Centro de Informática U·F·P·E

Pós-Graduação em Ciência da Computação

Ricardo Alexandre Afonso

Smartcluster: A Metamodel of Indicators for Smart and Human Cities

VIRTUS IMPAVIDA

Universidade Federal De Pernambuco
posgraduacao@cin.ufpe.br
http://www.cin.ufpe.br/~posgraduacao

RECIFE
2017

Ricardo Alexandre Afonso

Smartcluster: A Metamodel of Indicators for Smart and Human Cities

A Ph.D. Thesis presented to the Informatics Center of Federal University of Pernambuco in partial fulfillment of the requirements for the degree of Philosophy Doctor in Computer Science.

Supervisor: Vinicius Cardoso Garcia

RECIFE
2017

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

A257s Afonso, Ricardo Alexandre
Smartcluster: a metamodel of indicators for smart and human cities / Ricardo Alexandre Afonso. – 2017.
164 f.: il., fig., tab.

Orientador: Vinicius Cardoso Garcia.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2017.
Inclui referências.

1. Banco de dados. 2. Ontologia. I. Garcia, Vinicius Cardoso (orientador). II. Título.

025.04 CDD (23. ed.) UFPE- MEI 2017-137

Ricardo Alexandre Afonso

Smartcluster: A Metamodel of Indicators for Smart and Human Cities

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Aprovado em: 13/03/2017.

____
Orientador: Prof. Dr. Vinicius Cardoso Garcia

BANCA EXAMINADORA

____
Prof. Carlos André Guimarães Ferraz
Centro de Informática / UFPE

____
Prof. Kiev Santos da Gama
Centro de Informática / UFPE

____
Prof. Julieta Maria de Vasconcelos leite
Departamento de Arquitetura e Urbanismo / UFPE

____
Prof. Alexandre Álvaro
Departamento de Computação/ UFSCar

____
Prof. Jones Oliveira de Albuquerque
Departamento de Estatística e Informática / UFRPE

Acknowledgements

To my family, composed of honest people, friendly and hardworking even far was present for the gifts granted to me and for the love devoted to my creation. To my angel, Gabriel, one day will understand (and forgive me) for the reasons that kept me from your daily living to achieve goals a far away land. To my friends (almost brothers), whose words will never be enough to measure how important were their academic and personal examples: Joel da Silva, Ricardo Ramos Argenton, Alexander Alvaro, Vinicius Cardoso Garcia and many others who cultivated these years of study. To my mentor teacher: Vinicius Cardoso Garcia for leading me in this process where a timely and accurately pointed out the best way forward. To collegiate course of Computer Science UFAL, who took part of my workload to make possible the execution of this research, and Raquel da Silva Cabral my special thanks. To FACEPE for financial support, the UFPE for its excellence, UFAL for its facilities and the INEP / MEC for the opportunity to meet academic realities so different. To Estela, who became a star and taught me a path of peace…

“If it was easy to find the stepping stones,
many stumbling blocks would not be bad”
(Humberto Gessinger)

Abstract

Currently, there are several works on smart cities and the advances offered to the routine of its inhabitants and optimization of resources, however, there is still no consensus on the definition of the term "Smart Cities", nor their domains and indicators. The lack of a clear and widely usable definition, as well as the delimitation of domains and indicators makes it impossible to compare or measure cities in this context. It is very common for governments and private companies to redefine the concepts of Smart Cities and create models that meet only their interests. These models become isolated initiatives or serve as success cases for few domains of use. This work presents a proposal for a metamodel called **SmartCluster**, which was developed to allow uniformity in intelligent city models so that they can be used in any context and can be expanded at any time. The use of this metamodel will allow indicators drawn from public databases to serve to assist municipal managers in measuring, comparing and managing resources of smart cities.

Key-words: Smart Cities. e-Government. Metamodel. Ontology.

Resumo

Atualmente existem vários trabalhos sobre cidades inteligentes e os avanços oferecidos à rotina de seus habitantes e otimização de recursos, entretanto, ainda não existe um consenso sobre a definição do termo “Cidades Inteligentes”, nem de seus domínios e indicadores. A falta de uma definição clara e amplamente utilizável, bem como a delimitação de domínios e indicadores impossibilita comparar ou medir cidades nesse contexto. É muito comum que governos e empresas privadas redefinam os conceitos sobre Cidades Inteligentes e criem modelos que atendam somente aos seus interesses. Estes modelos acabam se tornando iniciativas isoladas ou servem como casos de sucesso para poucos domínios de uso. Por isso, esse trabalho apresenta uma proposta de um metamodelo chamado **SmartCluster,** que foi desenvolvido para permitir uma uniformidade nos modelos de cidades inteligentes de forma que possam ser utilizados em quaisquer contextos e possam ser ampliados a qualquer momento. A utilização deste metamodelo vai permitir que indicadores extraídos de bases de dados públicas possam servir para auxiliar os gestores municipais a medir, comparar e gerenciar recursos das cidades inteligentes.

Palavras-chaves: Cidades Inteligentes. Governo Eletrônico. Metamodelo. Ontologia.

List of Figures

List of Tables

Acronyms

DL	Description Logic
OIL	Ontology Inference Layer
OWL	Web Ontology Language
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RDQL	RDF Data Query Language
UML	Unified Model Language
W3C	World Wide Web Consortium
HCA	Hierarchical cluster analysis
PCA	Principal component analysis
CA	Cluster Analysis
SCM	Smart City Model
EBM	Evidence-based Medicine
EBSE	Evidence-based Software Engineering

Contents

1

INTRODUCTION

An old friend once told me something that gave me great comfort,
something he read. He said that Mozart, Beethoven and Chopin
had never died, they had just become music.

— Robert Ford, WestWorld

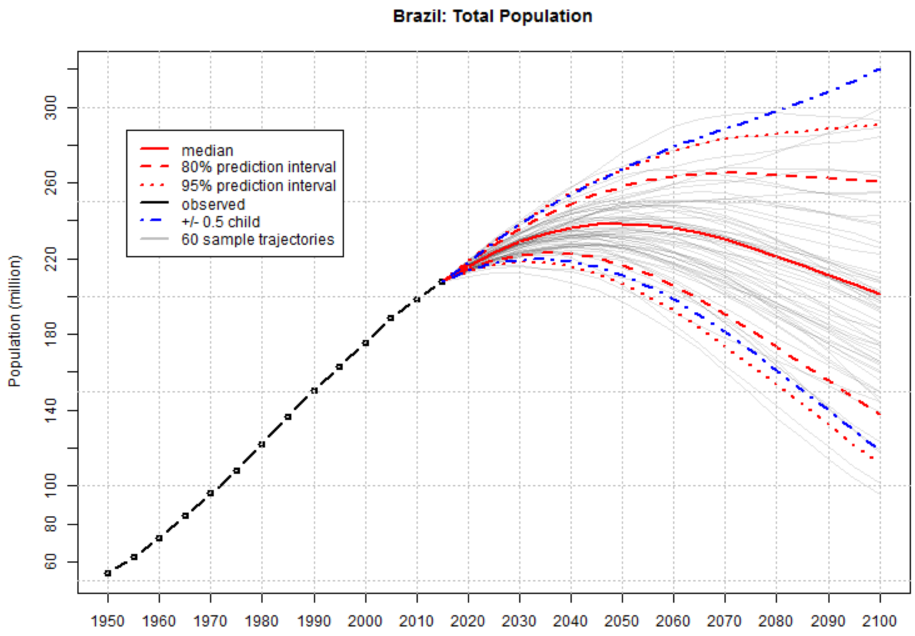
This chapter aims to present an overview of a thesis. It is presented a work proposal that aims to define a problem, is a question of research and contribution to a work developed by a defined methodology. Finally, a structure of the chapters of the thesis and a summary of the chapter is described.

1.1 Motivations

According to United Nations estimates, presented in the report "Perspectives of the World Population: The Review of 2015", the current world population of 7.3 billion people will reach the mark of 8.5 billion by 2030, and 9.7 billion in 2050. At this rate, the planet is expected to reach 2100 with 11.2 billion human beings, a growth of 53% compared to the scenario found today. Figure 1.1 shows the UN estimate for the Brazilian population. (UN, 2015)

Today, the five largest cities in the world is more populous than several countries on the globe. Migration from rural areas to urban centers is a geographic population migration, where whole cultures search for more alternatives that is viable to existence, and with this, cities double their population in a matter of decades, without having time to plan the necessary resources. The way these cities is managed can serve as a model for smaller cities that may in the future benefit from the results of successful administrations and thus avoid the problems faced in large cities today. In order to obtain these results, it is necessary to measure the way the city is being administered under the most diverse areas, and among them, Infrastructure, Services and Management is mentioned as the most cited areas in the studies of Smart Cities (Chapter 2 - Section 2.5).

Figure 1.1: Estimation of growth of the Brazilian population



Source: Department of Economics ans Social Affairs (UN, 2015)

1.2 Background

The concept of Smart Cities refers to the idea of optimizing resources and improving the supply of public services. Still, the definition of Smart Cities remains very wide, as its domains and areas and even the concept itself depends on a holistic view and often difficult to measure practice. Among other concepts, such as Intelligent City (HOLLANDS, 2008), City of Knowledge (ERGAZAKIS et al, 2004), Virtual City

(DONATH, 1997), Digital City (VAN DEN BESSELAAR, 1998) in this thesis, will be adopted the term Smart City, which may have originated the term "Smart Growth" which was proposed to evoke new policy practices for better urban planning (BOLLIER, 1998).

The motivation initially based on the paper of Kiev (GAMA, 2012) for this work is to clearly define the indicators, concepts and domains of Smart Cities, and proposes the conceptualization and formalization of a metamodel capable of covering existing models and allowing the mining of data obtained in big public dates. This data will serve to group cities by similarities of indicators and thus, to allow in the future that the municipal managers can be based on local solutions to optimize resources and to extend public policies.

1.3 Problem Statement

The key problem addressed in this thesis is create and validate a Metamodel (named SmartCluster) compatible with existing models and at the same time compatible with indicators and data from other international models that have the same purpose. For this, the main challenges that is proposed for this work are:

- How to assess how smart a city can be?
- How to get and process data of Smart Cities?
- How to provide a data extraction and visualization environment?
- How to create and validate a metamodel compatible with existing templates?

The relation between the theme of this thesis and the science of computing consists of:

a) Obtain, process and recover public data of Smart Cities;

b) Create a metamodel based on domain ontology;

c) Write extensive bibliographical research to extend the systematic review of literature and

d) Use evidence-based software engineering to validate the proposed model.

These problems will serve to define our Research Questions and consequently will define the structure of this work to solve each of these problems.

1.4 Research Questions and Goals

The problems identified in the previous section of this thesis gave rise to five Research Questions. The first Question **(RQ1)** refers to the need to construct a model to measure the intelligence indicators.

The second question **(RQ2)** aims to find a valid set of domains and indicators for Smart Cities that is compatible with the Brazilian reality. The third question **(RQ3)** aims to understand where data will be obtained and how it will be transformed into indicators.

The fourth question **(RQ4)** related to the concern with extracting and visualizing the data, and finally, the fifth question **(RQ5)** attacks the core of this thesis, which is to present a metamodel compatible with the other existing models and adequate to the Brazilian reality.

1.4.1 RQ1: How to evaluate how smart can a city be?

The literature review and Grounded Theory of Smart Cities have detected models used around the world that use domains to measure cities from specific indicators. The use of these domains and indicators is disparate and so this research question looks for better definitions about the term Smart City itself. The best way to answer this research question is to create a model capable of evaluating and comparing Smart Cities.

1.4.2 RQ2: Which indicators are appropriate to Brazilian reality?

Currently there is academic, commercial and even an ISO37120 (ISO, 2014) standard that advocate domains and indicators for Smart Cities. However, is these models adequate to the Brazilian reality? Unattainable indicators may invalidate models not oriented to developing countries, such as Brazil and its cities. Even Brazilian capitals are still lacking data and information on specific domains, and therefore, there is a need to create an evaluation model with indicators whose data can be mined in public databases.

1.4.3 RQ3: How to obtain and process data of Smart Cities?

It is known that in many developing countries there is no guarantee of effectiveness in the services of electronic governance and transparency of public data, as is the case in most developed countries. Brazil has a reasonable range of public data available through government agencies that may allow the consultation of a series of basic indicators. Getting the data will be part of the answer to this research question.

1.4.4 RQ4: How provide data extraction/visualization environment?

Categorizing Smart Cities, defining domains and indicators and processing the data is of paramount importance for research in this area. However, if there is no means of querying and viewing the data, the search becomes inappropriate for the intended purpose. This work searches for data extraction and visualization tools that complements the whole framework for obtaining and processing these data.

1.4.5 RQ5: How to create a metamodel compatible with atual models?

The objective of creating the metamodel is to reduce the distance between the evaluation and comparison of Smart Cities performed in academia and industry, using Evidence-Based Software Engineering (EBSE), which is represented in this work by the application of a formal approach for the identification of evidence (secondary studies) in the literature (Kitchenham et al., 2007).

1.5 Outline of Contributions

Thus, the main contribution of this doctoral thesis is to provide a meta-model of comparison and measurement of Smart Cities capable of identifying the particularities of each group of cities, and thus serve as a tool for municipal management based on indicators close to the Brazilian social reality.

Table 1.1: Relationship between Research Questions, Contributions and Goals

Research Questions	Outline of Contributions	Primary Goals
RQ1	**OC1, OC2**	Conduct a literature review of Smart Cities based on a set of indicators and areas appropriate to Brazilian reality and compatible with other models adopted in the world.
RQ2	**OC2, OC3**	Formally establish one domain ontology containing indicators and areas of Smart Cities.
RQ3	**OC3**	Set the levels and parameters of comparison to enable cataloging similarity between Smart Cities using multivariate analysis.
RQ4	**OC4, OC5**	Use tools to extract public data, analyze and generate Smart Cities clustering dendrogram for decision making for public managers.
RQ5	**OC2, OC3, OC4**	Create a metamodel containing the indicators and domains of Smart Cities compatible with the existing models and validate this metamodel.

Source: Made by author

The following are the contributions that will be achieved to the detriment of the primary objectives:

OC1 – Definition of Smart Cities: the bibliographic survey carried out in conjunction with Grounded Theory will allow a better understanding of the nuances of the Smart Cities, and thus, enable a better understanding of the term, domains and indicators;

OC2 – Creating a taxonomy of indicators: The creation of a taxonomy of indicators will combine the indicators used in the main models of Smart Cities with the specific needs of Brazilian cities. The creation of regional indicators makes the model proposed in this thesis more adherent to the reality that is faced in our cities;

OC3 – Setting comparison metrics: The cities that will be the object of case studies will go through a process of data comparison that will allow a better understanding of groupings by similarity and, consequently, how to apply the definition of Smart Cities can be applied to a more equitable set of municipalities;

OC4 – Creation of a meta-model for grouping and comparison of indicators: Perhaps the greatest contribution of this doctoral thesis is the construction of a metamodel capable of using public data converted into indicators to group cities by similarity and thus include them in a ranking similar to those developed in other parts of the world;

OC5 – Case study with Brazilian cities: To formally represent Brazilian cities through a model of Smart Cities and Humans will in the future optimize resources consider the most important indicators for strategic decision making for the development of cities.

Is presented in Table 1.1 the relationship between research questions (RQ), the Outline of Contributions (OC) of this work and how the Primary Goals expected to be achieved.

1.6 Negative Scope

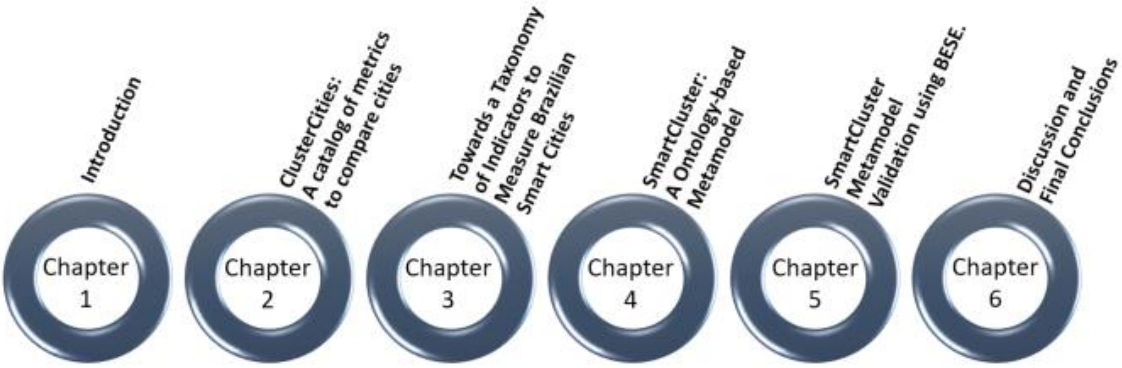
What not to expect from this work? This thesis has as main objective the proposal of a metamodel capable of serving as a model for the composition of models for Smart Cities that use the most varied types of indicators and domains. However, it is not the intention of this work to propose a single model to be adopted, but rather to serve as a subsidy for the creation and development of other models for indicators and rankings.

1.7 Methodology

The initial stage of this work consisted in a bibliographical survey on indicators capable of measuring Smart Cities, following the concept of Systematic Review of Literature, (KITCHENHAM AND CHARTERS, 2007). In order for cities to be compared, criteria for grouping by similarity were defined based on Demographic (Density, Population and Area) and Human (Income, Education and Longevity) characteristics (AFONSO et al., 2015).

The next step was to propose a taxonomy for indicators of Brazilian smart and Humanities Cities, whose data were compatible with other European and American models, and could measure and compare the local cities with any other cities in the world. In order to create this taxonomy these aspects (LI, 2012) were observed. Once the concepts and indicators have been defined, a metamodel called "SmartCluster" was proposed to group cities, compare indicators and allow analysis, interpretation and comparison with current models of Smart Cities. For this purpose domain ontologies were created for each of the models studied, to generate the necessary evidences for the creation of the proposed metamodel.

Figure 1.2: Thesis construction Steps


Introduction
Chapter 1
ClusterCities: A catalog of metrics to compare cities
Chapter 2
Towards a Taxonomy of Indicators to Measure Brazilian Smart Cities
Chapter 3
SmartCluster: A Ontology-based Metamodel
Chapter 4
SmartCluster Metamodel Validation using BESE.
Chapter 5
Discussion and Final Conclusions
Chapter 6

Source: Made by author.

To validate this metamodel, data from Brazilian cities were used. Figure 2 presents the steps developed throughout this work, characterized by the creation of a systematic review of the literature (Chapter 2), creation of metrics for city comparison (Chapter 3), definition of a taxonomy (Chapter 4), development of the metamodel "SmartCluster"(Chapter 5) and validation of the metamodel through the application of multivariate analysis and evidence-based software engineering (Chapter 6).

In accordance with the principles established (MARCONI AND LAKATOS, 2004) that define the concepts about research methodology, this work is positioned in a **pragmatic philosophical** way with an **inductive approach**, using **methods of bibliographic research.**

With a systematic review of the literature, proposal to create a metamodel, and application of the proposal through evidence-based software engineering with public data.

Both the nature of the variables used, as well as the procedures related to the systematic review of the literature and the proposal of the metamodel employ a quantitative analysis. Under these principles, this work is classified as descriptive and exploratory, in order to identify the relationship between the variables obtained to establish a data pattern. A summary of these principles is given in Table 1.2.

Table 1.2: Methodological research summary

Philosophical Positioning	Pragmatism
Approach Method	Inductive
Nature of Variables	Quantitative and Qualitative
Method of Procedure	Systematic Review of Literature Grounded Theory
About the Goal	Descriptive and Exploratory
Scope	Field study

Source: Made by author

The next section details the organization of the chapters of this thesis and how the tasks will be conducted.

1.8 Thesis Organization

This thesis presents six chapters in their totality, accompanied by a section dedicated to the references. Respecting a logical sequence the work has this introductory chapter that presents the motivation, desired objectives, questions of research, contributions and the methodology defined through of the six chapters that is described below:

- Chapter 2 (ClusterCities: A catalog of metrics for comparing cities): While the previous chapters present the theoretical concepts and the necessary steps for constructing the concept and indicators, this chapter presents the way to compare and group cities by territorial and social similarities;

- Chapter 3 (Towards a Taxonomy of Indicators to Measure Brazilian Smart Cities): It is in this chapter that the indicators to be used to represent the specificities of Brazilian cities is defined, without neglecting the models in use in other countrie**s**;
- Chapter 4 (SmartCluster: An Ontology-based Metamodel): This chapter presents the proposal to create a metamodel for model development for Smart Cities;
- Chapter 5 (Metamodel Validation using an Evidence-based software engineering): This chapter presents the validation of the meta-model (SmartCluster) to measure the groupings of Brazilian Smart Cities
- Chapter 6 (Discussion and Final Conclusions): This chapter closes the paper presenting the final considerations on the use of the proposed metamodel and brings an analysis of the clusters found.

1.9 Summary

In this chapter, it was presented the motivation, the problem statement, the research questions, and the outline of contributions, the methodology used, and how this thesis was organized.

2

CLUSTERCITIES: A CATALOG FOR COMPARING CITIES

I learned from my father that the only way to prove anything is real is to travel there.

— Bjorn Lothbrok, Vikings

This chapter details the design of a catalog of metrics used to compare cities. The comparison cities considers three dimensions (Territory, Population and Development) to create clusters cities by indicators of similarities (Section 2.2). For this, will be detailed dimensions and variables that represent these indicators (Section 2.3), then will be applied this catalog metrics using multivariate analysis technique to these public open data (Section 2.4).

2.1 Introduction

It is estimated that by 2050, almost 85% of the world population will live in cities, which account for a significant share of GDP. In this sense, much of the government investment should be channeled to the cities. The perception of the effectiveness of services and quality of life is a useful tool in managing budgets, enabling informed decision-making (CARDOSO, 2015).

At the beginning of this work, the research questions, **RQ3** mentions the mining and processing of data on smart cities, **"How can we perform the mining and data processing of cities?"**. To answer this question, it is necessary to perform a comparison process between cities, so that its features are respected in order to create real comparisons. Brazil, for example, with its continental dimension has cities with diverse cultural, social and economic characteristics and involves different issues related to demography, topography, climate and even cultural.

It is possible affirm that when comparing cities with similar socioeconomic characteristics and policies, the use of democratic management of resources depends directly on the characteristics of each region, so that the space interferes with the development of these cities. (SANTOS, 2006).

The analysis developed in this chapter presents the search for understanding of the relationship between the Territory, Population and Development, from analytical indicators that articulate these three dimensions aiming to find a pattern of similarity between the evaluated municipalities. This analysis is based on empirical evidence from data collected in a set of data involving clusters of Brazilian cities by similarity variables.

This chapter is organized as follows: the next section presents a conceptual reference that is the basis for the study of the relationship between the dimensions established for comparison cities. Section 2.3 presents the metrics to compare cities based on three areas (Territory, Population and Development).

Section 2.4 presents the behavior of the data sample, applying the set of Brazilian cities, in relation to the variables used, and the methods of factor analysis are then applied, in order to reduce the dimensions of analysis to factors underlying. In the last section (2.5) final considerations are presented.

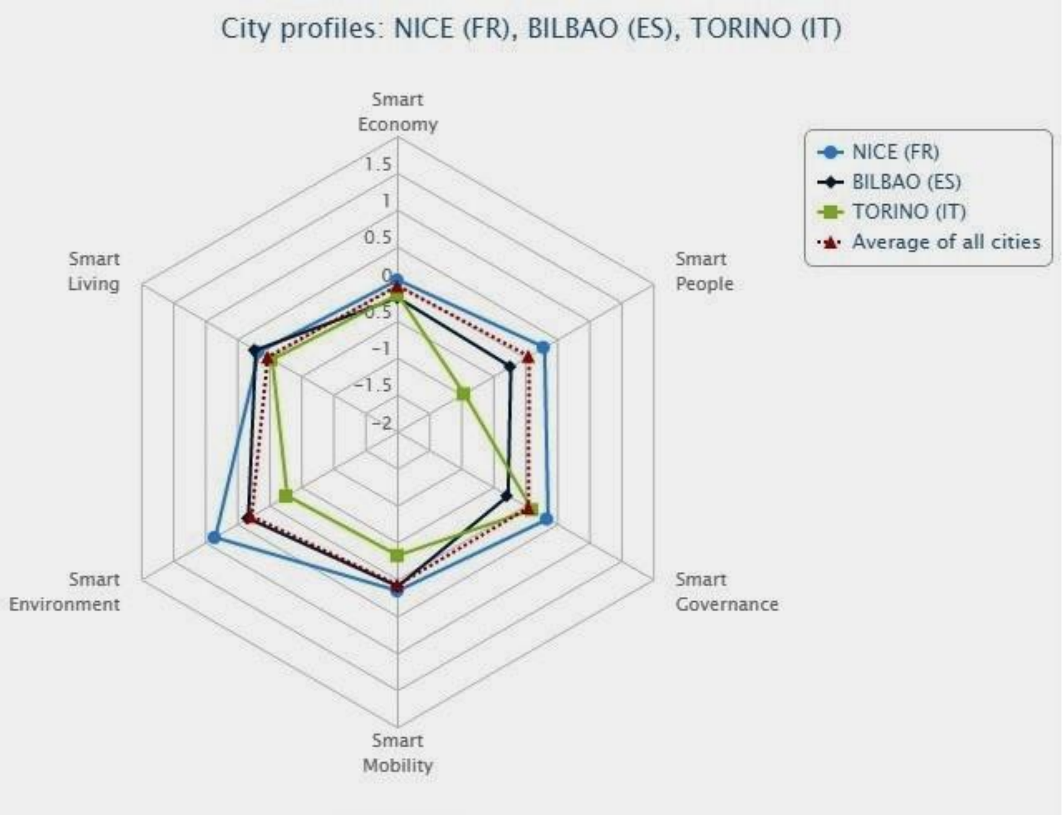
2.2 Conceptual referential on Cities and its dimensions

Around the world, some Smart City Models (SCMs) catalog these cities according to indicators, domains, and specific areas. However, few consider a "fair" comparison between cities that have the same characteristics based on population and geographic metrics. Creating a Metamodel for Smart Cities consists of incorporating the best features of these models, and adding the possibility of comparing cities in a more equivalent way.

This chapter considers the current models, and presents some considerations about the positives and negatives of these comparisons.

In Europe, the project entitled "European Smart Cities" is an initiative that brings together seventy European cities around a comparative model grounded in six distinct characteristics, divided into 33 levels. It is represented in Figure 2.1 to compare cities: Nice (FR), Bilbao (ES) and Torino (IT).

Figure 2.1: Comparison between European cities.



Source: Made by author

This project envisions the data of mid-sized cities, having in mind that these cities live the majority of the European population and they face greater challenges in terms of equipment and public resources, organizational capacity and in terms of competitiveness (WEISS, 2014).

Is presented in Table 2.1 the criteria for inclusion of cities in this project. These criteria were defined by the European model of comparison of medium-sized cities.

Table 2.1: Inclusion criteria of Cities Project Intelligent European.

Dimension	***Criterion***
Educational System	At least one university
Population	Between 100.000 and 500.000 inhabitants
Capture Area	Less than 1.5 million inhabitants

Source: Made by author

Define what are smart cities and measure the levels of sustainability, quality of life and well-being are still not consensus areas. Despite several existing indices and rankings, usually developed by companies and institutions, there is no standardization of indicators that establish what are, after all, smart cities. (CARDOSO, 2015). The publication of ISO 37120: 2014 is the first ISO framework with indicators for the cities, measuring the ability to provide services and quality of life. According to the guidelines of this, any city can use the new standard to measure its performance and compare it with other cities, with a view regardless of size, location or level of development.

The main problem in these comparison models is the absence of metrics that prior to the comparison of "intelligence" between cities and put them in without equal standing before check and compare their individual characteristics. In comparison presented in Figure 2.1, for example, are compared cities with HDI (Human Development Index) similar, however, the other indicators are very different, which somehow contaminate the sample data comparison between cities.

This problem in the comparison between cities becomes more evident when viewed indicators in Table 2.2, where for example, the city with the best HDI Brazil (São Caetano do Sul) can be compared to European human development indicators, however, its area reaches representing ten percent of the area of the city of Torino.

The city of Codó (BR) reveals the difference found in these data, even in comparison with other cities. If compared to other cities because of its large area and low population density makes it inconsistent, which somehow has a negative impact on its HDI. This is just one among many comparisons that make it impossible to compare cities based only on indicators that do not consider these individual characteristics.

Even in a comparison of European cities, where the city (Torino-IT) whose population represents a total almost three times higher than other (Nice-FR) will necessarily deviate in the comparison of public data such cities because of their different structures, services and management. Table 2.2 shows the variables (in green and red) with greater distance from the average of other municipalities.

Table 2.2: Data comparison between cities

City (Country)	***Area (km^2)***	***Density (inhab/km^2)***	***Population***	***HDI***
Nice (FR)	71,92	2.773,40	342.304	0,872
Bilbao (ES)	41,30	8.514,02	351.629	**0,878**
Torino (IT)	130,00	6.596,00	**907.704**	0,872
Codó (BR)	**2.364,49**	**27,05**	**118.072**	**0,595**
São Caetano do Sul (BR)	**15,33**	**9.736,03**	149.263	0,862

Source: Made by author

To fill this gap of inadequate comparison between cities, the next section presents a set of metrics, whose purpose is to serve as a step prior to the comparison of the intelligence indicators of cities for before, put them in the same real comparison plan and variable dimensions.

2.3 The metrics for Cities Comparison

To allow a real comparison between cities and their individual characteristics, three dimensions were created (Territory, Population and Development), with their respective variables (Table 2.3):

(A) Territorial Dimension: this dimension uses the variables Territorial Area and Population Density. This dimension aims to compare the territory of each municipality, since the total area may have a direct influence on how they will be implemented the infrastructure strategies and public services at the expense of extension of this municipality.

(B) Population Dimension: this dimension the data on Population Urban and Rural variables are presented. The comparison of these variables, it is possible to understand aspects related to urbanization of cities and the factors that can prevent a comparison with cities of different sizes characteristics.

(C) Development Dimension: this dimension of municipal human development variables are detailed in income, education and longevity. This dimension is more complex comparison between the municipalities, because it includes variables that though distinct composes a formula (HDI) responsible for indicating the development of the municipality.

Table 2.3: Details of the Domains and comparison variables

Dimension	Variable	Type	Data Source
Territory	Territorial area	km^2	(IBGE, 2012)
	Demographic density	Inhab./km²	
Population	Urban population	Inhabitant	(IBGE, 2010)
	Rural population		
Development	Income	Idh-r	(PNUD, 2000)
	Longevity	Idh-l	
	Education	Idh-e	

Source: Made by author

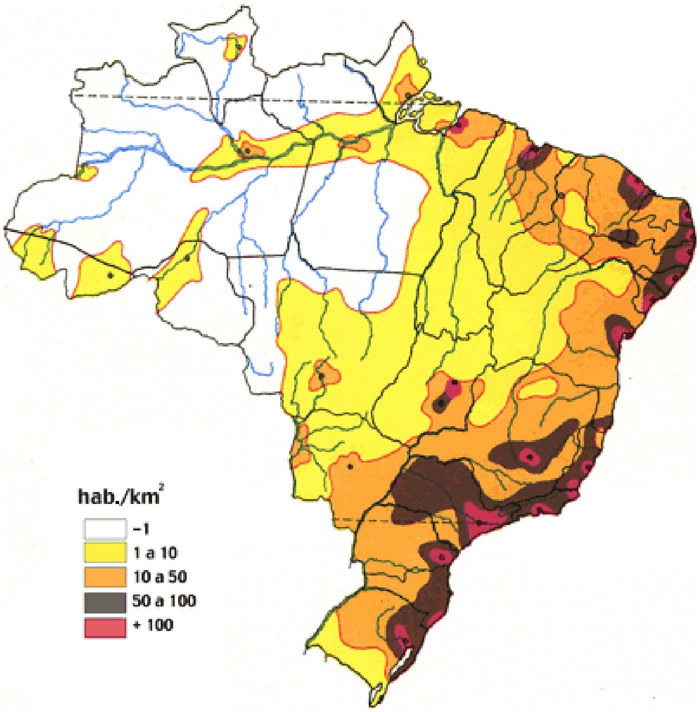
The following sections detail each of the three dimensions and their respective variables to be used in comparing cities. These dimensions will be identified by the respective colors Territory (green), Population (blue) and Development (orange).

2.3.1 Dimension: (A) Territory

This dimension is characterized by variable demographic area and density of the municipalities. The population density is a measure calculated by the relationship between population and land area, and reflects the number of inhabitants per square kilometer.

According to statistics from the Brazilian Institute of Geography and Statistics (IBGE, 2012), Brazil has a population of 202,768,562 inhabitants, distributed in an area of 8,515,767.049 square kilometers, resulting in a population density of 22.8 inhabitants per square kilometer (Figure 2.2).

Figure 2.2: Population density Brazilian



Source: (IBGE, 2012)

Because of the economic history of Brazil, the highest population density rates in Brazil are in the Southeast, followed by the South and the Northeast (Figure 2.1), and respectively (in inhabitants / km²): Southeast (67.77), South (38.38), Northeast (27.33), Central West (5.86) and North (2.66).

2.3.2 Dimension: (B) Population

In this dimension are explored variables concerning the number of inhabitants in urban and rural areas in the three cities. Currently, it is a feature of most municipalities, the population living in rural areas is much lower than that lives in urban areas, although until the 1950s, the rural population has been considerably higher (Table 2.4) than the urban (IBGE, 2010).

Table 2.4: Household situation in Brazil in recent decades.

By household situation (%)	*Urban*	*Rural*
1980	67,70	32,30
1991	75,47	24,53
1996	78,36	21,64
2000	81,23	18,77
2010	84,36	15,64

Source: IBGE, Population census 1980, 1991, 2000 e 2010.

Martins et al. (Martins, 2007) shows in his work a comparison of quality of life, which are revealed data from different perceptions of urban and rural residents in Brazilian cities. As the results shown in Table 2.5, the indices in the social and psychological domains were higher in the rural environment than in the urban environment, as in the physical and environmental fields, the inhabitants of urban areas showed higher levels.

Table 2.5: Quality of life in function of the environment.

Domains	**Average**			**Standard deviation**			**t (Rural x Urban)**
	Urban	**Rural**	**Avg**	**Urban**	**Rural**	**Avg**	
Material	3,41	3,38	3,40	0,43	0,48	0,45	0,41; p>0,05
Psychological	3,40	3,42	3,41	0,53	0,45	0,50	-0,30; p>0,05
Social	3,84	3,91	3,87	0,61	0,63	0,62	-0,96; p>0,05
Environment	3,42	3,30	3,37	0,61	0,49	0,56	1,74; p>0,05

Source: Adapted from (Martins, 2007)

The next dimension explores the indicators obtained with the HDI (Human Development Index) and explores the sub-indicators that compose it, and serve to locally measure (municipality) improvements in the living conditions of the population.

2.3.3 Dimension: (C) Development

The Human Development Index (HDI) is a summary measure of long-term progress in three basic dimensions of human development: income, education and health. The purpose of the HDI was created to offer a counterpoint to another widely used indicator, the Gross Domestic Product (GDP) per capita, which considers only the economic dimension of development. Created by Mahbub ul Haq in collaboration with the Indian economist Amartya Sen, Nobel Prize in Economics in 1998, the HDI is intended to be a general and synthetic measure that despite broaden the perspective on human development, not cover exhausts all aspects of development (UNDP, 2015).

Since 2010, when the Human Development Report completed 20 years, new methodologies were incorporated to calculate the HDI. Currently, the three pillars that make up the HDI (health, education and income) are measured as follows:

a) A long and healthy life (health) is measured by life expectancy;

b) Access to knowledge (education) is measured by: i) mean years of adult education, which is the average number of years of education received during the life of people from 25 years; and ii) the expected years of schooling for children at the age of starting school life, which is the total number of years of schooling a child at the age of starting school life can expect to receive if prevailing patterns of specific enrollment rates by age remain the same throughout the child's life and

c) The standard of living (income) is measured by the Gross National Income (GNI) per capita expressed in purchasing power parity (PPP) constant in US dollars, with 2005 as reference year.

Thus, the HDI is obtained by the geometric mean of the three previous standard indices (Equation 2.1), and is given by:

$$HDI = \sqrt[3]{LExEIxII} \qquad (2.1)$$

Where respectively, are: (a) LE (Life expectative), which in Brazil is about eighty-three; (b) EI (Education index), which considers mean years of schooling (MYSI) and expected years of schooling (EYSI); (c) II (Income Index) this income is calculated based on the gross domestic product per person with parity by purchasing power, indexed by the dollar, the calculated location.

The next section presents an application of this metric catalog to compare cities, and how the use of multivariate analysis could assist in the grouping of cities by similarities.

2.4 Application of the Metrics Catalogue

The purpose of this section is to present the application of metrics catalog using Multivariate Analysis and Cluster Analysis for a comparison of cities based on areas previously established. This practice allows to select cities with similar characteristics, so that at a later stage can be compared from the aspect of Smart Cities. The intent of this Doctoral Thesis is precisely cities present a comparison method based on these two steps: comparison of variables and indicators.

2.4.1 Multivariate Analysis

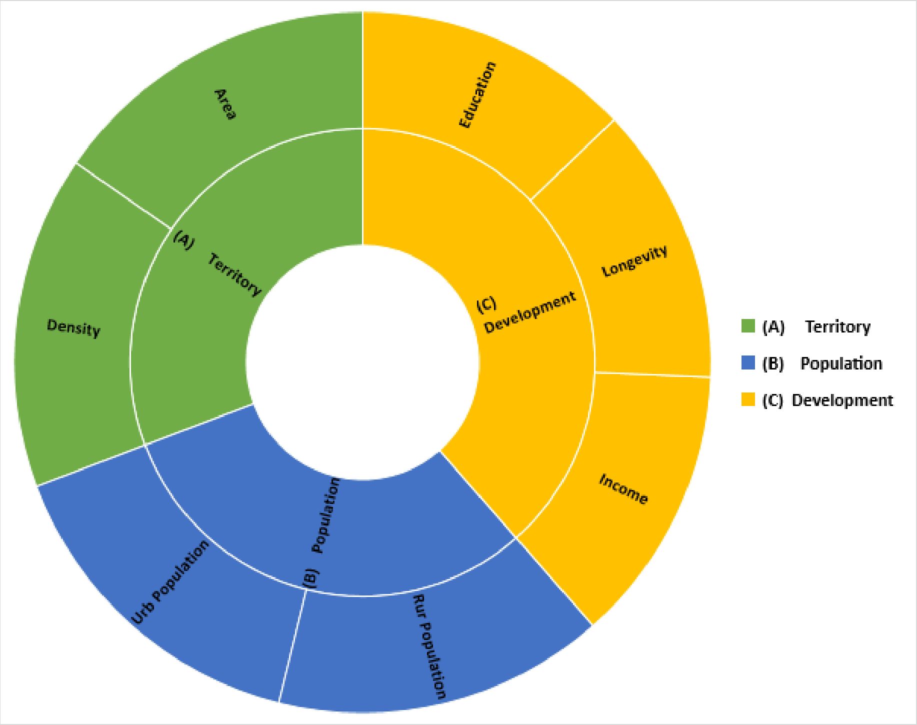
Multivariate analysis refers to all statistical methods that simultaneously analyze multiple measurements on each individual or object under investigation. Any simultaneous analysis of more than two variables may be, in a way, considered as multivariate analysis (Field, 2009).

The result of these analyzes may be the grouping of individuals by similarity. It can be achieved through the use of discriminant analysis to establish classification engines of new groups, considering the previously identified patterns (MILONE, 2009).

The combination of selected indicators makes it possible to obtain evidence about the similarity of municipalities before they are unranked by equality indicators for smart cities. Based on these indicators, it was applied the Multivariate analysis procedures, seeking, at first, through factor analysis, reducing the dimensions of analysis.

This process is essential to identify distinct patterns (area, population and development) (Figure 2.3) in clusters, since it allows the development of a cluster analysis for the set of analyzed clusters.

Figure 2.3: Dimensions and variables to compare cities



Source: Made by author

2.4.2. Clusters Analysis

The cluster analysis is an analytical technique to find significant subgroups of individuals or objects. Specifically, the goal is to classify a sample of entities (individuals or objects) in a small number of mutually exclusive groups. In cluster analysis, unlike the discriminant analysis, the groups are not predefined. Instead the technique is used to identify the groups. (Levine, 2008)

The cluster analysis typically involves two steps. The first is the measurement of some form of similarity or association between the entities to determine how many groups are, actually, the sample and that will be called in this analysis **Step 1**: (Creating clusters).

