       10.133871        0.0063017

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                3        9.178237        0.0270126
Koenker-Bassett test              3        9.176754        0.0270308

SPECIFICATION ROBUST TEST
TEST                             DF        VALUE           PROB
White                             9       17.409754        0.0426728

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1        5.601684        0.0179432
Robust LM (lag)                   1        2.454003        0.1172256
Lagrange Multiplier (error)       1      203.604921        0.0000000
Robust LM (error)                 1      200.457240        0.0000000
Lagrange Multiplier (SARMA)       2      206.058924        0.0000000
```

Figure 37 : Ordinary Least Square Output Example.

-Error Model

Error model has a pseudo R-square value instead of the R^2 coefficient. This value is the correlation coefficient betweem the dependent variable (y) and the predicted values for the model, and representes an indicator of how right the model is (eventhough it is not equivalent to R^2). The error model in this case has a pseudo-R^2 value of approximatley 45%. Consulting the Lagrange Multiplier presented before, the error model is the most adequate model for this set of variables, nonetheless, the lag model will also be calculated for the example purpose.
The regression coefficient is statistically significant and the equation that expresses the model is presented below.

```
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED LEAST SQUARES (HET)
-----------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var               Number of Observations:         278
Mean dependent var  :      0.8741               Number of Variables   :           4
S.D. dependent var  :      0.0350               Degrees of Freedom    :         274

Pseudo R-squared    :    0.450153

------------------------------------------------------------------------------------
      Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
      CONSTANT       0.7687694       0.0107421      71.5657088       0.0000000
      taxas_1o       0.1544132       0.0896460       1.7224767       0.0849832
     taxas_enSu      0.1305289       0.0393099       3.3205095       0.0008985
      taxas_ref      0.3191200       0.0245583      12.9944039       0.0000000
        lambda       0.1383939       0.0088499      15.6379815       0.0000000
------------------------------------------------------------------------------------
```

Figure 38 : Error Model Output Example.

$$\mathbf{Y = 0.77 + 0.14\ (\ 0.15\ x\ A + 0.13\ x\ B + 0.32\ x\ C\ ) + \varepsilon\ ,} \qquad \mathbf{(16.)}$$

Where:
A = Rate of employed people in the primary sector
B = Rate of people with superior education
C = Rate of retired people
And $\varepsilon = 0.14 \times W + \xi$
(W is the weight given to the variable and ξ represents the residuals)

- Lag Model

The lag model presents a similar value to R^2 as the error model, however, coeficients and their statistic significan vary accordingly. The most suitable model for this specific case is the error model, the equation that expresses the model is presented below.

```
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
-----------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var              Number of Observations:         278
Mean dependent var  :      0.8741              Number of Variables   :           5
S.D. dependent var  :      0.0350              Degrees of Freedom    :         273

Pseudo R-squared    :      0.4619
Spatial Pseudo R-squared:  0.4563

------------------------------------------------------------------------------------
      Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
      CONSTANT       0.7562560       0.0105200      71.8875689       0.0000000
     W_dep_var       0.0018794       0.0010495       1.7908566       0.0733163
      taxas_1o       0.0986719       0.0779021       1.2666137       0.2052935
    taxas_enSu       0.1545028       0.0493858       3.1284876       0.0017571
     taxas_ref       0.3440082       0.0247377      13.9062566       0.0000000
------------------------------------------------------------------------------------
Instrumented: W_dep_var
Instruments: W_taxas_1o, W_taxas_enSu, W_taxas_ref

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Anselin-Kelejian Test              1      172.011786       0.0000000
```

Figure 39 : Spatial Lag Output Example.

$$Y= 0.002 \times 0.76 \times Y + (0.10 \times A + 0.15 \times B + 0.34 \times C)+ \varepsilon \qquad (17.)$$

Where:
A = Rate of employed people in the primary sector
B = Rate of people with superior education
C = Rate of retired people

Eventhough this model is only inteded to be used as an example for this document, it has values that are quite interesting from an anlysis point of view. Though the model discards multicolinearity and heteroskedasticity is managed by appling an error model that consideres hetroskedasticity in the non-spatial regression, it also has a statistically significant Jaque-Bera index, that points to normal distribution of the residuals. This situation often requires the adition or redefinition of the variables to be included in the model, suggesting that there is a key variable missing. Overall, the conclusion points to an incomplete model, requiring extra variables, and the rate of people employed by the primary sector is not statiscally significant and should, as a result of it be removed from the model. Aditionally, the error model is the

best fit for this case (as revealed by the Langrange Multiplier and further along confirmed by the obtained coefficients).

5.3. Report implementation

The report functionality was implemented recurring to reportlab library in python. All the elements gathered by elements.py function are received by report.py, and are printed in a pdf file. The visual graphics (plots, tables and map images are assured in each of the modules and inserted directly into the elements array, storing all the data to be printed in the user's report.
The final output is presented on section 4, in the attachments. It includes a heading and all the required spatial analysis along with the dataset defined by the user.

5.4. Chapter Summary

The present chapter resumes all the functionalities added to the basic prototype system developed over the previous chapter.
Login and Admin functionalities were included on the website as well as the forms that will accept and process the required information to execute the spatial analysis methods proposed.
Furthermore, an extensive overview over the python modules built as well as their inputs and outputs are presented in a schematic approach that reveals the logic process associated with the module building within this project. The final result of the project is also presented along with the implemented report, originated by the report module. To the date of this report, the PDF file generated had still some alterations to be done concerning design, as it is still a prototype version, implicating the functionalities but yet to contemplate designing issues.

6. Conclusions and Recommendations

6.1. Summary of Research

The present project had several interesting twists as it was being implemented. It had in general three main goals bounded together by their results, they were: exploring viable architecture options to build a platform of significant stability and with a spatial extent, understanding the dimensions of the open source community and the application of this kind of project in both the academic and the business world and dealing with spatial data in a different environment, exploring existing tools and the known methods for spatial analysis.

There is a very solid conclusion to be drawn out of a considerable amount within the bibliography with a general content (Zhang et al., 2010; She et al., 2012; Kwakkel et al.; 2012) and also among the bibliography of applied case studies (Carneiro & Santos, 2003; Mateu & Uso, 1998, Beale et al., 2010) which point to a gap between spatial information and its application concerning tools.

Furthermore, the generalization of the web as a mean of communication, sharing and exploration of knowledge and data, carries several implications that have to be carefully considered and discussed in order to make this kind of project available to the general public. It is important, and it was also mentioned by some authors, to promote the awareness of spatial data in several applications as it was explored in (Beale et al., 2010), but data issues such as conceptualization and data quality which were documented in chapter 2 should never be underestimated.

The conclusion of this project was a step forward in a beginning of a journey that will take spatial data and its application to a whole new level, promoting sharing and exploration of data online, inexpensively and within a community environment.

So as this project came to an end, it is important to revisit most of its components, to understand their differences, potentials and limitations and to explore a little further the possibilities arisen by this research as well as future expectations within this area.

To begin this last chapter, the problem under consideration will be revised and dissected according to the project's results. The second topic will revise the methodology chosen and

bring the most important aspects to consider in this subject when building this type of system prototype. Whilst the third topic will revisit problems and obstacles encountered during this project.

Practical implications in the sector will be presented in the fifth item and further explored in the following topic where perspectives to future work will be considered and discussed.

6.2. Conclusions

In general, as most projects, this project revealed itself to be much more extensive than it was thought to be in the begging. There are still functionalities that are not fully implemented due to technical obstacles such as database communication. Though this document explains most of the process and introduces all the required mechanisms it does not contemplate configuration situations that would be required in order to make the prototype fully functional. As it was mentioned in the first chapter, the focus of this project was directed to the exploration of the related technology, spatial analysis methods and open source environment. As an example of this situation is the security limitations that were imposed for the client to upload data to the database, being the database hosted in a remote server.