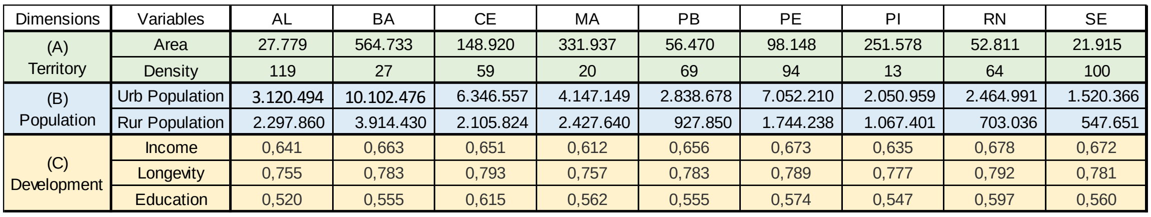
The **Step 2** (Analyzing clusters) is to define the profile of the variables in order to determine its heterogeneous composition, although dependent on each other. This step may be accompanied by the application of discriminant analysis to groups identified by cluster technique. (Britto et al, 2007)

For data analysis was used descriptive statistics and multivariate data analysis. According to Levine et al. (Levine, 2008), descriptive statistics aims to collect, summarize and present data, which in this study will be demonstrated by the frequency and average data. In this chapter, to analyze the profile of cities grouped, descriptive statistics will be applied as the multivariate analysis allows simultaneous checking samples of data characterized by correlated variables (Table 2.3 and Figure 2.2).

2.4.2.1 Step 1: (Creating Clusters)

This step will be considered dimensions (Territory, Population and Development) and their respective variables to allow the grouping of cities by similarity of these variables. To carry out the implementation of these metrics were mined data from all of the Brazilian Northeast cities, being respectively: Alagoas (102), Bahia (417), Ceará (184), Maranhão (217), Paraíba (223), Pernambuco (185), Piauí (224), Rio Grande do Norte (167) and Sergipe (75). They are represented in Figure 2.4 these metrics synthesized by states.

Figure 2.4: Dimensions and variables of northeastern states



Dimensions	Variables	AL	BA	CE	MA	PB	PE	PI	RN	SE
(A) Territory	Area	27.779	564.733	148.920	331.937	56.470	98.148	251.578	52.811	21.915
	Density	119	27	59	20	69	94	13	64	100
(B) Population	Urb Population	3.120.494	10.102.476	6.346.557	4.147.149	2.838.678	7.052.210	2.050.959	2.464.991	1.520.366
	Rur Population	2.297.860	3.914.430	2.105.824	2.427.640	927.850	1.744.238	1.067.401	703.036	547.651
(C) Development	Income	0,641	0,663	0,651	0,612	0,656	0,673	0,635	0,678	0,672
	Longevity	0,755	0,783	0,793	0,757	0,783	0,789	0,777	0,792	0,781
	Education	0,520	0,555	0,615	0,562	0,555	0,574	0,547	0,597	0,560

Source: (IBGE, 2010)

To carry out the implementation of these metrics will be performed Steps 1 and 2 to the state of Alagoas, which will serve as an example for the adoption of this catalog. In assessing the state of Alagoas and its one hundred and two municipalities can create three types of grouping, which are divided by size.

The first obtained cluster refers to the size of territory, and uses the variables Area and Population Density (Table 2.6). To this end, the variables were normalized using the analysis of variance technique, which consists in evaluating statements average populations (MILONE, 2009).

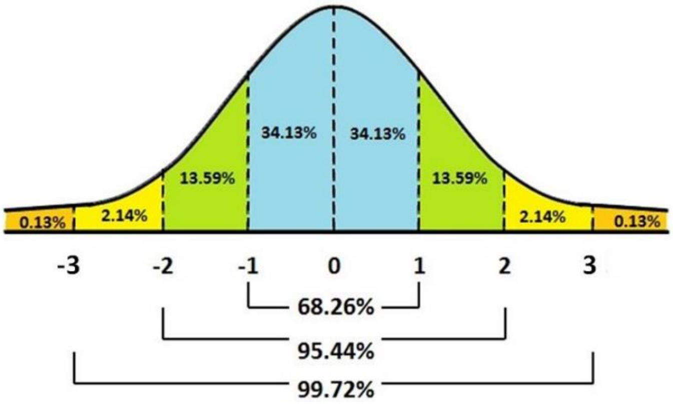
In statistics, the standard score is the (signed) number of standard deviations an observation or datum is above the mean. Thus, a positive standard score indicates a datum above the mean, while a negative standard score indicates a datum below the mean. It is a dimensionless quantity obtained by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. This conversion process is called standardizing or normalizing (CARROLL, 2002).

Table 2.6: Descriptive statistics of the variables used (N=102).

Dimension	***variables***	***Average***	***Variance***	***Standard deviation***
(A) Territory	Area	249	0,303	184
	Density	70	0,121	159
(B) Population	Pop. Urban	7054	0,112	93322
	Pop. Rural	5653	0,302	6797
(C)Development	HDI-r	0,542	0,253	0,040
	HDI-l	0,743	0,302	0,032
	HDI-e	0,435	0,303	0,054

In this application, there is the z-score when, for example, the dimensions of the variables measured values deviate from the average in terms of standard deviations. When the z-score is positive it indicates that the data is above the average and when it is negative means that the data is below the average. This is illustrated in Figure 2.5 the distribution graph of values in a sample which oscillates between -3 <Z <+3, representing 99.72% of the values within the so-called normal distribution.

Figure 2.5: Distribution of values for a z-score



In the example of Alagoas to calculate the z-score of the area of all the municipalities has used to Equation 2.2:

$$Z = \frac{x - \mu}{DP_{pop}} \tag{2.2}$$

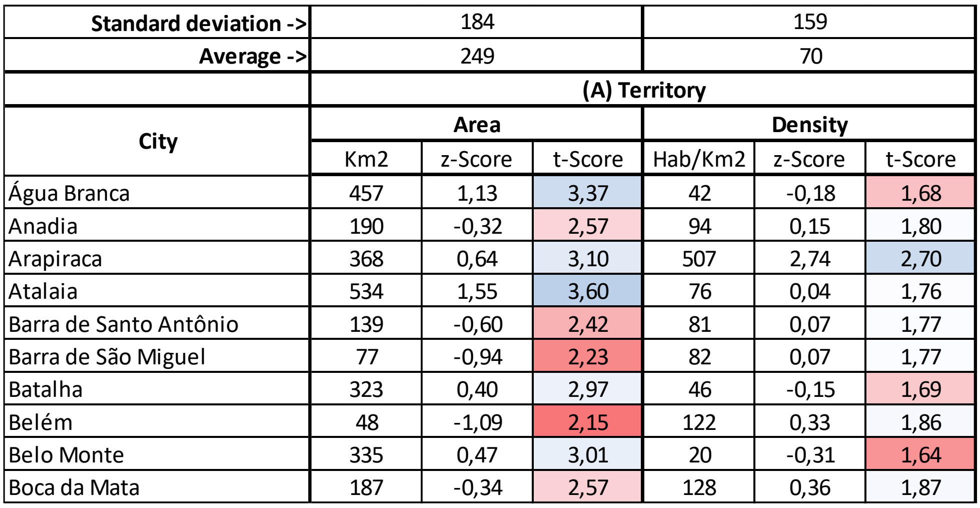
Where μ represents the average value of all the areas (in the state of Alagoas, the average is 249 km2), which is subtracted from the individual value of the area of each city and divided by the standard deviation value (which is given by 184 km2).

After defining the z-score is possible to identify those municipalities whose area away from the middle, and with that, the standard deviation value relative to the z-score is accentuated. However, for a better visualization of data possible, was chose to create the T-score using a range from 0 to 5 to allow further grouping of these municipalities according to the similarity values of this range (Equation 2.3).

$$X_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{2.3}$$

The combination of standardization of data, standardization in groups (0-5), allowed the creation of heatmaps in color scales that govern the display of the distribution of mean values and cities groupings. Figure 2.6 shows a partial view of the dimension (A) Territory of Alagoas municipalities.

Figure 2.6: Partial view of the territorial dimension.

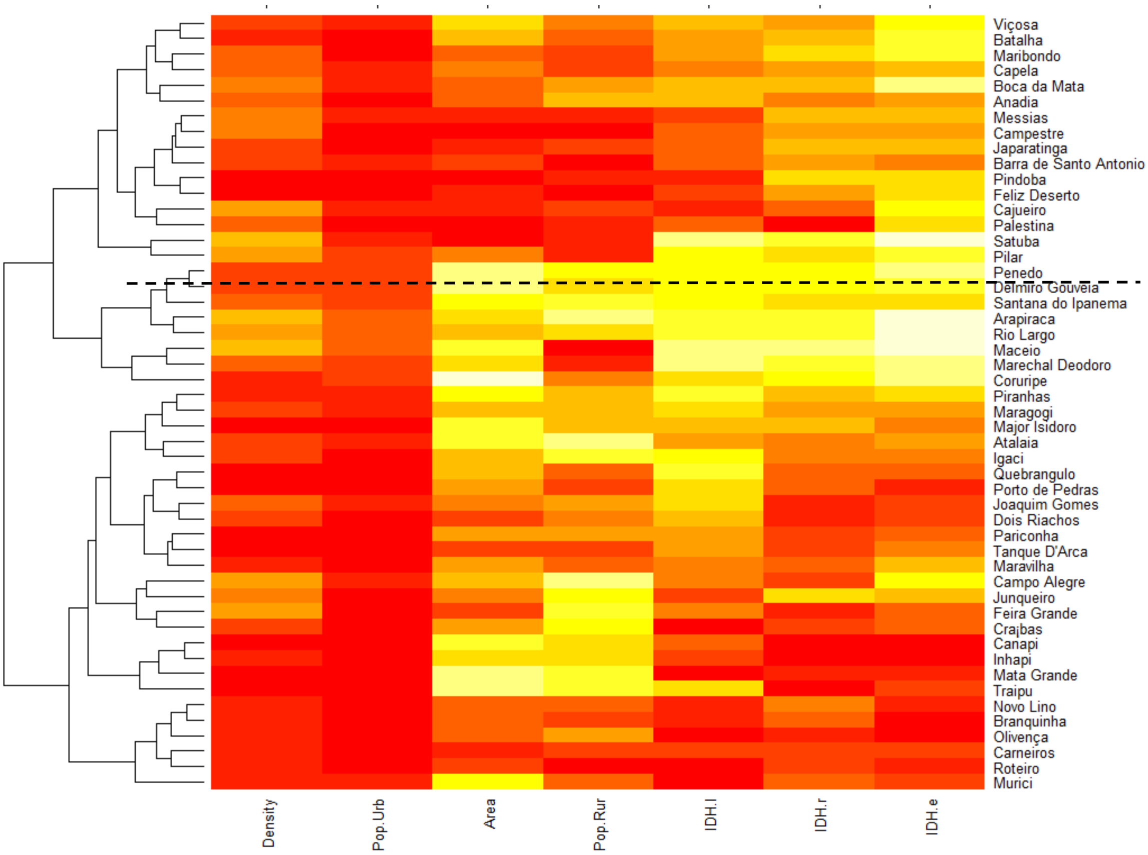


Standard deviation ->	184			159		
Average ->	249			70		
	(A) Territory					
City	Area			Density		
	Km2	z-Score	t-Score	Hab/Km2	z-Score	t-Score
Água Branca	457	1,13	3,37	42	-0,18	1,68
Anadia	190	-0,32	2,57	94	0,15	1,80
Arapiraca	368	0,64	3,10	507	2,74	2,70
Atalaia	534	1,55	3,60	76	0,04	1,76
Barra de Santo Antônio	139	-0,60	2,42	81	0,07	1,77
Barra de São Miguel	77	-0,94	2,23	82	0,07	1,77
Batalha	323	0,40	2,97	46	-0,15	1,69
Belém	48	-1,09	2,15	122	0,33	1,86
Belo Monte	335	0,47	3,01	20	-0,31	1,64
Boca da Mata	187	-0,34	2,57	128	0,36	1,87

Source: Made by author

When performing random selection of 50 cities in this data set, it was possible to obtain a total collation of all sizes (Figure 2.7).

Figure 2.7: Total cluster of cities and dimensions.



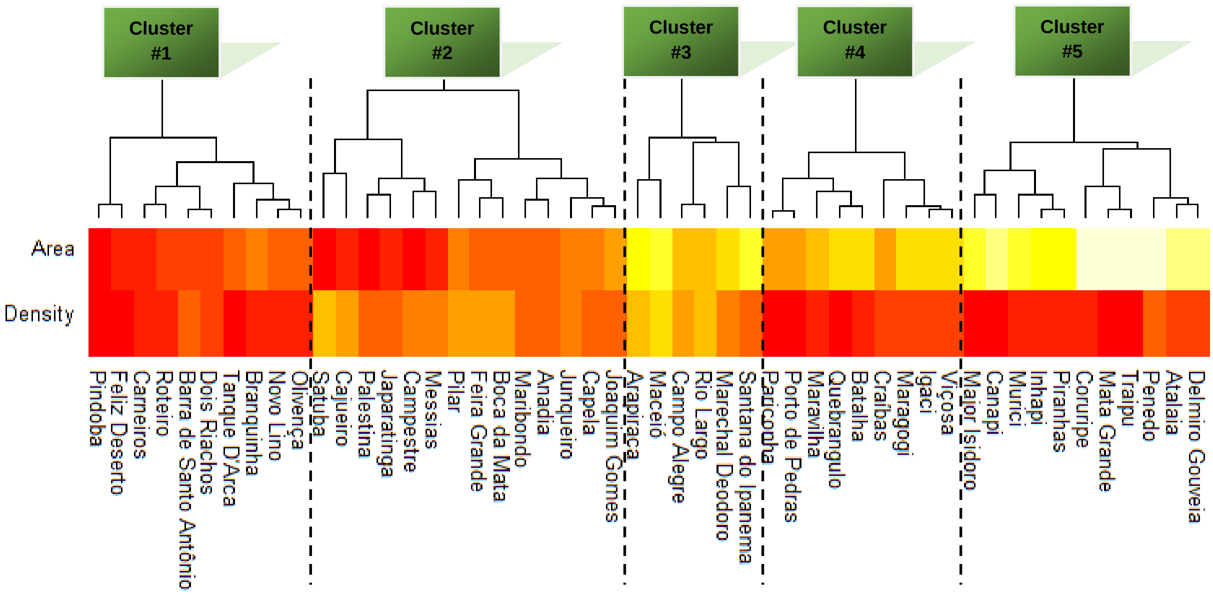
Source: Made by author

Can be observed left side to the image separation in groups of cities, bounded by a dashed line, which represents two large data clusters. This separation could in the first instance represent an interpretation of similarity indicators for these cities, but the next step presents a greater detail of these groups by cluster analysis.

2.4.2.2. Step 2: (Analyzing Clusters)

A way to better detail and conduct further analysis of clusters of cities by similarity indicators is to separate them into groups by size. Figure 2.8 shows the group of 50 cities Alagoas according to Spatial dimension, thus allowing to compare the cities with the size and characteristics more similar density.

Figure 2.8: Clustering cities by the Territory Domain.



Source: Made by author

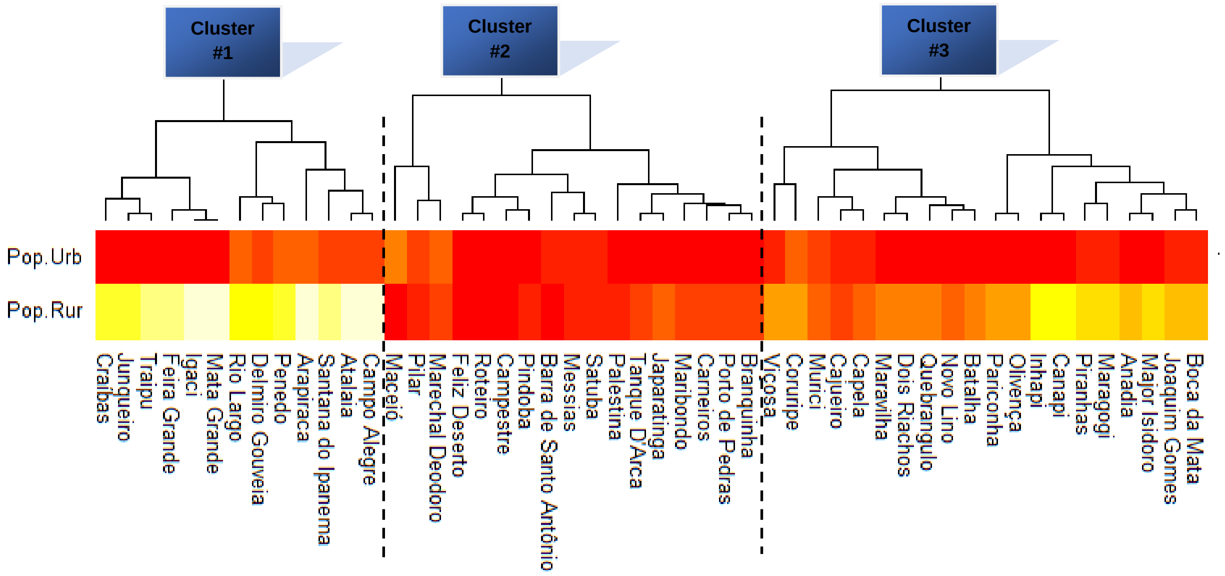
The comparison between these groups enables the creation of statistical data tables (Table 2.7) which are presented each of the groups and the cities ranked in each of them. Each group (cluster) is indicated by its number (Cluster # 1 to Cluster # 5), the average obtained between the variables (area and density) and the maximum variation of the data within the groups.

Table 2.7: Average and variation of Territory Cluster

Cluster #1 – Average 2,07 ± 0,139
Pindoba, Feliz Deserto, Carneiros, Roteiro, Barra de Santo Antônio, Dois Riachos, Tanque
Cluster #2 – Average 2,15 ± 0,117
Satuba, Cajueiro, Palestina, Japaratinga, Campestre, Messias, Pilar, Feira Grande, Boca da Mata, Maribondo, Anadia, Junqueiro, Capela, Joaquim Gomes
Cluster #3 – Average 2,84 ± 0,841
Arapiraca, Maceió, Campo Alegre, Rio Largo, Marechal Deodoro, Santana do Ipanema
Cluster #4 – Average 2,32 ± 0,393
Pariconha, Porto de Pedras, Maravilha, Quebrangulo, Batalha, Craíbas, Maragogi, Igaci, Viçosa
Cluster #5 – Average 2,76 ± 1,346
Delmiro Gouveia, Atalaia, Penedo, Traipu, Inhapi, Murici, Canapi, Major Isidoro

Source: Made by author

Figure 2.9: Clustering cities by the Population Domain.



Source: Made by author

Unlike the first group, in this case only three groups characterize municipalities with similar data on these variables. It is noticed that the variation between the data from municipalities is higher (Table 2.8), to group more municipalities with different characteristics. This grouping can be reduced to smaller groups for greater granularity of data, and thus, greater similarity between the variables. The clearer the indicative color, the higher the measured indicator, as can be observed in the city of Maceió, which has urban population superior to the rural one.

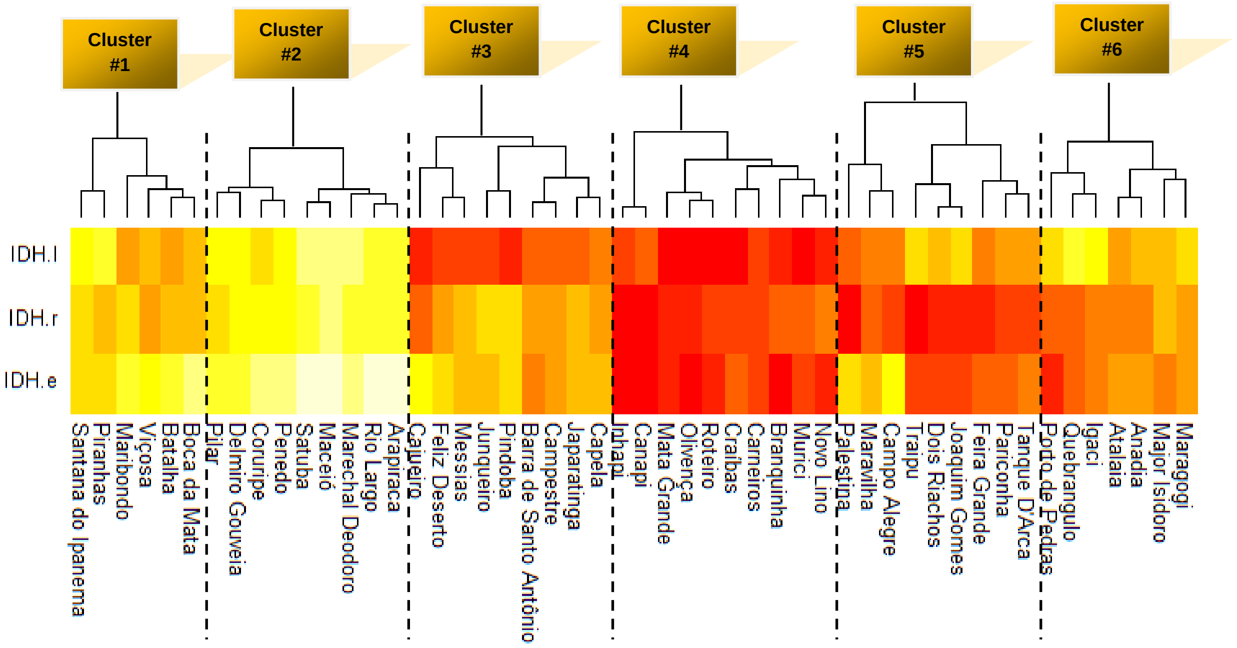
Table 2.8: Average and variation of Population Cluster.

Cluster #1 – Average 2,29 ± 0,378
Boca da Mata, Joaquim Gomes, Major Isidoro, Anadia, Maragogi, Piranhas, Canapi, Inhapi, Olivença, Pariconha, Batalha, Novo Lino, Quebrangulo, Dois Riachos, Maravilha, Capela, Cajueiro, Murici, Coruripe, Viçosa
Cluster #2 – Average 2,18 ± 0,397
Branquinha, Porto de Pedras, Carneiros, Maribondo, Japaratinga, Tanque D'Arca, Palestina, Satuba, Messias, Barra de Santo Antônio, Pindoba, Campestre, Roteiro, Feliz Deserto, Marechal Deodoro, Pilar, Maceió
Cluster #3 – Average 2,81 ± 1,185
Campo Alegre, Atalaia, Santana do Ipanema, Arapiraca, Penedo, Delmiro Gouveia, Rio Largo, Mata Grande, Traipu, Junqueiro,Craíbas

Source: Made by author

The third and last group refers to data related to the development dimension, which is composed of the variables: education, longevity and Employment (Figure 2.10).

Figure 2.10: Clustering cities by the Development Domain.



Source: Made by author

As mentioned in the previous grouping, this grouping granularity variable is higher, and thus, the similarity between municipalities becomes more evident as the variation among the data is smaller.

It is observed from Table 2.9 that the Cluster # 5 stands out with the highest average, and in this case, is the municipalities with better variables that indicate areas of better development in the state. This group included the state capital (Maceió) and the second largest city (Arapiraca).

Table 2.9: Average and variation of Development Cluster.

Cluster #1 – Average 2,78 ± 0,081
Maragogi, Major Isidoro, Anadia, Atalaia, Igaci, Quebrangulo, Porto de Pedras
Cluster #2 – Average 2,58 ± 0,132
Tanque D'Arca, Pariconha, Feira Grande, Joaquim Gomes, Dois Riachos, Traipu, Campo Alegre, Maravilha, Palestina
Cluster #3 – Average 2,09 ± 0,075
Novo Lino, Murici, Branquinha, Carneiros, Craíbas, Roteiro, Olivença, Mata Grande, Canapi, Inhapi
Cluster #4 – Average 2,67 ± 0,102
Capela, Japaratinga, Barra de Santo Antônio, Pindoba, Junqueiro, Messias, Feliz Deserto, Cajueiro
Cluster #5 – Average 3,62 ± 0,210
Arapiraca, Rio Largo, Marechal Deodoro, Maceió, Satuba, Penedo, Coruripe, Delmiro Gouveia, Pilar
Cluster #6 – Average 3,06 ± 0,063
Boca da Mata, Batalha, Viçosa, Maribondo, Piranhas, Santana do Ipanema

Source: Made by author

The next section presents the final remarks on this chapter.

2.5. Summary

This chapter aims to present a catalog of metrics comparing Brazilian cities, especially the northeastern cities, considering it to three dimensions (Territory, Population and Development).

The data from these cities were grouped according to the similarity of variables in each domain and the resulting data were presented in form of dendrograms.

The aim of this chapter to create the layer preceding the comparison of data smart cities, making use these pre-established domains (Territory, Population and Development). Thus, the comparison of cities will attend a standard more equalitarian comparison considering so, regional differences, population profile and development of the region in which the municipality is located.

3

TOWARDS A TAXONOMY TO MEASURE BRAZILIAN SMART CITIES

Bizarre is a good thing.
The common has thousands of explanations.
The bizarre has hardly any.

— Gregory House, M.D. House

This chapter presents the definition of a taxonomy of indicators compatible with the reality of Brazilian cities. The previous chapter explored models of smart cities around the world, and based on this Systematic Literature Review (SLR) were able to identify specific needs that will be solved with the taxonomy proposed in this chapter

3.1 Introduction

Cities and regions are facing new management challenges, ranging from the rationalization and optimization of resources to the implementation of technologies integrated into the daily lives of citizens to create the Smart Cities. The implementation of new technologies requires the use of large public data repositories (Open

Government Data), and these in turn give origin to specific indicators. After performing a detailed SLR on smart cities (Chapter 2), it was possible to understand that there are still many unanswered questions about the definition and use of indicators, which require both a better understanding of how much significant definition on the issue. Some of these questions on indicators for Smart Cities are:

- **Q1:** How the indicators are described, modeled and implemented?
- **Q2:** What is the origin of the data to feed these indicators?
- **Q3:** How a particular indicator is decomposed into other indicators?
- **Q4:** What indicators can be used by Brazilian cities?
- **Q5:** What is the relationship between indicators and public databases?
- **Q6:** How to evaluate the quality and usability of an indicator?
- **Q7:** How the indicators are divided into Domains and Areas?

To help answer these questions, this chapter proposes to create a taxonomy of indicators for smart cities based on the formal development of a domain ontology. For this, in Section 3.2 will be presented the methodology used for the definition and taxonomy of indicators. In Section 3.3 the Indicators features of smart cities will be described. It fell to Section 3.4 present the set of indicators compatible with the Brazilian reality. In Section 3.5 the implementation of the taxonomy will be detailed with examples for the Brazilian capitals and finally, in Section 3.6 describes a summary of this chapter.

3.2 Methodology of Establishing the Taxonomy

This chapter proposes the creation of a taxonomy for smart cities indicators, and according to the definition given by W. S. Judd (2007) taxonomy is configured as a definition of groups of individuals, based on common characteristics. This concept used in biological areas consists of grouping these individuals into sets and these, in turn, form new larger clusters, thus creating a hierarchical classification that designates each cluster with its respective rates.

There are different definitions for Taxonomy, but generally the steps for its creation are the same: design, naming and classification of groups. In this Chapter will be used the definition that describes a non-biological taxonomy as a field of science (and main component of systematic) which includes identification, description, nomenclature and classification (Simpson, 2010).

To create the formal taxonomy a Smart Cities domain ontology will be developed based on Maedche (A. Maedche and S. Staab, 2000) methodology that converts a generic ontology in a computable model; are specified text and obtains domain concepts from available sources; remove with generic concepts, and in the end only the domain remains (Kaon, 2001). The main intention is to use an ontology with allow the end of stages of construction of this ontology is possible to obtain a low maintenance taxonomy because the incremental and interactive creation process consisting of the following steps: specification, conceptualization, delivery and deployment.

3.2.1 Specification

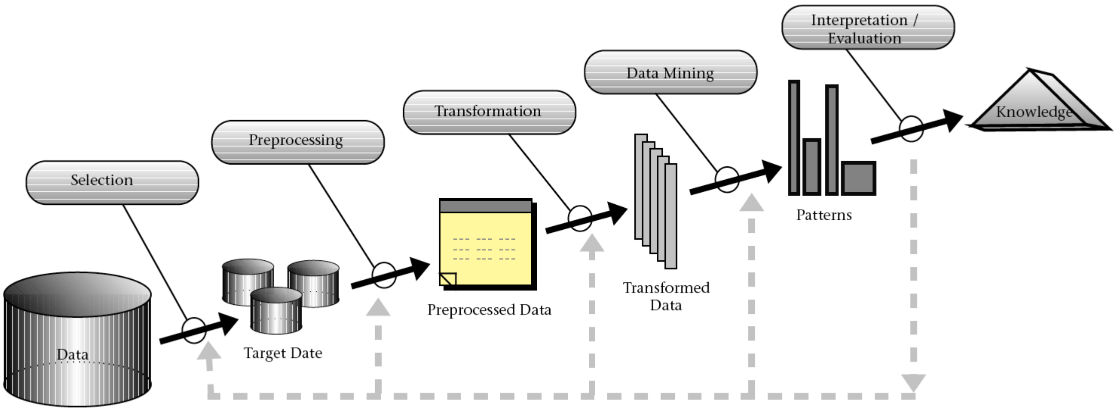
The first development stage of a domain ontology is the specification, and it is necessary to seek knowledge that will be used by this ontology. The knowledge acquisition (KA) for knowledge-based systems (KBS) in synthesis can be defined as a process where the knowledge gained from various sources (documented or not organized and better explained). The acquisition of knowledge is the total process of learning about content and download it to the computer representing it in a format usable by the machine (Regoczei and Hirst, 1994).

Another very common approach to acquiring knowledge in large volumes of data is the methodology KDD (Knowledge Discovery in Databases) which allows through steps clearly defined data extraction and transformation of these data into information. According to Fayyad (Fayyad, 1996) KDD it is a set of steps that process the data according to the following order:

1) **Selection:** Collect and search for data in the database;
2) **Pre-processing:** Treat the data, special characters and text encoding;
3) **Transformation:** Set data pattern and correct spelling errors;
4) **Mining:** Apply data analysis and interpretation of information and
5) **Interpretation:** Transforming information into charts.

Figure 3.1 shows the steps of the KDD, which were used in this work to carry out the collection and interpretation of the data used in the construction of ontology that will serve as a concept to define the taxonomy proposed.

Figure 3.1: KDD steps (Knowledge Discovery in Databases)



Source: https://goo.gl/xXKcfJ

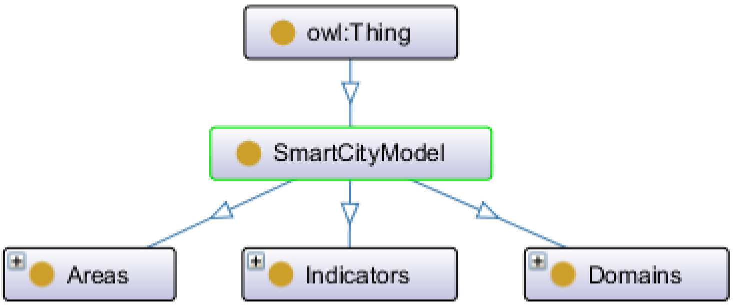
Data sources with knowledge of Smart Cities and their indicators are drawn from academic articles when running the SLR (Chapter 2). Through this step was specified structure that includes the use of indicators, domains and areas and will be detailed in Section 3.2.4

3.2.2 Conceptualization

Structure the domain knowledge in a conceptual model is the main task of this stage (Fernandez-Lopez, 1999). Therefore, it is important to understand that the main distinction between ontologies and knowledge bases is: an ontology provides a framework that serves as a foundation for building a knowledge base at a higher level. The ontology is composed of a set of concepts and terms that describe a domain, while the knowledge base uses this term structure to describe an environment. When this environment changes, a knowledge base also changes, however, an ontology does not change until its links are changed.

Nardi and Falbo (2015) presents an ontology for defining software requirements where it is possible to understand the formalization of axioms, classes and relations between them. The ontology for Smart Cities (SmartCluster) will use the same concept thereby ensuring the use of first-order logic can create a satisfactory set of restrictions on the indicators presented.

Figure 3.2: Smart Cities Diagram with “Areas”, “Domains” and “Indicators”



Source: Made by author

The use of a conceptual model facilitates the stage of formalization of the conceptual model, as it allows a higher abstraction scheme represented by the OWL language, leaving aside the technical details of structuring and placing the focus on building the class hierarchy, properties and relationships domain ontology. Considering that the "**Indicator**" classes, "**Domain**" and "**Area**" are respectively represented by the letters: "**In**", "**Do**" and "**Ar**", the following ontological axiom (A.1) is proposed:

$$(\forall\ \mathbf{Ar, Do, In})\ \mathbf{partOf\ (Ar, In)} \land \mathbf{request\ (Do, In)} \rightarrow \mathbf{request\ (Do, Ar)} \qquad (\text{A.1})$$

Thus, it is possible to assert that a particular indicator (**In**) is part of an area (**Ar**) and is required for a Domain (**Do**), then an area (**Ar**) requires a Domain (**Do**).

Similarly the following axioms can respectively introduce the notation for the concept of whole-part and sub-type classes. The concept of all-part (A2) illustrates the relationship between classes "**Indicator**", "**Domain**" and "**Area**".

$$(\forall\ \mathbf{Do,In,Ar})\ \mathbf{partOf\ (Do,Ar)} \land \mathbf{partOf\ (Ar,In)} \rightarrow \mathbf{partOf\ (Do,In)} \qquad (\text{A.2})$$

The axiom (A.3) shows the notation sub-type for the class "**Indicator**" and his "HDI" element. The HDI (Human Development Index) is obtained through a calculation to estimate the quality of life of the inhabitants of a particular location, and is used in this ontology as one of the indicators used by the domain "**Health**", in the "**Infrastructure**".

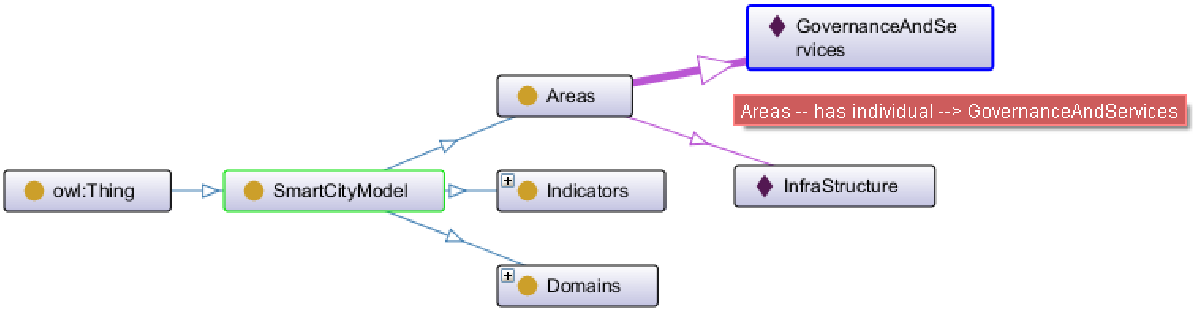
(∀ In,IDH,Do) subTypeOf(In,IDH) ∧ subTypeOf(Do,IDH) → subTypeOf(In,IDH) (A.3)

3.2.3 Delivery

The delivery stage transforms the conceptual model into a formal model or semi-computable (Fernandez-Lopez, 1999). Hence the definition of an ontology shows the existence of categories and elements which have meanings which may vary with regard to the terminology adopted by the area uses. So Gruber (Gruber, 1993) cataloged five components as fundamental to the creation of an independent ontology of the domain to which it is:

- **Classes and concepts:** express anything on which there may be a way to group them as tasks, functions, actions, strategies, among others. Figure 3.3 shows an example where the "**Area**" class is divided into two possible groups: **a)** “**Infrastructure”** and **b)** “**GovernanceAndServices**”.

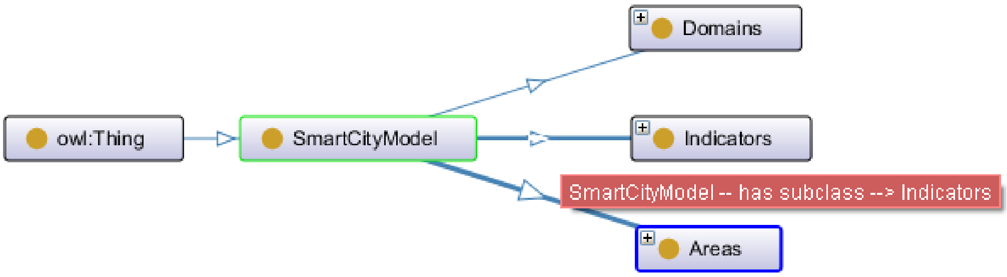
Figure 3.3: Class “Areas”,“Infrastructure”,”Services”,”Governance”



Source: Made by author

- **Relationships:** it symbolizes the interaction between classes and your domain, or how they connect. As an example, Figure 3.4 shows the use of the term "has subclass" in order to maintain a hierarchical relationships between classes;

Figure 3.4: Class, SubClass and their relationships



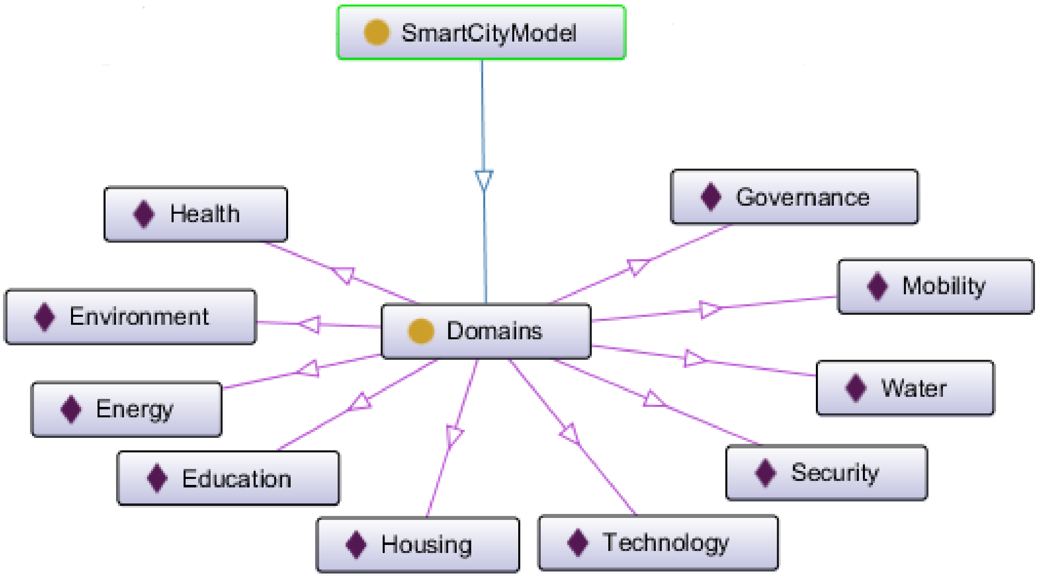
Source: Made by author

- **Functions:** A function determines how relationships are quantified. These functions are based on stocks quantifiers and universal based on axioms and mathematical expressions. Currently the ontology development tools are based on the creation of axioms and logical descriptions.

- **Axioms:** Are composed of a sentence, a proposition, a statement or a rule that allows the construction of a formal system and are classified in structural and non-structural.

- **Instances:** represent elements of an ontology, instances represent concepts and relationships that were created in the ontology. In Figure 3.5 Domains class instances are stated: “Health”, “Energy”, “Water”, and seven more (i.e. as discussed in Chapter 2).

Figure 3.5: Instances of an ontology



Source: Made by author

The creation of ontologies has the support tools such as: OntoEdit (Maedche et al, 2000) and Protégé (Noy et al, 2000) among others. This study used the Protégé which served to assist the process of defining the classes, relationships and axioms, and allows the axiomatic construction of relationships between classes.

3.2.4 Deployment

Dean (Dean et al., 2003) noting the need to represent knowledge in Artificial Intelligence initially created a language based on HTML (called SHOE). Later came the XML-based language (called taxol which later became OIL), and several other languages based on frames and knowledge acquisition approaches such as RDF (Resource Description Framework) (Lassila, 1999) OIL (Ontology Inference Layer or Ontology Interchange Language) (Fensel et al, 2001), the DAML + OIL (DARPA agent markup language) (Horrocks et al, 2001) and finally the OWL (Web Ontology Language).

It takes into consideration also the fact that OWL is currently a standard used almost all the developed ontologies. Table 3-2 shows that data collection about languages for ontologies.

Table 3.1: Ontology languages

Language	Description	Pros	Cons
RDF	Developed with the objective of representing knowledge through semantic networks. (Lassila, 1999)	It has a formal semantics which uses the vocabulary based on concepts located in URIs (RDF-S, containing the concepts of class, subclass) and an XML-based syntax.	It is a language not very expressive, allowing only the representation of concepts, taxonomies of concepts and binary relations.
DAML+OIL	DAML + OIL is a semantic markup language for the Web that features extensions to languages like DAML (DARPA Agent Markup Language). (Horrocks et al., 2001).	Broader than RDF can represent concepts, taxonomies, binary relations and instances.	Its usability has been significantly reduced by the establishment of the OWL language that keeps its features and incorporates new features.
OWL	OWL is designed to be used by applications that process the content of information instead of just presenting it to humans. (Dean et al, 2003)	It has most Web content playability by machines in relation to languages such as XML, RDF and RDFS.	Graphics are required development environments such as OWLViz or extension of the editing features.

Source: Made by author

Dean et al (2003) further states that the OWL language can be represented in three ways:

- **OWL Full:** this sub-language lets extend the predefined vocabulary written ontology in RDF or OWL. However, it is hardly compatible with inference software as these can not completely bear the features of OWL Full, or as OWL DL imposes restrictions on the use of RDF OWL Full allows mixing OWL RDF making it possible, for example, a class is both a class and an individual;

- **OWL DL:** ensures the concept of computability (all conclusions are computable) and decidability (all computations have a predetermined time to terminate). DL abbreviation refers description logic (description logics), and determines particularly to a restriction structure is more complex than Lite sub-language setting such a class can be subclassed to many classes, without it being another instance class;

- **OWL Lite:** It is a sub-language OWL DL using only some characteristics of the OWL language, and therefore is more limited than OWL DL and OWL Full. It is theoretically used in applications that simply require a hierarchy and simple constraints as restrictions supported by it are only those related to cardinality 0 or 1;

The choice for OWL-DL language is guided in the fact that ontology development tools such as Protégé (Noy et al., 2000) (with Plugin OWL), enables users, in addition to developing ontologies, finding errors and detect inconsistencies. The fact that this tool is able to generate corresponding classes in Java from an ontology in OWL was taken into account, as well as the fact of the OWL language is recommended as a standard language for developing ontologies according to W3C (World Wide Web Consortium).

3.3 Indicators Features of Smart Cities

Examining issues previously developed in Section 3.1, it is possible to correlate them with the necessary characteristics to describe the Smart Cities indicators, which are presented in Table 3.2.

Table 3.2: Indicators Features

Features	**Description**	**Questions**
Name	Represents the standardization of specific public data in order to measure and compare the predetermined areas.	Q1
Data source	Indicates the source of public information, and the origin of this source is public or private.	Q2, Q5
Type of Indicator	The types of indicators can be classified into city, state or country.	Q4, Q6
Domain	Outlines the domains that represent the concept of smart cities, which may be: Water, Energy, Mobility, Environment, Safety, Health, Education, Technology and Management.	Q3, Q7
Area	Indicates which of the three areas of public management belong to the fields and their respective indicators Infrastructure, Services and Governance.	Q3, Q7

Source: Made by author

In a simplified way, it can be stated that indicators are standardized representations of data selected to compose specific domains (Q3). These indicators have characteristics that differentiate them from being able to add new indicators if necessary extend the comparison and measurement possibilities (Q4, Q6).

Currently some Smart City models (see Chapter 2 of this thesis) using specific and not standardized, which creates the necessity of develop a taxonomy to catalog them and organize them. Creating this taxonomy specifies to which domain and area these indicators belong (Q3, Q7).

3.4 Set of Selected Indicators for the Brazilian Scenario

For the UN (2015), the minimum number of inhabitants to be a considered and to become city, have to be a human group with more than 20,000 inhabitants. The definition of the size for the cities was given by the International Conference of Statistics in 1887 and is maintained by the International Statistical Institute (ISI) (ISI, 2015).

According to the ISI, cities with a population over 100,000 inhabitants are considered large cities. However, establish criteria defining only based smart cities in the number of inhabitants can lead to misconceptions by not considering regional characteristics, political, social and economic of these cities.

Other less technical views present concepts about smart cities that are based on citizen relationships, quality of life, sustainability and behavioral aspects. The paper of HERNANDEZ-MUÑOZ (2011) contextualizes the Smart City on two levels (infrastructure and services), but sets it on a holistic level, cities are "systems of systems", and this could be the simplest definition of the term. Therefore, this work considered different papers that raised areas and areas of Smart Cities around the world, that they might be suitable to the Brazilian reality.

According to the Global Index ranking of Open Data 2015, produced by the Open Knowledge (OKF, 2015) Brazil occupies the twelfth position among the 122 places that have adopted the philosophy of sharing open public data. In the future, a paper to be written will deal with results found in the comparison of data among BRICS countries, which respectively occupy positions: Brazil (12), Russia (61), India (17) China (93) and South Africa (54).

However, among the countries of this group, Brazil was chosen because it has more public transparency initiatives, such as the one created by Transparency Brazil (TB, 2014). Thus, was be arrived to a model consisting of 10 domains called "Domains Basic" where each domain has its respective "Basic Indicator".

The next section presents the origin of these indicators and why they were chosen.

3.4.0. Where did these indicators come from?

The main objective of these domains and basic indicators is to understand the scenario in which the city is inserted, and thus understand the structural weaknesses that need further attention to the city to be comparable to a Smart City. Table 3.3 shows these domains and their basic indicators. All papers (P_n) and data sources (DS_n) mentioned hereafter.

Table 3.3: Basic Domains and their Indicators

Domains	**Basic Indicators**	**Papers**	**Data Sources**
A - Water	Piped water	[P01,P03,P20]	[DS04,DS23,DS27]
B - Education	HDI–Education	[P10,P15,P17,P19]	[DS04,DS26,DS9,DS6]
C - Energy	Access to energy	[P03,P08,P16,P18]	[DS04,DS28]
D - Governance	HDI/Employment	[P02,P06,P19]	[DS05,DS22,DS25]
E - Housing	Private residence	[P02,P03,P15]	[DS04,DS33,DS6]
F - Environment	Waste collected	[P07,P06,P15]	[DS14]
G - Health	HDI – Health	[P21,P12,P13]	[DS04,DS6,DS24,DS31]
H - Security	Homicides/1000	[P09,P16,P19]	[DS04,DS30,DS32]
I - Technology	Computers/home	[P02,P15,P20]	[DS04]
J - Mobility	Mass transport	[P03,P06,P19]	[DS15,DS29]

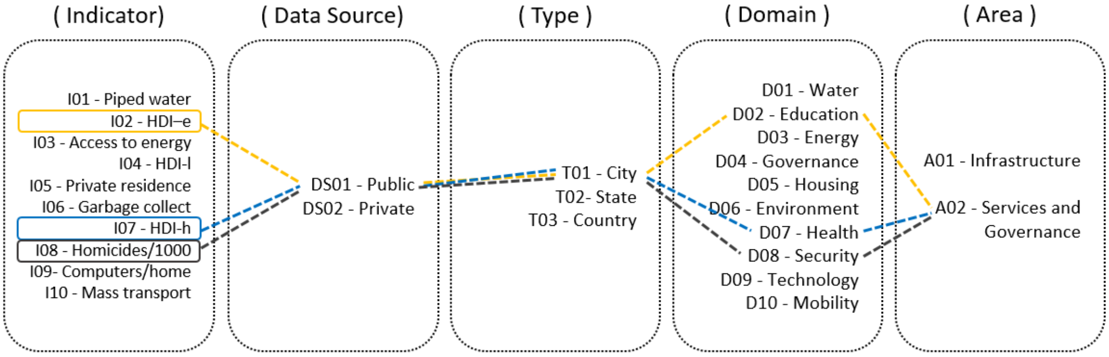
Source: Made by author

To develop the domains and their indicators were considered the studies surveyed both in Section 2 and in the field of Smart Cities. The intersection of the domains presented various papers made it possible to create this list containing 10 domains. To create the list of indicators for each of the domains presented in Table 3.3, were considered three important factors:

1. **The data representativity**: The domains and indicators are cited in the most relevant papers either dealing with models for smart cities, theoretically or using public data (See Chapter 2).
2. **The data availability:** Data to measure and compare cities are available in public databases and can be mined to be transformed into indicators.
3. **The data compatibility:** Data sets mined from public databases and transformed into indicators need to be compatible with international data sets to ensure that new Smart City models can use equivalent indicators.

These indicators are called basic indicators and for each of them, there are also two secondary indicators associated to the domain. Although mathematically, there is no difference in the calculation of ranking among the cities, this chapter presents only the basic indicators and the definition of secondary indicators is being created according to the criteria used for the primary indicators.

Figure 3.6: Features selected to measure Brazilian Capitals



Source: Made by author

Figure 3.6 shows each of the areas, their indicators and how these indicators are calculated to compose the concept of Smart City proposed by this thesis. This indicators has a unique code (token) to identify this sources, composed by a set of features (i.e: Water - Indicator **I01**, DataSource **DS01**, Type **T01**, Domain **D01** and Area **A01**), and will be presented in the next Sections.

3.4.1. (A) Water – Piped water (I01DS01T01D01A01)

The Water domain was appointed in the works of [P01, P03, P20] as essential to the understanding of Smart Cities. However, the reality found in Brazilian cities can become unviable, because among the 5570 Brazilian municipalities, only three have all households supplied by piped water and sanitation while 2147 municipalities had an index less than 90% of residential supply (IBGE, 2010).

Compare this scenario with the reality of European or North American countries may hinder the planning of public policies to expand the network of water and sewage in the country. Currently, according to the IDEC (Brazilian Institute of Consumer

Protection) for failures in governance, mega cities like São Paulo and Rio de Janeiro face problems such as waste and lack of water supply (IDEC, 2015).

As stated in the estimation of water from UNICEF (UNICEF, 2015) report, access to piped water increased from 83% in 1990 to 92% in 2010, while access to sanitation increased from 71% to 75%. So this domain will make use of indicator that quantifies the percentage of households served with piped water in the municipality assessed [DS04,DS23,DS27].

3.4.2. (B) Health – HDI Health (I07DS01T01D07A02)

The papers [P10,P15,P17,P19] refer to the area of smart health, and for this Thesis the following data sources were used: [DS04,DS6,DS24,DS31]. To calculate the indicator that represents the health of a city, was be used the HDI (Human Development Index) as an indicator. This index developed in the 90s and has been used by UN member countries, which are classified as developed, developing or underdeveloped according to the Human Development Report (HDR).

This index was rebuilt in 2010 and started to use a new method of calculation that is based on the calculation of three different variables. The first variable is the result of the equation obtained with (LE) life expectancy, which in Brazil is about eighty-three. The second variable is the result of Education Index (EI), which considers mean years of schooling (MYSI) and expected years of schooling (EYSI). The third variable considered the result of the equation obtained with the income index. This income is calculated based on the gross domestic product per person with parity by purchasing power, indexed by the dollar, the calculated location. Finally, in possession of these three variables, HDI is calculated as the arithmetic average of the values obtained.

This domain will therefore make use of the HDI indicator to measure the quality of municipal health evaluated because it is an indicator of international reach and used for both municipalities and countries. The papers of [P21,P12,P13] make use of this same indicator in their work, and in Brazil, these [DS04,DS6,DS24,DS31] are the sources of data from which the data can be mined.

3.4.3. (C) Education – HDI Education (I02DS01T01D02A02)

From the social point of view, Education can be seen as responsible to increase many other indicators, therefore, to the extent that a society becomes more educated, it also becomes more healthy and safe.

In Brazil, it has been several sources of data [DS04,DS26,DS9,DS6], of which the MEC (Ministry of Education) (MEC, 2015) uses different quality measuring instruments of education depending on the educational level that need to be measured. One is the IDEB (Basic Education Index) was created in 2007 is responsible for providing data on the quality of basic education. The index is measured every two years and the aim is that the country, from the reach of state and local targets will achieve a grade equal to 6, which corresponds to the quality of basic education in developing countries (IDEB, 2015).

Like other indicators from this work, the IDEB was chosen because it has similar tools used in other countries [P10,P15,P17,P19], thus allow comparisons with the targets achieved by other cities and countries around the world.

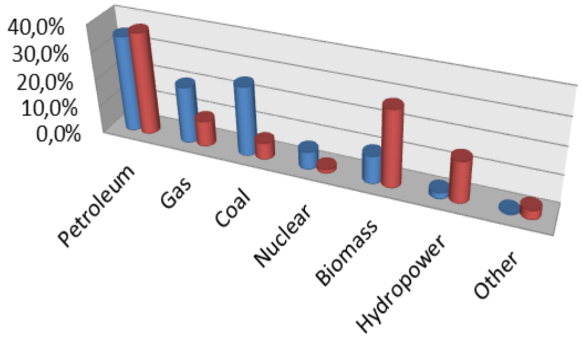
3.4.4. (D) Energy – Access to energy (I03DS01DS03T01D03A01)

The papers [P03,P28,P16,P18] shows the primary energy supply as main Smart Cities indicators when they have ways to manage their resources and optimize their use. Initiatives such as the SGMM (Smart Grid Maturity Model) developed by SEI (Carnegie Mellon University's Software Engineering Institute) (SEI, 2015) points to a model that consist eight domains, which contains incremental indicators with intelligent network features that represent the strategic aspects of the organization, implementation and operation of these networks.

In Brazil, the Ministry of Mines and Energy (MME, 2015) is responsible for managing the data on the electricity distribution services in cities. In a survey conducted by the ministry, was explicit the need in country to seek new forms of renewable energy production to balance its energy matrix. This type of national survey can be used locally as a good indicator for growth and financial and political municipalities involved in energy generation projects.

The comparison of the Brazilian energy matrix with the world can be seen in Figure 3.7 and reveals a large dependency on hydropower. As described in the water domain, Brazil is going through a water crisis that has direct impact on power generation.

Figure 3.7: Comparison of the energy mix of Brazilian and global cities.



	Petroleum	Gas	Coal	Nuclear	Biomass	Hydropower	Other
🟦 World	35,0%	20,7%	25,3%	6,3%	10,0%	2,2%	0,5%
🟥 Brazil	37,4%	9,3%	6,0%	1,4%	27,8%	14,9%	3,2%

Source: (MME, 2015)

The indicator used for this work considers the percentage of households served by the distribution of electricity in the city. It is important to remember that the population density in some regions of Brazil is very low, given the characteristics of terrain and vegetation. The data sources used for this work can be mined in: [DS04,DS28].

3.4.5. (E)Governance–HDIIncome/Employment(I04DS01T01D04A02)

The models proposed by [P13,P10,P12] use the Governance among the policies for the definition of Smart City. With different definitions and indicators [DS05,DS22,DS25], Governance can be summarized as a set of processes, policies, laws, regulations and institutions that regulate the way that public resources and services are managed.

One way to measure the quality of governance of cities is to measure its gross domestic product (GDP) and thus know the economic activity of a region. GDP is the sum (in monetary terms) of all finished goods and services produced during a given period.

Models that aims to define Smart Cities use management and policy issues to assess the governance capacity of a municipality, so if a municipality does not have its GDP growth, this may be a clear indicator that action is needed to resumption of growth.

The consulting firm PricewaterhouseCoopers [P12] developed a ranking of the 100 cities and / or its richest metropolitan areas in the world by GDP. From this list, Brazil has five cities: São Paulo, Rio de Janeiro, Brasilia, Porto Alegre and Belo Horizonte respectively occupying the positions: 10, 31, 57, 89 and 91 of this ranking. Previous studies show that cities such as Porto Alegre, Rio de Janeiro, Recife and Brasilia seek to achieve targets for building smarter cities (Macadar, 2013), (Brito et al, 2014).

However, the production capacity of a municipality is tied directly to their ability to manage and optimize their productive resources. Thus, this work is not limited to only measure the nominal GDP, but considers its growth over previous periods, favoring this way also the smaller municipalities.

3.4.6. (F) Housing – Private residence (I05DS01T01D05A01)

The Housing Domain was set from a basic indicator commonly used by local governments: the own homes index. In Brazil, this index is measured by the IBGE and published by sites such as IPEA (IPEA, 2015) and the Portal MDGs (ODM, 2015).

Among the studies conducted municipalities [P02,P03,P15], smaller feature significanter facilities in the acquisition of own residence ranging from encouraging residential own credit policies until federal government grants to purchase homes. (CEF, 2015)

The results of a survey released by the Applied Economic Research Institute (IPEA) show a reduction of the housing deficit in the country. Based on the National Survey by Household Sample (PNAD-2012), the study shows that 10% of the total deficit of Brazilian households recorded in 2007 dropped to 8.53% in 2012, representing 5.24 million of homes (IPEA, 2015). The datasouces used from this indicator was: [DS04,DS33,DS6].

3.4.7. (G) Environment – Waste treatment (I06DS01T01D06A01)

Although the term Smart Cities have multiple definitions, it is almost a consensus that is included the related domain to the environment. [P07,P06,P15]

One way to measure the impact of the cities on the environment is to assess whether the city has mechanisms to neutralize the production of damaging effects on nature.

According to the Ministry of the Environment (MMA, 2015) the municipalities are responsible for the daily production of approximately two hundred thousand tons of waste per day, totaling seventy two million per year of household waste.

To compose this indicator was considered the percentage of households served by the collection and treatment of household waste service [DS14]. The percentage found in both large and small cities almost in its entirety rates reach more than ninety-five percent of waste collected and treated.

3.4.8. (H) Security – Homicides per thousand (I08DS01T01D08A02)

By choosing the Security indicator was determined by calculating the number of deaths per thousand inhabitants, this indicator is called the Homicide risk. According to WHO (World Health Organization) studies in Homicide risk can be classified by age, gender or race. (FBS, 2015), (OMS, 2015).

The Homicide risk is an index used specifically to measure violence in the cities. The number is collected dividing the deaths caused by third parties by the population of the studied area; afterwards they made their equivalence per 100 thousand inhabitants.

Among the 50 most violent cities in the world, 16 are held in Brazil (Table 3.4), according to the ranking made by experts from non-governmental Mexican organization Citizen Council (CCSPJP, 2015).

Safety based on the figures of world cities in homicides over 300 thousand inhabitants. In spite of the northeast region of Brazil concentrates small towns whose indices are well above the national average which demand urgently control violence policies.

Table 3.3: Ten most violent cities in Brazil.

Position	City	Index
5	Maceió	79,76
7	Fortaleza	72,81
9	João Pessoa	66,92
12	Natal	57,62
13	Salvador	57,51
14	Vitória	57,39
15	São Luís	57,04
16	Belém	48,23
25	Campina Grande	46,00
28	Goiânia	44,56

Source: (CCSPJP, 2015)

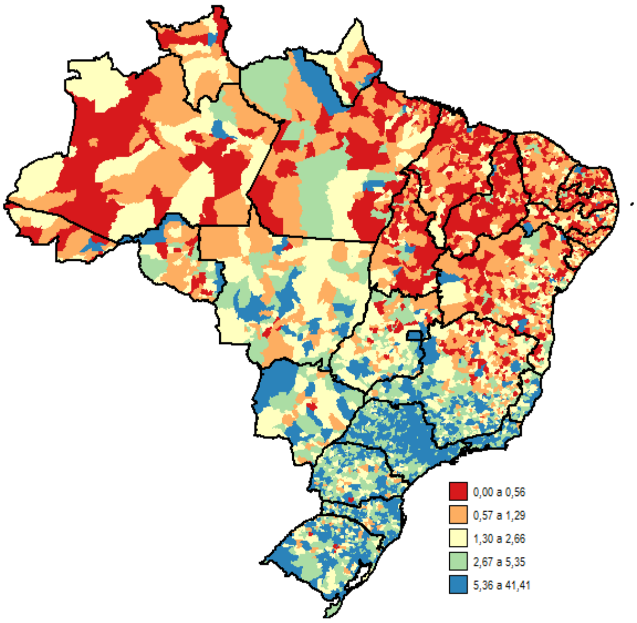
In Brazil there is a significant difficulty in obtaining these data on safety, because there is no institution to centralize state data. Thus, each state board (in the country are 27 departments) they responsible for collect and disseminate data on public safety. Thus, each State Security Bureau determines how and when to make the data available [DS04,DS30,DS32], hence there is a clear need to establish a Big Data with centralized information about public safety.

3.4.9. (I) Technology - Computers at home (I09DS01T01D09A01)

The work of [P10] and (Giffinger, 2007, 2010) considers technology as a key factor for the development of Smart Cities. Both point to technology as necessary domain, to act integrating the other domains and also to achieve the expected results for smart cities.

A city can be considered technologically advanced when it makes use of computational in order to improve their processes and manage their resources optimally. Make use of technology will create a better environment to the citizens enabling them to become part of the processes and monitoring the optimization resources. To compose the domain score studies consider the number of households with computer. The indicator described above is most recent one was created ten years ago [P02,P15,P20].

Figure 3.8: Percentage of households with computer in Brazil



Source: Made by author

Apparently the north and northeast regions of Brazil have a technological deficit in relation to the south and southeast regions, as related to in Figure 3.8, which have attracted many digital inclusion projects in these regions.

3.4.10. (J) Mobility - Mass transportation (I10DS01T01D10A01)

Urban mobility of cities is generally associated with the ability of production flow and mass public transport supply [P03,P06,P19]. Thus, it appears that the most developed municipalities also have the best transport indicators.

According to the Brazilian Ministry of Transport (MT, 2015), road transport prevailing in the country, which has 1.03 kilometers of paved road per capita and 7.35 km of unpaved road.

The Midwest region stands out in this indicator, having, respectively, 1.74 and 13.85. Mato Grosso do Sul state has significantest indicator for both paved roads and for unpaved: 2.56 and 33.18 respectively, because it has low density housing with a vigorous economy. Among the regions, the highlight was the South Region, which has 1.46 and 10.68 respectively.

In addition to the road indicators, it is necessary to measure the municipality's ability to manage the daily mass transportation. The National Land Transportation Agency (ANTT) [26] is the competent body for the award and monitoring of permits and authorizations for the public transport service operation. For this indicator, the data made available by this agency [DS15, DS29] confirms that road transport by bus is the main mode in the collective movement of users in Brazilian cities were used.

3.5 Application of the Taxonomy

To apply the developed taxonomy, have been followed three steps, as follows: (**1) Selecting data sources, (2) Normalizing data** and **(3) Defining and visualizing Indicators**. The following subsections detail these steps:

3.5.1. Selecting data sources

The indicators for each domain follow two simple criteria: the availability of public data for measurement and the comparability of these local data with the same set of data collected in other cities around the world. Based on public data collection work, created SLR (Chapter 2) and scientific papers.

After defining the ten domains and indicators, a search for public data sources to the process of mining of these data was performed. This process of data mining was done manually and should in future make use of tools to automate the search data.

Although these data are statistical and are released at intervals of two years, yet require higher speed in obtaining supplies.

3.5.2. Normalizing data

The available indicators to measure domains specified for Smart Cities have different methods of calculation. Thus, it was necessary to standardize the data to make it possible to compare them within their respective domains. Normalization was necessary to keep the data the same order of magnitude. (Equation 3.1).

$$X_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad (3.1)$$

Where Xi is each of the indicators measured in the standard range (x = 0) to the maximum value (xmax = 5). In statistics, the standard score is the represented by a number of standard deviations an observation or datum is above the mean.

Thus, a positive standard score indicates a datum above the mean, while a negative standard score indicates a datum below the mean. It is a dimensionless quantity obtained by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. This conversion process is called standardizing or normalizing (however, "normalizing" can refer to many types of ratios; see normalization (statistics) for more). (Kreyszig, 1979)

Standard scores are also called z-scores, where the use of "Z" is because the normal distribution is also known as the "Z distribution". They are most frequently used to compare a sample to a standard normal deviate, though they can be defined without assumptions of normality.

In this work normalize the data transforming their values in a z-score notation, and then turn it into a T-score for the indicator values conform to a range of values from 0 to 5.

3.5.3. Defining and Visualizing Indicators

Among the ten predefined domains in the previous section, this study used the mined indicators in public databases **(A) Education, (B) Health** and **(C) Security** to exemplify the use of the taxonomy proposed. Are represented in Figure 3.9 the indicators, their respective domains and standardized data to enable a comparison between these indicators to all Brazilian capitals.

The simplest way to reproduce this experiment in which cities are compared is:

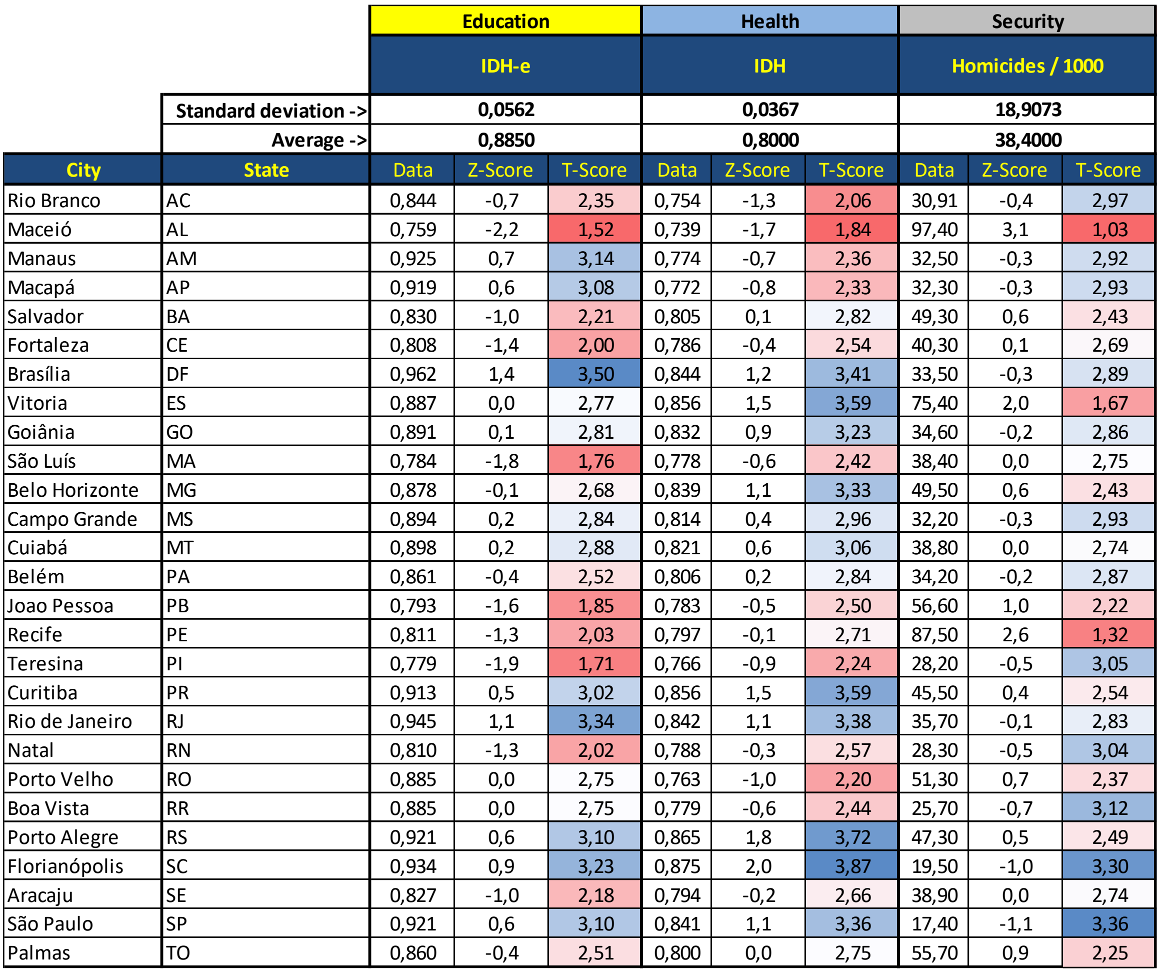
A) Define which indicators (IDH, Homicides, etc) to compare;

B) Search for public databases that have these indicators;

C) Perform mining, normalization (z-score) and standardization (z-score);

D) Generate text files with standardized data for graphing tools (R-Studio).

Figure 3.9: Heatmap of domains:
(A) Education, (B) Health and (C) Security.



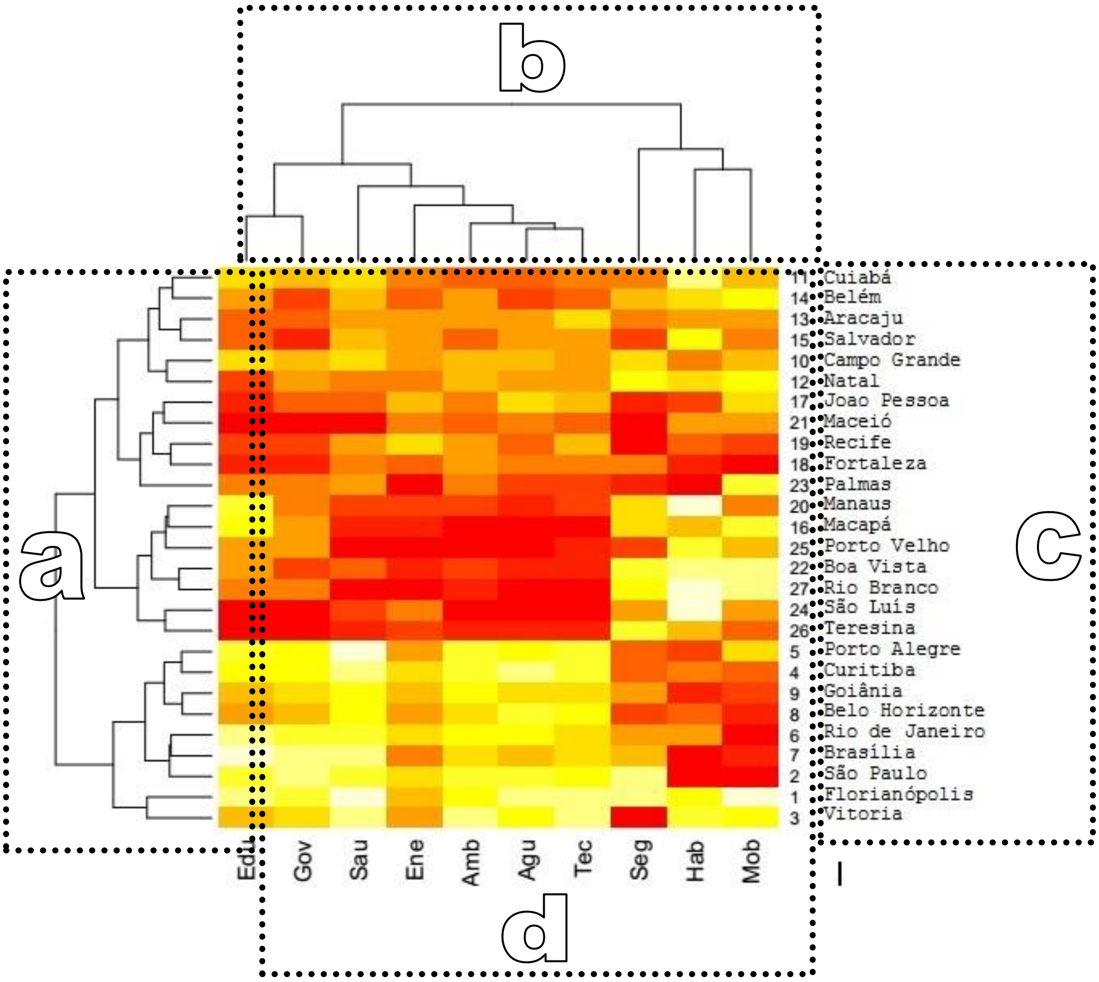
		Education			Health			Security		
		IDH-e			IDH			Homicides / 1000		
	Standard deviation ->	0,0562			0,0367			18,9073		
	Average ->	0,8850			0,8000			38,4000		
City	State	Data	Z-Score	T-Score	Data	Z-Score	T-Score	Data	Z-Score	T-Score
Rio Branco	AC	0,844	-0,7	2,35	0,754	-1,3	2,06	30,91	-0,4	2,97
Maceió	AL	0,759	-2,2	1,52	0,739	-1,7	1,84	97,40	3,1	1,03
Manaus	AM	0,925	0,7	3,14	0,774	-0,7	2,36	32,50	-0,3	2,92
Macapá	AP	0,919	0,6	3,08	0,772	-0,8	2,33	32,30	-0,3	2,93
Salvador	BA	0,830	-1,0	2,21	0,805	0,1	2,82	49,30	0,6	2,43
Fortaleza	CE	0,808	-1,4	2,00	0,786	-0,4	2,54	40,30	0,1	2,69
Brasília	DF	0,962	1,4	3,50	0,844	1,2	3,41	33,50	-0,3	2,89
Vitoria	ES	0,887	0,0	2,77	0,856	1,5	3,59	75,40	2,0	1,67
Goiânia	GO	0,891	0,1	2,81	0,832	0,9	3,23	34,60	-0,2	2,86
São Luís	MA	0,784	-1,8	1,76	0,778	-0,6	2,42	38,40	0,0	2,75
Belo Horizonte	MG	0,878	-0,1	2,68	0,839	1,1	3,33	49,50	0,6	2,43
Campo Grande	MS	0,894	0,2	2,84	0,814	0,4	2,96	32,20	-0,3	2,93
Cuiabá	MT	0,898	0,2	2,88	0,821	0,6	3,06	38,80	0,0	2,74
Belém	PA	0,861	-0,4	2,52	0,806	0,2	2,84	34,20	-0,2	2,87
Joao Pessoa	PB	0,793	-1,6	1,85	0,783	-0,5	2,50	56,60	1,0	2,22
Recife	PE	0,811	-1,3	2,03	0,797	-0,1	2,71	87,50	2,6	1,32
Teresina	PI	0,779	-1,9	1,71	0,766	-0,9	2,24	28,20	-0,5	3,05
Curitiba	PR	0,913	0,5	3,02	0,856	1,5	3,59	45,50	0,4	2,54
Rio de Janeiro	RJ	0,945	1,1	3,34	0,842	1,1	3,38	35,70	-0,1	2,83
Natal	RN	0,810	-1,3	2,02	0,788	-0,3	2,57	28,30	-0,5	3,04
Porto Velho	RO	0,885	0,0	2,75	0,763	-1,0	2,20	51,30	0,7	2,37
Boa Vista	RR	0,885	0,0	2,75	0,779	-0,6	2,44	25,70	-0,7	3,12
Porto Alegre	RS	0,921	0,6	3,10	0,865	1,8	3,72	47,30	0,5	2,49
Florianópolis	SC	0,934	0,9	3,23	0,875	2,0	3,87	19,50	-1,0	3,30
Aracaju	SE	0,827	-1,0	2,18	0,794	-0,2	2,66	38,90	0,0	2,74
São Paulo	SP	0,921	0,6	3,10	0,841	1,1	3,36	17,40	-1,1	3,36
Palmas	TO	0,860	-0,4	2,51	0,800	0,0	2,75	55,70	0,9	2,25

Source: Made by author

After the choice and definition of the indicators used, the next step is to transform these data into information, it is the process of consolidation of a large amount of data in a graphical simplification for better understanding of the scenario presented.

The use of spreadsheets with heatmaps is one of a data visualization example and another example is the use of dendrograms (Figure 3.10), which group these heatmaps similarity indicators, thus making didactically more understandable information.

Figure 3.10: Dendrogram of Brazilian Capitals



Source: Made by author

To read and analyze the information in this chart, this dendrogram is divided into 4 parts, which are:

a. **Cluster of Brazilian capitals (by similarity)**: the capitals were grouped considering the similarity between the data of the ten indicators of each capital.

b. **Cluster of indicators**: grouped indicators represent the strongest colors (red) the smallest values, and lighter colors (yellow) represent better grades.

c. **List of Brazilian capitals**: cities were listed according to the similarity and clustering of data of the indicators.

d. **List with the ten indicators:** set of data used to cluster capitals.

The next section briefly summarizes this chapter.

3.6 Summary

This chapter introduces the definition of a taxonomy of indicators for smart cities based on the Brazilian reality. This taxonomy was formally implemented by building an ontology domain and data visualization was made through the selection of public data and the construction of heatmaps and dendrograms.

The main contribution of this chapter to this doctoral thesis was to serve as a basis for the collection and systematization of data on Brazilian cities, so that in future such data can be cataloged and compared generating graphical information for decision-making by public administrators.

4

SMARTCLUSTER AN ONTOLOGY-BASED METAMODEL

Sometimes the only way to gain the respect of your superior is to challenge him.

— Frank Underwood, House of Cards

Based on the previous chapter, where a taxonomy was proposed to measure and compare Brazilian Intelligent Cities, this chapter presents a study about the current SCM, and how a metamodel can encompass the characteristics of these existing models. The objective of this chapter is to present the metamodel called SmartCluster, its characteristics and how it can be reused and expanded to meet the creation of new Brazilian and global SCM (SCM).

4.1 Introduction

In the last two decades, with the emergence of model-oriented software engineering techniques in which data, models and mappings between models are represented as data, metadata plays an important role and becomes central concepts manipulated by these techniques. They are used in the definition of computer-based solutions for interoperability, data exchange, software reuse, model transformation, and so on. However, in the case of SCM, no metamodel repositories have been proposed and there is no attempt in the literature to combine their advantages into a single repository, or vocabulary.

The Ontology Definition Metamodel is an Object Management Group (OMG) specification to make the concepts of Model-Driven Architecture applicable to the engineering of ontologies. Hence, it links Common Logic (CL)[1], the Web Ontology Language (OWL)[2], and the Resource Description Framework (RDF)[3].

OWL and RDF were initially defined to provide an XML-based machine to machine interchange of metadata and semantics. ODM now integrates these into visual modeling, giving a standard well-defined process for modeling the ontology, as well as, allowing for interoperability with other modeling based on languages like UML[4], SysML[5] and UPDM[6].

To achieve the process of building common vocabularies that unify the currently fractured state of SCM representations it was used ontologies, defined as a formal explicit specification of a shared conceptualization within a domain of interest, to implement shared vocabularies that help represent and organize the SCM (Studder, 1998). Specifically, the ontology implementation technology use Web Ontology Language (OWL (McGuinness, 2017) and Semantic Web Technologies (Miller, 2017) OWL-DL is used to formally define patterns and relationships between patterns.

To achieve this goal, this chapter is divided as follows: This Section 4.1 presents the introduction to the subject and structure of this chapter; Section 4.2 presents an ontology-based approach to intelligent city models; In Section 4.3 is defined the ontology-based methodology for SmartCluster; Section 4.4 presents the necessary steps for the construction of SCM ontologies; And this chapter ends with section 4.4.

[1] **Common Logic (CL):** avaliable at https://goo.gl/uR9tGT

[2] **Web Ontology Language (OWL):** available at https://goo.gl/DtZDkQ

[3] **Resource Description Framework (RDF):** available at https://goo.gl/JJesXv

[4] **UML** Available at https://goo.gl/EMeeVk

[5] **SysML** available at https://goo.gl/LVzTbP

[6] **UPDM** available at https://goo.gl/E3slyo

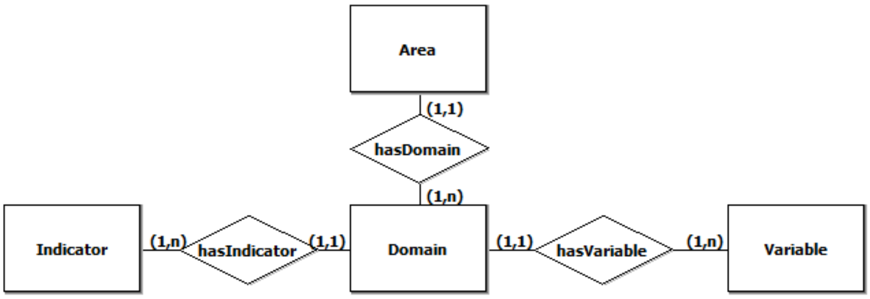
4.2. Ontology-Based Approaches to Smart Cities Models

The use of ontologies to represent SCM is the combination of two methodologies that can be complementary. One goal of SCM is to provide the means for public managers and agents to use a common vocabulary about indicators, domains, and areas of smart cities. The Semantic Web and its technologies, such as OWL, implement techniques to formally define ontologies through shared vocabularies, axiomatic definitions, and formal logic to further support machine-based automated thinking (Henninger, 2006). The importance of the use of ontologies as support for applications to smart cities is also described in other papers. (Komninos, 2016)

4.2.1 OWL Description Logics

The Web Ontology Language (OWL) relies on the Resource Description Framework (RDF) and RDF Schema to create a framework-based language representation language with axiomatic constructs for logic-based expressivity. OWL includes vocabulary for describing properties and classes. OWL constructs allow the construction of class taxonomies and properties act as predicates representing a triple RDF between two classes. Figure 4.1 presents the main elements of this ontology model as a simplified UML model, with their respective entities, relationships and cardinalities.