There is still a long way from this prototype to the real potential of the platform itself and even though this first approach may lack in processing capabilities, the concept that generated it has all the research and most tools necessary to build a fully implemented web platform for spatial data sharing and analysis within an open source environment.

The open source environment, even though it's becoming increasingly more explored and navigated, is still a territory with little support and directed to a very specific population of knowledgeable people within the programming extent. Great efforts towards complete documentation are being made but it will definitely represent more time spent on the project if there is no immediate support provided (as it is a free service). Depending on the project's dimension, all the advantages and disadvantages must be weighted in order to make the wisest choice, being aware that open source tools will require more time and effort that commercial ones while commercial tools will represent a significant extra budget and no flexibility for results at hand. Nonetheless, if there is a particular project to which one believes to have the expertise to contribute, there must always be active people on these projects and since they

are open and free, there is always the need for contributions. As in most areas the life, the key is communication.

Generally speaking, from the open source projects used (PostgreSQL- PostGIS, Django, Geoserver, Python – OSGEO, GDAL, Pysal, MatlplotLib, ReportLab, Numpy), it was an overall interesting experience of exploration, and most of the documentation is well structured, being the most challenging part creating such an environment in the machine (prerequisites, missing libraries and other time consuming obstacles can be rather demotivating when beginning and generally along the journey).

The first idea for this project was to build a sharing data platform in which spatial data could be shared among communities (especially scientific), with reliable metadata and the possibility of contacting the user to complete any missing information about the data. This idea had to be set aside for technical reasons related with the storage of the data, but it would certainly represent an interesting tool for scientists and people in general.

As for the quality of the data, statistical data form INE was chosen because it was considered to be more appealing to a greater amount of users, carrying useful information for several applications. The data uploaded by the user would be dependent on the application that the user wants to make of it, being the user responsible for consulting the documentation provided on the website and exploring the data adequately according to the objective proposed. It is important to quote Anselin (2006) on this subject to underline that 'The importance of the statistical insights lies in the quantification of the uncertainty associated with various estimates and in exploiting the spatial characteristics of this uncertainty in the decision process.'

The Spatial Analysis module has an interesting amount of options. However, (Beale et al., 2010) presents a table with all the existent spatial analysis modules in Pysal and there is still a wide variety of methods that can be considered when performing spatial analysis. The methods that were applied were mostly presented by Rey ((Rey et al., 2005), (Rey, 2007), (Rey, 2013) and Anselin (Anselin et al., 2012) and represent the most frequent appearances in the bibliography. Once again it is important to refer that the idea of this project was not only to build a specific platform but also to provide unspecialized people with the option of using these tools.

6.3. Conclusions regarding the followed methodology

There were, as referred in the first part of this chapter, three main focuses to this project, to which the conclusions will be presented in the following paragraphs.

Regarding the architecture, there are plenty of sophisticated open source tools available within the Open Source community, making it possible to create this type of infrastructures within any environment (academic, commercial, personal..) nonetheless, the documentation towards Open Source projects is not always as consistent as desirable and often a look at the code can make a difference for a coder (which will definitely be of no use for someone whom had no contact with code before).

The database option – PostgreSql with a PostGIS extension is, at the moment and as referred by the bibliography exposed in chapter 3 (section 3.2.1.) the most mature open source database that has an adequate spatial extension. Projects have to have a focus and PostgreSQL has, from an early stage, targeted the spatial features, making it very interesting for this purpose.

Python is definitely a practical language and combined with web frameworks such as Django can have impressive results, however, python is known to be rather slower than other languages which can pose an obstacle in some situations. Among all the options considered, python presented itself as the most valuable language, having an incredible amount of libraries that can be combined towards a specific goal (some of them were presented in chapter 3).

Django, although being an interesting framework has a complex structure, resulting in a frustrating learning curve and sometimes in a confusing organization for the applications. It does, however perform its task adequately, and having python as an advantage for this specific case.

Spatial analysis, which was to be the main focus of this project has become the less challenging part as the methods are quite well documented in the bibliography (Rey, 2007; Anselin et al., 2012; Anselin & Rey, 2012; Smith et al., 2013; Rey et al., 2005; Haining, 2004). Although there are innumerous methods to be applied in spatial statistics, Pysal has some of the most common options, which made it easier to apply.

When doing the research for this topic it became clear that most of the users to this kind of tools are often from other areas and many times unaware of its existence or afraid of applying them incorrectly (Beale et al., 2010).

6.4. Research contributions

The proposed project can have innumerous applications. The dataset made available is still only applicable to the Portuguese continental territory and is still quite limited to the statistical data from the Portuguese census. However, if transferred to a bigger server and remodeling the database this project can be applied to a great amount of areas as defined in the bibliography (Rangel et al., 2010; Rosenberg & Anderson, 2011; She et al., 2012) such as: ecological, epidemiology, geology, geography, mathematics, environmental science, hazard and risk assessment ..

In fact, the present platform prototype intends to allow users to perform guided special analysis in their own spatial data or in a selected dataset without the need of installing or downloading any specific software, using an open source resource.

Furthermore, as suggested by Zhang et al (2010), this is a further effort to promote spatial intelligence adaptable to users across interdisciplinary fields with different decision levels, by suggesting new advances in the availability of spatial analysis tools.

Modules included in this spatial analysis platform comprehend visualization of the data, spatial autocorrelation and regression contemplating most of the most common spatial analysis methods that are usually applied for the mentioned objectives.

A Cyber Infrastructure (CI) referred in (Anselin & Rey 2012) is a high-performance computing infrastructure , allowing access to distributed data and sensor information, visualization, data analysis, and the establishment of collaborative networks of scientists. Considering this is a very large step to be taken on such a limited time and resources, the approach to this definition of CI seems to be getting almost complete, being the rest of the components mentioned included in the perspectives of future work in this area (chapter 6.6.).

The prototype suggested has a limited space and limited capabilities due to the limited amount of time and resources, however, the application of the prototype to a larger scale, and if

considering the perspectives and proposals made in the sixth part of this chapter it becomes quite clear that the potential for such a tool is tremendous in some different branches that will be further along explored in this chapter.

There are several examples of possibilities of application of such a tool, which are briefly described below:

- Public safety: to share maps of roads, bridges, electrical grids, water systems, buildings, ect. And to better plan for and respond to emergencies and disasters.

- Public health community: to share location based information securely and track pandemics, analyze trends and monitor population health

- Local population community to connect people and communities, map the future and realize opportunities and environment and sustainable development to better manage land and water assesses the environment and monitor ecosystems.

6.5. Limitations

Data consistency in spatial data collections is still a big obstacle in this type of data's management. The amount of spatial data available is increasing, as documented in the first chapter leading to a need to standard definition. Though it is beyond the scope of this project, it is important to mention the INSPIRE directive. The referred directive addresses this question, by proposing a European Union (EU) spatial data infrastructure. The objective is to enable spatial information to be shared within public sector organizations and facilitate public access to said data across Europe (European Comission 2007).

There are problems of different areas involved in this kind of project. There are always conceptual problems – related with the data storage and management, punctual problems and technical problems.

As in most of the projects involving technologies it is very common to deal with several obstacles related with software setup and utilization. It is important to notice that most of the software used during the project is Open Source, therefore technical problems have no specific technical support. Most of the cases have to be investigated within a programmers'

community such as stack overflow(stack exchange inc 2014b), Gis Stack exchange (stack exchange inc 2014a) or similar.

Most of the obstacles related with software occurred in its installation or setup. There is always a solution in this kind of situation and a list of possible options to be taken if problems such as the problems mentioned above arise:

- Change environment: Windows can have the tendency to arise problems related with libraries and often there are collisions with different versions of the same library. Creating a virtual environment to use a different operating system (for example Linux) may be advisable.

- Change machine: Sometimes allocating all the process to the server may be an interesting idea. First and foremost because it doesn't use your resources and secondly because most of the conflict problems are already dealt with (these services are often in unix). For developing, however, it is always better to have at least the python and python libraries locally, otherwise it may get difficult to work online all the time. Hosting servers are available with a range of different approaches, space capacity and installed software. It will always depend on the resources and the specific requirements, but it is always an interesting option.

Data input also revealed itself to be an obstacle in this project. In the case of remote servers, there are often security protocols that do not allow introducing data in the database from files that are introduced online. This was a really interesting idea that ended up being abandoned because of this problem (that would probably be solved easily if it was being developed locally or with different settings).
Even though data management and conceptualization was a challenging step of the process, the most significant problems arisen were related with software and architecture, mainly technical questions.
All the libraries are well documented and the theory related to each of the segment has extensive bibliography to support it, making this part the less turbulent part of the process.

6.6. Recommendations for future work

As mentioned before, there is still a lot to be developed in this direction that due to time and technical constraints was not possible to achieve in this project. In the next few paragraphs suggestions will be presented as to future approaches to this subject that can represent interesting projects:

- Sharing environment – technical issues have made the idea of sharing data not applicable in this prototype. The suggestion to consider here is adding a form to allow the users to introduce a file of their own in the common database. To this file there must be associated features such as metadata, the author, the method of acquiring, the associated precision, the data of acquisition and other interesting features that may be of use to future users. Besides the data made available in the database, the direct contact to the author of the data must be assured – weather inside the platform or providing an e-mail to future users of the same dataset. The login element will make this situation simpler, since it can automatically insert the user's data into the database.

- Standardization – Data standardization has also been an interesting topic to approach. OGC standards, as mentioned in Chapter 3 have been becoming increasingly more common and the tendency towards interoperability is often mentioned in the bibliography (Ballatore et al., 2011) as a way of facilitating the design of interoperable open source GIS tools. Whereas standardization is often mentioned by the bibliography an urgent data for spatial data, it is not very common to find references to INSPIRE, which can be found to be rather intriguing. That would be, nonetheless, other interesting suggestion towards future developments of this project.

- Virtual environment – As this platform is based on the server side, most of the computational work, and storing capacity is being held by the server. This type of architecture carries innumerous advantages as the computational capacity may be a lot more flexible than if it was stored on a traditional machine. Virtual machines (which would probably have to be hired to this end or have some sort of agreement to provide some flexible space), have the advantage of managing their space in order to be close to infinite or re-oriented towards a specific task at a certain moment. This is definitely

one of the greatest advantages of producing a web platform: it can have as much space as defined by the server.

- Integration of further methods – Spatial analysis concerns a wide range of data and of fields. There are several interesting methods that can be applied to all types of data such as surface and field analysis, network and location analysis and geocomputational methods and modeling are some of the areas that may be of interest to explore in the geospatial context. Data mining is also a very interesting area, which in spite of being out of the scope of this particular project may make sense if integrated in a wider idea of a data exploration platform.

References

Amrita A. Manjrekar, R.V.M., 2012. *Implication of Image Processing in GIS and Remote Sensing.* Conference: National Conference on Recent Advancement in Engineering, Volume: 3

Anselin, L., 2008. *Global spatial autocorrelation*. Spatial Analysis Course.

Anselin, L., 2006. *How (not) to lie with spatial statistics*. American journal of preventive medicine, 30(2 Suppl), pp.S3–6.

Anselin, L., 1995. *Local indicators of spatial association — LISA. Geographical Analysis*, Geographical Analysis Volume 27, Issue 2, pages 93–115

Anselin, L., 2009. *The Future of Spatial Analysis in the Social Sciences.* Geographic Infomation Sciences: A Journal of the Association of Chinese Professionals in Geographic Information Systems, 2(5), pp.67–70.

Anselin, L., Amaral, P. V & Arribas-bel, D., 2012. *Technical Aspects of Implementing GMM Estimation of the Spatial Error Model in PySAL*. , pp.1–20.

Anselin, L., Kim, Y.W. & Syabri, I., 2004. *Web-based analytical tools for the exploration of spatial data.* Journal of Geographical Systems, 6(2), pp.197–218.

Anselin, L. & Rey, S.J., 2012. *Spatial econometrics in an age of CyberGIScience.* International Journal of Geographical Information Science, 26(12), pp.2211–2226.

Azavea, 2012. *Azavea - Beyond dots on a map. Advanced GIS solutions.* Available at: http://www.azavea.com/ [Accessed May 15, 2014].

Ballatore, A., Tahir, A. & Mcardle, G., 2011. *A comparison of open source geospatial technologies for web mapping*. Int. J. Web Engineering and Technology, pp.1–21.

Beale, C.M., Lennon, J. J., Yearsley, J.M. , 2010. *Regression analysis of spatial data.* Ecology letters, 13(2), pp.246–64.

Berry, B.J.L. & Marble, D.F., 1968. *Spatial analysis: a reader in statistical geography*, Englewood Cliffs, New Jersey: Prentice-Hall.

Borges, K.A.V., Davis, C.A. & Laender, A.H.F., 2001. OMT-G: An Object-Oriented Data Model for Geographic Applications. *GeoInformatica*, 5(3), pp.221–260. Available at: http://link.springer.com/article/10.1023/A:1011482030093 [Accessed September 22, 2014].

Buja, A., Cook, D. & Swayne, D.F., 1996. *Interactive High-Dimensional Data Visualization.* Journal of Computational and Graphical Statistics, 5(1), pp.78–99.

Buttenfield, B. & Beard, M., 1994. *Graphical and geographical components of data quality*. Visualization in geographic information systems, 150-157

Cabral, P. da C.B., 2001. *SISTEMAS ESPACIAIS DE APOIO À DECISÃO O Sistema de Apoio ao Licenciamento da Direcção Regional do Ambiente do Alentejo*, Lisbon. Master Thesis

Cagnacci, F. & Urbano, F., 2008. *Managing wildlife: A spatial information system for GPS collars data.* Environmental Modelling & Software, 23(7), pp.957–959.

Carneiro, E.O. & Santos, R.L., 2003. *Análise espacial aplicada na determinação de áreas de risco para algumas doenças endêmicas (calazar, dengue, diarréia, D.S.T. - doenças sexualmente transmissíveis e tuberculose), no bairro de campo limpo - Feira de Santana (BA).* Sitientibus, (28), pp.51–75.

Cloud, G., 2012. *GIS Cloud :: It's about the Apps, not the Maps*! Available at: http://www.giscloud.com/ [Accessed May 15, 2014].

D'Amore, F., Cinnirella, S. & Pirrone, N., 2012. *ICT Methodologies and Spatial Data Infrastructure for Air Quality Information Management*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5(6), pp.1761–1771.

Developers, Numpy., 2013. *NumPy — Numpy*. Available at: http://www.numpy.org/ [Accessed May 17, 2014].

Django, 2014. *GeoDjango*. Available at: http://geodjango.org/ [Accessed May 17, 2014].

Druck, S. et al., 2004. Spatial analysis of geographic data. Available at: http://www.cabdirect.org/abstracts/20053018692.html;jsessionid=99F253D37F3792E2E71CEDE32151BDC1 [Accessed September 15, 2014]. 81

ESRI, 2013. ArcGIS Help 10.1 - What is a z-score? What is a p-value? Available at: http://resources.arcgis.com/en/help/main/10.1/index.html#/What_is_a_z_score_What_is_a_p_value/005p00000006000000/ [Accessed May 15, 2014].