Figure 4.1: Class and relationships of SmartCluster Metamodel



Source: Made by author

OWL properties are predicates that operate on subjects (domains) and map to an object (range). Range values can be constrained through several axiomatic class construction operators (Figure 4.1). For example, if it is necessary to state that the "Area" superclass is the set of default instances that are a subclass of a Domain pattern and all values of the "hasIndicators" property come from the "Indicators" and "Variable" classes, so would be declared:

$$\textbf{‘Area’} \sqsubseteq \textbf{‘Domain’} \sqcap \forall \textbf{hasIndicators.(‘Indicators’} \sqcup \textbf{‘Variable’)}$$

OWL relies on this infrastructure with richer types of properties, relationships between classes, class constructors, enumerated classes, cardinality, equality and characteristics of qualified cardinalities, and general axiomatic definitions of class members by means of complex expressions.

4.2.2 OWL and Smart Cities Models (SCM)

Using OWL DL can use the resources that the Semantic Web provides in creating an infrastructure capable of inferring automated reasoning for SCM:

- **Distributed representations**: Since OWL is built on RDF and XML (Klein, 2001), Uniform Resource Indicators (URIs) are used to support common vocabulary in distributed files. A URI defines a unique namespace (using the same syntax as URLs) and the concept within the namespace, thus ensuring that two patterns using the same URI are referring to the same OWL element. This is important to prevent namespaces from having duplicate references to concepts to which the represented city models refer.
- **Well-defined semantic:** The description logic defined by OWL-DL allows precise definitions of concepts, as described in the previous item. Patterns should, in this case, allow understanding by both humans and machines. It is hoped that in the future, indicators can be read and processed autonomously by devices connected to public databases.
- **Rule-based search and semantics:** In addition to specifying pattern attributes and relationships, rules can be used to accurately specify matching criteria. Smart search based on semantic relationships is also possible, increasing the ability to find patterns that meet the needs of city managers. In the SCM examples presented in this chapter, some search for indicators

will be related, and how semantics can contribute to the understanding of these concepts.

- **Heterogeneous representations**: A current problem with SCMs is that they have too many sets of attributes, descriptions, and terminology used to describe them. The idea of designing SmartCluster is to facilitate more homogeneous representations, using the OWL axioms, which provide a number of constructions to establish the equivalence of concepts and properties.

4.3. Ontology-Based Methodology for SmartCluster

To support the evolution of metamodels, the OMG proposal consists of introducing a fourth higher level: the meta-metamodel. Thus, new metamodels are created as instances of this higher level. The main objective of this Section is the integration of SCM variants through the use of a general ontology (SmartCluster) and the representation of models expressed with these variants in terms of the ontology.

The goal of a metamodel is to define the basic building blocks and rules for building well-formed models within some domain of interest (OMG, 2002). A metamodel for SCM will therefore provide the basic building blocks for creating domains, areas, and indicators for comparison and measurement.

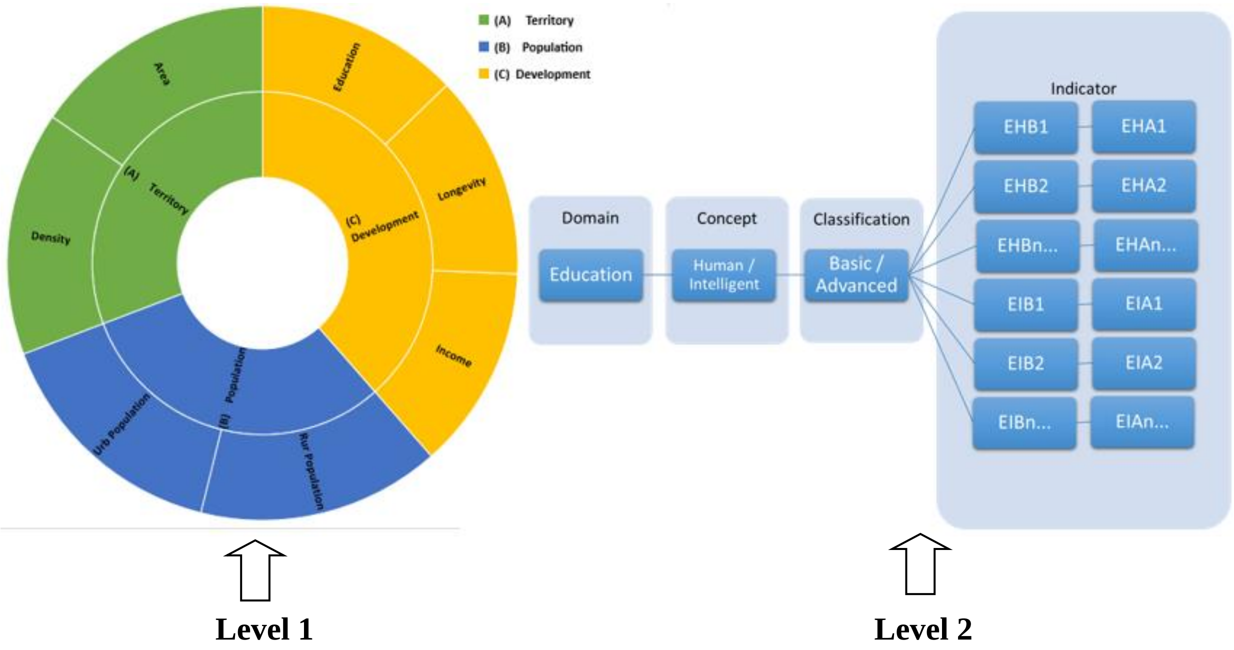
Thus, to allow the understanding and integration of these models, the methodology proposed by Karen (Najera, 2013) is divided into four specific objectives::

1. The development of an ontology called SmartCluster that represents the construction of a Metamodel (umbrella) that contains all the structures of the other existing models;
2. The development of a methodology to integrate SCM constructs through the use of ontologies from the SmartCluster ontology;
3. The application of the methodology to the different models with their concepts and indicators, thus generating an ontology with these concepts and integrated indicators and
4. The transformation of a model represented with one of the variations of models integrated in an ontology derived from the concepts of the ontology with the integrated models.

4.3.1. SmartCluster: Levels of Ontology (Metrics and Domains)

In this section will be presented the ontology that serves as the basis for the construction of SCM based on the fusion of preexisting models. The ontology called SmartCuster is based on a set of indicators and variables (Figure 4.2) classified into two levels of comparison of values, being: Level 1 (Metrics for city comparison) and Level 2 (Smart City domains).

Figure 4.2: Levels of SmartCluster Metamodel



Source: Made by author

Level 1 (Figure 4.2) is divided into three Metrics and their respective variables: Territory (Area, Density), Population (Urban, Rural) and Development (Income, Longevity, Education). These metrics are basic and serve to create a step earlier than the comparison of smart cities, clustering cities by their similar characteristics.

In Level 2, after grouping the most geographically similar cities, it is time to classify the cities according to the domains and areas of smart cities. At this level each indicator can be classified as "basic" or "advanced" to become SCM compliant using the concept of two-level indicators.

The SmartCluster metamodel differentiates between variables and indicators, as variables are used to compare cities, and indicators to measure and compare smart cities. Basically the basis of this metamodel is divided into two levels, one of which is responsible for comparing statistical data on cities and the second for intelligence indicators.

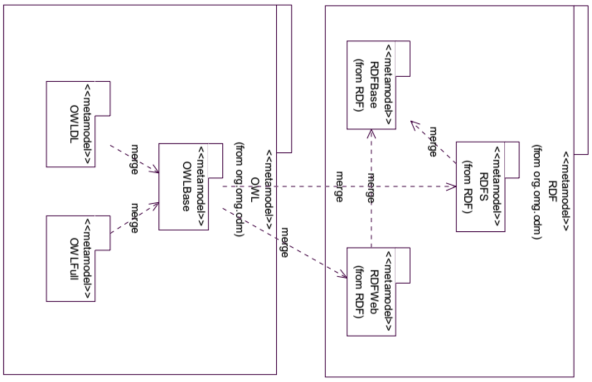
4.3.2. SmartCluster: levels of MOF (Meta Object Facility)

To develop the SmartCluster metamodel, it was used the Model Driven Architecture (MDA) paradigm that was created by OMG (Object Management Group) in 2000, according to which models can be used to create software.

Its process provides a similar life cycle to the conventional one. In the analysis phase, are create models that can be quickly interpreted by a computer program (for example, an ontology in OWL).

Figure 4.3 shows the integration of the metamodel architecture in RDF and OWL that allows the construction of ontology models with metadata storage.

Figure 4.3: SmartCluster architecture in RDF


<<metamodel>> OWLDL
<<metamodel>> OWLFull
merge
merge
<<metamodel>> OWLBase
<<metamodel>> OWL (from org.omg.odm)
merge
merge
merge
merge
<<metamodel>> RDFBase (from RDF)
<<metamodel>> RDFS (from RDF)
<<metamodel>> RDFWeb (from RDF)
<<metamodel>> RDF (from org.omg.odm)

Source: Made by author

The development of metamodels foresees the construction of different levels, which are specified by the MOF (Meta Object Facility), that is, an abstract language and a framework for specification, construction and management of independent metamodels of implementation technology. The MOF has a set of rules for implementing repositories, which manipulate metadata described by metamodels.

Table 4.1 shows the four levels that characterize the MOF structure. At the level M3 establishes concepts that compose a language, such as class, attribute, association, among others, being thus called meta-level metamodel. At the level M2 establishes the language and its structural items, such as OWL (classes, relationships, instances, etc.), being called the metamodel level. The M1 level is the level of the model itself, such as a model of an intelligent city, which contains classes, such as the Energy, Water, or Transport classes. And the M0 level constitutes the instances of a model, such as the Transport (object) class instance "Bus", "Subway" or "Car".

Table 4.1: Levels of MOF (Meta Object Facility)

Level	Modeling Level	Examples
M3	Meta-metamodel	A MOF class, attribute, association, package, operation
M2	Metamodel	OWL Language, Class, Relationship, Attribute
M1	Model	City model contains a Class "Health"
M0	Object	An instance of "Health" such as "HDI"

Source: Made by author

The SmartCluster metamodel must satisfy two requirements: (1) allow modification of the metamodel level and (2) ensure that these modifications correspond to the semantics of the embedded models: an instance of a metamodel (level called M2 in the MOF) defines a valid model (M1) that are expected to represent population instances (M0). These two requirements are fulfilled as follows:

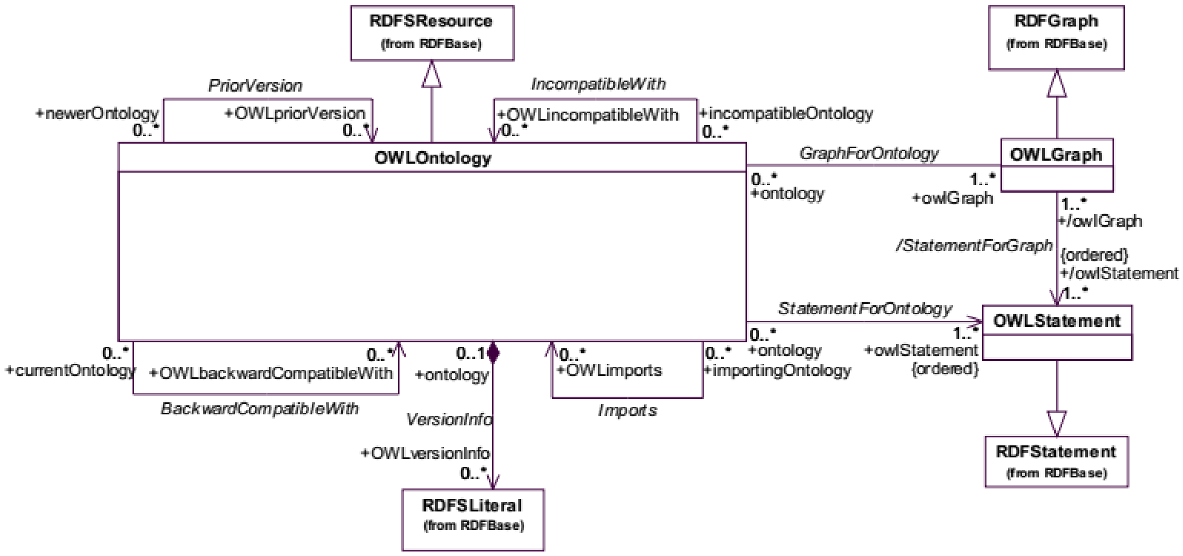
(1) The instantiation of this metamodel (Smart Cluster) elevate the level of the existing model, guaranteeing the flexibility and compatibility of the metamodel with different levels of models.

(2) The metamodel level consists of a predefined core model associated with an operational semantics. Is described this basic model in the next section and then show how it can be extended.

Following OMG recommendations, an OWL ontology should contain a sequence of annotations, axioms, and facts. Annotations on OWL ontologies are used to record authorship and other information associated with an ontology, including import references to other ontologies. In our work the annotations are important to understand how the SCM were assimilated by the SmartCluster ontology.

The main content of OWLOntology is realized in its axioms and facts, which provide information about classes, properties and individuals in the ontology. Ontology names are used in abstract syntax to convey the meaning associated with the publication of an ontology on the Web. The intention is that the name of an ontology in abstract syntax is the URI where it can be found, although this is not part of the formal meaning of OWL (Figure 4.4).

Figure 4.4: Metamodel Core



Source: Adapted from "The OWL Metamodel" (OMG, 2016)

The attributes of this metamodel are not presented here, but the associations are detailed below:

- **owlGraph**: OWLGraph [1..*] in association GraphForOntology links an ontology to one or more graphs containing the statements that define it. (e.g., most of the Figures presented in this thesis used this class to allow expressing the association between the classes of SCM)
- **currentOntology**: OWLOntology [0..*] in association BackwardCompatibleWith - links an ontology to zero or more other ontologies it has backwards compatibility with. (i.e., SCMs should point

to SmartCluster as their zero ontology, just as "Thing" is the zero ontology for SmartCluster)

- **OWLbackwardCompatibleWith**: OWLOntology [0..*] in association BackwardCompatibleWith - links an ontology to zero or more other ontologies it has backwards compatibility with.

- **importingOntology**: OWLOntology [0..*] in association Imports - links an ontology to zero or more other ontologies it imports.

- **OWLimports**: OWLOntology [0..*] in association Imports - links an ontology to zero or more other ontologies it imports

- **incompatibleOntology**: OWLOntology [0..*] This association creates links with the zero ontology (or higher class, or superclass) indicating that there is no compatibility with this class. In this case it is important to define when an SCM cannot be compatible with SmartCluster for semantic differences (and it is necessary to redefine SmartCluster).

- **OWLincompatibleWith**: OWLOntology [0..*] in association IncompatibleWith - links an ontology to zero or more other ontologies it is not compatible with. (e.g., the "Dimension" and "Area" classes of the SmartCluster ontology may create an incompatibility with the SCM if identical semantic concepts are not defined).

- **newerOntology**: OWLOntology [0..*] in association PriorVersion - links an ontology to zero or more other ontologies that are earlier versions of the current ontology. (i.e., allows the inclusion of new SCMs in the SmartCluster ontology)

- **OWLpriorVersion**: OWLOntology [0..*] in association PriorVersion - links an ontology to zero (“SmartCluster”) or more other ontologies that are earlier versions (“Thing”) of the current ontology.

- **OWLversionInfo**: RDFSLiteral [0..*] in association VersionInfo - links an ontology to an annotation providing version information. (e.g., “SmartCluster version 2.15”)

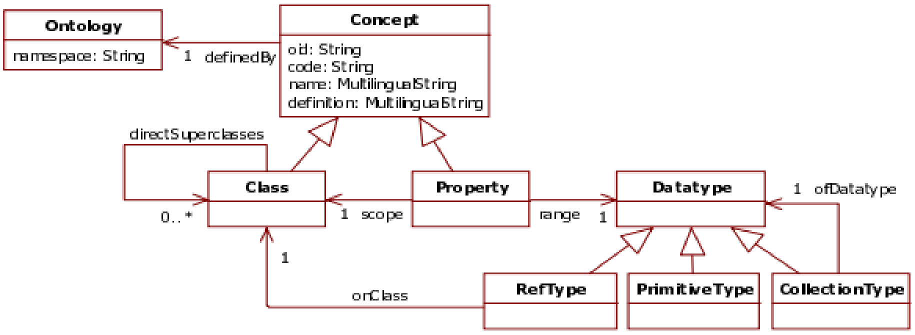
- **OwlStatement**: OWLStatement [1..*] in association StatementForOntology links an ontology to one or more ordered statements it contains. (i.e., an SCM may be out of date with the current number of indicators.)

In the next subsection will be presented the core metamodel of the SmartCluster model, its main components and how this ontology can be extended.

4.3.3. SmartCluster: Core Metamodel

Generally the metamodel corresponds to the ontological model used to define ontologies. In the SmartCluster metamodel was used the shared constructors of the main ontology models in the application domain (Smart Cities): RDFs and OWL (Figure 4.5).

Figure 4.5: Shared constructors of the main ontology models.



Source: Adapted from (Jean, 2010)

Figure 4.5 presents the main elements of this ontology model as a simplified UML model with its entities, classes and properties. The main elements of this model according to Stéphane (Jean, 2010) are the following:

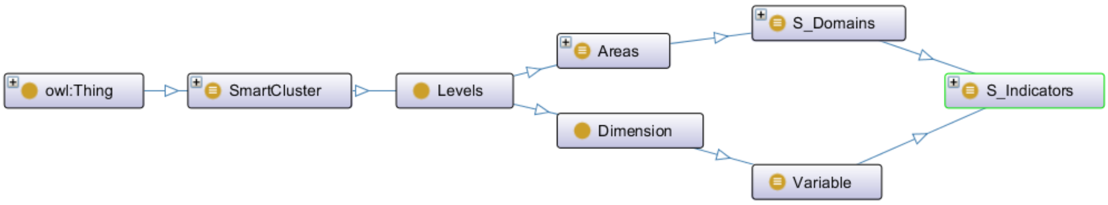
- An ontology (**Ontology**) introduces a unique namespace also called Namespace. Defines concepts that are classes and properties.
- A class (**Class**) is the abstract description of one or many similar objects. It has an underlying system specific identifier (oid) and an identifier independent of it (code).
- Properties (**Property**) describe instances of a class. As classes, properties have an identifier and a textual part. Each property must be

defined for the class of instances it describes (scope). Each property has an interval (range) to restrict its value domain.

- The Datatype of a property can be a simple type (**primitiveType**), such as integer or string. A property value can also be a reference to an instance of a class (**refType**). Finally, this value can be a collection whose elements are either of the simple type or type of reference (**CollectionType**).

This core metamodel contains all the specific characteristics of ontologies (namespaces, multilingualism, universal identifier). The most important feature of this work is that a Metamodel based on this architecture is able to manage ontologies and semantically describe other models that are integrated with it (Figure 4.6).

Figure 4.6: Core Ontology of SmartCluster



Source: Made by author

The next subsection will present the methodology used to develop the ontologies of intelligent city models, and how these ontologies will be integrated into the meta-model SmartCluster.

4.4. Development Methodology of ontology for SCM

The proposed methodology provides guidelines for integrating SCM ontology constructs into a more comprehensive ontology called SmartCluster. The ontology SmartCluster allows the assimilation of other ontologies to SCM, and is based on Model-Driven Engineering (MDE), respecting the specifications for creation of metamodels (OMG, 2002) and taking into account the ontology integration methodology proposed by Karen (Najera, 2013). The methodology proposed here consists of two phases: 1) the development of an ontology for each variation of SCM desired for integration and 2) the integration of the ontologies of the model variations generating the SmartCluster ontology.

Chapter 2 of this Doctoral Thesis presented a bibliographical survey on intelligent cities and their respective models. For this chapter it will be used the model of evaluation of Ranking of Smart Cities proposed by Rudolf (Giffinger, 2010), and that is based on Table 4.2.

Table 4.2: Dimensions for analysis of city rankings

Dimension	Analyzed indicator (examples)
Authorship and publication	Author(s) and sponsor(s) / Type of publishing
Data base	Time scale of used data Published source of data and/or raw data available Method of calculation of overall-ranking
Use of indicators	Number of indicators Method of calculation Use of standardized values
Spatial dimension	Size of city sample / Selection criteria for cities
Elaborateness of results	Overall-ranking; Results for selected topics and cities Results available for free/liable to pay costs

Source: Made by author

The use of this comparison and evaluation table allows us to affirm that most of the proposed models cite two authors and their works that evolved the concepts of ranking and comparison of intelligent cities (Chourabi, 2012 and Cohen, 2012), thus giving rise to A ranking of European cities (Giffinger, 2007) and currently a model of ISO standards for sustainable cities (ISO, 2014). With these four works it is possible to evaluate SCM from the concept of intelligent cities to the way in which these indicators will be mined and compared.

Chourabi (2012) proposed a framework (Integrative Framework) that uses eight factors (Management and Organization, Technology, Governance, Political Context, People and Communities, Economy, Built Infrastructure, Natural Environment) with 47 indicators called strategies.

The Smart Cities Wheel proposed by Cohen (2012) points out six key factors for the definition of Smart Cities, Smart Economy, Smart Government, Smart Living and Smart Mobility with 100 specific indicators.

Then, the Center of Regional Science (Giffinger, 2007) elaborated a ranking that is divided in six characteristics (Competitiveness, Social Aspects, Participation of the population in the making of decisions, Quality of life, Transport and Human resources) and these characteristics are based in the direct comparison of 90 indicators for medium-sized cities.

Currently ISO 37120 created by the International Standards Organization (ISO, 2014) was developed containing 17 themes, with 46 core indicators and 54 supporting indicators that can help in the definition of public policies.

The next sections will present the methodology used to develop the ontologies based on this model (Phase 1).

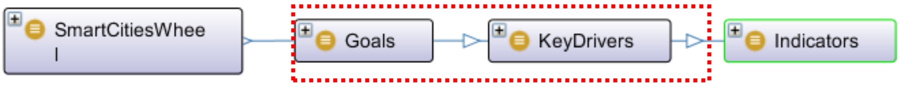
4.4.1. Phase 1) Development of an ontology for a specific Smart City Model

In this phase, the ontology for a specific model is generated and can be performed as many times as necessary until all domains, areas and indicators of this model are mapped to the ontology that should be compatible with SmartCluster. This phase has four steps:

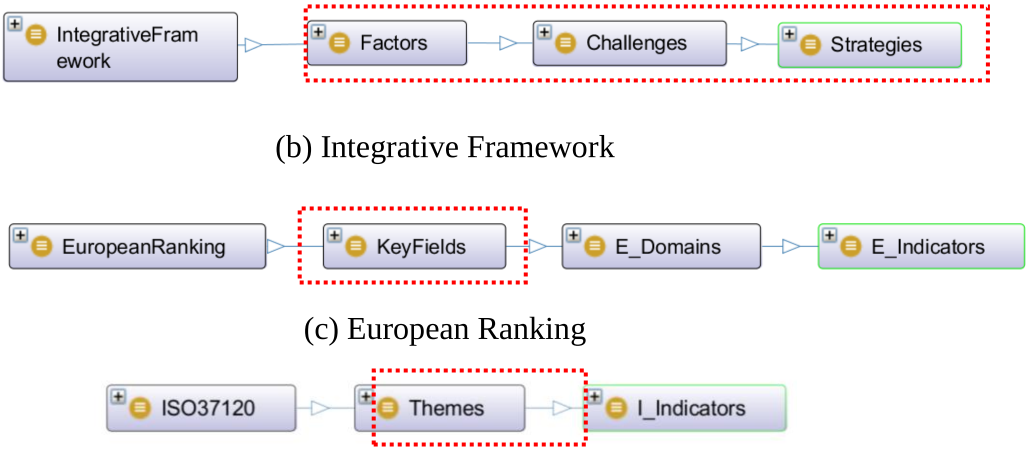
4.4.1.1. Identifying

In the first step, additional SCM constructs that are not part of the SmartCluster ontology are identified. A representation of the model created in UML can pass the idea of the evaluated construction and of which components are not present in the meta-metamodel. Figure 4.7 shows the four SCMs and their differences from the SmartCluster model. In addition to the UML model, these ontologies were created using the Protégé Tool (Protégé, 2001), which allows the creation of classes and relationships.

Figure 4.7: SCM and its differences for SmartCluster



(a) Smart Cities Wheel



(b) Integrative Framework

(c) European Ranking

(d) ISO 37120

Source: Made by author

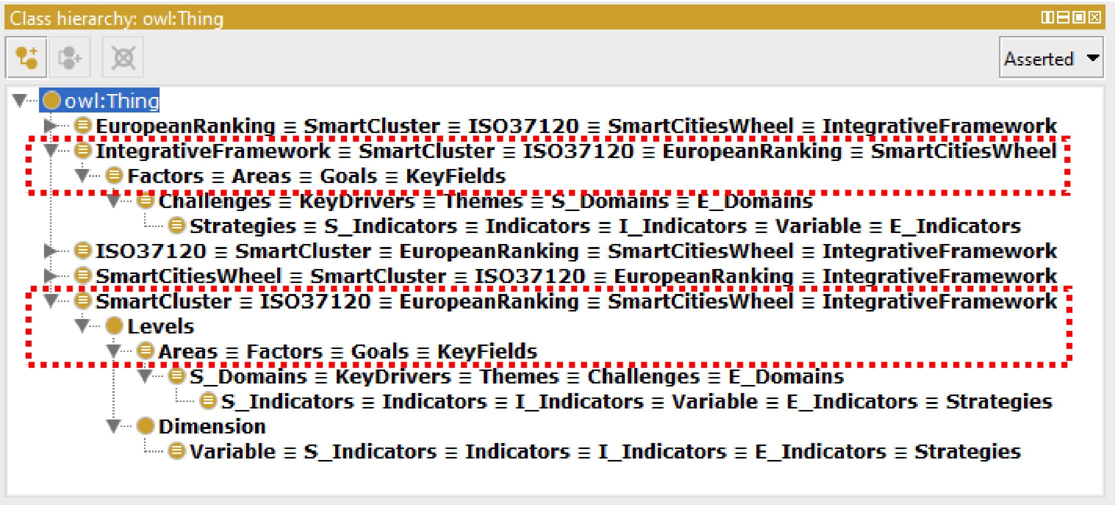
In Figure 4.7 it is possible to identify the semantic and conceptual differences between the models, which should be incorporated by the SmartCluster metamodel. In the model (a) the classes "Goals" and "KeyDrivers" do not exist in the SmartCluster metamodel, however the concepts are respectively compatible with the "Areas" and "S_Domains" classes of this metamodel. The other models that also present these conceptual and semantic differences were identified and their classes served as a basis for expanding SmartCluster semantics.

4.4.1.2. Categorizing

The second step categorizes the constructs and classifies them into: Concept (representation of something from the real world), Relation (relation of one or more concepts), Attribute (definition of property, value or characteristic of concepts or instances.) and Attribute value (Values or indicators categorized as attribute).

Figure 4.8 presents a view of the ontology development tool, where it is possible to identify two ontologies under development ("IntegrativeFramework" and "SmartCluster"). When developing the ontology it is possible to define the relation of equality between classes, and thus, to ensure compatibility between the SCM and the proposed metamodel. Specifically in this figure the "Factor" class of the "IntegrativeFramework" ontology and the equality relationship with the "Areas" class of the "SmartCluster" ontology are presented.

Figure 4.8: SCMs: “IntegrativeFramework” and “SmartCluster



Class hierarchy: owl:Thing

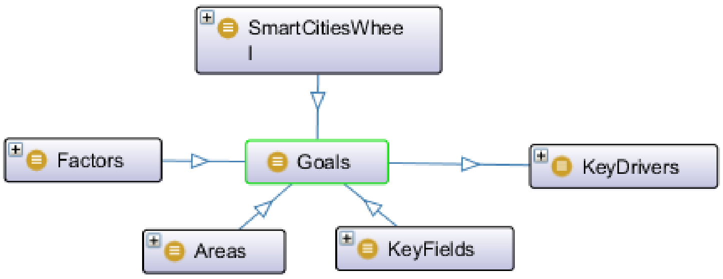
Asserted

```
owl:Thing
  EuropeanRanking ≡ SmartCluster ≡ ISO37120 ≡ SmartCitiesWheel ≡ IntegrativeFramework
  IntegrativeFramework ≡ SmartCluster ≡ ISO37120 ≡ EuropeanRanking ≡ SmartCitiesWheel
    Factors ≡ Areas ≡ Goals ≡ KeyFields
      Challenges ≡ KeyDrivers ≡ Themes ≡ S_Domains ≡ E_Domains
        Strategies ≡ S_Indicators ≡ Indicators ≡ I_Indicators ≡ Variable ≡ E_Indicators
  ISO37120 ≡ SmartCluster ≡ EuropeanRanking ≡ SmartCitiesWheel ≡ IntegrativeFramework
  SmartCitiesWheel ≡ SmartCluster ≡ ISO37120 ≡ EuropeanRanking ≡ IntegrativeFramework
  SmartCluster ≡ ISO37120 ≡ EuropeanRanking ≡ SmartCitiesWheel ≡ IntegrativeFramework
    Levels
      Areas ≡ Factors ≡ Goals ≡ KeyFields
        S_Domains ≡ KeyDrivers ≡ Themes ≡ Challenges ≡ E_Domains
          S_Indicators ≡ Indicators ≡ I_Indicators ≡ Variable ≡ E_Indicators ≡ Strategies
      Dimension
        Variable ≡ S_Indicators ≡ Indicators ≡ I_Indicators ≡ E_Indicators ≡ Strategies
```

Source: Made by author

The visualization of this equality relationship between classes of different models can be performed through a plugin called OntoGraf, which allows to verify the relationships between similar classes: "Goals", "Factors", "Areas", "KeyFileds" and "KeyDrivers "(Figure 4.9).

Figure 4.9: Relationship between different SCM classes



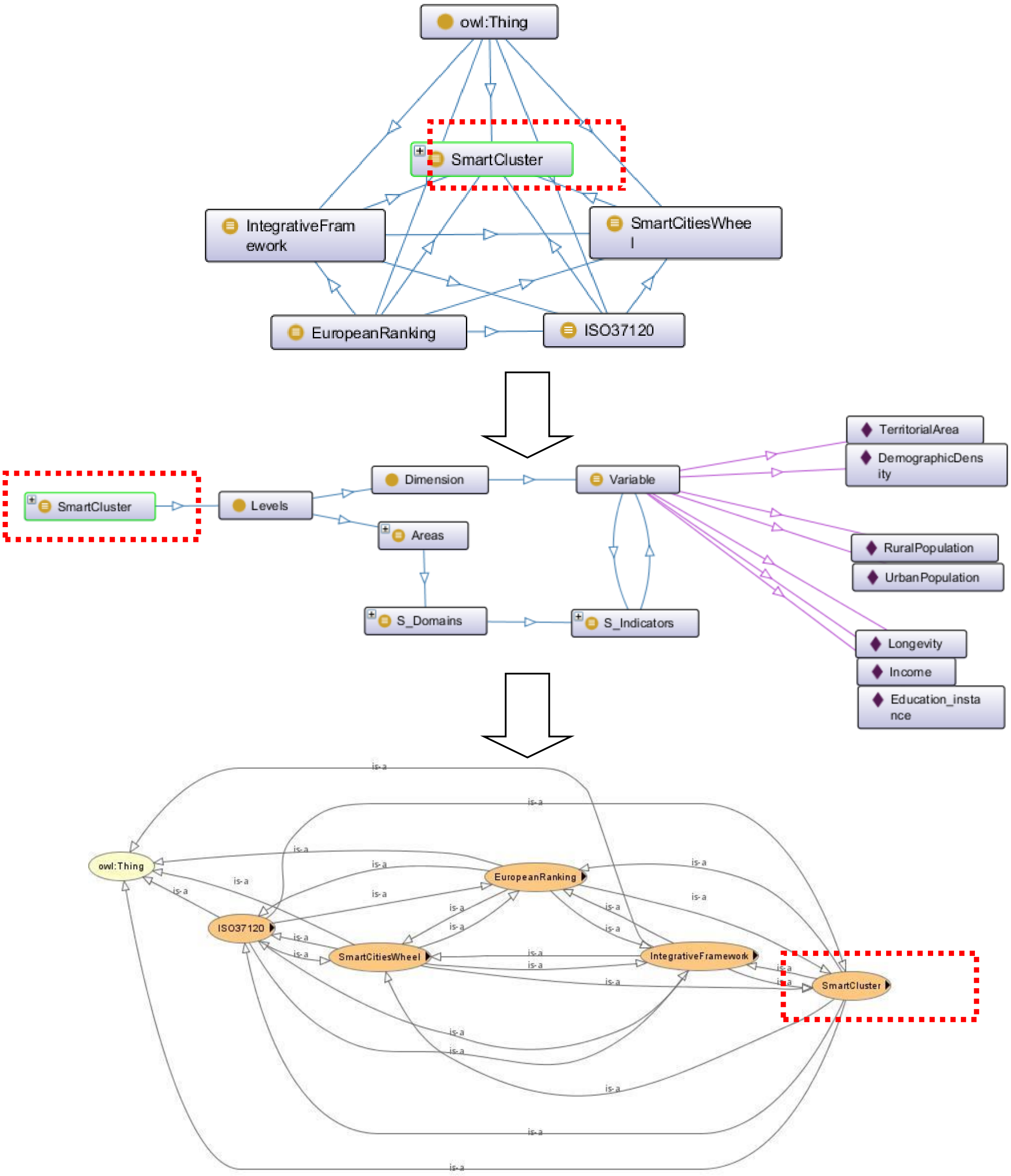
Source: Made by author

The next subsection presents the ontology itself, with the SCMs incorporated.

4.4.1.3. Transforming

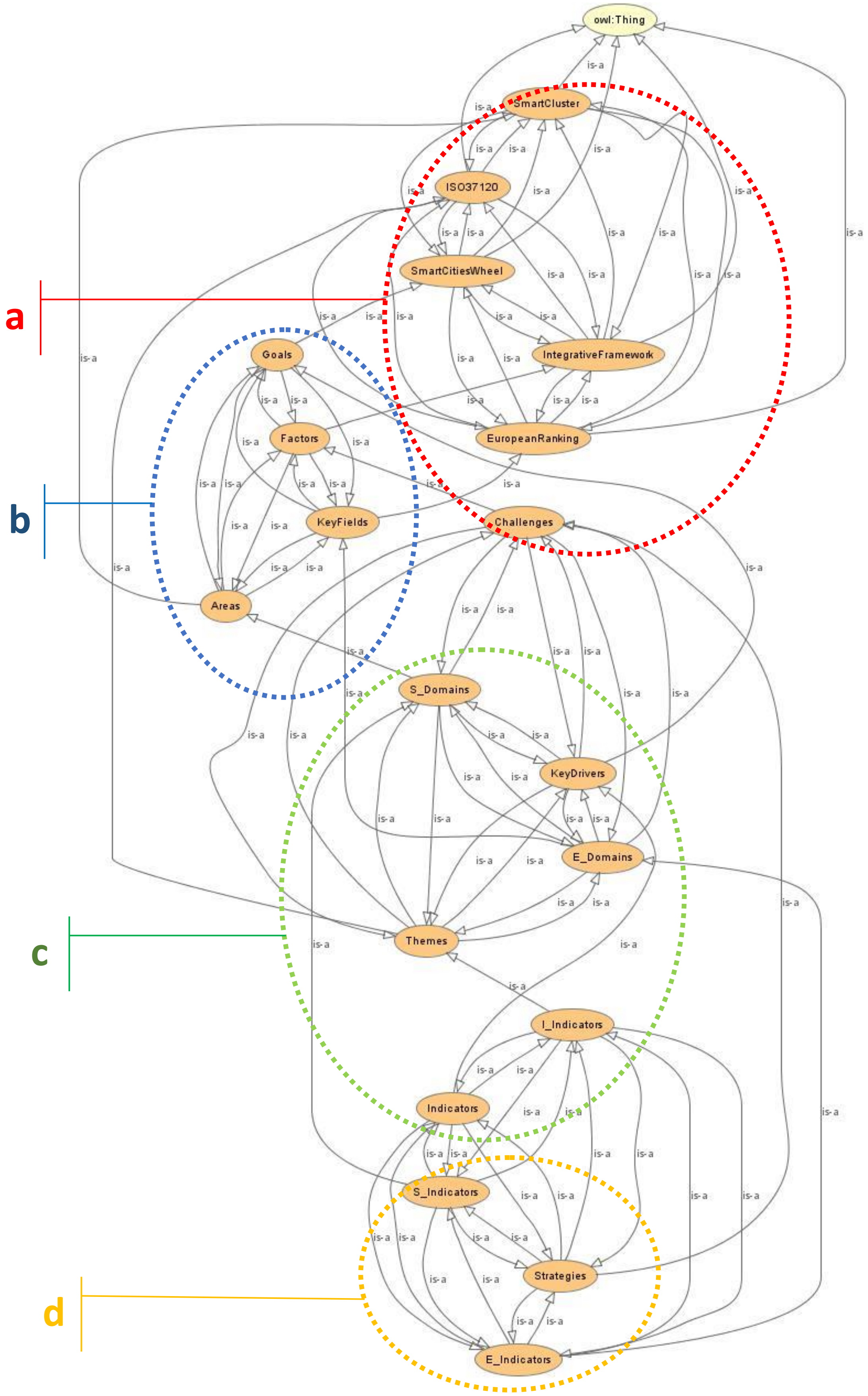
In the third step are built the actual ontologies, with their classes, properties and axioms. Figure 4.10 shows the construction of the SmartCluster ontology and the relationships with the other models previously presented.

Figure 4.10: Construction of the SmartCluster ontology.



Source: Made by author

Figure 4.11: Metamodel SmartCluster and SCMs



Source: Made by author

4.4.1.4. Classifying

These ontologies have their own classes, relationships, axioms, and a rich set of metadata. Using a tool to exploit this metadata allows to manipulate and define not only the set of information that the ontology brings together about Smart Cities, but also to redefine the ontological structure itself, adjusting it to the existing models.

To perform the SCM classification using a language such as OntoQL (Jean, 2006) can provide operators to define, manipulate and query ontologies from the SmartCluster metamodel. This metamodel is not static, which means that it can be extended and adapted to the needs of the new SCM models, and for that the use of a query language and manipulation of ontologies can be very useful. Thus, this data definition language creates, modifies, and deletes entities and attributes of the metamodel using a syntax similar to the manipulation of user-defined SQL types (CREATE, ALTER, DROP). In chapter 4 of this doctoral thesis, it was presented the Taxonomy of indicators to measure Brazilian cities, and from there, the basic metamodel proposed in this work was developed. From this metamodel it is possible to classify the SCMs and incorporate them into the SmartCluster.

Figure 4.11 shows the complete integration ontology between SmartCluster and the other SCMs studied in this chapter (ISO37120, SmartCityWheel, EuropeanRanking and IntegrativeFramework). (a - red group).

```
CREATE CLASS SmartCityModels EXTENDS "SmartCluster"(

  DESCRIPTOR (
    #name[pt,en] = (‘SmartCityModels’,‘SmartCityModels),
    #definition = ‘ontologia de modelos de cidades inteligentes’,
    #definition[en] = ‘SCM Ontology’)

  PROPERTIES (
    URI String,
    Name String,
    superClasses String,
    equivalentClasses String,
    disjointClasses String)
);
```

At this point in the work, the use of OntoQL helps to create update and delete concepts of an ontology (classes, properties, ...) and values of attributes (names, definitions, ...). Using Data Definition Language (DDL) makes it possible to use statements as related to below, which creates a class with a name "SmartCityModels".

Extends the class "SmartCluster" defines the metadata properties for the relationship with the other classes and subclasses.

As Data Definition Language (DDL) allows to create and update the concepts of the ontology and its classes, data manipulation language (DML) enables as in SQL3 (Eisenberg, 1999) to insert classes and their instances, as presented in the next instructions.

```
INSERT INTO SmartCityModels (URI, Name, superclasses, equivalentClasses,
disjointClasses) VALUES
(‘http://semanticweb.org/Ricardo/ontologies/2017/0/SmartCluster/#ISO37120’,‘ISO
37120’, ‘SmartCluster’, ‘SmartCityWeel, EuropeanRanking, IntegrativeFramework’,
‘’);
```

In the previous instruction the class representing SCM "ISO37120" was included as subclass of class "SmartCityModels", and together with it, the values of class name, its superclasses and equivalent classes were assigned. In Figure 4.11 these classes belong to group (a) and its subclasses to group (b). This type of inclusion of values using OntoQL allows checking the semantics of the ontology model and verifying that the meta-metamodel structure is compatible with the proposed SCMs and their respective subclasses.

In the following statement, the inclusion of a "Health" instance of the "S_Domains" class (Figure 4.11, group c - green) is shown, which in turn represents all domains of the "SmartCluster" class. By including this instance, also assumed its values that guarantee that this instance is of the same type of metadata as the other individuals in this Domain.

```
INSERT INTO S_Domains (URI, Name, sameIndividuals) VALUES
(‘http://semanticweb.org/Ricardo/ontologies/2017/0/SmartCluster/#Health’,‘Healt
h’, ‘Mobility, Transport, Security, Education’);
```

The set of classes and subclasses that make up the Group d (orange) of Figure 4.11 mention the last level of the SCM ontologies and the meta-model SmartCluster itself. At this level are stored the metadata for indicators that measure and compare smart cities. For each SCM there are different semantics of the use of the term "indicator", and therefore it is necessary to create an ontology where these terms have their guaranteed equivalence.