ESRI, 1998. ESRI Shapefile Technical Description. *Environmental Systems Research Intitute, Inc.* Available at: http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf [Accessed May 18, 2014].

European Comission, 2007. INSPIRE. Available at: http://inspire.ec.europa.eu/index.cfm/pageid/48 [Accessed July 15, 2014].

Fischer, M.M. & Getis, A., 2010. *Handbook of Applied Spatial Analysis* M. M. Fischer & A. Getis, eds., Berlin, Heidelberg: Springer Berlin Heidelberg. Available at: http://www.springerlink.com/index/10.1007/978-3-642-03647-7.

Foundation, D.S., 2014. The Web framework for perfectionists with deadlines | Django. Available at: https://www.djangoproject.com/ [Accessed May 15, 2014].

GDAL, 2014. GDAL: GDAL - Geospatial Data Abstraction Library. Available at: http://www.gdal.org/ [Accessed May 15, 2014].

Geary, R.C., 1954. The contiguity ratio and statistical mapping. In *The Incorporated Statistician*. pp. 115–45.

GeoServer, 2012. GeoServer. Available at: http://geoserver.org/ [Accessed May 15, 2014].

Gillenson, M.L., 2011. *Fundamentals of Database Management Systems, 2nd Edition*, John Wiley & Sons. Available at: http://books.google.com/books?id=-eYbAAAAQBAJ&pgis=1 [Accessed August 11, 2014].

Goodchild, M.F. & Haining, R.P., 2004. *GIS and Spatial Data Analysis: Converging Perspectives*. , Papers in Regional Science 83(1), pp.368–385.

Haining, R., 2004. *Spatial Data Analysis Theory and Practice*, Cambridge, United Kingdom: Cambridge University Press.

Holt, D. et al., 2010. *Aggregation and Ecological Effects in Geographically Based Data.* Geographical Analysis, 28(3), pp.244–261.

Holt, D., Steel, D.G. & Tranmer, M., 1996. *Area homogeneity and the Modifiable Areal Unit Problem*. Geographical Systems, 3(2-3), pp.181–200.

Hubert, L., Golledge, R. & Costanzo, C.M., 1981. *Generalized procedures for evaluating spatial autocorrelation.* Geographical Analysis, 13, pp.224–233. 82

Instituto Nacional de Estatística, 2011. *Instituto Nacional de Estatistica, Censos 2011*. Available at: http://censos.ine.pt/xportal/xmain?xpid=CENSOS&xpgid=censos2011_apresentacao [Accessed May 15, 2014].

Hunter, J., Darren, D., Firing, E., Droettboom, M., 2012. *matplotlib: python plotting — Matplotlib 1.4.0 documentation.* Available at: http://matplotlib.org/ [Accessed May 17, 2014].

Kraak, M.-J. & Ormeling, F., 2010. *Cartography: Visualization of Spatial Data* 3rd Edition., Routledge.

Kwakkel, J.H. , Carley, S., Chase, J., Cunningham, S. W., 2012. *Visualizing geo-spatial data in science, technology and innovation.* Technological Forecasting and Social Change, 81, pp.67–81.

Lane, D.M., 2007. *Online Statistics Education: A Free Resource for Introductory Statistics*. Available at: http://onlinestatbook.com/2/index.html [Accessed June 19, 2014].

Lee, K., 2009. *Technical architecture for land monitoring portal using google maps API and open source GIS.* 17th International Conference on Geoinformatics, pp.1–5.

Lu, Y., Zhang, M., Li, T., Guang, Y, Rishe, N., 2013. *Online spatial data analysis and visualization system.* Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics - IDEA ’13, pp.71–78.

Mao, L., 2005. *Web-based Information System for Land Management*. University of Calgary.

Mateu, J. & Uso, J.L., 1998. *The spatial pattern of a forest ecosystem*. Ecological Modeling 108, pp.163–174.

Moran PAP, 1948. The interpretation of statistical maps. *J Roy Stat Soc B*, 10(2), pp.243–251.

OpenLayers, 2014. *OpenLayers 3 - Welcome.* Available at: http://openlayers.org/ [Accessed April 22, 2014].

Ord, J.K. & Getis, A., 1995. *Local spatial autocorrelation statistics: distributional issues and an application.* Geographical Analysis, 27(4), pp.286–306.

OSGeo, 2014. *OSGeo4W.* Available at: http://trac.osgeo.org/osgeo4w/ [Accessed May 15, 2014].

Oussalah, M., Bhat, F., Challis, K., Schnier, T., 2013. *A software architecture for Twitter collection, search and geolocation services.* Knowledge-Based Systems, 37, pp.105–120. 83

Pascaul, M. , Alves, E., Almeida, T., França, G., Roing, H., 2012. *An Architecture for Geographic Information Systems on the Web - webGIS.* , GEOProcessing 2012 : The Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services An (c), pp.209–214.

PEP, 2011. *reformed Form Builder | FREE Online HTML5 Themeable Form Generator*. Available at: http://www.reformedapp.com/#saved_forms [Accessed June 19, 2014].

PostgreSQL, 2012. *PostGIS — Spatial and Geographic Objects for PostgreSQL*. Available at: http://postgis.net/ [Accessed May 15, 2014].

PostgreSQL, 2014. *PostgreSQL: The world's most advanced open source database.* Available at: http://www.postgresql.org/ [Accessed May 15, 2014].

Pysal, 2014. *Pysal — Python Spatial Analysis Library.* Available at: http://pysal.readthedocs.org/en/v1.8/search.html?q=pysal+developers&check_keywords=yes&area=default [Accessed May 19, 2014].

Rangel, T.F., Diniz-Filho, J.A.F. & Bini, L.M., 2010. *SAM: a comprehensive application for Spatial Analysis*, Macroecology. Ecography, 33(1), pp.46–50. Available at: http://doi.wiley.com/10.1111/j.1600-0587.2009.06299.x [Accessed July 18, 2014].

Reportlab, 2014. *ReportLab open-source PDF Toolkit - ReportLab.com*. Available at: http://www.reportlab.com/opensource/ [Accessed May 19, 2014].

Rey, S., 2013. *PySAL Documentation*. (Unpublished)

Rey, S., Duque, J.C. & Anselin, L., 2005. *Clustering Components of PySAL*. (Unpublished)

Rey, S.J., 2007. *PySAL : A Python Library of Spatial Analytical Methods*., The Review of Regional Studies, 37(1) , pp. 5 − 27.

Rosenberg, M.S. & Anderson, C.D., 2011. *PASSaGE: Pattern Analysis, Spatial Statistics and Geographic Exegesis.* Version 2. Methods in Ecology and Evolution, 2(3), pp.229–232.

Ruiz, T., 2007. *Free online web templates generator.* Available at: http://www.dotemplate.com/ [Accessed July 16, 2014].

Salgé, F., 1995. *Elements of Spatial Data Quality*, Oxford, Elsevier Science, 139(51)

Scholten, H.J., Velde, R. & Manen, N. eds., 2009. *Geospatial Technology and the Role of Location in Science*, Dordrecht: Springer Netherlands.

Schrader-patton, C., Bunzel, K. & Service, U.F., 2010. *GeoBrowser Deployment in the USDA Forest Service : A Case Study.* COM.Geo 2010, Washington DC, USA

Scott, L., 2009. *Regression Analysis – A Tutorial.* ESRI. Available at: http://arcscripts.esri.com/details.asp?dbid=16428 [Accessed June 15, 2014].

She, B., Zhu, X. & Xiao, W., 2012. *Building An Integrated WEB-Based Environment for Exploratory Spatiotemporal Data Analysis.* In ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 169–174.

SKE, 2012. *Ske Map - Ske Map Street Map City Map CBD Map Electronic Map Mobile Map Satellite Map.* Available at: http://ske.mapinhand.com/ [Accessed July 5, 2014].

Smith, M. de, Goodchild, M.F. & Longley, P.A., 2013. *Geospatial Analysis 4th Edition : A Comprehensive Guide to Principles, Techniques and Software Tools* 4th editio., Winchelsea, UK: Winchelsea Press. Available at: www.spatialanalysisonline.com.

stack exchange inc, 2014a. *Geographic Information Systems Stack Exchange*. Available at: http://gis.stackexchange.com/ [Accessed May 22, 2014].

stack exchange inc, 2014b. *Stack Overflow*. Available at: http://stackoverflow.com/ [Accessed May 22, 2014].

Tobler, W.R., 1970. *A computer movie simulating urban growth in the Detroit region.* Economic Geography, (46), pp.234–240.

Tukey, J.W., 1962. *The Future of Data Analysis*. The Annals of Mathematical Statistics, 33(1), pp.1–67.

Wang, Y. & Zhang, M., 2010. *Open GIS-Based Lightning Information System for Electric Power System.* 2010 International Conference on Intelligent System Design and Engineering Application, pp.1049–1052. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5743356 [Accessed July 31, 2014].

Weigel, M., Preuss, T. & Brüstel, J., 2010. *Generating Navigation Capable Maps from User Provided Data with Woodtracker.* 2010 International Conference on Complex, Intelligent and Software Intensive Systems, pp.544–548.

Westra, E., 2010. *Python Geospatial Development*, Birmingham: Packt Publishing Ltd.

Woodall, C.W. & Graham, J.M., 2004. *A technique for conducting point pattern analysis of cluster plot stem-maps.* Forest Ecology and Management, 198(1-3), pp.31–37.

Zhang, X, Bao, S., Zhu, X., Su, K., 2010. *Spatial Intelligence with Spatial Statistics.* Geoinformatics, 2010 18th International Conference. Beijing: IEEE, pp. 1–

Attachments

1. *Available DataSources for Portuguese data*

Table 5 - Available Data Sources for Portuguese Spatial Data.

Nome	Fonte	Ano	Formato	Observações	Link
CAOP	Direcção Geral do Território	2013	WMS	Área administrativa de Portugal continental com as regiões, distritos, munícipios, freguesias.	http://mapas.dgterritorio.pt/ows/caop/continente
Esperança média de vida à nascença	Direcção geral de Saúde	2010-2012	shp	−−−	http://www.geosaude.dgs.pt/websig/v5/portal2/public/index.php?par=geosaude&lang=pt
População residente	Direcção geral de Saúde	2009	shp	População por região.	http://www.geosaude.dgs.pt/websig/v5/portal2/public/index.php?par=geosaude&lang=pt
Índice de evelhecimento	Direcção geral de Saúde	2009	shp	−−−	http://www.geosaude.dgs.pt/websig/v5/portal2/public/index.php?par=geosaude&lang=pt
Densidade Populacional	Direcção geral de Saúde	2009	shp	−−−	http://www.geosaude.dgs.pt/websig/v5/portal2/public/index.php?par=geosaude&lang=pt
Índice de Poder de compra	Direcção geral de Saúde	2009	shp		http://www.geosaude.dgs.pt/websig/v5/portal2/public/index.php?par=geosaude&lang=pt
Geo-sítios	Laboratório Nacional de Energia e Geologia	2010	WMS	Inventário de património geológico e de locais com interesse geológico	http://geoportal.lneg.pt/ArcGIS/rest/services/Geositios/MapServer
Atlas eólico	Laboratório Nacional de Energia e Geologia	2013	WMS	Mapa do potencial eólico onshore de Portugal Continental	http://geoportal.lneg.pt/ArcGIS/rest/services/AtlasEolico/MapServer
CERAM	Laboratório Nacional de Energia e Geologia	1999	WMS	Sistema de Informação de Matérias Primas Minerais com	http://geoportal.lneg.pt/ArcGIS/rest/services/Ceram/MapServer

				Utilização na Indústria Cerâmica	
Cartografia geologica 1:1000000	Laboratório Nacional de Energia e Geologia	2010	WMS	Carta Geológica de Portugal, na escala 1:1 000 000 (2010)	http://geoportal.lneg.pt/ArcGIS/rest/services/CGP1M/MapServer
COMET	Laboratório Nacional de Energia e Geologia	2014	WMS	infra-estruturas mais rentáveis de transporte de CO2 de armazenamento geológico	http://geoportal.lneg.pt/ArcGIS/rest/services/COMET/MapServer
JazigosMinerais	Laboratório Nacional de Energia e Geologia	−−−	WMS	Jazigos Minerais Portugueses e informação sobre minérios	http://geoportal.lneg.pt/ArcGIS/rest/services/JazigosMinerais/MapServer
NEPS	Laboratório Nacional de Energia e Geologia	2013	WMS	horas anuais equivalentes à potência nominal (NEPs) de Portugal	http://geoportal.lneg.pt/ArcGIS/rest/services/NEPS/MapServer
OcorrenciasMinerais	Laboratório Nacional de Energia e Geologia	2010	WMS	informação geocientífica, técnica e económica relativa a ocorrências, recursos e reservas minerais e áreas com potencial mineiro.	http://geoportal.lneg.pt/ArcGIS/rest/services/OcorrenciasMinerais/MapServer
SONDABASE	Laboratório Nacional de Energia e Geologia	−−−	WMS	Base de Dados de Sondagens de recursos minerais	http://geoportal.lneg.pt/ArcGIS/rest/services/Sondabase/MapServer
RecursosHidro	Laboratório Nacional de Energia e Geologia	2014	WMS	Base de dados de Recursos Hidrogeológicos Portugueses	http://geoportal.lneg.pt/ArcGIS/rest/services/RecursosHidro/MapServer
RecursosGeotermicos	Laboratório Nacional de Energia e Geologia	−−−	WMS	Recursos Geotérmicos em Portugal Continental	http://geoportal.lneg.pt/ArcGIS/rest/services/RecursosGeotermicos/MapServer
Termalbase	Laboratório Nacional de Energia e Geologia	2006	WMS	Ocorrências termais do continente	http://geoportal.lneg.pt/ArcGIS/rest/services/Termalbase/MapServer
PGRH/PGRH_RC	Sistema Nacional de Informação de Ambiente	2009	WMS	Planos de Gestão de Região Hidrográfica - Recursos Hidricos	http://sniamb.apambiente.pt/ArcGIS/rest/services/PGRH/PGRH_RC/MapServer
AreasProtegidas	Sistema Nacional de Informação de Ambiente	−−−	WMS	Áreas protegidas	http://sniamb.apambiente.pt/ArcGIS/rest/services/SIDS-MapasTematicos/AreasProtegidas/MapServer

IDS Corine	Sistema Nacional de Informação de Ambiente	2006	WMS	Cobertura da terra (territorio urbano, áreas agricolas, massa de água, etc)	http://sniamb.apambiente.pt/ArcGIS/rest/services/SIDS-MapasTematicos/IDS_Corine/MapServer