For this, OntoQL allows the query of content, once this content is connected to the ontology, the query will not depend on any specific logical database model.

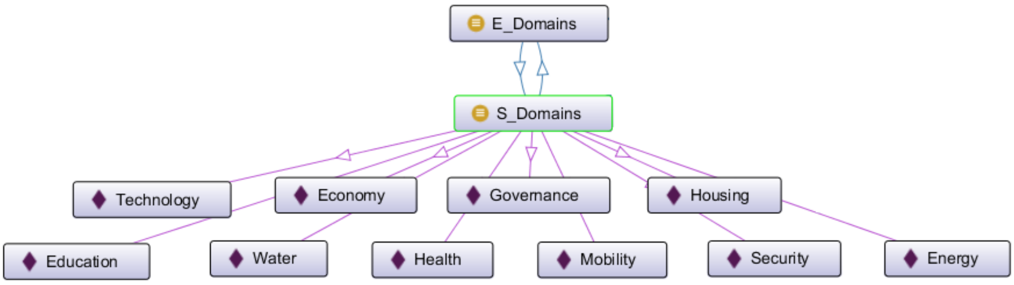
Technically, this allows different systems or applications to execute queries in SmartCluster even though the data is stored in different logical database models.

One way to perform these shared queries is to use the data query language (DQL), provided that some query rules are respected: Each instance has a unique identifier (oid), each instance has a basic class in the ontology, each instance is described by Values of the properties defined in the class extension and the ontology classes can be connected by an inheritance relation.

Figure 4.11 shows the classes of group (d) and how a DQL instruction can be used to perform queries on the indicators of the different SCM models. In the query below, all instances that do not have the same names in both SCMs are selected. This type of query helps to identify which indicators may be duplicated in different SCMs (Figure 4.12).

```
SELECT s.name, e.name
FROM s in S_Domains, e in E_Domains
WHERE s.name <> ALL (SELECT e.name FROM E_Domains)
```

Figure 4.12: Non-duplicate indicators in different SCMs



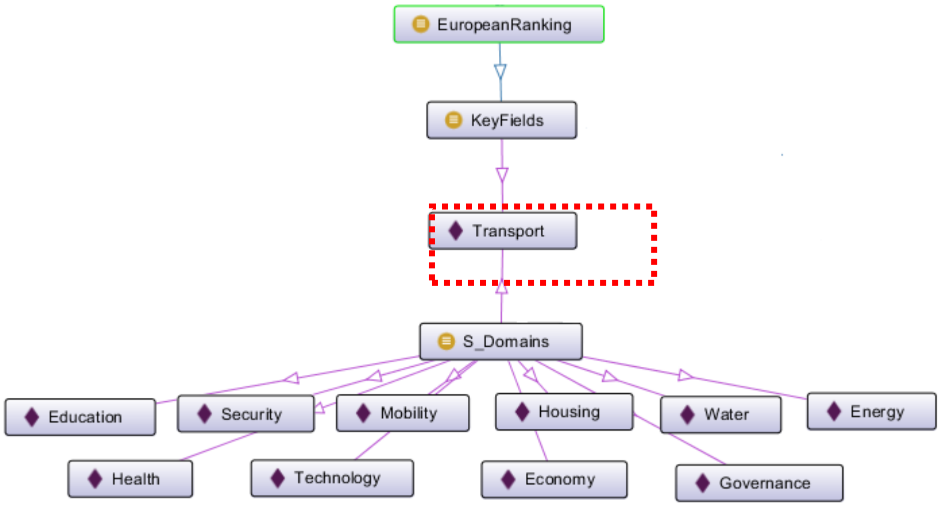
Source: Made by author

Like in SQL, a polymorphic search operator * allows you to query the instances of a particular class and all its subclasses. In the following statements the first one retrieves the names of instances whose class is "S_Domain", while the second returns the names of instances that have the same name in the two SCMs (Figure 4.13), which in this case is the "Transport" instance.

```
SELECT name FROM "S_Domains"

SELECT name FROM "KeyField" WHERE name == ALL (SELECT s.name FROM S_Domains)
```

Figure 4.13: Indicators in common between classes



Source: Made by author

After the classification stage of the ontologies, the next step presents the next phase, which corresponds to integration and verification of compatibility between SCM and SmartCluster.

4.4.2. Phase 2) Integration of the resulting ontologies to SmartCluster.

In this phase, the SCM ontologies (generated in phase 1) are merged interactively, resulting in the SmartCluster ontology already with the assured compatibility for the models detailed in the previous phase. The fusion or alignment can be done by evaluating the elements of Ontology that can include concepts, relationships and instances of these concepts (Euzenat and Shvaiko, 2007). Instance-based approaches are especially suited to scenarios where Ontology has many instances, such as in the case of SCM, where each model has at least 40 indicators that are formalized as instances of its ontologies.

In the proposal of (Xi-Juan et al, 2006), a first alignment is made using ontology labels and instances to find the preliminary concept and then the accumulated experiences are reused to modify the preliminary alignment. Then, a graph-based iteration process is performed. Finally, to decide the attribute matches a logical relationship mining approach that is based on instances is used.

In (Brauner, 2008) two approaches are presented for the alignment of instances: a) a priori, in which the discovery of the mappings is done before the implantation of the mediator, and b) adaptive, in which the discovery and adaptation of the mappings are performed Incremental form, using responses to user queries as evidence of mappings.

Other initiatives also aim to achieve a satisfactory level in the degree of similarity between ontologies and then align them. The paper of (Souza, 2010) makes use of an algorithm that uses different similarity functions and calculates the degree of similarity between concepts recursively, calculating the result of the similarity function between two concepts based on the degree of similarity between concepts that have close kinship.

The paper of (Alves et al, 2012) proposes the application of Data Mining techniques to improve the alignment between domain ontologies. Considering that alignment is not a deterministic task, it is interesting to consider techniques that respect the uncertainty of these ontology joining processes.

The process of knowledge discovery in databases (KDD) is a multi-step process, not trivial, interactive and iterative. These steps are aimed at identifying comprehensible, valid, new and potentially useful patterns from large datasets (Fayyad et al., 1996). KDD was used in chapter 4 of this doctoral thesis to enable the creation of the Taxonomy of Indicators to Measure Brazilian Smart Cities, and the results obtained will be the methodology used to develop the merger between SCM and SmartCluster.

In this work, four SCMs were used, and three processes of direct fusion were possible. It was decided to merge the variant ontologies directly with SmartCluster, thus avoiding mergers with repeated ontologies. The function of fusion is to gather all the constructions of two ontologies, taking into account that the duplicate constructions are only considered once in the final blended ontology.

According to Fayyad (Fayyad, 1996) KDD is a set of steps that process the date according to the following order:

1) **Selection:** Collect and search for data in the database;
2) **Pre-processing:** Treat the data, special characters and text encoding;
3) **Transformation:** Set data pattern and correct spelling errors;
4) **Mining:** Apply data analysis and interpretation of information and
5) **Interpretation:** Transforming information into charts.

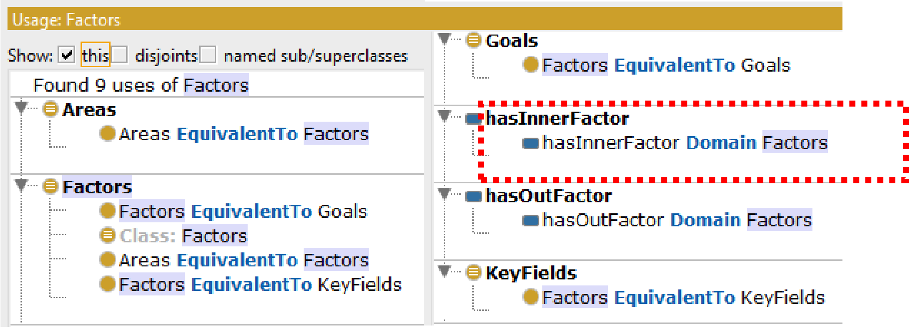
4.4.2.1. - Selection

In this step the data of the ontologies of origin and destination are collected. This collection can be done through queries based on OntoQL that uses the Ontology Query Language (OQL) to search the elements in both ontologies. Queries are similar to the DQL language, except that entities and properties are used instead of classes and properties (Figure 4.14).

```
SELECT #name[en], #allValuesFrom.#name[en]
FROM #OWLRestrictionAllValuesFrom
WHERE #onProperty.#name[en] = 'hasInnerFactor'
```

This query uses the same SQL3 concepts, and consists of a selection and a projection. The selection retrieves the constraints on the property named in English "hasInnerFactor." The path expression used in this selection consists of the onProperty attribute that retrieves the identifier of the property in which the constraint is defined and the name attribute that retrieves the English name of this property from its identifier (Figure 4.14).

Figure 4.14: Object Properties of Ontology



Usage: Factors

Show: ☑ this □ disjoints □ named sub/superclasses

Found 9 uses of Factors

Areas
- Areas EquivalentTo Factors

Factors
- Factors EquivalentTo Goals
- Class: Factors
- Areas EquivalentTo Factors
- Factors EquivalentTo KeyFields

Goals
- Factors EquivalentTo Goals

hasInnerFactor
- hasInnerFactor Domain Factors

hasOutFactor
- hasOutFactor Domain Factors

KeyFields
- Factors EquivalentTo KeyFields

Source: Made by author

The projection also applies the name attribute to retrieve the name of the constraint and the path expression composed of the attributes "allValuesFrom" and "name" to retrieve the name of the class in which the property implicit in the constraint must have its values (Jean, 2006).

4.4.2.2 - Pre-processing

In the preprocessing stage the SCM were compared from the concepts that define them, through the associations and reaching the properties of objects and their respective instances. Table 4.3 presents the comparison between SCMs, and how they were mapped to their respective ontology elements. Four domains are considered for this comparison: a) Each concept, concept relation and enumeration class is represented as a class in OWL; b) Each association is represented as an object property in OWL; c) Each class property is represented as axioms in OWL WL and d) Each enumeration element is represented as a class instance of the owner enumeration class in OWL

Table 4.3: Transformation Rules

ID Model	**(01) European Ranking**	**(02) Integrative Framework**	**(03) Smart Cities Wheel**	**(04) ISO 37120**	**(05) SmartCluster**
Author / Year	**Giffinger (2007)**	**Chourabi (2012)**	**Cohen (2012)**	**ISO (2015)**	**Afonso (2017)**
(a)	6 Key Fields	8 Factors	6 Goals	17 Themes	10 Domains
(b)	hasDomain	hasInnerFactor hasOutFactor	hasKeyDrivers	-	hasDomains
(c)	hasIndicators	hasStrategies	hasIndicators	hasCIndicators hasSIndicadotrs	hasIndicators
(d)	90 Indicators	42 Strategies	100 Indicators	46 C Indicators 54 S Indicators	40 Indicators (min)

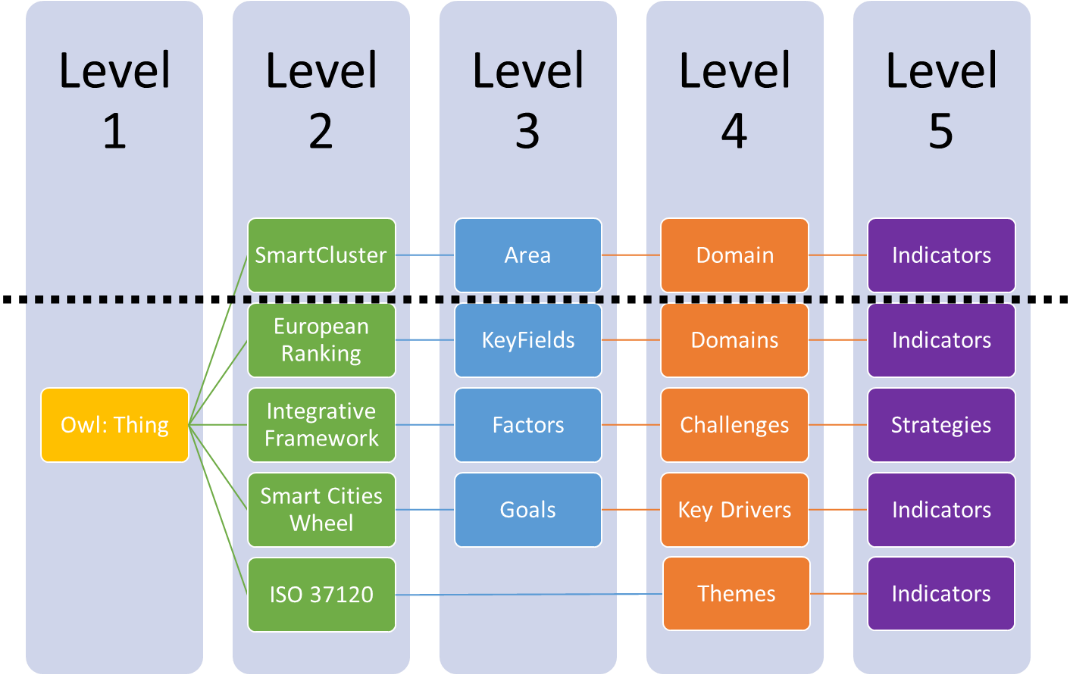
Source: Adapted from (Najera, 2013)

After the preprocessing stage, where the initial compatibility of each SCM is verified, it is also possible to identify the needs that the meta-model SmartCluster will need to extend its domains to cover all inherited classes, relationships and instances. The SmartCluster metamodel has 10 domains to represent the compatibility with other models, but the number of domains is unlimited as new models are added.

4.4.2.3. – Transformation

This step consists of transforming these models into a single ontology, thus defining a naming pattern for classes, relationships, instances, and axioms. In addition to maintaining a semantic standard, it is important that concepts are equivalent in ontologies so that their metadata can be effectively stored, queried, and altered if there is a need for such an expansion. Figure 4.15 shows the equivalence scheme between the classes of the SCM ontologies and the SmartCluster metamodel.

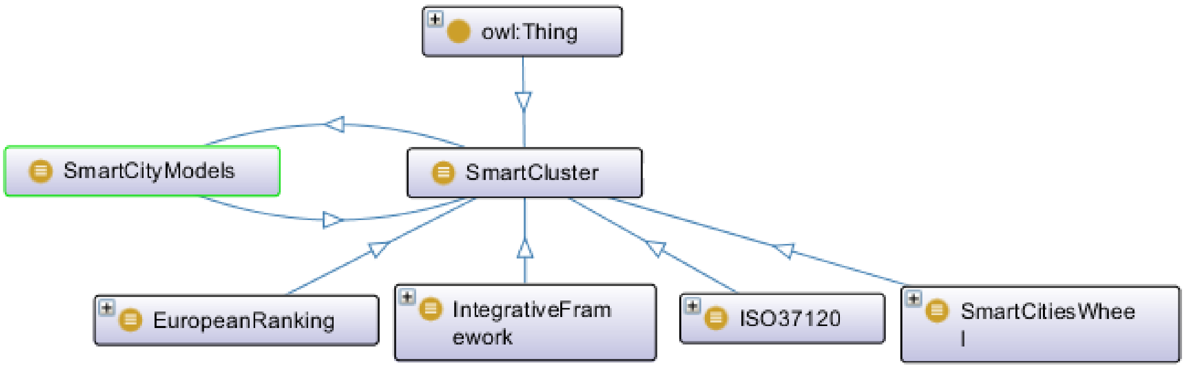
Figure 4.15: SCM and SmartCluster levels



Source: Made by author

The previous step facilitates this work by cataloging the relationships and properties required for each of the SCMs, which here, in this step, are transformed through the Protégé tool into an integrated ontology. (Figure 4.16)

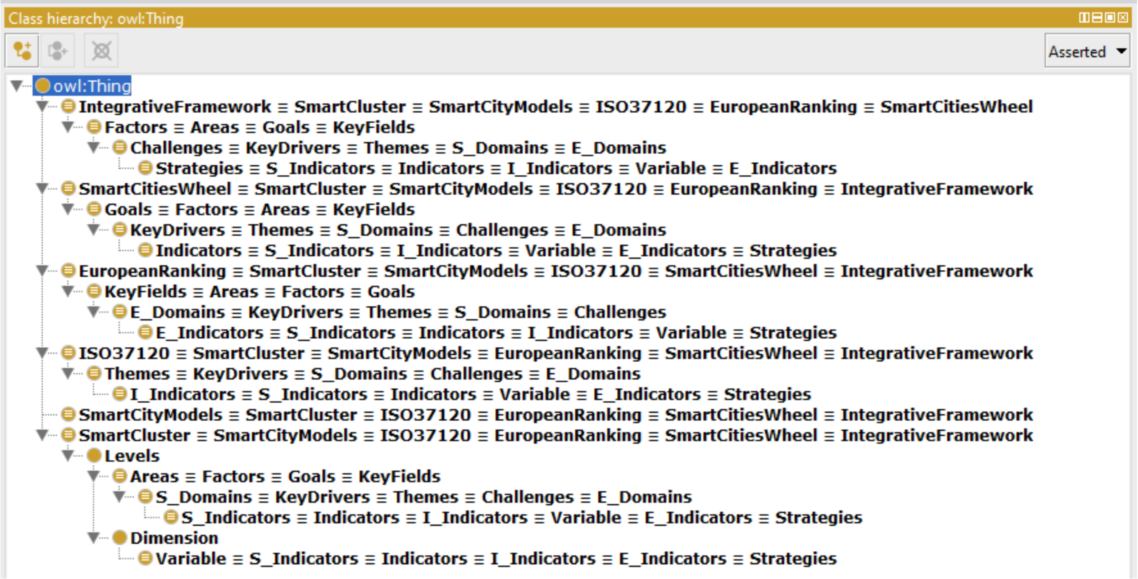
Figure 4.16: SCM integrated through an Ontology



Source: Made by author

The standardization of ontologies is necessary so that in the next steps it is possible to carry out the queries necessary to perform the queries, alignments and joins of ontologies. In simplified form, these ontologies are aligned respecting the same concepts and ontological levels. Figure 4.17 shows a view of the Protégé tool with the mapped ontologies.

Figure 4.17: Mapping SCM ontologies



```
owl:Thing
  IntegrativeFramework ≡ SmartCluster ≡ SmartCityModels ≡ ISO37120 ≡ EuropeanRanking ≡ SmartCitiesWheel
    Factors ≡ Areas ≡ Goals ≡ KeyFields
      Challenges ≡ KeyDrivers ≡ Themes ≡ S_Domains ≡ E_Domains
        Strategies ≡ S_Indicators ≡ Indicators ≡ I_Indicators ≡ Variable ≡ E_Indicators
  SmartCitiesWheel ≡ SmartCluster ≡ SmartCityModels ≡ ISO37120 ≡ EuropeanRanking ≡ IntegrativeFramework
    Goals ≡ Factors ≡ Areas ≡ KeyFields
      KeyDrivers ≡ Themes ≡ S_Domains ≡ Challenges ≡ E_Domains
        Indicators ≡ S_Indicators ≡ I_Indicators ≡ Variable ≡ E_Indicators ≡ Strategies
  EuropeanRanking ≡ SmartCluster ≡ SmartCityModels ≡ ISO37120 ≡ SmartCitiesWheel ≡ IntegrativeFramework
    KeyFields ≡ Areas ≡ Factors ≡ Goals
      E_Domains ≡ KeyDrivers ≡ Themes ≡ S_Domains ≡ Challenges
        E_Indicators ≡ S_Indicators ≡ Indicators ≡ I_Indicators ≡ Variable ≡ Strategies
  ISO37120 ≡ SmartCluster ≡ SmartCityModels ≡ EuropeanRanking ≡ SmartCitiesWheel ≡ IntegrativeFramework
    Themes ≡ KeyDrivers ≡ S_Domains ≡ Challenges ≡ E_Domains
      I_Indicators ≡ S_Indicators ≡ Indicators ≡ Variable ≡ E_Indicators ≡ Strategies
  SmartCityModels ≡ SmartCluster ≡ ISO37120 ≡ EuropeanRanking ≡ SmartCitiesWheel ≡ IntegrativeFramework
  SmartCluster ≡ SmartCityModels ≡ ISO37120 ≡ EuropeanRanking ≡ SmartCitiesWheel ≡ IntegrativeFramework
    Levels
      Areas ≡ Factors ≡ Goals ≡ KeyFields
        S_Domains ≡ KeyDrivers ≡ Themes ≡ Challenges ≡ E_Domains
          S_Indicators ≡ Indicators ≡ I_Indicators ≡ Variable ≡ E_Indicators ≡ Strategies
      Dimension
        Variable ≡ S_Indicators ≡ Indicators ≡ I_Indicators ≡ E_Indicators ≡ Strategies
```

Source: Made by author

The next step is the mining and analysis of the metadata obtained with the proposed ontology structure.

4.4.2.4 – Mining

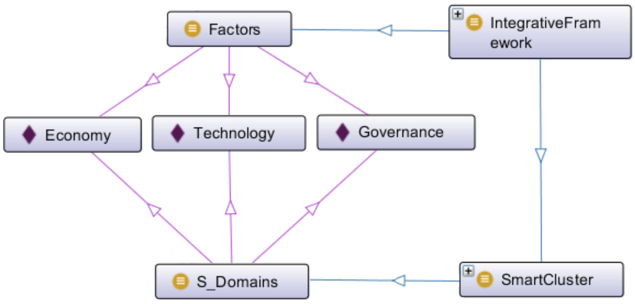
In the mining stage, the metadata and the structure of the SCM ontologies are analyzed so that it is possible to retrieve information about instances (indicators) and about the typing of the data of these instances. Direct mining in ontologies serves as a kind of test to verify that future queries to be performed by external applications and systems will meet the demand for information search.

Since ontologies, relationships and instances (URI) links are declared in the Ontology metamodel, OntoQL uses these URIs to query the structure of the ontology and the content of its classes. The next query returns the names of instances that belong to the "Factors" class of the SCM "IntegrativeFramework" that start with "Tech".

```
SELECT u.name
FROM Factors in #class, u in Factors*
WHERE u.#name like ‘Tech%’
```

The result of this query may include instances such as "techology" or other similar names that identify instances that may be duplicated in other SCMs, such as in this case where the "Technology" instance is part of the "Factors" class and the "S_Domains" (Figure 4.18).

Figure 4.18: Common instances between classes.



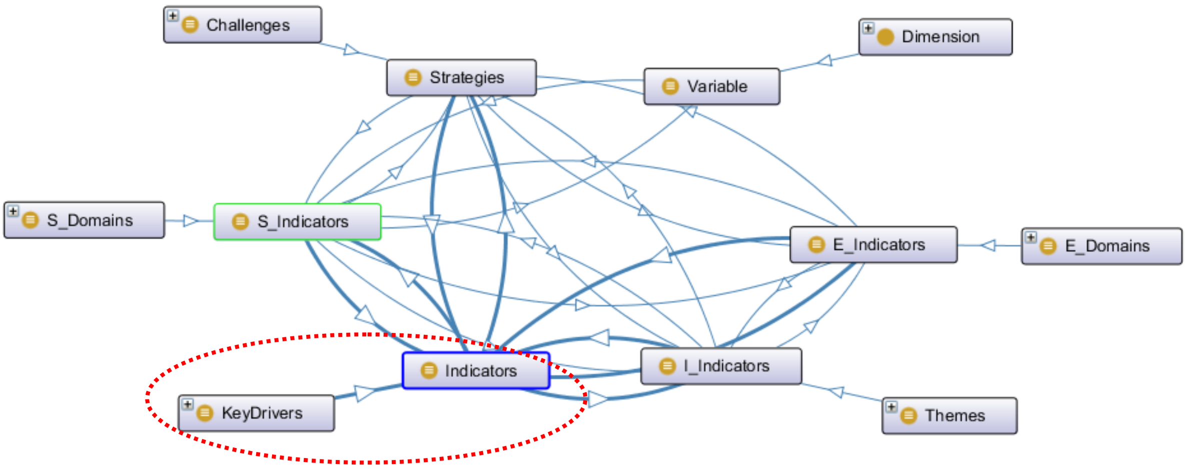
Source: Made by author

More elaborate queries can retrieve information about sets of instances and classes. In the following query the "Indicators" class is retrieved using the WHERE clause and identified by the "c" iterator. The UNNEST operator provides the "csup" iterator in class "c" superclasses. Finally, the names of superclasses are projected in the SELECT clause.

```
SELECT csup.#name
FROM #Class AS c,
UNNEST(c.#superclasses) AS csup
WHERE c.#name = 'Indicators'
```

The purpose of this query is to retrieve all the classes that are called "Indicators" and which superclasses they belong to. This type of query allows us to compare the structures of the ontologies and even better understand the concepts that derive from the initial concept of indicators and their variations in the other ontologies (Figure 4.19).

Figure 4.19: Class "Indicator" and its superclasses



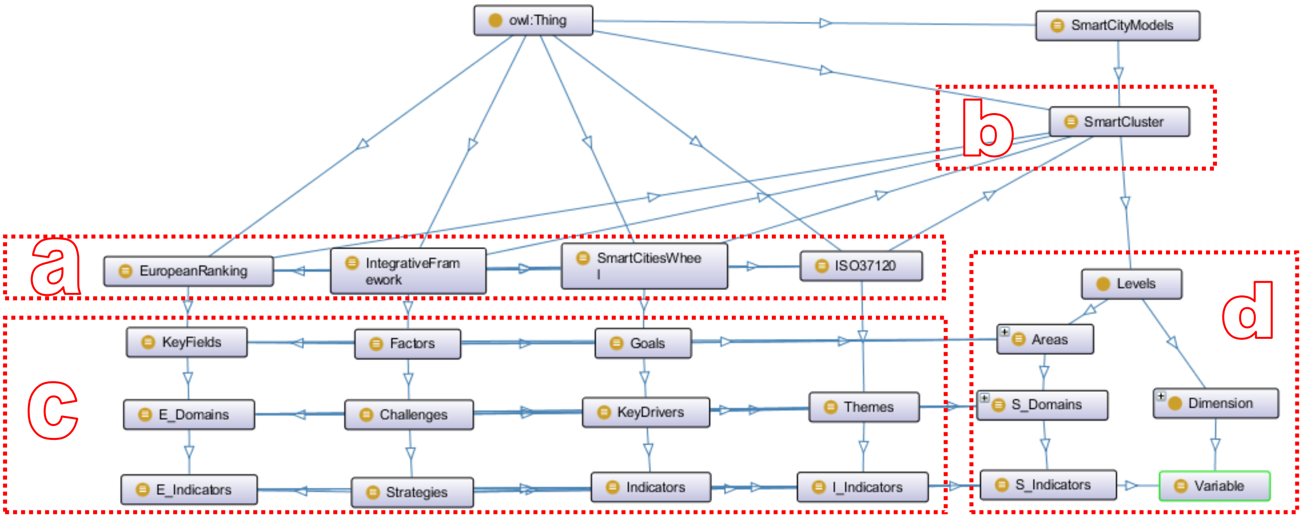
Source: Made by author

The final step consists of the interpretation of the ontologies generated and incorporated by the SmartCluster is the Interpretation, which will be discussed below.

4.4.2.5- Interpretation

The final step of integration of KDD-based ontologies is the interpretation phase, where information is generated to visualize the structure and metadata of the final ontology (Figure 4.20). In this Thesis four SCM were used to compose the SmartCluster metamodel. In Figure 4.20 the letter "a" presents these four models (i.e., EuropeanRanking, IntegrativeFramework, SmartCitiesWheel, ISO37120). The purpose of the letter "b" is to show that these models must be compatible with the "SmartCluster" metamodel.

Figure 4.20: SmartCluster Ontology Metamodel



Source: Made by author

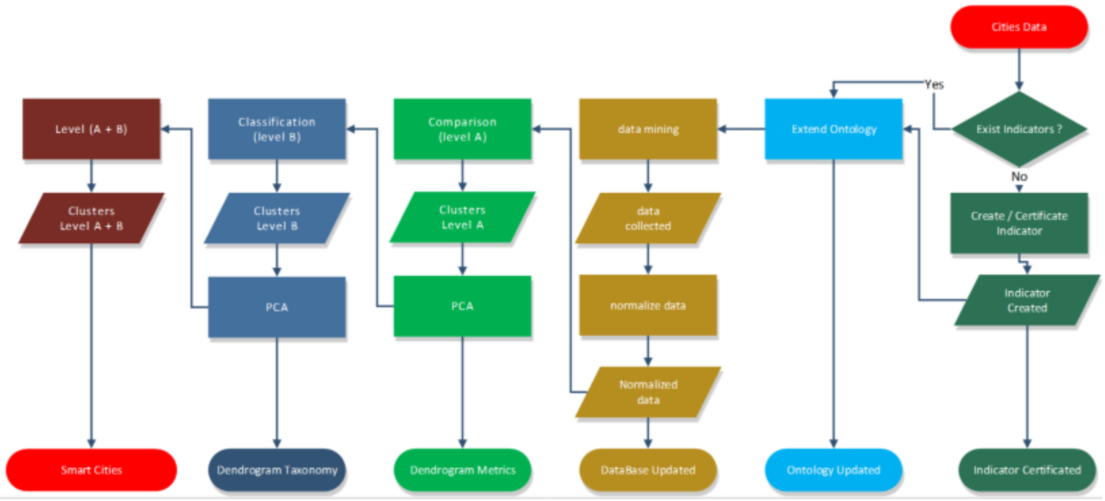
The structures of all SCMs have different concepts and nomenclatures, and through the visualization of these ontologies it is possible to create ways of presenting the levels and compatibilities between these concepts in a more didactic way (c). The inclusion of instances in both SCM and SmartCluster allows us to identify which classes of these SCMs are represented in the metamodel in a way appropriate to the original concepts (d).

4.5 Summary

Based on the previous chapter, where a taxonomy was proposed to measure and compare Brazilian Intelligent Cities, this chapter presents a study on the four models of smart cities most cited and used to create rankings of intelligent cities around the world.

Throughout this chapter these Intelligent Cities Models were called SCMs and the main objective of this chapter was to present a metamodel called SmartCluster, its classes, relationships, axioms and instances that proved to be compatible with the SCM compared (Figure 4.21).

Figure 4.21: SmartCluster flowcharter



Source: Created by author

The idea of creating a metamodel for SCM is precisely to allow the need to create new models that can be compatible with existing models and thus incorporate best practices, indicators and even visualization of results based on stored metadata.

5

METAMODEL VALIDATION USING EBSE

Never forget who you are. The rest of the world will not forget.
Use this as an armor, and this can never be used to hurt you.

—Tyrion Lannister, Game of Thrones

This chapter introduces the use of cluster analysis (C.A.) as a way to identify similarities between cities and compare them. The group stages that composes this particular analysis, is named: SmartCluster. This analysis consists of five steps (Select, Process, Transform, View and Interpretation) that allow visualization of data in Brazilian smart cities indicators like no other model proposed, respecting the particularities of each group of cities.

5.1 Introduction

The importance of applying an evidence-based methodology to scientific research can be illustrated by experience in medicine. For a long time, the medical area was full of revisions that did not use methods to identify, evaluate and synthesize information existing in the literature (Cochrane, 2003). At the end of the 1980s, studies conducted to evaluate the quality of medical publications drew attention to the low scientific quality (Cochrane, 2003).

It was the work of Kitchenham et al. (2004), the first to establish a parallel between Medicine and Software Engineering, with respect to the evidence-based approach. According to the authors, Evidence-Based Software Engineering should provide means by which better evidence from the research can be integrated with practical experience and human values in the decision-making process considering the development and maintenance of the software.

Thus, Barbara's paper (Kitchenham, 2004) makes it clear that EBSE serves to:

- Direct research to the needs of industry, academia and other groups;
- Base the decisions of industry professionals on the adoption of technology;
- Improve software reliability, thereby improving technology choice;
- Increase the acceptability of software that interacts with citizens and
- Create possibilities to define certification processes.

The purpose of evidence-based medicine (EBM) is "to integrate the best research evidence with clinical knowledge and patient values" (Sacket et al, 2000). The author further states that the purpose of evidence-based software engineering (EBSE) is: "to provide the means by which the best current evidence of research can be integrated with practical experience and human values in the decision-making process on the Development and maintenance of software.”

Sackett identifies in his work five steps that are necessary to practice evidence-based medicine and correlates them with the EBSE. The steps of EBM are shown in the second column of Table 5.1.

The author created a correlation between the EBM steps with the EBSE and thus presents them in column 3 of Table 5.1. The content of this column was adapted to meet the needs of the work proposed by this thesis.

Table 5.1: Five steps used in EBM and EBSE

Step	(EBM) Evidence-based Medicine	(EBSE) Evidence-based Software Engineering
1	Converting the need for information (about prevention, diagnosis, prognosis, therapy, causation, etc) into an answerable question.	Converting the need for information into concepts (about SCM, Measurement Models, Domains, Indicators, etc) into an answerable question.
2	Tracking down the best evidence with which to answer that question.	Tracking down the best evidence with which to answer that question.
3	Critically appraising that evidence for its validity (closeness to the truth), impact (size of the effect), and applicability (usefulness in our clinical practice).	Critically evaluate the evidence of its validity (proximity to truth), impact (size of effect) and applicability (practical use of Smart City models).
4	Integrating the critical appraisal with our clinical expertise and with our patient's unique biology, values and circumstances.	Integrate a critical assessment with our experience in Smart Cities Models, Areas, Domains, and Indicators.
5	Evaluating our effectiveness and efficiency in executing Steps 1-4 and seeking ways to improve them both for next time.	Evaluating our effectiveness and efficiency in executing Steps 1-4 and seeking ways to improve them both for next time.

Source: Adapted from (SACKET, 2000)

According to (Mafra, 2006) the EBSE is divided into two types of study:
primary studies and secondary studies:

- Primary studies are studies that characterize a concept in use within a specific context. In our case, It was searched for SCM and its classifications through the use of the methodology called Grounded Theory, which was developed in chapter 2 of this thesis.
- Secondary Studies: these are studies that identify, evaluate and interpret the results of a given research topic. The systematic literature review (SLR) is a type of secondary study (Biolchini et al., 2005) and served as a source for the creation of Chapter 2 of this

thesis, which consolidates the usefulness of this chapter to the base of secondary studies.

To achieve the objectives proposed by this chapter, the following contents are divided as follows: This Section 5.1 presents the introduction on the subject and the structure of this chapter; Section 5.2 presents the EBSE methodology applied to the SCM and the five steps adopted to practice the EBSE; In Section 5.3 the two layers that comprise the SmartCluster are presented, being one responsible for the Metrics (Variables) and the next layer for the Taxonomy (Domains and Indicators); In Section 5.4 a case study (experimentation) was conducted with a set of data referring to the cities of northeastern Brazil, in the state of Alagoas; Section 5.5 discussed the contributions of the SmartCluster metamodel and this chapter concludes with the final conclusions in Section 5.5..

5.2 Methodology for practicing EBSE for SCM

The starting point for this methodology indicated by (Mafra, 2006) is observed that (Sackett et al, 2000) considered EBM from the point of view of an individual physician who must decide how to treat a particular patient exhibiting a given set of symptoms. In this case, when using EBSE, it is important to make it clear that SCMs seldom choose the same set of indicators or concepts. The adoption of a particular SCM can often be decided by public managers according to the local interest in reaching particular indicators. As a result, existing SCMs are not standardized, do not meet data normalization standards and are almost never compatible with each other.

The following sections detail how the EBSE can be applied following the five steps proposed by (Sackett et al, 2000).

5.2.1 Step 1: Defining an answerable question

The first step in this methodology is to raise the questions that will guide the process of gathering the evidence. For a question to be well formulated (Sackett et al., 2000), it must be composed of three factors:

1. **Study factor** (intervention, diagnostic test or exposure);
2. **Population** (the group of indicators, domains, areas or models) and

3. Results.

As described in Table 1, EBM, health professionals are generally interested in broader processes such as prevention, diagnosis, prognosis, therapy, and causality and not necessarily the causes of diseases, which represent the lower granularity of these processes. This is because these doctors seek a broader analysis of the variation of the study factor in the populations studied.

In this case, in the EBSE, the study factor of this thesis are the SCM models and how they correlate. Therefore, as in EBM's view, to create the EBSE issues it is not necessary to specify a very deep level of abstraction (less granularity). With this, the focus of the study factor is directly related to the models, and on the necessity (or not) of these models to become adaptable and correlated with each other.

As for the population, in this case, there is a certain difficulty in determining the correct level of abstraction to specify the population of interest, since as stated earlier, these SCMs do not follow standards.

The population of interest of these SCMs often presents in the form of indicators, strategies or even objectives, which allows for varying classifications and reveal the need to create constraints so that the concepts are at least equivalent, thus preventing evidence being discarded for lack of conceptual uniformity.

In Chapter 2 of this thesis, the research questions that defined this work were correlated with the theories formulated through an extensive bibliographical survey (Table 5.2).

Table 5.2: Formulated theories to address the research questions.

ID	Formulated Theory	Research Questions
FT1	It´s possible evaluate smart cities using public data.	**RQ1**
FT2	There are specific characteristics that demand respective indicators.	**RQ1, RQ2**
FT3	A statistical formalism is required to compare smart cities.	**RQ2, RQ3**
FT4	Data visualization helps in strategic decision making.	**RQ4**
FT5	Smart City models are heterogeneous and require structural standardization.	**RQ2, RQ5**

Source: Made by author

Based on the results obtained with this correlation it is possible to identify some key words that require attention to formulate the question for which one wishes to seek evidence: "Smart Cities", "Public Data", "Indicators", "Statics Formalism", "Date Visualization "," Smart City Models "and" Standardization ". These words are also repeatedly found in the research questions: **RQ1**: How can a city be evaluated? **RQ2**: Which indicators are appropriate to the Brazilian reality ?; **RQ3**: How to get data about smart cities? **RQ4**: How to provide an extraction environment and data? And **RQ5**: How to create and validate a metamodel compatible with SCMs?. In the context of EBSE, the strategies to be considered when deciding which question to answer first, second (Dyba, 2005) include:

(a) What is the most important issue for your customers?

(b) What is the most relevant issue for your situation?

(c) What question is most interesting in the context of business strategy?

(d) What is the most likely issue to repeat in your practice?

(e) Can the question be answered in the time available?

Observing the correlation between research questions and found theories, it is possible to create the structure of the question (Table 5.3) that will be formulated to seek the evidence.

Table 5.3: Structure of the question answerable by EBSE.

Factors (Sackett et al, 2000)	**Source/Expected Chapter 2**	**Questions/Theory Chapter 1**	**Strategies (Dyba, 2005)**
(1) Study factor	Grounded Theory Systematic Literature Review	RQ1 -> FT1	(c), (d)
(2) Population	Smart Cities Models	RQ1, RQ2 -> FT2 RQ4 -> FT4	(b), (c), (d)
(3) Results (expected)	“Smart Cities Metamodel”	RQ2, RQ3 -> FT3 RQ2, RQ5 -> FT5	(a), (e)

Source: Made by author

Thus, the question to be answered with the methodology of Evidence-Based Software Engineering (EBSE) is:

```
“Is it possible that a Metamodel for Smart Cities can          (3)
group and allow to create SCMs semantically similar            (2)
using public data to measure and compare any cities?”          (1)
```

Given the question, the next step is to find the best evidence that there are possible answers to this question.

5.2.2 Step 2: Finding the best evidence

One of the reasons for asking the question is not only to help researchers and practitioners find all relevant studies, but also to find effective results from the high number of publications in events and journals.

Finding a response includes selecting an appropriate information resource and executing a search strategy that specifies a rich detail issue so that the search does not return millions of responses from which multiple filters will be required. Strategies for searching for scholarly works were defined in Chapter 2 of this thesis, however, it is common sense that in the scholarly works on Review of Literature (specifically in computer science) the sources of data are recurrent:

1. **IEEE Xplore** (http://ieeexplore.ieee.org) provides access to IEEE publications published since 1988 (and selected articles back to 1950) and to current IEEE standards. Access to abstracts and tables of contents is free. Access to full text requires IEEE membership, a subscription, or payment for individual articles.
2. **The ACM Digital Library** (www.acm.org/dl) provides access to ACM publications and related citations. Full access requires ACM membership and possibly a subscription; nonmembers can browse the DL and perform basic searches.
3. **Google Scholar** (http://scholar.google.com) indexes scholarly literature from all research areas, including abstracts, books, peer-reviewed papers, preprints, technical reports, and theses. Users can find scholarly literature from different publishers, professional societies, preprint repositories, and universities, as well as articles posted on the Web.

One of the reasons for asking the question is not only to help researchers and practitioners find all relevant studies, but also to find effective results from the high number of publications in events and journals. In Table 5.4, two search string structures are presented, in which the first one searches for results linked to the greater granularity of the subject (S001) and the second search in a more general way to answer the question for which evidence is sought (S002).