SNIAMB_RS	Sistema Nacional de Informação de Ambiente	–––	WMS	Localização de aterros de resíduos. Sistemas de Gestão de Resíduos Urbanos	http://sniamb.apambiente.pt/ArcGIS/rest/services/SNIAMB/SNIAMB_RS/MapServer
Atlas do ambiente	Sistema Nacional de Informação de Ambiente	–––	WMS	Ambiente Físico, Ambiente Humano, Ambiente Protegido, Ambiente Biofísico.	http://sniamb.apambiente.pt/ArcGIS/rest/services/atlas/MapServer
AAQualSup	Sistema Nacional de Informação de Recursos Hídricos	–––	WMS	Estações de Qualidade das Águas Superficiais - Redes de Monitorização	http://geo.snirh.pt/ArcGIS/rest/services/AAQualSup/MapServer
AASubterraneas	Sistema Nacional de Informação de Recursos Hídricos	–––	WMS	Águas subterrâneas: Estações de tratamento e bacias subterrâneas	http://geo.snirh.pt/ArcGIS/rest/services/AASubterraneas/MapServer
AASuperficiais	Sistema Nacional de Informação de Recursos Hídricos	–––	WMS	Águas superficiais: Estação de tratamento de águas, Rede hidrográfica (todos os cursos de água), albufeiras e barragens	http://geo.snirh.pt/ArcGIS/rest/services/AASuperficiais/MapServer
SNIAMB_LC	Sistema Nacional de Informação de Recursos Hídricos	2011	WMS	Visualização de informação georreferenciada relativa a licenciamento ambiental.	http://sniamb.apambiente.pt/ArcGIS/rest/services/SNIAMB/SNIAMB_LC/MapServer
SNIAMB_QA	Sistema Nacional de Informação de Recursos Hídricos	2008	WMS	Visualização de informação georreferenciada relativa a estações de monitorização da qualidade do ar, zonas de gestão e indices de qualidade do ar.	http://sniamb.apambiente.pt/ArcGIS/rest/services/SNIAMB/SNIAMB_QA/MapServer
Censos 2011	Instituto Nacional de Estatística	2011	shp	Importação dos principais dados alfanuméricos e geográficos, relativa aos censos 2011.	http://mapas.ine.pt/download/index2011.phtml

2. Statistical methods applied: Inputs and Outputs

- Visualization

- Map Classifiers

Available Methods

- Box_Plot
- Equal_Interval
- Fisher_Jenks
- Jenks_Caspall
- Jenks_Caspall_Forced
- Jenks_Caspall_Sampled
- Max_P_Classifier
- Maximum_Breaks
- Natural_Breaks
- Quantiles
- Percentiles
- Std_Mean

Inputs: Dataset, Variable, Method

Output: Elem – array of elements with the results of the operations appended.

- Weights

Spatial weights matrix expresses the potential for interaction between observations at each pair i,j of locations. It is calculated by the modules based on contiguity criteria, distance criteria or according to kernel weights.

Available methods (and associated inputs)

- queen_from_shapefile
- rook_from_shapefile
- knnW_from_array (k)
- adaptative_kernelW_from_shapefile (k)
- knnW_from_shapefile (threshold)
- threshold_binaryW_from_shapefile (threshold)
- threshold_continuousW_from_shapefile (k)
- kernel (k)
- kernelW_from_shapefile (k)
- min_thershold_dist_from_shapefile (threshold)

n : threshold (float) - distance band

k: int - the number of nearest neighbors

Inputs: Dataset, Variable, Method (k= 4 and n = 0,62 by default)

Output: Elem – array of elements with the results of the operations appended.

- Exploratory Spatial Analysis – Spatial Autocorrelation

- **Moran**

Inputs

Arguments	Type	Description
Y	array	original variable
W	W	Weights matrix

Outputs

Arguments	Type	Description
I	float	value of Moran's I
EI	float	expected value under normality assumption
VI_norm	float	variance of I under normality assumption
seI_norm	float	standard deviation of I under normality assumption
z_norm	float	z-value of I under normality assumption
p_norm	float	p-value of I under normality assumption (1-tailed)
VI_rand	float	variance of I under randomization assumption
seI_rand	float	standard deviation of I under randomization assumption
z_rand	float	z-value of I under randomization assumption
p_rand	float	p-value of I under randomization assumption (1-tailed)
Sim	array (if permutations>0)	vector of I values for permutated samples
p_sim	array (if permutations>0)	p-value based on permutations
EI_sim	float (if permutations>0)	average value of I from permutations

VI_sim	float (if permutations>0)	variance of I from permutations
seI_sim	float (if permutations>0)	standard deviation of I under permutations.
z_sim	float (if permutations>0)	standardized I based on permutations
p_z_sim	float (if permutations>0)	p-value based on standard normal approximation

- **Geary**

Inputs

Arguments	Type	Description
Y	array	original variable
W	W	Weights matrix

Outputs

Arguments	Type	Description
C	float	value of statistic
EC	float	expected value
VC	float	variance of G under normality assumption
z_norm	float	z-statistic for C under normality assumption
z_rand	float	z-statistic for C under randomization assumption
p_norm	float	p-value under normality assumption (one-tailed)
p_rand	float	p-value under randomization assumption (one-tailed)
sim	array (if permutations!=0)	vector of I values for permutated samples
p_sim	float (if permutations!=0)	p-value based on permutations
EC_sim	float (if permutations!=0)	average value of C from permutations
VC_sim	float (if permutations!=0)	variance of C from permutations
seC_sim	float (if permutations!=0)	standard deviation of C under permutations.
z_sim	float (if permutations!=0)	standardized C based on permutations
p_z_sim	float (if	p-value based on standard normal approximation

	permutations!=0)	from permutations

- **LOCAL MORAN**

Inputs

Arguments	Type	Description
Y	array	original variable
W	W	Weights matrix

Outputs

Arguments	Type	Description
I	float	value of Moran's I
Q	array (if permutations>0)	values indicate quadrat location 1 HH, 2 LH, 3 LL, 4 HL
sim	array (if permutations>0)	vector of I values for permutated samples
p_sim	array (if permutations>0)	p-value based on permutations
EI_sim	float (if permutations>0)	average value of I from permutations
VI_sim	float (if permutations>0)	variance of I from permutations
seI_sim	float (if permutations>0)	standard deviation of I under permutations.
z_sim	float (if permutations>0)	standardized I based on permutations
p_z_sim	float (if permutations>0)	p-value based on standard normal approximation from permutations

- Regression:

- **Diagonostics – Non spatial**

Statistic Variables	**Input**	**Output**
t-stat	Reg	fs_result (tuple)
Fstat	Reg	ts_result (tuple)
Zstat	Reg	zs_result(tuple)
r2	Reg	r2result
ar2	Reg	ar2result
log_likelihood	Reg	ll_result

reg - regression object

- **Diagonostics - Spatial**

Inputs

Lagrange multipliers testes	ols	w
Moran Residuals	ols	w

Moran Residual Output	Type	Description
I	Float	Moran’s I statistic
eI	Float	Moran’s I expectation
vI	Float	Moran’s I variance
zI	Float	Moran’s I standardized value

OLS
Outputs

Arguments	Type	Description
mean_y		
std_y		
Vm		
Utu		
Y	array	nx1 array of dependent variable
X	array	nxk array of independent variables (with constant added if constant parameter set to True)
betas	array	kx1 array with estimated coefficients
U	array	nx1 array of residuals

predy	array	nx1 array of predicted values
N	int	Number of observations
K	int	Number of variables (constant included)
name_ds	string	dataset's name
name_y	string	Dependent variable's name
name_x	tuple	Independent variables' names
r2	float	R squared
ar2	float	Adjusted R squared
sig2	float	Sigma squared
sig2ML	float	Sigma squared ML
f_stat	tuple	Statistic (float), p-value (float)
logll	float	Log likelihood
aic	float	Akaike info criterion
schwarz	float	Schwarz info criterion
std_err	array	1xk array of Std.Error
t_stat	list of tuples	Each tuple contains the pair (statistic, p-value), where each is a float; same order as self.x
mulColli	float	Multicollinearity condition number
jarque_bera	dictionary	'jb': Jarque-Bera statistic (float); 'pvalue': p-value (float); 'df': degrees of freedom (int)
breusch_pagan	dictionary	'bp': Breusch-Pagan statistic (float); 'pvalue': p-value (float); 'df': degrees of freedom (int)
koenker_bassett	dictionary	'kb': Koenker-Bassett statistic (float); 'pvalue': p-value (float); 'df': degrees of freedom (int)
white	dictionary	'wh': White statistic (float); 'pvalue': p-value (float); 'df': degrees of freedom (int)
lm_error	tuple	Lagrange multiplier test for spatial error model; each tuple contains the pair (statistic, p-value), where each is a float; only available if w defined
lm_lag	tuple	Lagrange multiplier test for spatial lag model; each tuple contains the pair (statistic, p-value), where each is a float; only available if w defined
rlm_error	tuple	Robust lagrange multiplier test for spatial error model;

		each tuple contains the pair (statistic, p-value), where each is a float; only available if w defined
rlm_lag	tuple	Robust lagrange multiplier test for spatial lag model; each tuple contains the pair (statistic, p-value), where each is a float; only available if w defined
lm_sarma	tuple	Lagrange multiplier test for spatial SARMA model; each tuple contains the pair (statistic, p-value), where each is a float; only available if w defined
moran_res	tuple	Tuple containing the triple (Moran's I, standardized Moran's I, p-value); only available if w defined
summary	string	Includes OLS regression results and diagnostics in a nice format for printing.

3. Instruction Section

Spatial analysis is defined by (Carneiro & Santos 2003) as a sequence of sequential procedures which's objective is to define a inferential model that explicitly considers the spatial relations present in the model.

These instructions are aimed to give a general and simple explanation regarding the methods considered in this application. For further development on the subject consult the bibliography present in the end of this page.

This website is still a prototype, which leads to it being limited to areal data that is delimited by closed polygons.

Presumptions:

- Homogeneity within the polygon (which is not always true and highly dependent on the criterion of definition of the areas)

Appliance

This type of analysis is often applied to aggregated units, where the event distribution is associated with a delimited area within polygons. Meaning that there is no exact value for the location but a value aggregated by area.

The geographical space to be studied is a region A, comprehended by a fixed set of spatial units.

Generally speaking, the distributive model considers a stochastic process {zi: i=1,..,n} composed by a set of aleatory variables where the goal is to build an approximation of the aggregated distribution of those variables z={z1,…,zn}.

Where each aleatory variable is associated to one of the areas and has a distribution to be estimated.

If the process is stationary the expected value of the region and the covariance depends exclusively on the distance of the structure of the neighborhood between areas.

The most common approach is to suppose that the areas are differentiates and each of them has its own 'identity'. Statically this means that we have a spatial discrete model. The other hypothesis is to suppose we have a spatial continued model building a surface.

1. Visualization:

   Data can be approached in a general way, without considering it's spatial component. By using box plots, scatter plots or histograms, the visualization of the

data can reveal data characteristics such as mean, extreme values, distribution of values.

In order to identify extreme values, graphical tools are used. Mapping according to classes allows the identification of extreme values.

2. Autocorrelation:

Autocorrelation aids in identifying the structure of spatial autocorrelation that better describes the data. By estimating the magnitude of autocorrelation between areas.

- Moran's I
- Gamma
- Gear's C

The index itself can reveal whether there is autocorrelation or not, but it is always necessary to certify that it is statistically valid.

There are two ways to check this:

- Associate to a statistical distribution: usually a normal distribution.
- Pseudo-significance teste: This test doesn't have presumptions about the data distribution. It generates different permutation of the associated attribute value, each permutation has a different spatial arrangement where the values are distributed amongst the areas.

3. Regression Models

A Regression model is a statistical tool that uses the relationship between two or more variables to explain the observed value.

When there is a spatial autocorrelation, the model estimation should incorporate this spatial structure.

A Linear regression analysis intends to quantify a linear regression between a dependent variable and a set of explaining variable:

$Y=x\beta + \varepsilon , \varepsilon \sim N(0, \sigma 2)$

Where Y is the variable to be explained and X are the dependent variables which will explain Y.

Objectives:

- Find a good adjustment between predicted values and observed values.

- Find the variables which have a more significant contribution to the variable.

Presumptions:
Observations are uncorrelated, which means that Residuals (ε) are independent and non correlated with Y and have a normal distribution with mean equal to zero.

Nevertheless, spatial data which has spatial dependence has often correlated observations, resulting in residuals with spatial autocorrelation.

This fact makes it necessary to investigate regression residuals in order to investigate spatial structure (running an autocorrelation index over the residuals), in order to include the interference caused by the said autocorrelation in the model.

Models with spatial effects

Captures the spatial correlations on a single parameter (added to the traditional model)

- SAR – Spatial autoregressive / Spatial lag model
  In this model spatial autocorrelation is ignored and attributed to the dependent variable (Y)
  $Y = \rho WY + X\beta + \varepsilon$
  Where ρ is a spatial autoregressive coefficient (non correlation means this coefficient is equal to 0)

Spatial Error Model / CAR – Conditional Autoregressive Model

Spatial effects are considered noise meaning they are something to be removed.
$Y = X\beta + \varepsilon$, where $\varepsilon = \lambda W + \xi$
Where λ is the autoregressive coefficient and ξ the residuals

These models assume that the spatial process underneath the analyzed data is stationary, allowing the patterns of autocorrelation to be captured in one parameter.

RESULTS:

- Map residuals – high concentration of positive reseals (or negative) are an indicator of spatial autocorrelation
- Moran's I – good indicator over residuals

- R2 – determinations coefficient, can be unsufficient due to spatial effects or omitted explaining variables.
- AIC – Akaik information criterion – maximizes log of similitude

Some Basic Regression Diagnostics

- The so-called *p-value* associated with the variable
    - For any statistical method, including regression, we are testing some hypothesis. In regression, we are testing the *null hypothesis* that the coefficient (i.e., slope) ? is equal to zero (i.e., that the explanatory variable is not a significant predictor of the dependent variable)
    - Formally, the p-value is the probability of observing the value of ? as extreme (i.e., as different from 0 as its estimated value is) when in reality it equals to zero (i.e., when the Null Hypothesis holds). If this probability is small enough (generally, $p<0.05$), we reject the null hypothesis of ?=0 for an *alternative hypothesis* of ?<>0
        - Again, when the null hypothesis (of ?=0) cannot be rejected, the dependent variable is not related to the independent variable.