Table 5.4: Search Strings.

Search	Search String
S001	(“Smart City” or “Intelligent City” or “Digital City”) and (“Indicator” or “Area” or “Domain”)
S002	(“Smart City” or “Intelligent City” or “Digital City”) and (“Model” or “Ontology” or “Ranking” or “Cluster”)

Source: Made by author

5.2.3 Step 3: Critically appraising the evidence

The critical evaluation of the evidence found when using the EBM method is based on a methodology that has undergone several improvements and adaptations over time and today several organizations have developed guidelines for systematic reviews and evaluation of evidence in addition to the medical journals that have been under pressure and oversight Research in order to improve the conduct and the way individual experience reports are presented (Moher, 2001). The work of (Dyba, 2005) presents a list of factors to be considered in the evaluation of an empirical study:

(1) Are there any vested interests?

a) Who sponsored the study?

b) Do the researchers have any interest in the results?

(2) Is the proof valid?

a) Was the study design appropriate to answer the question?

b) How were the tasks, subjects and scenario selected?

c) What data were collected and what were the methods for collecting the data?

d) What methods of data analysis were used, and were they appropriate?

(3) Are the tests important?

a) What were the results of the study?

b) Are the results credible, and if so, how accurate are they?

c) What conclusions have been drawn, and justified by the results?

d) Do the results have practical and statistical significance?

(4) Can evidence be used in practice?

a) Are the findings of the study transferable to industrial contexts?

b) Did the study evaluate all of the important outcome measures?

c) Does the study provide guidelines for practice based on results?

d) Are the guidelines well described and easy to use?

e) Do the benefits of using the guidelines outweigh the costs?

(5) Is the evidence consistent with the evidence in available studies?

a) Are there any good reasons for any apparent inconsistency?

b) The reasons for any misunderstandings were investigated?

Based on these factors, it developed Table 5.5, which compares the four SCMs used in this work in relation to the SmartCluster metamodel.

Table 5.5: Factors comparison

Factors (Dyba, 2005)	(SCM01) European Ranking	(SCM 02) Integrative Framework	(SCM 03) Smart Cities Wheel	(SCM 04) ISO 37120	(SCM 05) SmartCluster
(1)	**Yes**	**No**	**No**	**Yes**	**Yes**
(2)	**Yes**	**No**	**No**	**No**	**Yes**
(3)	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
(4)	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
(5)	**No**	**No**	**No**	**No**	**Yes**

Source: Made by author

The following section presents how this critical assessment integrates with the expertise of each SCM.

5.2.4 Step 4: Integrating the critical appraisal with SCM expertise

In this section will present the factors (Dyba, 2005) that were considered in the evaluation of the empirical studies about the SCM so that evidence could be found about the need for a Smart Cities metamodel.

For this, the four SCMs and the [(SCM01) European Ranking metamodel were considered; (SCM02) Integrative Framework; (SCM03) Smart Cities Wheel; (SCM04) ISO 37120 and (SCM05) SmartCluster)] and took into consideration the factors that were respectively based on the following questions:

5.2.4.1 (1) Are there any vested interests?

There are SCMs (SCM01) that are created by civil organizations, to better understand the scenario in which cities are inserted, and to compare their social indicators. Other SCMs (SCM04) can be created for commercial purposes to allow cities to adjust to the proposed indicators, and thus certify them within pre-defined standards. SCM02 presents itself as an academic initiative and conceptually explores the scenario of smart cities, thus having no apparent commercial interest.

Other SCM (SCM03) can be included in the category of "theoretical models", which are generally used to serve as the basis for the construction of new models if using the concepts mined in bibliographical surveys.

The metamodel (SCM05) inherits the common interest of existing models, and intends to serve as a knowledge base for the development of new models, at the same time as it can serve as a tool for certification and standardization of new indicators.

5.2.4.2 (2) Is the proof valid?

The theoretical models (SCM02, SCM03) are not concerned with producing evidence, since it proposes to serve as "knowledge bases" for the models to be developed following their strategies and objectives.

That is why civil and commercial models (SCM01, SCM04) are able to combine the theory of their concepts with practical (albeit preliminary) results. Specifically the model (SCM01) is in use for several years and is systematically comparing medium-sized smart cities, and already serves as a strategic reference for these cities to change their public policies for improvements in some services.

The metamodel (SCM05) used the indicators proposed by the SCMs and through mining in public databases already presents several results on comparison of indicators in several Brazilian cities. (Afonso, 2015).

5.2.4.3 (3) Is the evidence important?

Both the SCMs and the metamodel (SCM02, SCM03 and SCM05) present results that lead to the proof that metamodel and models are able to compare and measure smart cities. The evidence clearly indicates the need for standardization of indicators between these models and the use of methodologies that make compatible these existing models and those that are proposed in the future.

5.2.4.4 (4) Can evidence be used in practice?

Evidence of the need for a metamodel (SCM05) can be used in a very practical way, since the metadata of this architecture is compatible with all the models studied, allowing it to be fed with indicators obtained in public databases.

The proposed metamodel (SCM05) allowed not only to incorporate the concepts and definitions of the SCM but also served as a data visualization tool by using techniques of mathematical normalization and data clustering, creating information dendrograms. This type of data visualization may allow managers less familiar with the technology to obtain and translate this comparative information in a more didactic way.

5.2.4.5 (5) Is the evidence consistent with the evidence in available studies?

The evidences found in the tests and in the use of SmartCluster (SCM05) for Brazilian cities and capitals make this metamodel totally consistent with the other SCMs and also allows the expansion to be compatible with international models and metadata.

The main evidences (SCM01) still remain of comparisons or attempts to impose (SCM04) indicators for cities to fit the SCM, whereas in SmartCluster this happens in reverse: the model fits the strategic needs of the city.

As in the medical field, a very small percentage of technological solutions use the evidence for strategic decision making. According to Armando Lopes (senior vice president of Siemens Healthcare of Brazil), "technology has to be seen as an investment, not as a cost, and you have to pay, have a foot and a head and bring a return to those who invest, so you have to generate Evidence." (Folha, 2015)

The SmartCluster presents a crucial difference in evidence compared to other models: its division into comparison levels (Levels 1 and 2) prevents cities with very different characteristics from being compared generating distortions in ranking and comparison of data. Evidence of the need for a metamodel for smart cities is not directly linked to the practical proof of use of the metamodel, but the Section that presents data visualization (dendrograms) can still meet expectations regarding practical use.

5.2.5 Step 5: Evaluation of the process

Both in EBM and in EBSE, the final step consists in the reflection that must be made on the use of the methodology and if this use implied in change and improvement of the existing process. The use of an evidence-based methodology represented by systematic reviews and experimental studies contributes satisfactorily to the definition of new concepts and the safe application of these concepts.

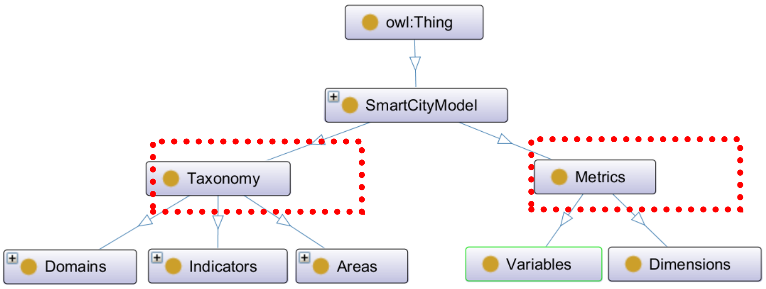
According to Mafra (2006), conducting a systematic review as a step in the initial definition step of a technology makes it possible to: (a) minimize risks and (b) accelerate the definition process. In turn, the execution of experimental studies allows to evaluate the application of the technology that is being defined without creating an immature technology in the industrial context.

From the point of view of data collection, this work made use of an extensive systematic review of the literature and merged it with the methodology of grounded theory in Chapter 2 of this thesis. The other experiments with the proposed model are presented in the following sections, which show how the SmartCluster metamodel can make use of public data to perform the same functions as the SCMs created, and extend the capabilities of these models to a more satisfactory level of comparison of Smart cities.

5.3 Applying Levels of Cluster Analysis (C. A.)

This section details the merging of two levels of city grouping through the SmartCluster metamodel, consisting of a **Level (A)** where the variables are composed of population, territory and development data, and another **Level (B)** composed of city indicators the previously proposed taxonomy. The two levels used in this grouping merge are shown in Figure 5.1.

Figure 5.1: Level A and Level B of the SmartCluster metamodel.



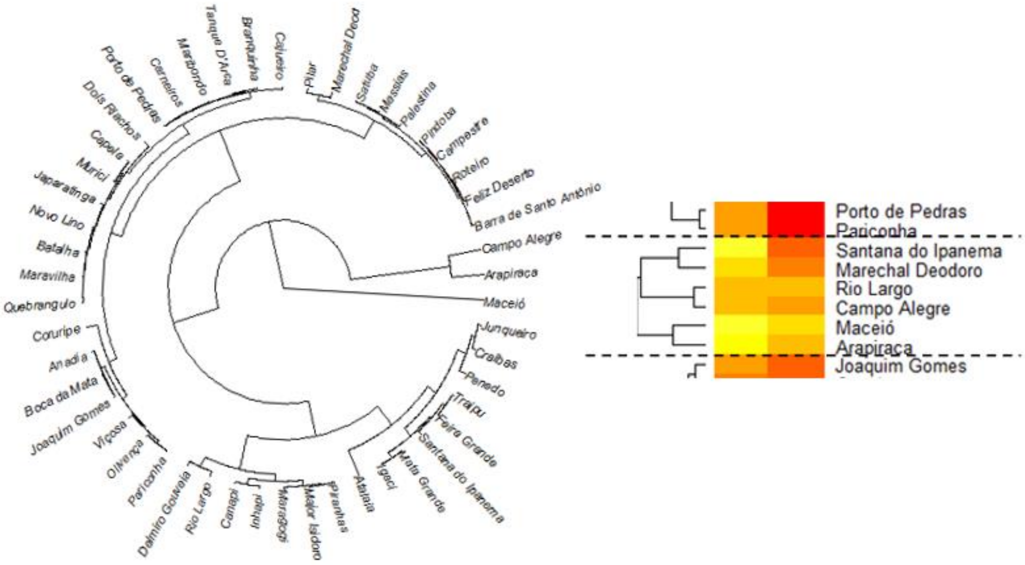
Source: Made by author

To allow visualization of city clusters at different levels, visualization plots of clusters called dendrograms will be used. A Dendrogram (dendro = tree) is a specific type of diagram or iconic representation that organizes certain factors and variables. It results from a statistical analysis of certain data, which employs a quantitative method that leads to groupings and their ascending hierarchical ordering - which in graphical terms resembles the branches of a tree that are divided successively in others. That is, it illustrates the clustering arrangement derived from the application of a "clustering algorithm". (Phipps, 1971)

The results of dendrogram groupings serve to represent the distance and similarity between the objects compared depending on the distance measure selected. When working with data clustering, distance and similarity metrics serve to accurately identify what leads each individual to belong to their respective groupings. The larger the differences between the values analyzed, the greater the distances between these individuals and consequently, less similar they will be.

In the most common representation, the rows or columns show the distance or similarity between the rows and the nodes that each line belongs to as a result of agglomeration. Figure 5.2 shows the grouping of cities by similarity of values in two types of dendrogram graphs. In this work will be used the "fan" type to represent city groups from now on.

Figure 5.2: Example of cities cluster dendrogram.



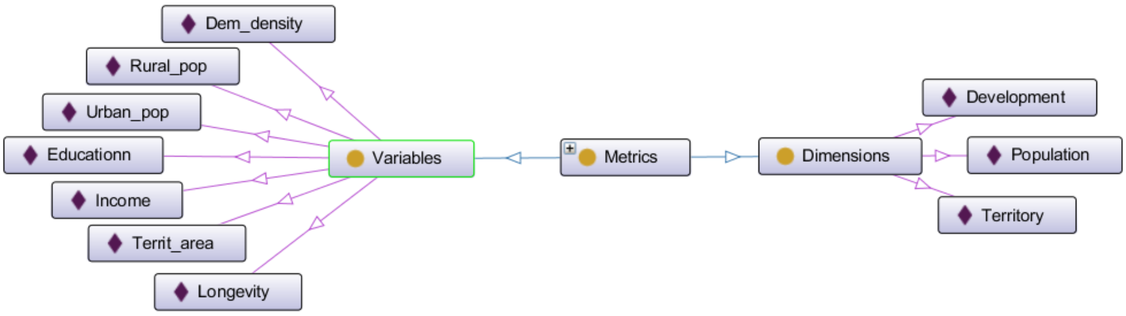
Source: Made by author

Perhaps this represents the biggest difference between the SmartCluster metamodel and the other SCMs studied, since no model predicts the level separation of these data comparisons. This lack of respect for differences in demographic characteristics for example can generate very large distortions in the comparison between medium and large cities. The details and the respective dendrograms of each of the levels cited in this Section will be presented in the following sections.

5.3.1 Level A: Metrics Cluster Analysis

This first level of city grouping is the main differential over other models of smart cities and serves as a prerequisite for comparing cities respecting their particularities. At this level (Figure 5.3) three dimensions (Population, Development and Territory) and their respective variables (Territorial area, Demographic density, urban population, rural population, Income, Longevity and Education) are compared.

Figure 5.3: Domains and Level A Variables (Metrics).

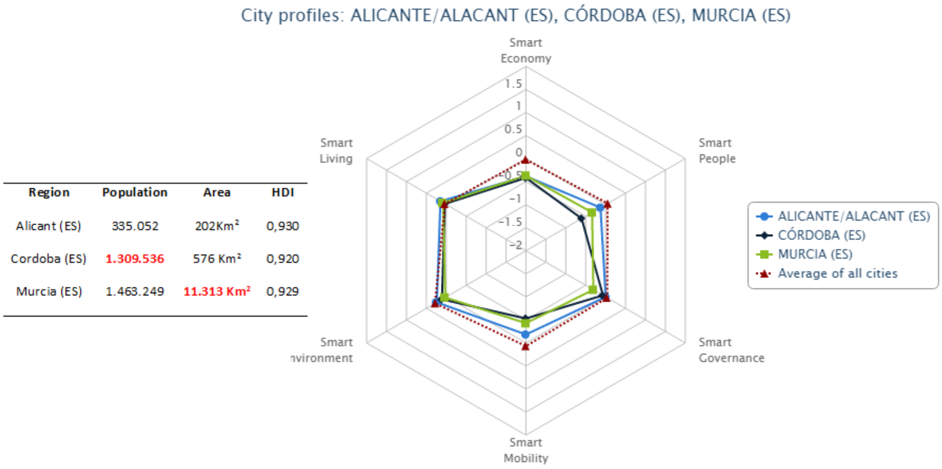


Source: Made by author

An example of grouping of this level was presented in Chapter 4 (Section 4.4.2), where it was possible to observe a set of cities grouped by the Domain mentioned here, thus making it possible to compare cities with similar Area and Demographic Density variables. This level aims to group municipalities using all variables and thus to allow cities to be grouped by technical indicators before a comparison on indicators of intelligence that will be proposed in the next Section.

Current SCMs do not distinguish cities according to their characteristics, and thus only SCM01 considers a specific type of cities for their data comparisons, midsize cities. Nevertheless, as in the following example, two regions (Provinces) are compared, although they are from the same country, have very different populations (Figure 5.4), which makes it impossible to compare the six characteristics of this SCM since it does not consider the difference between these realities.

Figure 5.4: Comparison between Spanish Provinces (SCM01)

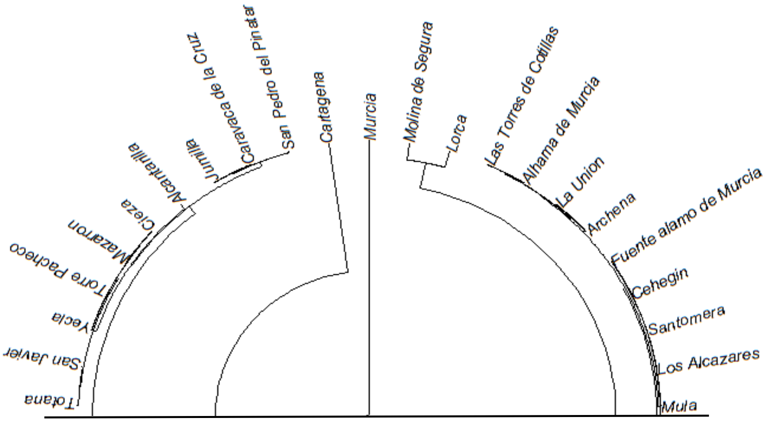


Region	Population	Area	HDI
Alicant (ES)	335.052	202Km²	0,930
Cordoba (ES)	1.309.536	576 Km²	0,920
Murcia (ES)	1.463.249	11.313 Km²	0,929

Source: Made by author

This type of comparison that does not consider the level of grouping by similarities can put in the same comparison group entire Provinces and even municipalities. In Figure 5.5 it is possible to see that through the method proposed by the SmartCluster metamodel, there is a grouping by similarity for municipalities within the Province of Murcia that is not treated by SCM01.

Figure 5.5: Clustering of Murcias Municipalities (ES) (SmartCluster).



Source: Made by author

The way of visualizing the clusters of this level by means of a dendrogram took into account the Euclidean distance (Ed) between the samples in the sample space. As the values obtained for Territory (t), Population (p) and Development (d), the Euclidean distance between any two cities ("City$_a$" and "City$_b$") can be calculated using Equation 5.1, where t, p and d represent the coordinates of any city, for the 3 variables in question.

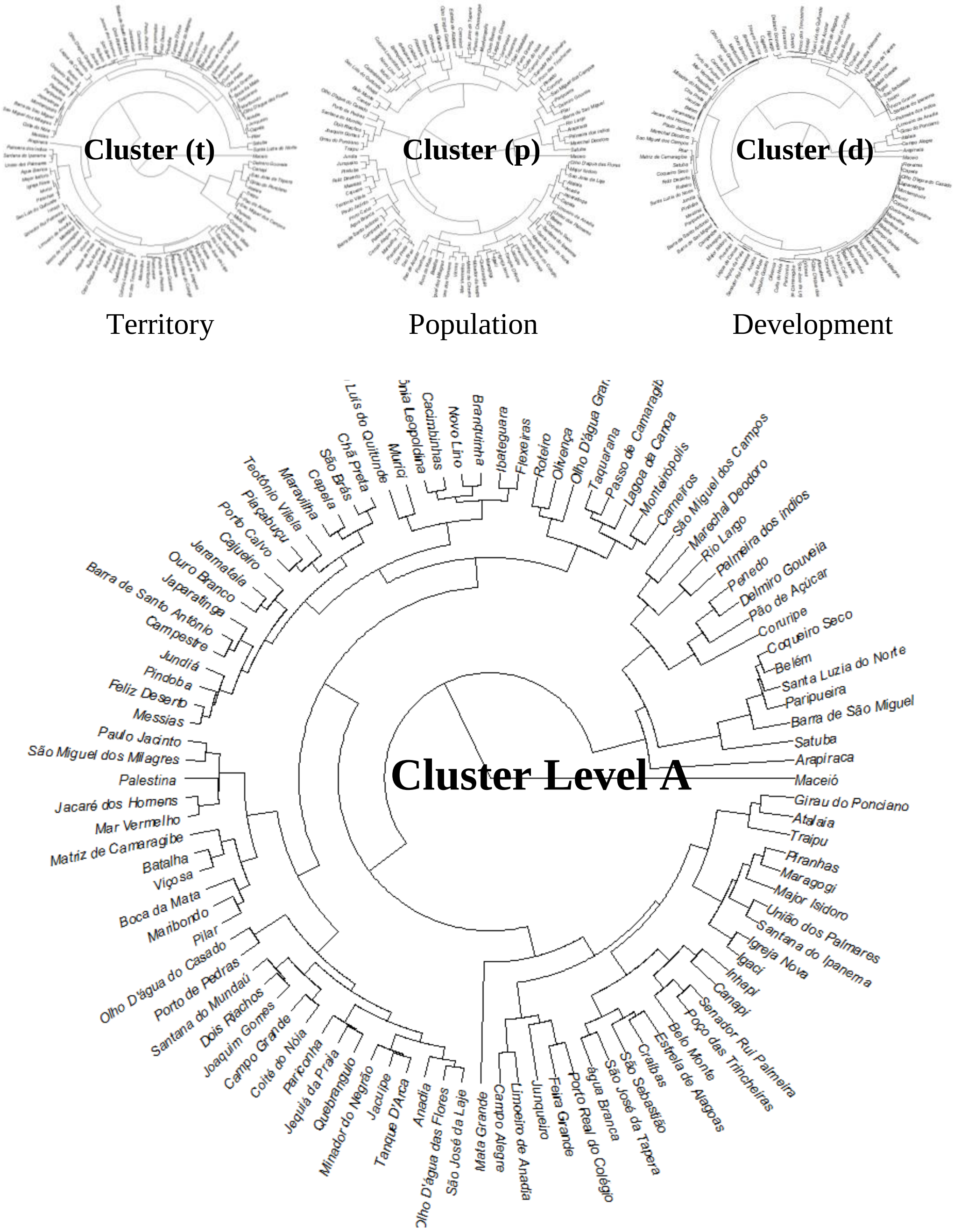
This calculation was performed for all cities of the three clusters, which merged together form the: **Cluster Level A** (Figure 5.6).

$$Ed(City_a City_b) = \sqrt{t(City_a - City_b)^2 + p(City_a - City_b)^2 + d(City_a - City_b)^2} \tag{5.1}$$

For the purpose of data visualization, Figure 5.5. Represents the 102 municipalities of the state of Alagoas. These municipalities were grouped into three categories (Territory, Population and Development) which allowed three different constructions of data dendrograms. In these dendrograms it is possible (at first sight) to identify that the capital of the state (i.e., Maceió) does not group with the other municipalities, since it presents indicators very distant from the others.

In the clusters (t and d) the clusters are less dense, which indicates a smaller variation between the distance of indicators between the municipalities, whereas the cluster p presents several levels of clusters, since the number of inhabitants between the municipalities varies a lot.

Figure 5.6: Cities of Alagoas clustered by Metrics of Level A.



Source: Made by author

5.3.2 Level B: Taxonomy Cluster Analysis

At this level, it will be understood that a grouping of cities will be carried out based on similarity of indicators that were pre-determined through a bibliographic survey (Chapter 2) in models of smart cities.

Using a Taxonomy of indicators presupposes the possibility of meeting the characteristics of cities that meet the demand and availability of public data to carry out such measurement. In addition to the availability of data, to create this taxonomy were considered the most basic indicators that make it possible to gauge the offer of public services, municipal management and infrastructure. As in Level A, at this level the distances calculated between cities of the same Cluster show a great variation (0.412 to 1.007). As the dendrogram in Figure 5.5 groups cities with such disparate indicators (Infrastructure; Services and Management), it is possible to see that larger distances imply in very different cities. In this way, the calculated distances can help in the search for similarity (or dissimilarity) between the cities, being easy to verify that the Cluster (i) is much less fragmented than the Cluster (s).

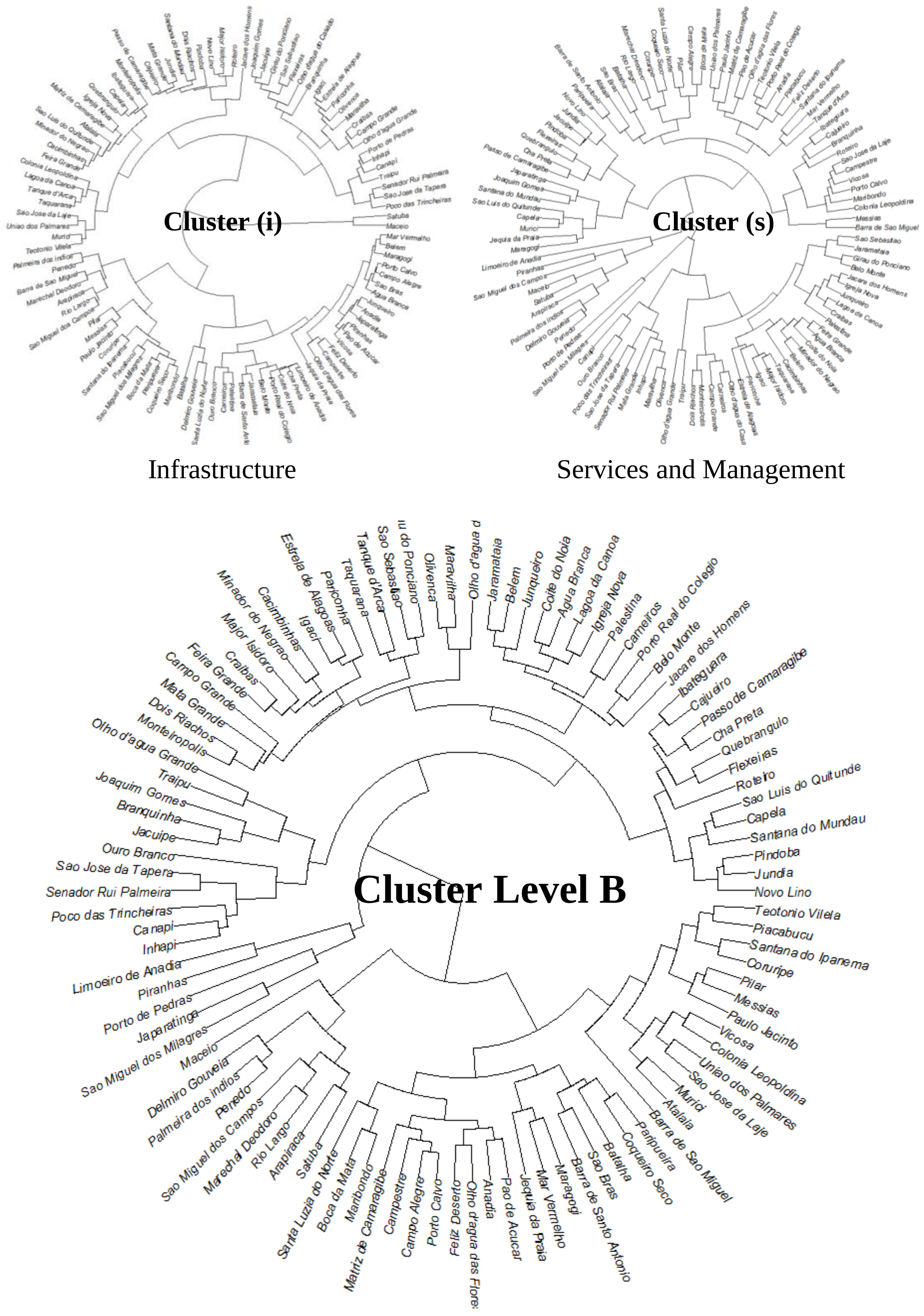
The calculation of the similarity index follows Equation 5.2 and was performed after all cities were grouped, with d($City_a City_b$) being the distance calculated between any two cities ("$City_a$" and "$City_b$") and dmax the largest calculated distance between cities. The advantage of using the similarity index as a scale instead of distance is that it always varies between 0 (if $d_{CityaCityb} = d_{max}$) and 1 (when cities are identical).

$$S(City_a City_b) = 1{,}0 - \frac{d_{(City_a City_b)}}{d_{(\max)}} \qquad (5.2)$$

A quick assessment of Cluster Level B (Figure 5.7), considering 1.077 (result of Equation 5.2 and tests with RStudio GraphView) as the threshold value for the similarity index, shows that the cities compared can be grouped into groups with a high index of similarity. One possible alternative to improve the discriminatory power of cities is the inclusion of more indicators in multivariate treatment. For this reason, and to make the inclusion of indicators a less expensive task, Chapter 3 presented the possibility of creating a Taxonomy of indicators, which can be easily changed or included to make cities more adherent to the type of Cluster that one wishes create.

In this Figure (5.7) it is possible to visualize that the city of Maceió now groups with other municipalities, since the dataset uses values of ten indicators instead of only three as in Level A. This makes the groupings more meticulous and thus, requires the threshold value quoted in the previous paragraph.

Figure 5.7: Alagoas Cities grouped by taxonomy of Smart Cities (Level B).



Source: Made by author

5.4 SmartCluster: Cluster analysis of Brazilian Smart Cities

The sections that compose grouping analysis (C.A.) using the SmartCluster metamodel will be presented in this Section, which describe in detail how the data were selected, processed, transformed, visualized and then interpreted.

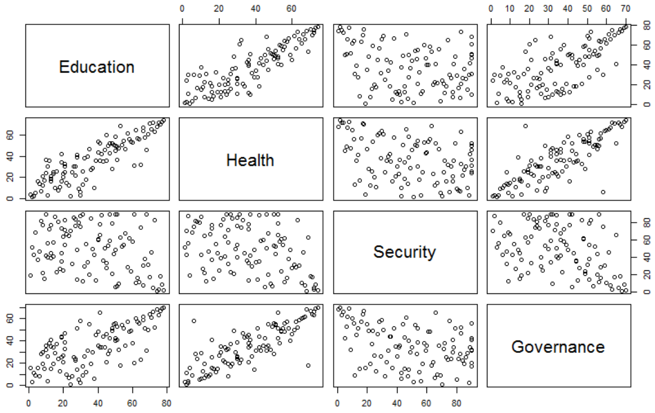
5.4.1 Select (Objectives and Smart Cities Indicators)

Problem Definition

It is intended to investigate the degree of similarity between indicators of smart cities and to define which common indicators can be improved to the detriment of solutions adopted by these cities in common. Thus, it is necessary to classify the city database into homogeneous groups according to the selected indicators. Once this classification was created, the study could be restricted to a specific group of cities, obtaining more varied and less costly results. The first difficulty that arose was how to treat cities with such disparate characteristics similarly. None of the SCMs found in the literature approached this classification, which makes the results of comparisons very distant from the socio-political reality of these municipalities.

Therefore, two levels of city grouping were created, the first one focused on the characteristics of comparison based on variables (Territory, Population and Development) and the second level is based on domains and their indicators (Water, Energy, Transportation, Health, Education , Etc.). Figure 5.8 it can be seen from the dispersion graph that the cities of Alagoas (Level B - Services and Management) follow a more homogeneous grouping pattern in relation to the Education and Health indicators, while the other variables present a broad distribution profile.

Figure 5.8: Cities of Alagoas grouped by Metrics



Source: Made by author

Among several benefits of using dispersion diagrams as a quality tool, one of particular importance is the possibility of inferring a causal relationship between variables, helping to determine the root cause of problems. In practice will be oftened need to study the relationship of correspondence between two variables. Dispersion Graphs can relate to and be interpreted as:

- **Positive correlation:** when an increase of x leads to an increase in y, so if will be controled x, y will also be controlled.
- **Possible positive correlation:** when an increase of x leads to an increase in y, so if will be controled x, y will also be controlled. However, there may be other factors that influence the behavior of variables.
- **Negative correlation:** when an increase of x leads to a decrease in y, so if will be controled x, y will also be controlled.
- **Possible negative correlation:** when an increase of x leads to a decreasing trend in y, so if will be controled x, y will also be controlled. But there may be other factors.
- **No correlation:** when one variable does not relate to the other.

What this thesis intends to show is that there is a positive correlation between these indicators. Clusters have shown that when there is an increase in one indicator, this leads to an increase in others, so when adopting a strategy to improve one indicator, others benefit. Scatterplots can be used as a tool for visualizing data correlation, but will be chosen to use dendrograms that provide a clearer and more didactic view of city grouping, although there is nothing to prevent a combination of the two being used in the future types of charts.

In this way, the main objective is to group the cities respecting the groupings in two levels, thus allowing an interpretation of the data of cities with a similar degree of similarity, and thus, to create a merge of dendrograms (Level A + Level B).

Therefore, it can be said that the objective of this section is to define the problem, and our problem is to prove that it is possible to measure and compare cities based on indicators obtained in public databases, and that these indicators influence each other. Once the problem has been defined, the next section presents a way to obtain and process this data.

5.4.2 Process (Acquisition and data processing to create Scores)

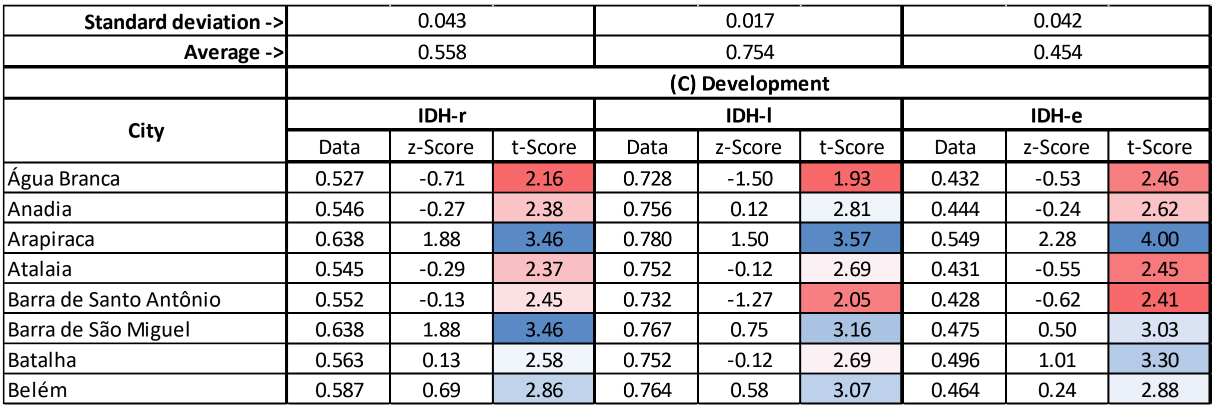
The three tasks (**Acquisition**, **Processing** and **Normalization**) that represent the basis for this Section dealing with the transformation of this data and the choice of the Cluster Analysis technique will be presented at this stage. In order to choose the ideal Cluster analysis technique, it is necessary to obtain the data and classify them in such a way that their values represent the same amplitude of measurement.

Acquisition

The data used for this example were divided into two levels (A and B), and the source of these data is basically the data that gave rise to levels A and B were obtained from public databases (Chapter 2), Having as main data source the data sources of IBGE (Brazilian Institute of Geography and Statistics).

These data were mined in these data sources, and fed spreadsheets on all the cities of all the states of the northeastern Brazilian region, so that they could be measured and compared. Figure 5.9 shows some of the cities in the state of Alagoas, with their respective data on the Development of this region.

Figure 5.9: Data on Development of some cities of Alagoas.



Standard deviation ->	0.043			0.017			0.042		
Average ->	0.558			0.754			0.454		
	(C) Development								
City	IDH-r			IDH-l			IDH-e		
	Data	z-Score	t-Score	Data	z-Score	t-Score	Data	z-Score	t-Score
Água Branca	0.527	-0.71	2.16	0.728	-1.50	1.93	0.432	-0.53	2.46
Anadia	0.546	-0.27	2.38	0.756	0.12	2.81	0.444	-0.24	2.62
Arapiraca	0.638	1.88	3.46	0.780	1.50	3.57	0.549	2.28	4.00
Atalaia	0.545	-0.29	2.37	0.752	-0.12	2.69	0.431	-0.55	2.45
Barra de Santo Antônio	0.552	-0.13	2.45	0.732	-1.27	2.05	0.428	-0.62	2.41
Barra de São Miguel	0.638	1.88	3.46	0.767	0.75	3.16	0.475	0.50	3.03
Batalha	0.563	0.13	2.58	0.752	-0.12	2.69	0.496	1.01	3.30
Belém	0.587	0.69	2.86	0.764	0.58	3.07	0.464	0.24	2.88

Source: Made by author

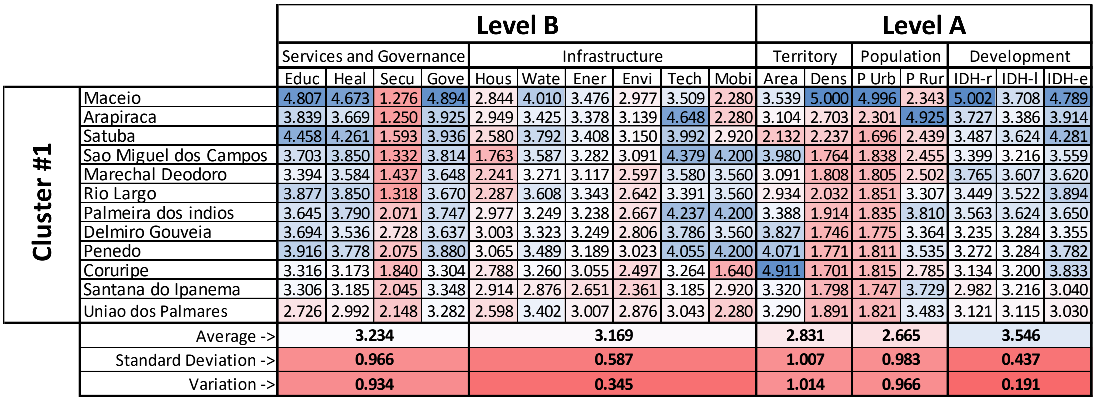
This fragment of the spreadsheet reveals the data referring to the HDI of these cities, of only one among the eight northeastern cities. The next task was to normalize these data, and turn them into heatmaps that served to define the city clusters.

Processing and Standardization

This task consisted of processing the data of these worksheets, and thus standardizing them so that they served the same order of magnitude. In Figure 5.7, in addition to the raw data obtained on the HDI of these cities, it is possible to observe the transformation of these data into a z-Score, and then to a t-Score that ranged from 1 to 5 (Hair et al., 2005). Thus, the standardization of the variables was the adequate procedure that allowed to minimize the effect of different scales of measures of the variables and indicators, making all the data have equivalent importance in the definition of groups (Barroso & Arties, 2003; Corrar et al., 2007).

This normalization of data by the z-score method together with the multivariate analysis model for the standardization of variables and indicators has proved quite feasible in the context of this work since these methods presuppose the use of a high number of strongly related observations. Figure 5.10 shows the data already standardized, standardized and grouped, which allows a first comparison between the smart cities of Alagoas.

Figure 5.10: Alagoas Smart Cities grouped by Metrics (Level A) and Taxonomy (Level B)



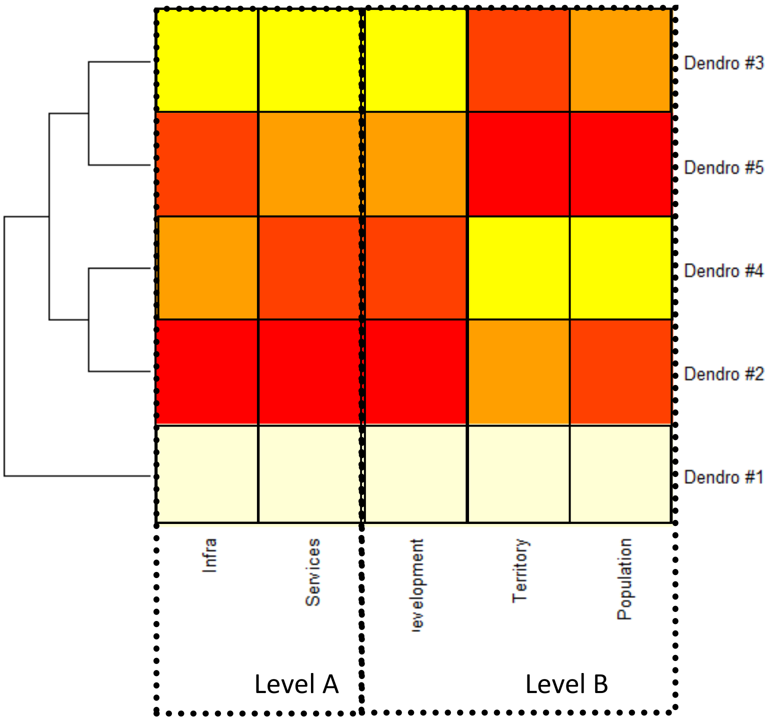
		Level B										Level A						
		Services and Governance				Infrastructure						Territory		Population		Development		
		Educ	Heal	Secu	Gove	Hous	Wate	Ener	Envi	Tech	Mobi	Area	Dens	P Urb	P Rur	IDH-r	IDH-l	IDH-e
Cluster #1	Maceio	4.807	4.673	1.276	4.894	2.844	4.010	3.476	2.977	3.509	2.280	3.539	5.000	4.996	2.343	5.002	3.708	4.789
	Arapiraca	3.839	3.669	1.250	3.925	2.949	3.425	3.378	3.139	4.648	2.280	3.104	2.703	2.301	4.925	3.727	3.386	3.914
	Satuba	4.458	4.261	1.593	3.936	2.580	3.792	3.408	3.150	3.992	2.920	2.132	2.237	1.696	2.439	3.487	3.624	4.281
	Sao Miguel dos Campos	3.703	3.850	1.332	3.814	1.763	3.587	3.282	3.091	4.379	4.200	3.980	1.764	1.838	2.455	3.399	3.216	3.559
	Marechal Deodoro	3.394	3.584	1.437	3.648	2.241	3.271	3.117	2.597	3.580	3.560	3.091	1.808	1.805	2.502	3.765	3.607	3.620
	Rio Largo	3.877	3.850	1.318	3.670	2.287	3.608	3.343	2.642	3.391	3.560	2.934	2.032	1.851	3.307	3.449	3.522	3.894
	Palmeira dos indios	3.645	3.790	2.071	3.747	2.977	3.249	3.238	2.667	4.237	4.200	3.388	1.914	1.835	3.810	3.563	3.624	3.650
	Delmiro Gouveia	3.694	3.536	2.728	3.637	3.003	3.323	3.249	2.806	3.786	3.560	3.827	1.746	1.775	3.364	3.235	3.284	3.355
	Penedo	3.916	3.778	2.075	3.880	3.065	3.489	3.189	3.023	4.055	4.200	4.071	1.771	1.811	3.535	3.272	3.284	3.782
	Coruripe	3.316	3.173	1.840	3.304	2.788	3.260	3.055	2.497	3.264	1.640	4.911	1.701	1.815	2.785	3.134	3.200	3.833
	Santana do Ipanema	3.306	3.185	2.045	3.348	2.914	2.876	2.651	2.361	3.185	2.920	3.320	1.798	1.747	3.729	2.982	3.216	3.040
	Uniao dos Palmares	2.726	2.992	2.148	3.282	2.598	3.402	3.007	2.876	3.043	2.280	3.290	1.891	1.821	3.483	3.121	3.115	3.030
	Average ->	**3.234**				**3.169**						**2.831**		**2.665**		**3.546**		
	Standard Deviation ->	**0.966**				**0.587**						**1.007**		**0.983**		**0.437**		
	Variation ->	**0.934**				**0.345**						**1.014**		**0.966**		**0.191**		

Source: Made by author

This process of normalizing the raw data mining allows to compose the Clusters (#1 to #5) and, thus, to compose new heatmaps that indicate that the hotter (red), the worse the grouping indicators, and the colder (White), the better these indicators are. It is identified visually that **Cluster#1** is the one whose indicators are better, which makes it difficult to be grouped with the others, since this cluster is composed of the cities that obtained the best values for the indicators measured. Therefore, although no method of interpretation of these data has been used, it is possible for the municipal manager to have an idea of which groups of cities are closest to the reality in which he is inserted.