        - The rejection of a null hypothesis (i.e., when $p <0.05$) indicates that the independent variable is a statistically significant predictor of the dependent variable
    - One p-value per independent variable
- The *sign* of the coefficient of the independent variable (i.e., the slope of the regression line)
    - One coefficient per independent variable Indicates whether the relationship between the dependent and independent variables is positive or negative
    - We should look at the sign when the coefficient is statistically significant (i.e., significantly different from zero)
- *R-squared* (AKA Coefficient of Determination): the percent of variance in the dependent variable that is explained by the predictors

- In the single predictor case, R-squared is simply the square of the correlation between the predictor and dependent variable
- The more independent variables included, the higher the R-squared
- Adjusted R-squared: percent of variance in the dependent variable explained, adjusted by the number of predictors
- One R-squared for the regression model

4. Report

Spatially: Spatial Analysis Report

Alexandra Dias

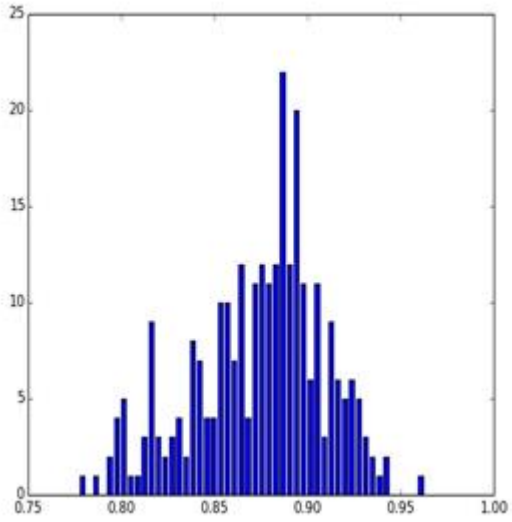
http://www.twink.pythonanywhere.com

alexandra.cordeiro.dias@gmail.com

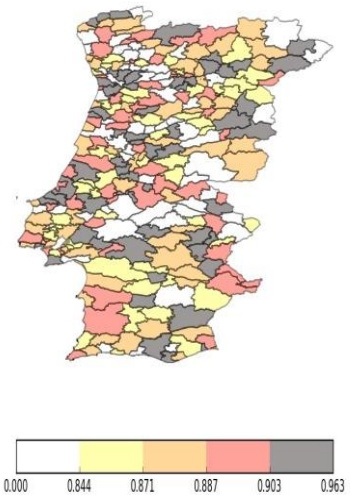
Abstract

This report is built upon the data stored in the web platform made available. The purpose of this platform is to provide Open data to be analyzed within a OpenSource environment. Please be conscious that all the analysis provided depend upon the data stored in the database and therefore it is crucial to select the most adequate data, preform the suitable analysis and interpret the results accordingly. This platform is a tool, use it wisely

Data Vizualization: histogram - Variable: taxas_fam histogram, variable = taxas_fam



Map Classification:quantiles Method, k=4, variable: taxas_fam



Global Autocorrelation :Geary Index, Variable: taxas_fam

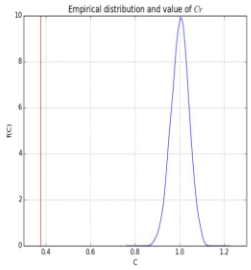
Geary's C: 0.376137654345

Geary's C expected value: 1.0

Average value of C from permutations: 0.998228716618

Variance of C from permutations: 0.00172139433889

Empirical distribution and value of Geary's C, variable = taxas_fam



Global Autocorrelation :Gamma Index, Variable: taxas_fam

Global Autocorrelation :Moran Index, Variable: taxas_fam

Moran's I: 0.608529731716

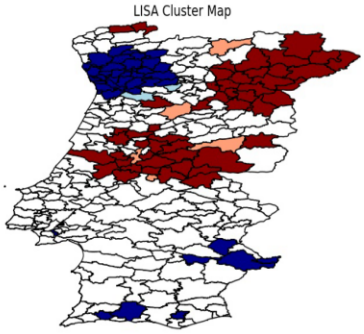
Moran's I expected value: -0.00361010830325

Average value of I from permutations: -0.00389226847746

Variance of I from permutations: 0.00140192629875

Average value of I from permutations: -0.00389785176686
Variance of I from permutations: 0.218106838121
Standard deviation of I from permutations: 0.46701909824

LISA Cluster Map, variable = taxas_fam



Regression: base
Dependent Variable: taxas_fam
Independent Variable(s): ['taxas_enSu', 'taxas_1o', 'taxas_ref']

REGRESSION

SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES

Data set : unknown
Weights matrix : unknown
Dependent Variable : dep_var Number of Observations: 278
Mean dependent var : 0.8741 Number of Variables : 4
S.D. dependent var : 0.0350 Degrees of Freedom : 274
R-squared : 0.451485
Adjusted R-squared : 0.4455
Sum squared residual: 0.186 F-statistic : 75.1769
Sigma-square : 0.001 Prob(F-statistic) : 1.66e-35
S.E. of regression : 0.026 Log likelihood : 621.582
Sigma-square ML : 0.001 Akaike info criterion : -1235.163
S.E of regression ML: 0.0259 Schwarz criterion : -1220.653

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	0.7654200	0.0093472	81.8874477	0.0000000
taxas_1o	0.1077903	0.0790519	1.3635397	0.1738317
taxas_enSu	0.1473941	0.0500595	2.9443754	0.0035136
taxas_ref	0.3436237	0.0251556	13.6599352	0.0000000

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER 13.820521

TEST ON NORMALITY OF ERRORS
TEST DF VALUE PROB
Jarque-Bera 2 10.133871 0.0063017
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST DF VALUE PROB
Breusch-Pagan test 3 9.178237 0.0270126
Koenker-Bassett test 3 9.176754 0.0270308
SPECIFICATION ROBUST TEST
TEST DF VALUE PROB
White 9 17.409754 0.0426728
DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST MI/DF VALUE PROB
Lagrange Multiplier (lag) 1 149.638703 0.0000000
Robust LM (lag) 1 1.584170 0.2081610
Lagrange Multiplier (error) 1 195.917373 0.0000000
Robust LM (error) 1 47.862840 0.0000000
Lagrange Multiplier (SARMA) 2 197.501543 0.0000000
================================ END OF REPORT
====================================

Regression: error
Dependent Variable: taxas_fam
Independent Variable(s): ['taxas_enSu', 'taxas_1o', 'taxas_ref']

REGRESSION

SUMMARY OF OUTPUT: SPATIALLY WEIGHTED LEAST SQUARES

Data set : unknown
Weights matrix : unknown
Dependent Variable : dep_var Number of Observations: 278
Mean dependent var : 0.8741 Number of Variables : 4
S.D. dependent var : 0.0350 Degrees of Freedom : 274
Pseudo R-squared : 0.450471

Variable Coefficient Std.Error z-Statistic Probability

CONSTANT 0.7760427 0.0102147 75.9727614 0.0000000
taxas_1o 0.1368465 0.0796501 1.7180949 0.0857793
taxas_enSu 0.1215843 0.0427108 2.8466853 0.0044177
taxas_ref 0.3069803 0.0262645 11.6880493 0.0000000
lambda 0.7372591

================================ END OF REPORT
====================================

Regression: lag
Dependent Variable: taxas_fam
Independent Variable(s): ['taxas_enSu', 'taxas_1o', 'taxas_ref']

REGRESSION

SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES

Data set : unknown
Weights matrix : unknown
Dependent Variable : dep_var Number of Observations: 278
Mean dependent var : 0.8741 Number of Variables : 5
S.D. dependent var : 0.0350 Degrees of Freedom : 273
Pseudo R-squared : 0.5505
Spatial Pseudo R-squared: 0.4548

Variable Coefficient Std.Error z-Statistic Probability

CONSTANT 0.6271102 0.0958349 6.5436542 0.0000000
W_dep_var 0.1725020 0.1190628 1.4488326 0.1473843
taxas_1o 0.0927068 0.0720920 1.2859518 0.1984599
taxas_enSu 0.1292465 0.0468781 2.7570783 0.0058320
taxas_ref 0.3048106 0.0351137 8.6806861 0.0000000

Instrumented: W_dep_var
Instruments: W_taxas_1o, W_taxas_enSu, W_taxas_ref
DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST MI/DF VALUE PROB
Anselin-Kelejian Test 1 11.186138 0.0008241
========== END OF REPORT
==========