Figure 5.11 shows this difference between clusters (**Clusters#1** to **#5**) according to Levels A and B. It is possible to notice that there is a clear difference between **Cluster#1** and **Cluster#2**, which is represented by colors which indicate that the cities belonging to Cluster#2 obtained lower values than **Cluster#1**.

Figure 5.11: Heatmap that originated the individual Dendrograms by Clusters.



Source: Made by author

The first step of this methodology was to define the problem, and this step showed how to obtain and normalize the data. The next section (more technical and formal) shows how this data can be mathematically grouped to transform a set of normalized data into pointers that influence each other and allow you to create Smart City clusters. The following section deals with the criteria used to define the groupings and the subsequent implementation of Group Analysis (C.A).

5.4.3 Transform (Similarity criteria and implementation of a C.A.)

This phase is important to define the homogeneity criteria of the groups, thus, different criteria lead to different homogeneous groups, and the type of homogeneity depends on the objectives to be achieved. In Section 5.3.1. It was defined that the main objective is to group the cities according to the junction of the levels A (Metrics) and Level B (Taxonomy) indicators.

The mechanism used to perform the grouping is quite simple and follows the steps proposed in (Lattin, 2011). It starts with each type of city in its own isolated grouping, that is, "n" size groupings 1. At each stage of the process, two "closer" groupings are found and the two groups are joined together. This step is repeated until a cluster of size "n" remains. The following steps were followed:

Step 1: All objects were grouped separately into five large groups of similarity represented respectively by C_1, C_2, C_3, C_4 and $C_{5.}$ In this step, the distance between two clusters was defined as ($d_{CiCj} = d_{ij}$) and (t = 1) was defined as the index of the iterative process;

Step 2: The shortest distance between the clusters was defined by: (C_i and C_j);

Step 3: The distance (Ci and Cj) was combined to form new clusters called (C_n + 1);

Step 4: The distance between the new clusters was defined by (C_n + t) and all clusters C_k as follows: ($d_{C\ (n+t)}\ C_{(k)} = \min\{d_{C(i)}), DC_{(j)}\ C_{(k)})$;

Step 5: New clusters were added ($C_{n\ +\ t}$) removing the above (C_1 and C_3). At each new interaction, the index (t = t +1) and

Step 6: With each new iteration, steps 4 to 6 were executed again.

The classification methods can be defined in two main types: (a) Hierarchical, where objects are assigned to groups that are arranged in groups, as in a dendrogram and (b) Non-hierarchical, where objects are assigned to groups. The methods are also classified as: (c) Agglomeration, where the analysis starts from the objects joining them, or from (d) Division, where all objects begin as members of a single group and this group is repeatedly divided. For computational and presentation reasons hierarchical-agglomerative methods are the most popular, and so this work makes use of this method and thus, defines the main components (SNEATH and SOKAL, 1973).

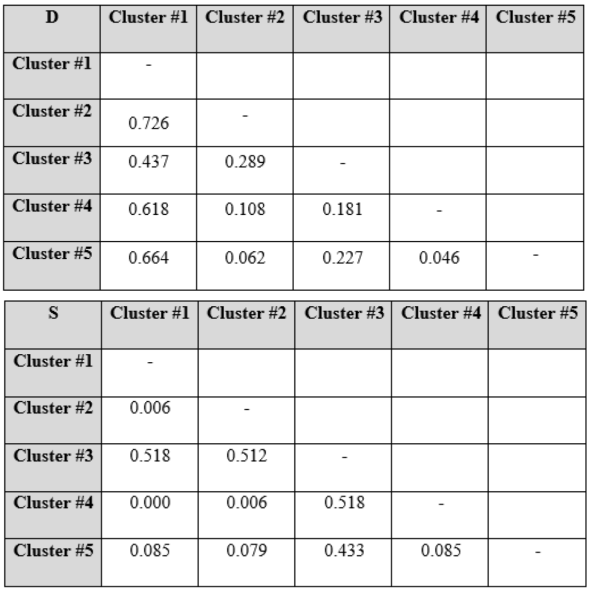
For this work the hierarchical method was chosen, which allowed the data to be partitioned successively, producing a hierarchical representation of the groupings, and thus, facilitated the visualization on the groupings, as well as the perception of the degree of similarity between them. This type of method is widely used because it does not require grouping number definitions while offering the facility to deal with any measure of similarity used. (BERKHIN, 2002).

To apply the principal component analysis, you must follow a few steps until you get the final result. Initially, the matrix S is calculated and it is checked if the variables are correlated in relation to each other. It is important that the variables used to give rise to the components of the cluster have been normalized.

The criterion for defining the final number of groups (Stop Criterion) can be restricted when reaching a certain number of groupings or when some type of stop condition occurs. Such a criterion requires a distance matrix between the groupings, called the similarity matrix (JAIN AND DUBES, 1988).

In Section 5.3.2 the standardization of the indicators was presented so that this model of Cluster Analysis could be implemented. The result of the initial clustering can be observed through the distance and similarity matrix (Figure 5.12).

Figure 5.12: Distance and similarity matrices



D	Cluster #1	Cluster #2	Cluster #3	Cluster #4	Cluster #5
Cluster #1	-				
Cluster #2	0.726	-			
Cluster #3	0.437	0.289	-		
Cluster #4	0.618	0.108	0.181	-	
Cluster #5	0.664	0.062	0.227	0.046	-

S	Cluster #1	Cluster #2	Cluster #3	Cluster #4	Cluster #5
Cluster #1	-				
Cluster #2	0.006	-			
Cluster #3	0.518	0.512	-		
Cluster #4	0.000	0.006	0.518	-	
Cluster #5	0.085	0.079	0.433	0.085	-

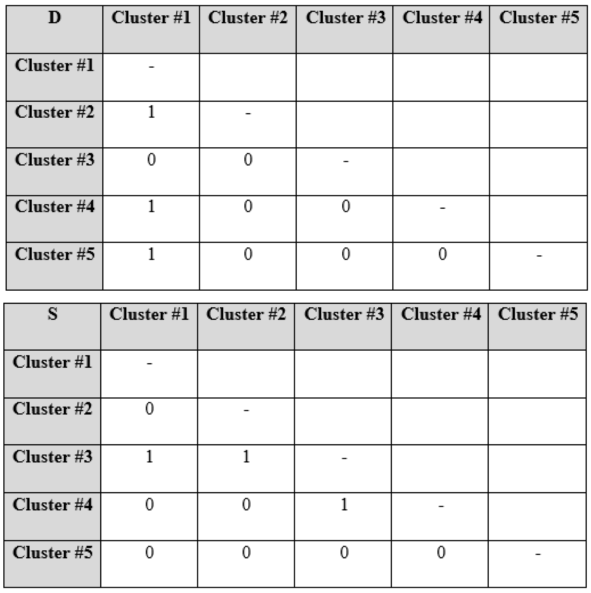
Source: Made by author

The cophenetic distance (or similarity) between two objects x1 and x2 are defined by the level of distance (or similarity) between these objects within a group. (Jain & Dubes, 1988). The distance calculation between these objects can be used to compose a cophenetic matrix (Sokal & Rohlf, 1962).

In the next step, it is decided by the total number of components that will best explain the set of original variables. The components were selected using the criterion suggested by Kaiser (Kaiser, 1960), which consists of including only components whose values are greater than 1 (Figure 5.13).

This criterion has as main characteristic, to include few components when the number of original variables is less than twenty and, in this case, as the sample used is of 102 municipalities, it was possible to reach a cumulative variance of around 70%.

Figure 5.13: Cophenetic matrices of distance and similarity



D	Cluster #1	Cluster #2	Cluster #3	Cluster #4	Cluster #5
Cluster #1	-				
Cluster #2	1	-			
Cluster #3	0	0	-		
Cluster #4	1	0	0	-	
Cluster #5	1	0	0	0	-

S	Cluster #1	Cluster #2	Cluster #3	Cluster #4	Cluster #5
Cluster #1	-				
Cluster #2	0	-			
Cluster #3	1	1	-		
Cluster #4	0	0	1	-	
Cluster #5	0	0	0	0	-

Source: Made by author

Ward method

According to Hair et al. (2005), the Ward method consists of a hierarchical grouping procedure in which the measure of similarity used to join groupings is calculated as the sum of squares between the two groupings made on all variables. This method tends to result in approximately equal size groupings due to their minimization of internal variation.

At each stage, will be combined the two clusters that present the smallest increase in the global sum of squares within the clusters. The use of this method explains, for example, the great value found in the variation of **Cluster#1**, which indicates that this cluster contains elements whose averages represent great distance from each other.

Table 5.6: Averages and Standard Deviation of Clusters# 1 to #5.

Variables	Cluster#1		Cluster#2		Cluster#3		Cluster#4		Cluster#5	
	A	S	A	S	A	S	A	S	A	S
Level A										
Services	3,234	0,966	2,440	0,424	2,860	0,268	2,555	0,357	2,711	0,410
Infra	3,196	0,609	2,479	0,616	2,870	0,594	2,655	0,537	2,468	0,634
Level B										
Territory	2,831	1,007	2,289	0,674	2,349	0,672	2,352	0,809	2,196	0,525
Population	2,665	0,983	2,309	0,685	2,532	0,962	2,481	0,909	2,188	0,534
Development	3,546	0,437	2,300	0,412	2,857	0,349	2,385	0,384	2,647	0,448

*A = Average, S = Standard Deviation

Source: Made by author

The mean values in Table 5.1 range between 2.1 and 3.2 and the variation values are between 0.3 and 0.9. Regarding variation, the closer to one, the better the representation, and the closer to zero will be worse.

It is a consensus among the models of dendrograms studied, that a coefficient around 0.8 already can be considered a good fit for the generation of clusters. **Cluster#1** was maintained with high variation values to exemplify the difference between it and the other clusters. In the next section are presented the characteristics related to the visualization of these groupings, which was done with the help of the graphs in the form of Dendrograms.

This third step was responsible for the mathematical formalization that groups the municipalities by similarity and proximity of indicators and practically ends the process of acquisition, normalization and formalization of Clusters. The next step is to generate the visualization of these clusters and perform the analysis of these clusters.

5.4.4 View (Visualizing clusters)

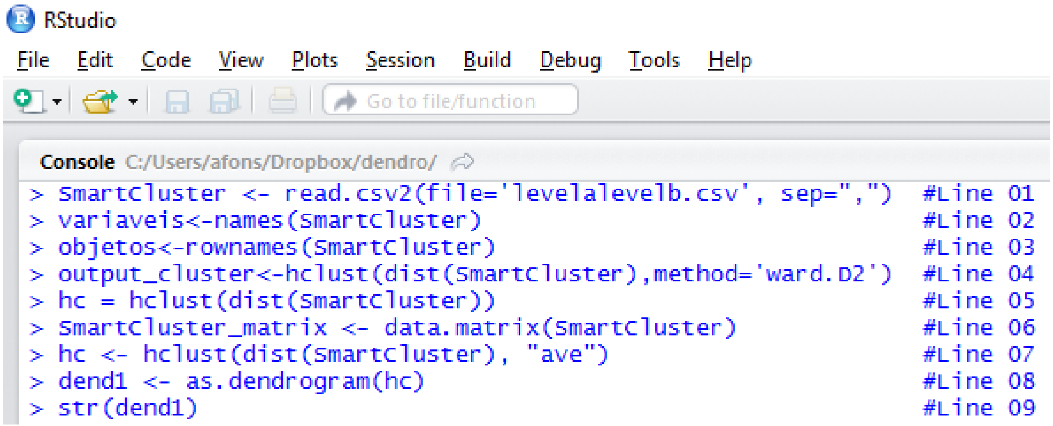
In the previous steps, calculations were performed to obtain matrices of distance and similarity between the cities of each group, and after several calculations in the post-processing stage, a complex visualization through dendrograms occurs to try to identify patterns and, consequently, to obtain some Knowledge about these groupings.

The process of generating graphs of clusters can meet the properties of: density, variance, size, shape and separation. Based on these properties, there are clusters that can be classified as hyperspherical, elongated, curvilinear or may have more differentiated structures (ALDENDERFER and BLASHFIELD, 1984).

The clusters created in this work respected the hierarchical agglomerative configuration taxonomy because of the adopted CA strategy and the desired objectives, making it clear that different taxonomies (hierarchical divisive, iterative partitioning, density analyzing, factor analytic, clumping and graphtheoretic) when applied to a Data, generate different results (EVERITT, 2001) and (BERKHIN, 2002).

Using the hierarchical taxonomy, it was possible to implement the methods with the features of the programming language R, in version 3.2.3, (Venables, 2005) using RStudio (Raccine, 2012) as a graphical interface. As an illustrative example, Figure 5.14 shows a fragment of the source code used to generate the dendrograms used for city comparison in this work.

Figure 5.14: RStudio graphical interface



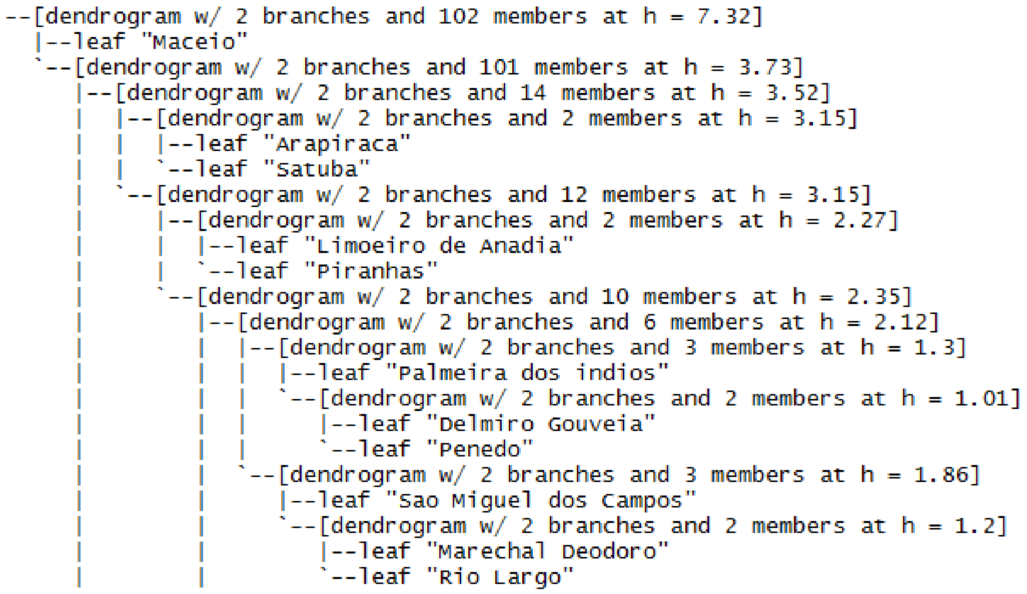
```
> SmartCluster <- read.csv2(file='levelalevelb.csv', sep=",")  #Line 01
> variaveis<-names(SmartCluster)                                #Line 02
> objetos<-rownames(SmartCluster)                               #Line 03
> output_cluster<-hclust(dist(SmartCluster),method='ward.D2')   #Line 04
> hc = hclust(dist(SmartCluster))                               #Line 05
> SmartCluster_matrix <- data.matrix(SmartCluster)              #Line 06
> hc <- hclust(dist(SmartCluster), "ave")                       #Line 07
> dend1 <- as.dendrogram(hc)                                    #Line 08
> str(dend1)                                                    #Line 09
```

Source: Made by author

The source code for generating this cluster is started by reading a CSV (Comma Separated Values) file that is converted into a data matrix called SmartCluster (Line

01). This file is the compilation of city data that has been normalized and standardized to obtain the averages of each of the variables and indicators. The following steps in this source code transcribe the creation of the structure (Line 04) required for the matrix that will be used to create the dendrograms. Figure 5.15 shows the tree composed of the cities, their values (Line 06) in the dendrogram tree and the "leaves" structure (Line 09) that makes up the tree.

Figure 5.15: Structure of Cluster#1 of smart cities.

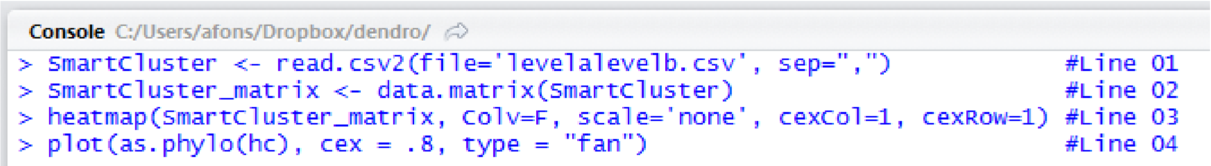


```
--[dendrogram w/ 2 branches and 102 members at h = 7.32]
  |--leaf "Maceio"
  `--[dendrogram w/ 2 branches and 101 members at h = 3.73]
     |--[dendrogram w/ 2 branches and 14 members at h = 3.52]
     |  |--[dendrogram w/ 2 branches and 2 members at h = 3.15]
     |  |  |--leaf "Arapiraca"
     |  |  `--leaf "Satuba"
     |  `--[dendrogram w/ 2 branches and 12 members at h = 3.15]
     |     |--[dendrogram w/ 2 branches and 2 members at h = 2.27]
     |     |  |--leaf "Limoeiro de Anadia"
     |     |  `--leaf "Piranhas"
     |     `--[dendrogram w/ 2 branches and 10 members at h = 2.35]
     |        |--[dendrogram w/ 2 branches and 6 members at h = 2.12]
     |        |  |--[dendrogram w/ 2 branches and 3 members at h = 1.3]
     |        |  |  |--leaf "Palmeira dos indios"
     |        |  |  `--[dendrogram w/ 2 branches and 2 members at h = 1.01]
     |        |  |     |--leaf "Delmiro Gouveia"
     |        |  |     `--leaf "Penedo"
     |        |  `--[dendrogram w/ 2 branches and 3 members at h = 1.86]
     |        |     |--leaf "Sao Miguel dos Campos"
     |        |     `--[dendrogram w/ 2 branches and 2 members at h = 1.2]
     |        |        |--leaf "Marechal Deodoro"
     |        |        `--leaf "Rio Largo"
```

Source: Made by author

The transcription of this tree to the dendrogram through command lines is given by converting the data into a computable array (Line 02) that is formatted in a Heatmap (Line 03) and thus, the printed version of the dendrogram (Line04) is generated. These commands are described in detail in Figure 5.15.

Figure 5.16: Commands for Dendrogram generation.

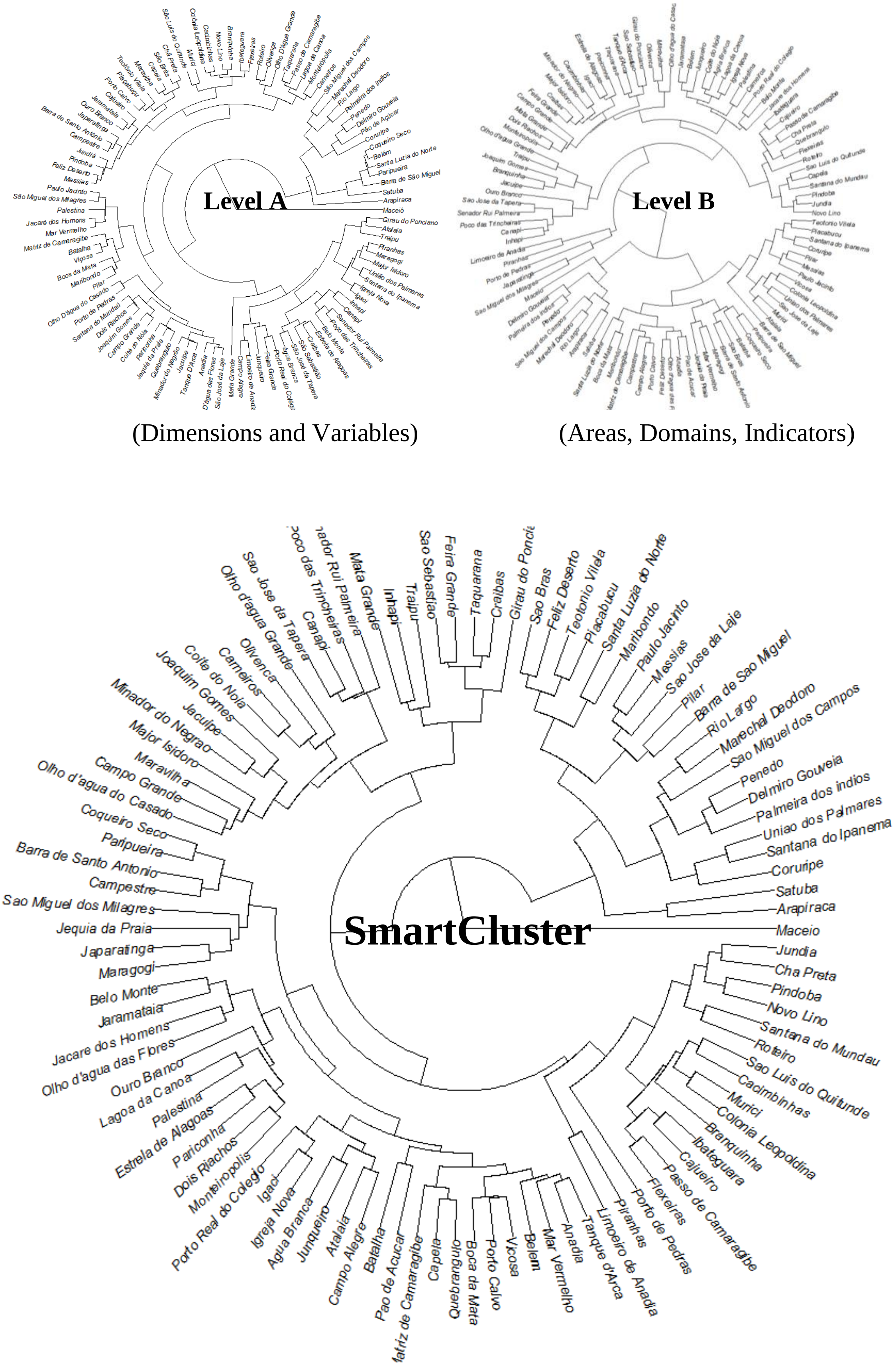


```
> SmartCluster <- read.csv2(file='levelalevelb.csv', sep=",")          #Line 01
> SmartCluster_matrix <- data.matrix(SmartCluster)                      #Line 02
> heatmap(SmartCluster_matrix, Colv=F, scale='none', cexCol=1, cexRow=1) #Line 03
> plot(as.phylo(hc), cex = .8, type = "fan")                            #Line 04
```

Source: Made by author

The dendrogram resulting from the union of Level A and Level B Clusters is shown in Figure 5.17 and named SmartCluster.

Figure 5.17: SmartCluster: Smart Cities of Alagoas



(Dimensions and Variables)

(Areas, Domains, Indicators)

Source: Made by author

5.4.5 Interpretation (Dendrograms Analysis)

It is possible to extract many interpretations based on the relation of individuals (Cities) that form the clusters, so many which depend on the type of approach, the purpose of the analysis and the particular vision of the proposed study. The interpretation of these clusters cannot be asserted as an exact science, nor is there so little to know, how exactly the set of interpretations can be extracted from one or several clusters. Valentim (Valentim, 2000), recommends the use of three rules for the interpretation of dendrograms, being:

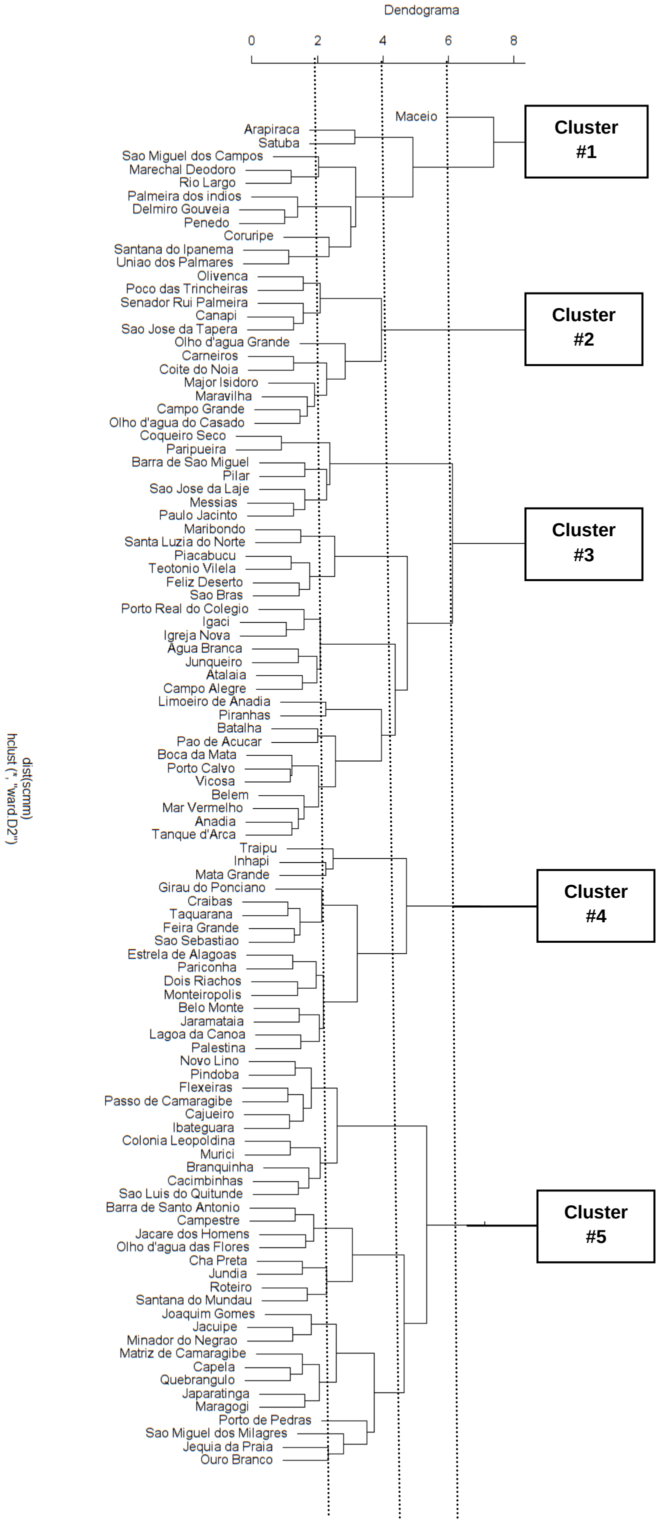
1) Detail in the dendrogram produced, for each grouping, its characteristics and aspects of similarity between individuals and aspects of dissimilarity in relation to individuals from other groups;
2) Perform the reading of the dendrograms data starting from the lowest similarity values for the largest ones, thus, the groupings with more individuals will be initially interpreted, making it possible to formulate hypotheses about the smallest groupings (which may be the most complex);
3) When possible, develop, in parallel, with the same data, a sort analysis, which will show the factors responsible for the groupings.

The number of groups defined in the dendrogram constitute a proposition about the basic and unknown organization of the data and generally the clustering algorithms do not present solution to determine the ideal number of these groups, so, one way to determine the number of groups is by examining the Dendrogram. The dendrogram is a tree-shaped graph that reveals the changes in similarity levels for the successive stages of grouping where the vertical axis represents the level of similarity and the horizontal axis the individuals. One way to read the similarity of individuals is to see if the vertical lines starting from the grouped individuals have corresponding height.

Due to the inexistence of a method to select the best grouping technique, it is important to evaluate the degree of fit of the grouping, and for this, will be used the co-optic correlation coefficient (CCC), proposed by Sokal & Rohlf (1962). It states that the higher the CCC, the better the result of the grouping, provided that a CCC of less than 0.7 indicates an inadequate grouping method (Rohlf, 1970). Figure 5.18 presents a

Dendrogram of the groupings of the cities of Alagoas by Levels A and B presented in Sections 5.2.1 and 5.2.2.

Figure 5.18: SmartCluster: Five Clusters of Smart Cities of



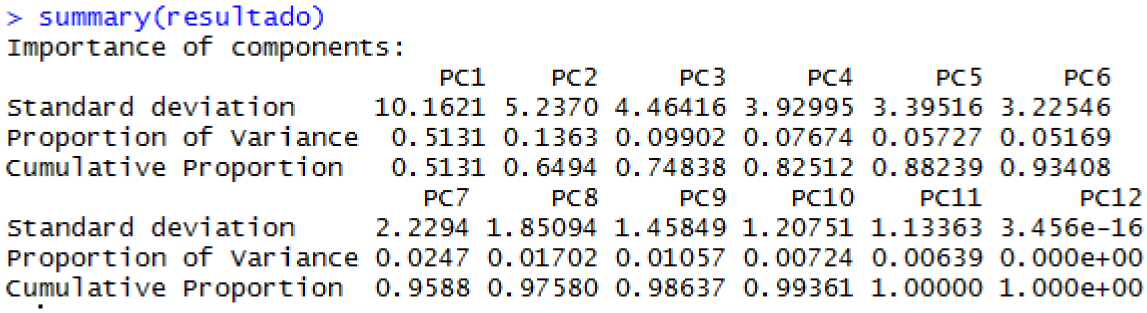
Source: Made by author

The number of variables and indicators in the Smart City pattern recognition studies is high, and the graphical representation of the complete data set facilitates the interpretation of the results. Some algorithms were used to elaborate and interpret graphs that represent the greatest possible amount of information about these cities. Among them, the hierarchical grouping analysis (HCA) and the principal component analysis (PCA) stand out. (Beebe, 1997; Sharaf, 1986))

The HCA and PCA analyzes allow the graphical visualization of the entire data set, and especially of these Clusters, by containing a high number of cities and meeting the main objective of increasing the comprehension of the data set by examining the presence or absence Of natural groupings between cities. Sharaf (Sharaf, 1986) classifies both as exploratory or unsupervised, since information on the names of cities is not considered, but rather the values referring to indicators and variables

In Figure 5.19 will be presented the Principal Components Analysis of **Cluster#1** with their respective variation values within the Cluster, where the NCPs (Number of dimensions) represent respectively the cities: (PC1) Maceió, (PC2) Arapiraca, (PC3) Satuba, (PC4) São Miguel dos Campos, (PC5) Marechal Deodoro, (PC6) Rio Largo, (PC7) Palmeira dos índios, (PC8) Delmiro Gouveia, (PC9) Penedo, (PC10) Coruripe, (PC11) Santana do Ipanema e (PC12) União dos Palmares.

Figure 5.19: Cluster Core Components Analysis Values # 1.



```
> summary(resultado)
Importance of components:
                           PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     10.1621 5.2370 4.46416 3.92995 3.39516 3.22546
Proportion of Variance  0.5131 0.1363 0.09902 0.07674 0.05727 0.05169
Cumulative Proportion   0.5131 0.6494 0.74838 0.82512 0.88239 0.93408
                          PC7     PC8     PC9    PC10    PC11      PC12
Standard deviation     2.2294 1.85094 1.45849 1.20751 1.13363 3.456e-16
Proportion of Variance 0.0247 0.01702 0.01057 0.00724 0.00639 0.000e+00
Cumulative Proportion  0.9588 0.97580 0.98637 0.99361 1.00000 1.000e+00
```

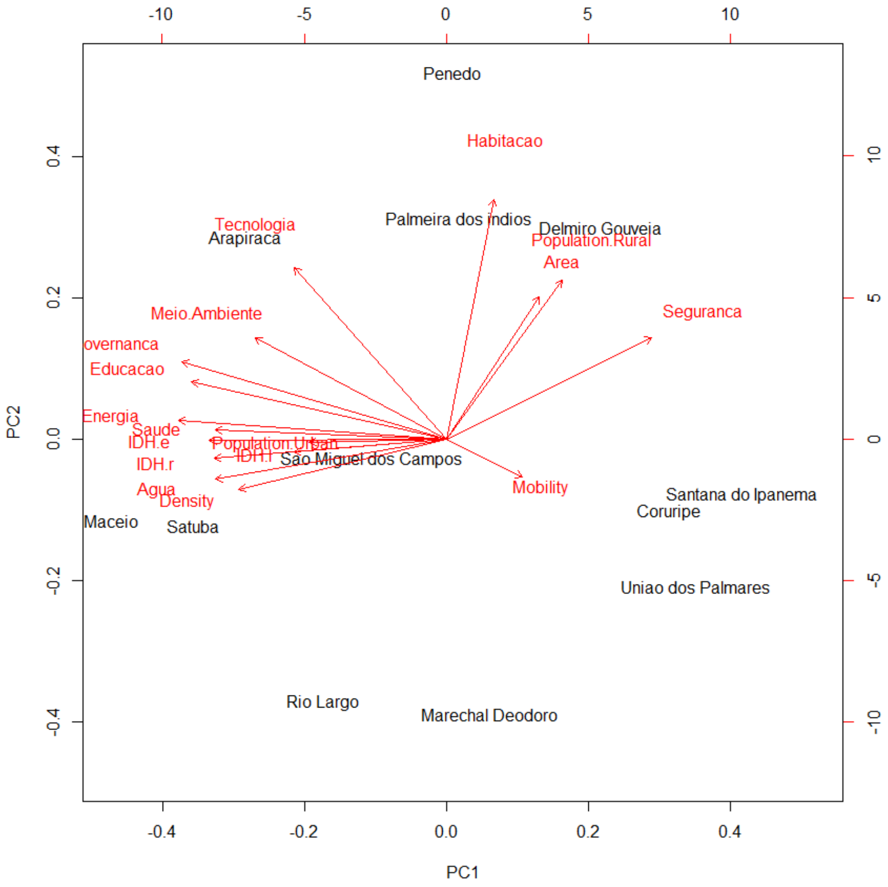
Source: Made by author

This analysis allows for example to understand the differences between the proportions of variation of each of the components within its Cluster, and thus, it is possible to group cities with similar variations, thus obtaining more homogeneous groups.

The grouping by similarity of characteristics is the object of desire of this thesis, so that it can be affirmed that certain cities can use solutions to solve their problems, according to other solutions adopted by similar cities.

The HCA grouped the cities into clusters, based on the similarity of the standard values respecting the set of variables (seven) and indicators (ten) of the data set (Figure 5.20), while the PCA reduces the size of the original data set, preserving the largest amount of information (variance) possible. This reduction is obtained through the establishment of new variables, called main components (PCs). (Christie, 1995).

Figure 5.20: Scatter plot of PCA Analysis of Cluster#1.



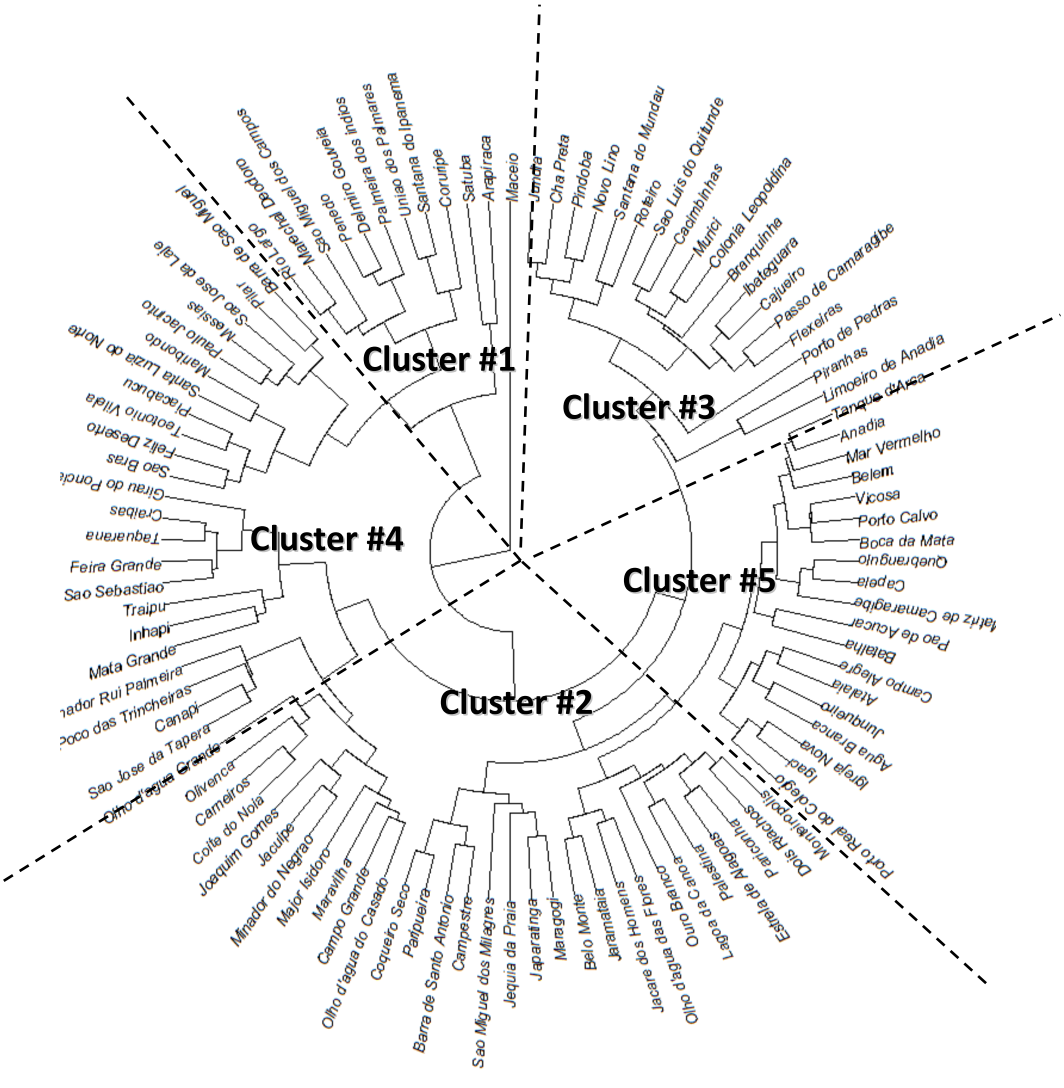
Source: Made by author

Both HCA and PCA allow the multivariate interpretation of large and complex datasets by means of bi or three-dimensional graphs. These graphs present information that expresses the interrelationships that may exist between the variables, facilitating the multivariate interpretation of the behavior of the samples. (Correia, 2007)

In general, two types of PCA are constructed: the covariance PCA and the correlation PCA. In this specific case, the correlation PCA is more appropriate because it contains variables that were measured in different units and mainly because the variance of each variable is very different from each other. (Beebe, 1997)

Using these analyzes, the reading of Smart City dendrograms becomes clearer in that it is understood that the separation by groups attends to predefined models and calculations to generate these groupings by similarity, which is not accomplished by the currently existing models, Which insist on the comparison of "pure" indicators, which as explained in Chapter 3, can generate serious misinterpretations of these data sets and cities. Figure 5.21 shows the dendrogram with the visual demarcation of these groupings, which will be better detailed in the tables that will follow.

Figure 5.21: SmartCluster: Clusters of Alagoas



Source: Made by author

Table 5.7: Average and variation of SmartCluster Cities

Cluster #1 – Average 3,089 ± 0,690 Maceió, Arapiraca, Satuba, São Miguel dos Campos, Marechal Deodoro, Rio Largo Palmeira dos Índios, Delmiro Gouveia, Penedo, Coruripe, Santana do Ipanema, União dos Palmares
Cluster #2 – Average 2,863 ± 0,330 Olivença, Canapi, Poço das Trincheiras, Senador Rui Palmeira, Campo Grande, Carneiros Coite do Noia, Major Isidoro, Maravilha, Olho d'agua do Casado, Olho d'agua Grande, São Jose da Tapera
Cluster #3 – Average 2,652 ± 0,301 Barra de São Miguel, Coqueiro Seco, Feliz Deserto, Maribondo, Messias, Paripueira, Paulo Jacinto, Piaçabuçu, Pilar, Santa Luzia do Norte, São Bras, São Jose da Laje, Teotônio Vilela, Agua Branca, Anadia, Atalaia, Batalha, Belem, Boca da Mata, Campo Alegre, Igaci, Igreja Nova, Junqueiro, Limoeiro de Anadia, Mar Vermelho, Pão de Açúcar, Piranhas, Porto Calvo, Porto Real do Colegio, Tanque d'Arca, Viçosa
Cluster #4 – Average 2,471 ± 0,377 Belo Monte, Craíbas, Dois Riachos, Estrela de Alagoas, Feira Grande, Girau do Ponciano, Inhapi, Jaramataia, Lagoa da Canoa, Mata Grande, Monteirópolis, Palestina, Pariconha, São Sebastiao, Taquarana, Traipu
Cluster #5 – Average 2,425 ± 0,246 Barra de Santo Antônio, Branquinha, Cacimbinhas, Cajueiro, Campestre, Cha Preta, Colonia, Leopoldina, Flexeiras, Ibateguara, Jacare dos Homens, Jundia, Murici, Novo Lino, Olho d'agua das Flores, Passo de Camaragibe, Pindoba, Sao Luis do Quitunde, Capela, Jacuipe, Japaratinga, Jequia da Praia, Joaquim Gomes, Maragogi, Matriz de Camaragibe, Minador do Negrao, Ouro Branco, Porto de Pedras, Quebrangulo, Roteiro, Santana do Mundau, Sao Miguel dos Milagres

Source: Made by author

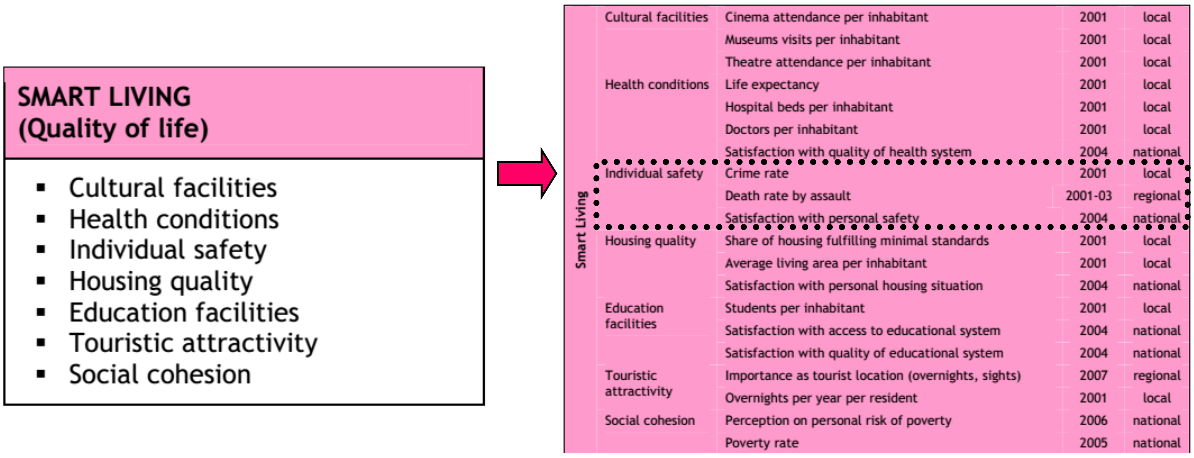
In **Cluster#1**, which is composed of 12 cities, the difference between city profiles is quite visible and this difference can be confirmed through the variance value (0.690) which is the highest among all the clusters. This was the cluster used in the previous examples and reveals the differences between municipalities that need to be taken into consideration before any comparisons.

If other models of smart cities were used, this grouping would not allow, for example, to compare the cities of Satuba and Arapiraca, since respectively they occupy in the ranking of number of inhabitants the positions of second and thirty third.

When these cities are ranked, Maceio (the state capital) appears as the best city ranked in the indicators of Education, Health and Governance, and yet, as the second most violent in the Security indicator, with this, this Cluster is considered the safest among the others.

In the SCM01 there is the domain "Smart Living" that deals with the aspects related to the security of the individual, and for this domain are considered the indicators: "Crime rate", "Death rate by robbery" and "Satisfaction with personal safety" (Figure 5.22).

Figure 5.22: SCM (01) Safety Indicators



SMART LIVING (Quality of life)

- Cultural facilities
- Health conditions
- Individual safety
- Housing quality
- Education facilities
- Touristic attractivity
- Social cohesion

Smart Living	Cultural facilities	Cinema attendance per inhabitant	2001	local
		Museums visits per inhabitant	2001	local
		Theatre attendance per inhabitant	2001	local
	Health conditions	Life expectancy	2001	local
		Hospital beds per inhabitant	2001	local
		Doctors per inhabitant	2001	local
		Satisfaction with quality of health system	2004	national
	Individual safety	Crime rate	2001	local
		Death rate by assault	2001-03	regional
		Satisfaction with personal safety	2004	national
	Housing quality	Share of housing fulfilling minimal standards	2001	local
		Average living area per inhabitant	2001	local
		Satisfaction with personal housing situation	2004	national
	Education facilities	Students per inhabitant	2001	local
		Satisfaction with access to educational system	2004	national
		Satisfaction with quality of educational system	2004	national
	Touristic attractivity	Importance as tourist location (overnights, sights)	2007	regional
		Overnights per year per resident	2001	local
	Social cohesion	Perception on personal risk of poverty	2006	national
		Poverty rate	2005	national

Source: Made by author

These indicators would not be adequate for the Brazilian reality, since the deaths caused by robbery represent only one of the indicators of violence and prisons in the country (Strazza, 2006), while drug trafficking and armed robbery (without deaths) represent more than 50% of police incidents. Therefore, the international indicators are incorporated by SmartCluster, however, new indicators that meet local needs are also used by the metamodel (Table 5.3).

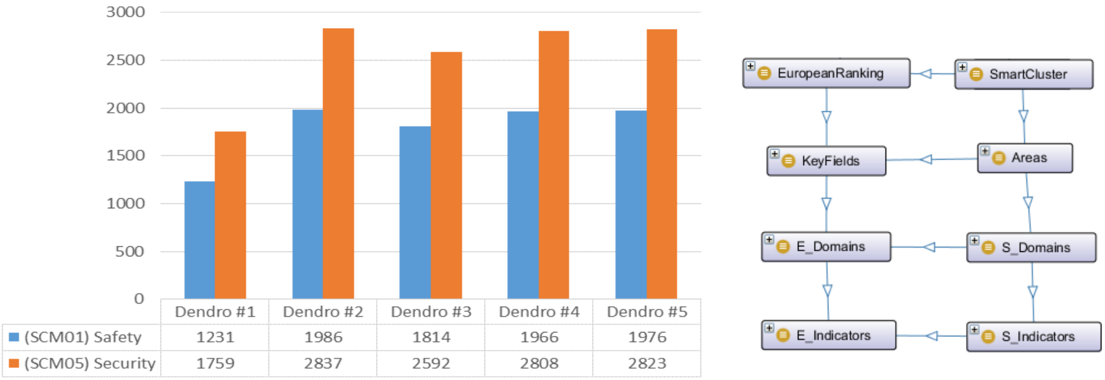
Table 5.8: (SCM01) versus (SCM05) Safety Concepts

European Ranking (SCM01)		**SmartCluster (SCM05)**	
KeyField	"Smart Living"	**Area**	"Services"
E_Domains	"Individual Safety"	**S_Domains**	"Security"
E_Indicators	• Crime rate • Death rate by robbery • Satisfaction with personal safety	**S_Indicators**	• Death rate (robbery, drug trafficking, Woman, Children) • Crime rate against equity • Police officers killed on duty

Source: Made by author

When, for example, SCM01 is used to measure and compare **Cluster#1**, a 30% difference in death rate is recorded, making comparison of types of crimes misleading. It is still necessary to consider that the crimes committed in the interior of the state have different characteristics from those committed in large and medium-sized cities (Figure 5.23), which again leads to the evident need for a prior equalization between cities of similar characteristics. (Level 1 of SmartCluster).

Figure 5.23: Safety and Security average Indicator



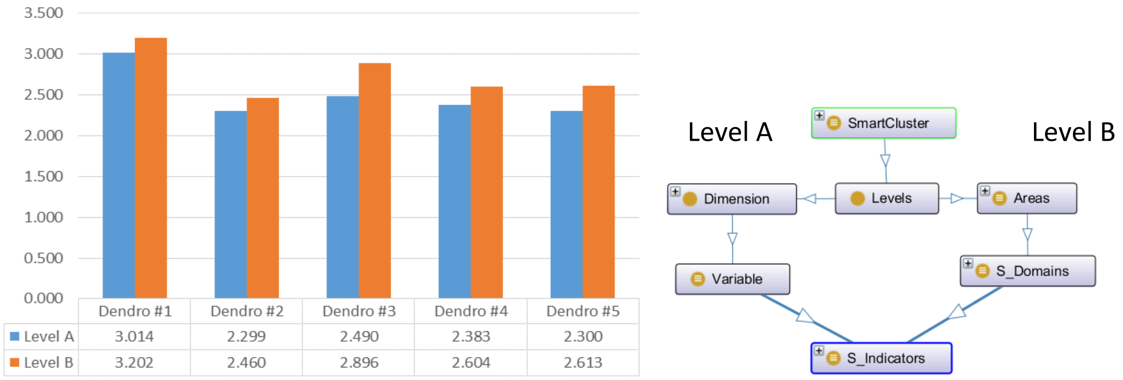
(a) Differences between crime SCM01 and SCM05 (b) Compatibility between SCMs

Source: Made by author

Cluster#2 (12 cities) has a much smaller variation than **Cluster#1** (0.330), and with this, the characteristics of the cities in this group are much more similar. The variation between the indicators of Levels A and B are very small, and reveal a hegemony of the variables related to the HDI of these cities. Therefore, from the managerial point of view, it could be said that solutions to the areas of Health, Education and Employment could be equally adopted.

In this group cities with larger numbers of inhabitants than **Cluster#1** cities have insufficient indicators to compare them with other Clusters, which reveals that the Level A "filter" allowed them to be grouped, and that Level B acted by separating In a group with less variation of indicators (Figure 24).

Figure 5.24: Comparison of SmartCluster metamodel levels.



(a) Level A and Level b comparison. SmartCluster

(b) Ontology of Metamodel

Source: Made by author

Cluster#3 has more than double (31) municipalities of previous Clusters, and yet the variation (0.301) between the indicators is smaller than the previous clusters. Infrastructure indicators in this cluster also only lose to **Cluster#1** and are better than **Cluster#2** indicators compared to Service and Governance indicators, although **Cluster#2** clearly contains larger, denser cities. This type of comparison makes it possible to prove that cities with greater population density are being grouped in this cluster so that they are not only considered the number of inhabitants. In Brazil, an initiative entitled "ConnectedSmartCities Ranking" (ConnectedSmartCities, 2017) approaches the concept presented by the SmartCluster metamodel and also categorizes cities by population band, however, only using this indicator to create categories of cities may prevent a comparison between cities of Medium or small in order to respect their regionalities.

Another example of a national ranking of cities is proposed by the newspaper “Folha de São Paulo” (REM-F, 2017) and proposes to show which municipalities in Brazil (among 5281) that achieve better health, education and sanitation results by spending less. In this ranking it is clear the distortion when comparing municipalities based on few indicators, which in Table 5.9 are presented and in almost all indicators are quite different.

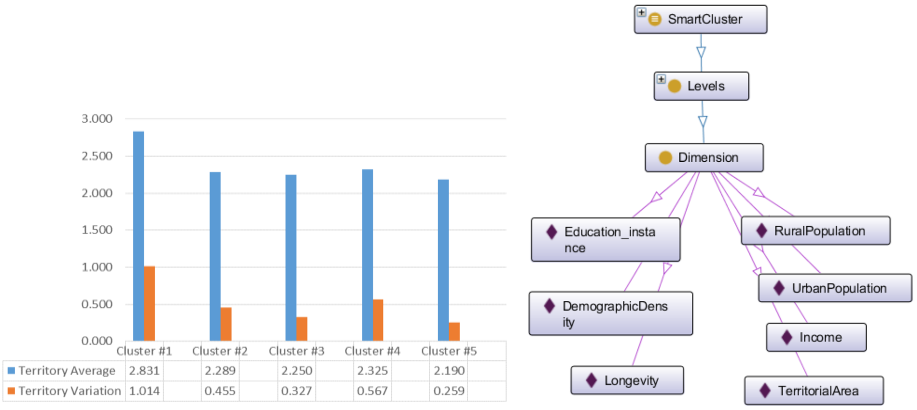
Table 5.9: City Efficiency Ranking

	Birigui (SP) 258th place **0.558** **Efficient**	Codó (MA) 3,810th place **0.405** **Low efficiency**
Population (2015)	117.143	119.962
Area	530,03 km²	4.361,34 km²
HDI (2010)	0,780	0,595
GDP (2013)	R$ 2,5 billion.	R$ 778,8 million
Children 0/3 years at school	39%	17%
Doctors/1000inhabitants -2014	0,5	0,3
Water treatment	96%	78%
Waste treatment	97%	10%

Source: Adapted from (REM-F, 2017)

For this reason, the SmartCluster metamodel (Figure 5.25) considers the use of three basic dimensions with their respective seven indicators to perform the grouping of similar cities: Territory (Territorial Area, Demographic density), Population (Urban population, rural population) and Development, Longevity, Education).

Figure 5.25: Territory average and Variation between Clusters



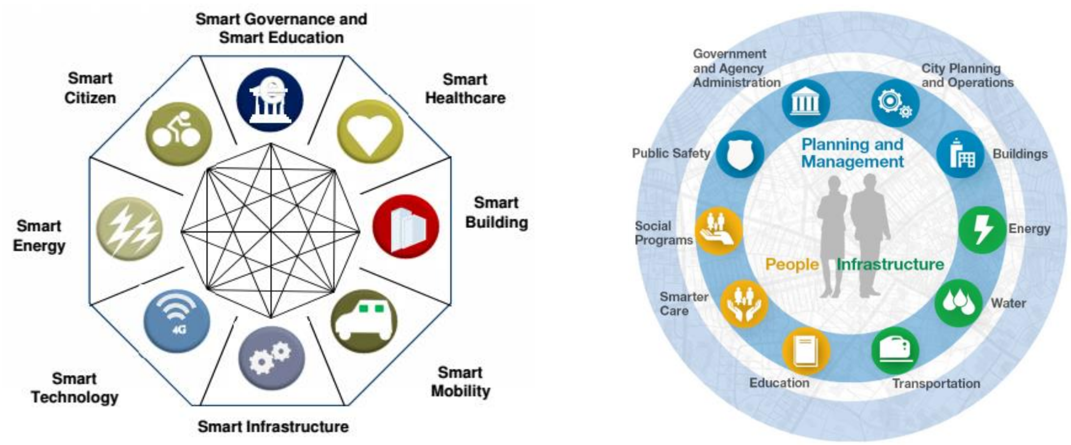
(a) Average and Variation of Territories. **(b) Dimensions and Variables**

Source: Made by author

Cluster#4 represents a range between **Cluster#3** and **Cluster#5** averages, and it has the characteristic of clustering approximately the same number of cities as the initial Clusters, only 16 cities. Although there is such difference between the variations, this grouping is similar to the previous groupings by the good indicators obtained in Level A. With this, it is possible to affirm that the cities in this group have similar metric characteristics, but with very different taxonomies of intelligence. The current SCM intelligence taxonomies vary widely, and depending on the purpose of the model, indicators can also target strategies for public problem solving, or simply the acquisition of hardware equipment.

In Figure 5.26 two SCMs from private initiatives are presented whose indicators used for the Governance domain converge to the provision of systems and technology for: Electronic Government, Electronic Education and Disaster Management.

Figure 5.26: Smart Cities commercial models

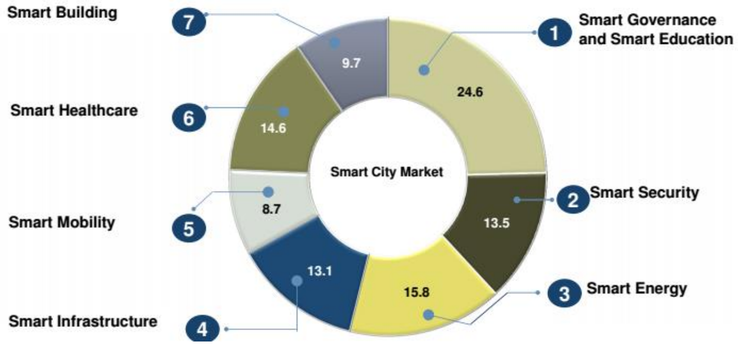


(a) (Frost & Sullivan, 2017) **b) (IBM, 2017)**

Source: Made by author

When comparing the Infrastructure and Governance indicators between these SCMs (whether commercial or not), this difference is visible even in city comparison heatmaps. The Technology indicator is one of these that presents individual averages well below those observed in previous clusters and deserves attention, since the increment of this indicator generally makes possible the provision of new Governance services. Still second (Frost & Sullivan, 2017), Figure 5.27 represents the market share of each segment in the market for smart cities. According to this consultancy, the commercialization of Intelligent Infrastructure includes sensor networks and governmental, educational and water and electric resources management systems.

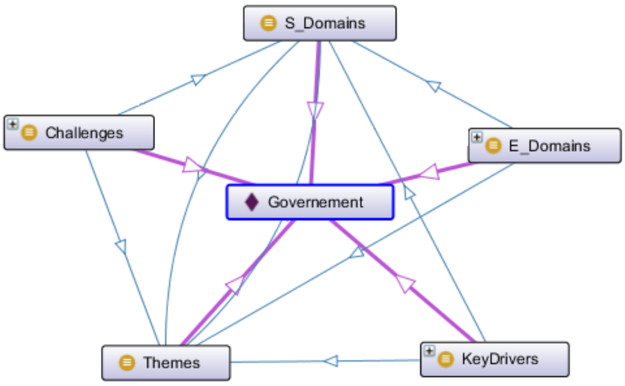
Figure 5.27: Smart Cities Market by segments.



Source: Made by author

While there is a concern of the private and industrial sectors to provide solutions to the trends of smart and sustainable cities, these solutions depend on academic initiatives that encourage the search for solutions, indicators and metrics so that cities can be audited by following some kind of pre-convention established. And in this case of e-governance indicators, there seem to be many suggestions for private initiatives, but with few indications of comparable indicators. Within the proposed metamodel, the domain of Governance (or electronic government) is inherited from the other SCM, since there is a consensus among all models that this is a fundamental item for the management of smart cities (Figure 5.28).

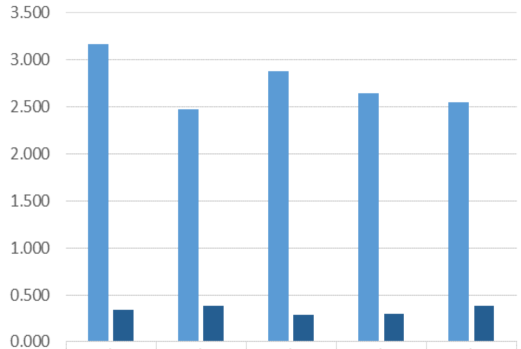
Figure 5.28: SmartCluster MetaModel and SCMs instance for Government.



Source: Made by author

In the last grouping, **Cluster#5**, the set of cities (31) contained the smallest variation of indicators (0.246) among the other groups and contains cities with a very present characteristic in the sample: smaller cities with smaller density than the other Clusters (Figure 5.29).

Figure 5.29: Infrastructure Variation between Clusters

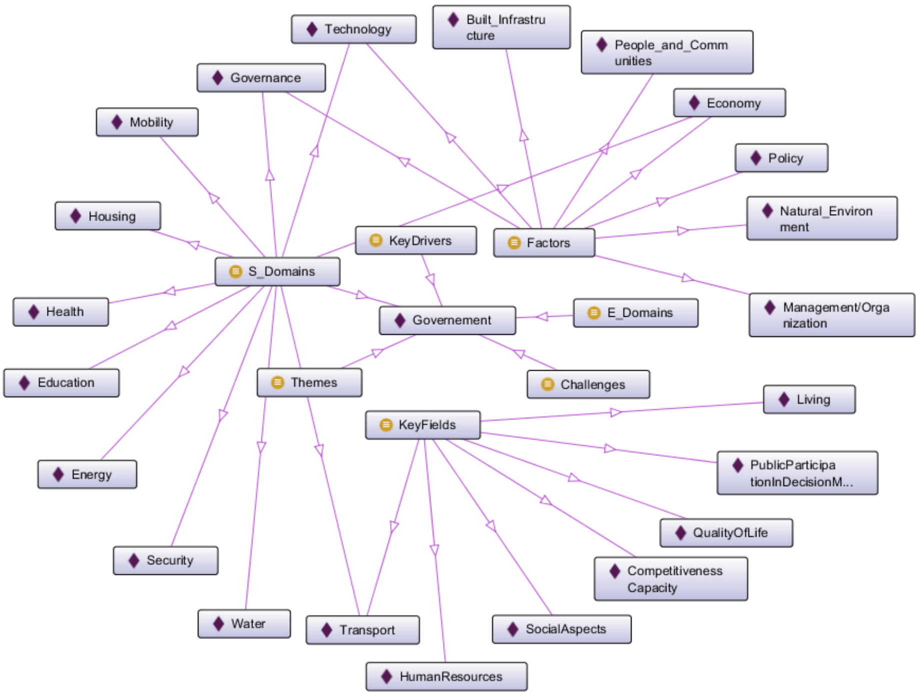


	Dendro #1	Dendro #2	Dendro #3	Dendro #4	Dendro #5
■ Infrastructure Average	3.169	2.479	2.885	2.644	2.545
■ Infrastructure Variation	0.345	0.379	0.283	0.295	0.388

Source: Made by author

Even so, the variation between the indicators of infrastructure is the largest among the groups, which reveals that although these cities have great similarity of size and population, nevertheless very different strategies are being adopted to offer public services and because of this, The evidence is clear that there is a need to standardize and identify what these services are and how to measure their supply to the population. In the proposed metamodel were included the indicators most commonly used by the municipal management for the service offer. (Figure 5.30).

Figure 5.30: Governance in the different SCMs



Source: Made by author

The following section concludes this chapter and presents the final discussions about the SmartCluster metamodel and the evidence found in this work.

5.5 Discussions about Study and Experimentation of SmartCluster

In general, the cophenetic correlation coefficient (CCC) obtained from the combination of variables (Level A) and indicators (Level B) using the measure of Euclidean clustering and Ward's clustering method showed that this metamodel titled SmartCluster proved feasible To simulate Smart City clusters using public data and indicators compatible with other SCMs.

The treatment of the analytical data concerning the samples of smart cities was carried out in two stages. Initially, the metric characteristics of these cities (Territory Population and Development) were considered in order to introduce the multivariate approach of data analysis through hierarchical grouping (HCA) and principal component analysis (PCA). The computational package for the elaboration of the comparison dendrograms was the RStudio.

Afterwards, all the analytical results obtained with the heatmaps and dendrograms were used, consequently, unsupervised methods of pattern recognition were used to evaluate in a multivariate way the complete data set of two levels (A and B) and seventeen indicators and variables.

The Smart City cluster analysis is a set of minimum indicators and variables, which can be extended according to the needs of adopting strategies to meet the demands of the inhabitants. As new indicators and variables are inserted in this analysis, new clusters will emerge, and cluster analysis will allow new insights into particular domains.

From the point of view of the EBSE methodology, the experiments performed through the clustering of data of this Chapter combined with the grounded theory of Chapter 2 provided elements that relates to the need of a metamodel for smart cities.

5.6 Summary

The objective of this chapter was to present a clustering analysis using the SmartCluster metamodel, proving through the EBSE methodology that there is sufficient evidence about the need for a metamodel for SCM that normalizes concepts, metrics and indicators in order to respect the regionalisms of each city and Even so, tally the indicators compatible with other models around the world.

For this it was presented the Level A (Metrics) and Level B (Taxonomy) layers that make up the SmartCluster; Then the chapter presented the steps to carry out the EBSE and its stages; In the study stage, the results obtained with a extensive systematic review of literature were evaluated and included the concepts for the extension of the secondary studies; Next, this chapter presented the experimentation of the SmartCluster metamodel, presenting the steps (select, process, transform, view and interpretation) to create and analyze the Smart City groupings from Alagoas as a case study. This chapter concludes with the sections devoted to a brief discussion of SmartCluster and Summary.

6

SUMMARY

I'm not in danger. I'm the danger.

— Walter White, Breaking Bad

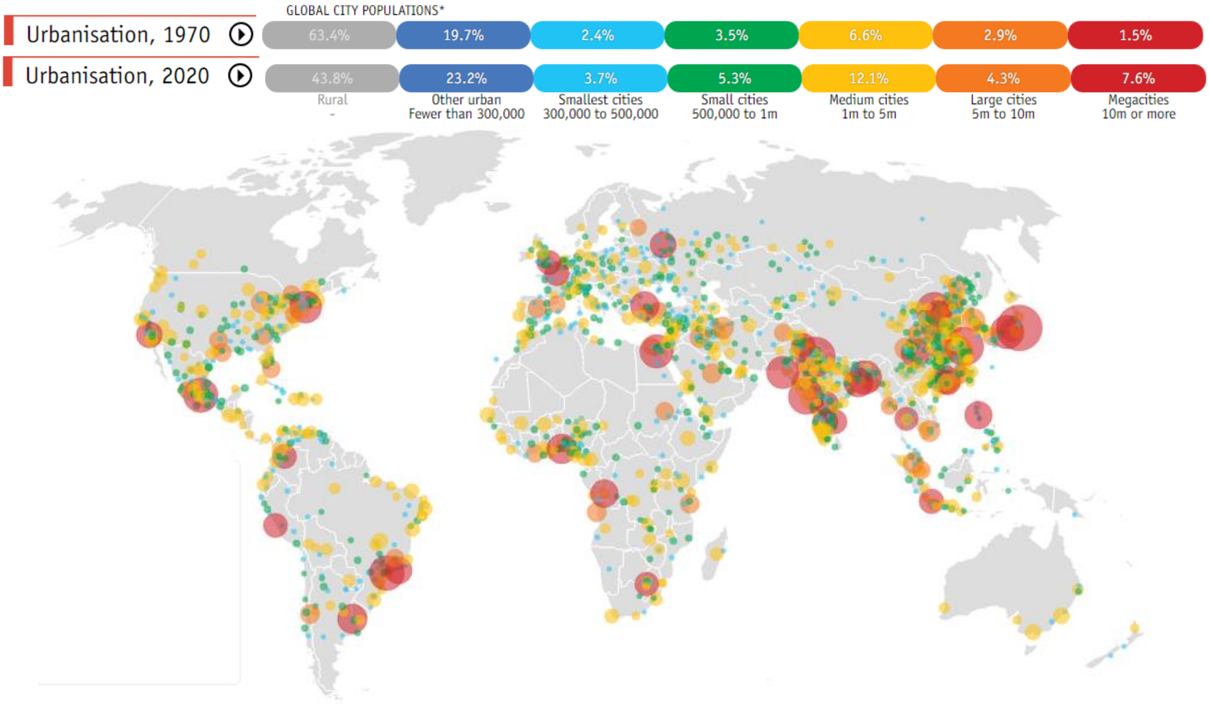
This chapter concludes the study presenting the final considerations on the use of a model for smart cities that is adherent to the reality of Brazilian cities and indicates the prospects for future work can be developed complementing and extending the current stage of research in this area.

6.1 Research Contributions

According to Magazine The Economist (2015), the population of London has already surpassed the mark of eight million people by 2015 and thus exceeded its previous population peak of 1939, which was just over 5 million.

According to the magazine, the task for other municipal managers will be even more "frightening", as around 9% of the world population will live in only 41 megacities (with more than 10 million inhabitants) by 2030 (Figure 6.1).

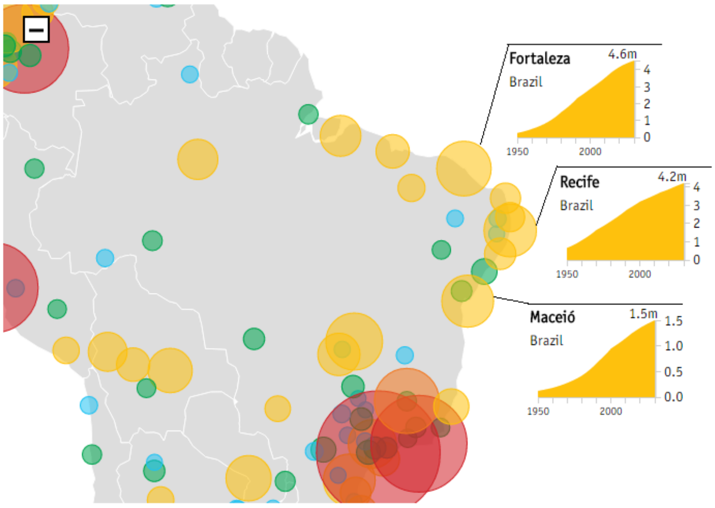
Figure 6.1: Urbanization process around the world



Source: Made by author

In the coastal northeast of Brazil, these estimates indicate that the cities of Fortaleza, Recife and Maceió will have respectively 4.6, 4.2 and 1.5 million inhabitants by 2030, which together correspond to 10 million inhabitants (Figure 6.2).

Figure 6.2: Urbanization of Northeastern coastal cities



Source: Made by author

The way these megacities are being managed by the public sector can indicate solutions to be adopted by medium and small cities. These same solutions can also serve as a roadmap for Smart Cities to better manage their resources and optimize service delivery. It is in this way that this work follows, indicating a way to ensure that the best practices obtained with the models of smart cities are assimilated by a metamodel.

Therefore, the most important contribution of this thesis was to propose and create a metamodel for smart cities that is compatible with the existing models and that allows to be expandable according to the needs of models that will be created, respecting a normalization of data and the correct comparison of cities respecting their regional characteristics.

For this, this thesis made use of data mining, development of ontologies, evidence-based software engineering and expanded the literary review of smart cities.

The creation of a taxonomy of indicators for smart cities used a domain ontology, and so, as is own ontologies, this taxonomy can be reused, merged with other ontologies and above all, expanded according to the need to use new indicators. By making use of multivariate analysis, dendrogram if data clusters, are proposing new ways of knowledge discovery and data visualization for strategic decision making by municipal managers (Table 6.1).

Table 6.1: Chapters and contents

(I) Chapter	***(II) Contents***
(2) Grounded Theory	SLR, Grounded Theory
(3) Metrics	Multivariate and Cluster Creation
(4) Taxonomy	Data Mining and Normalizing
(5) SmartCluster	Ontology, Meta Objects,
(6) Metamodel Validation	EBSE, Cluster analysis

Source: Made by author

Following the methodology of work proposed in the initial chapter of this thesis, this work was positioned in a pragmatic and philosophical way using an inductive approach, referencing methods of bibliographic research. Thus, this descriptive and exploratory thesis sought to identify a relationship between the proposed indicators to establish a data standard called SmartCluster that could be extended to any Smart Cities Models.

This work has as its starting point the use of Computer Science combined with technologies and processes (Table 6.1) to obtain answers to the research questions raised in Chapter 1 and which will be detailed in Section 6.5.

6.2 Publications

6.2.1 Written books

Afonso, R. A.; Albuquerque, A. C. R. ; Paula, A. M. T. ; Sousa, I. S. ; Costa, J. R. ; Menezes, J. M. S. ; Silva, L. L. J. ; Faustino, M. H. A. F. ; Farias, M. M. ; Rios, R. D. M. ; Magalhaes, T. B. S. . **Smart Cities: A Comprehensive Systematic Literature Review**. 1. ed. Amazon, 2015. v. 1. 96p.

6.2.2 Papers in Journals

Afonso, R. A. ; Brito, K. S. ; Nascimento, C. H. ; Costa, L. C. ; Alvaro, Alexandre ; Garcia, V. C. . **(Br-SCMM) Brazilian Smart City Maturity Model: A Perspective from the Health Domain.** Studies in Health Technology and Informatics, v. 216, p. 983-983, 2015.

Afonso, R. A.; Pereira, C. F. **MaTUTO: adaptação da metodologia de aprendizagem baseada em problemas aplicada ao ensino de ontologias**. AtoZ: novas práticas em informação e conhecimento, v. 2, p. 34,

Afonso, R. A.; Cabral, R. S.; Garcia, V. C.; Alvaro, Alexandre. **DendroIDH: Agrupando Cidades Por Semelhança de Indicadores.** Journal of Health Informatics, v. 8, p. 907-914, 2016.

Afonso, R. A.; Costa, L. C.; Alvaro, Alexandre ; Garcia, V. C. . **SCiAl: Usando Dados Públicos para Agrupar Cidades Alagoanas.** Gestão.Org, v. 13, p. 331-339, 2015.

6.2.3 Papers in Conferences

Afonso, R. A.; Cabral, R. S.; Garcia, V. C.; Alvaro, Alexandre. **'DendroIDH: Agrupando Cidades Por Semelhança de Indicadores'**. In: XV Congresso Brasileiro de Informática em Saúde - CBIS 2016, 2016, Goiânia - GO. Informática Transformando a Saúde. Goiânia - GO, 2016.

Afonso, Ricardo Alexandre; Dos Santos Brito, Kellyton; Do Nascimento, Clóvis Holanda; Garcia, Vinicius Cardoso; Álvaro, Alexandre. **Brazilian Smart Cities: Using a Maturity Model to Measure and Compare Inequality in Cities**. In: the 16th Annual International Conference, 2015, Phoenix. - dg.o '15, 2015. p. 230.

Afonso, R. A. ; Lima, L. C. ; Costa, L. C. ; Alvaro, Alexandre ; Garcia, V. C. . **Mapeamento Sistemático de Informática Médica para Cidades Inteligentes**. In: CBIS (Congresso Brasileiro de Informática em Saúde), 2014, Santos / SP. CBIS 2014 - XIV Congresso Brasileiro de Informática em Saúde. Santos / SP: SBIS (Sociedade Brasileira de Informática em Saúde), 2014.

Afonso, R. A.; Alvaro, Alexandre; Nascimento, C. H.; Garcia, V. C. **SmartCluster: Utilizando Dados Públicos para Agrupar Cidades Inteligentes por Domínios**. In: SBSI 2015 XI Brazilian Symposium on Information System, 2015, Goiânia / GO. v. 1. p. 699-702.

Afonso, R. A. ; Brito, K. S. ; Nascimento, C. ; Alvaro, Alexandre ; Garcia, V. C. . **Br-SCMM - Brazilian Smart City Maturity Model: A Perspective from the Health Domain**. In: 15th World Congress on Health and Biomedical Informatics, 2015, São Paulo. MEDINFO 2015: eHealth-enabled Health. São Paulo/SP: SBIS, 2015.

Afonso, R. A. ; Silva, W. M. ; Tomas, G. H. ; Gama, K. ; Lima, A. O. ; Alvaro, Alexandre ; Garcia, V. C.. **Br-SCMM: Modelo Brasileiro de Maturidade para Cidades Inteligentes**. In: IX Simpósio Brasileiro de Sistemas de Informação (SBSI), 2013, João Pessoa/PB. v. V1. p. 511-516. **(Awarded as best paper)**

Afonso, R. A.; Alvaro, Alexandre; Garcia, V. C. **Scial: Usando Dados Públicos Para Agrupar Cidades Inteligentes Alagoanas**. In: IV Simpósio Brasileiro de Tecnologia da Informação (SBTI), 2015, Aracaju / SE. Internet das Coisas (Internet of Things). Aracaju / SE, 2015.

Afonso, R. A.; Garcia, D. S.; Garcia, V. C.; Alvaro, Alexandre. **e-PID: Electronic Professional Identity Document**. In: VII Congresso Tecnológico InfoBrasil 2014, 2014, Fortaleza - Ceará. VII Congresso Tecnológico InfoBrasil 2014. Fortaleza - Ceará: Editora InfoBrasil, 2014

6.2.4 Awards

Best Paper (Br-SCMM: Modelo Brasileiro de Maturidade para Cidades Inteligentes) of the Track: “S.I. e os Desafios do Mundo Aberto”, SBSI (2013).

6.3 Limitations and future work

Although the design of a metamodel for smart cities compatible with existing and extensible models has been elaborated, there are still some paths to be followed and for a matter of time and scope they have not yet been.

The main limitations of this study may be cited:

- The initial design of a maturity model for smart cities was not possible due to lack of time and the absence of professionals who could guide the creation of this model.

- One limitation that may be cited is the lack of insistence on suggesting a single model for measuring and comparing cities. This is because this work intends to serve as a metamodel for the cited models, and thus complement existing models, without losing the existing characteristics.

To continue this model of creating the proposal for smart cities, future research and improvements can be considered:

- Creation of a graphical autonomic environment for data visualization without the need for mining, convert and display data in statistical tools;

- Creation of mining tools and automated data extraction;

- Expansion of the metamodel and the creation of new levels for these indicators, according to the needs of each city;

- Although this metamodel has been presented in a technical way, the social and welfare characteristics of each proposed model will be adequate and incorporated by SmartCluster to allow the generation of more smart models. An initiative being developed in Brazil entitled RBCIH[1] (Brazilian Network of Intelligent and Human Cities) uses some of these metamodel concepts, including in the context of using levels for prior comparison of cities.

[1] RBCIH, available at: http://redebrasileira.org/

6.4 Summary

The key issue addressed in this thesis was how to create a metamodel for smart cities capable of enabling compatibility with existing models and becoming expandable to new model proposals.

For this, the main challenges faced in this work and their solutions were:

- **(RQ1) How to evaluate how smart a city can be?**
  The best way found in this thesis was to compare the current models of smart cities **(Chapter 2, 5 and 6)** and propose a metamodel that used variables **(Chapter 3)** and indicators **(Chapter 4)** compatible with the existents SCMs to allow new models.

- **(RQ2) Which domains and indicators are appropriate to the Brazilian reality?**
  This thesis presented a way to filter and sort the cities using two levels: the first (Level A) by similarity variables **(Chapter 3)**, and then (Level B) by indicators based on public data **(Chapter 4)**. The use of these two levels is configured as a differential in relation to other models of smart cities and allows for more accurate comparisons between the three cities.

- **(RQ3) How to obtain and process data on smart cities?**
  The data on smart cities were obtained from open public data sources. The processing of these data was the normalization of indicators and variables, and then we used the Hierarchical Cluster method (HCA) to create clusters of cities by similarity. Finally, we used the Principal Component Analysis technique (PCA) to identify a possible level of pattern recognition **(Chapter 5 and 6)**.

- **(RQ4) How to provide an environment to extraction and data visualization?**
  Data visualization can influence strategic decisions about investments in different areas of Smart Cities. After mining, cataloging and processing of data of similar cities, this information was disposed in dendrograms of smart cities clusters. The grouping of smart cities identified patterns in indicators of cities, and thus allow managers to choose which ways will be adopted to improve processes and resource optimization **(Chapter 3, 5 and 6)**.

- **(RQ5) How to create and validate a metamodel compatible with existing models?**
  The objective of creating the metamodel **(Chapter 6)** is to reduce the distance between the evaluation and comparison of smart cities performed in academia and industry, using Evidence-Based Software Engineering (EBSE), which is represented in this work by the application of a formal approach for the identification of evidence (secondary studies) combined with experiments with the metamodel applied to the reality of Brazilian Northeastern cities.

In summary, this thesis presents a solution to the problem of multiple models for smart cities that are not compatible or do not meet regional demands. The creation of this metamodel took into consideration the existing models and proposed a division into levels for the categorization of cities, which makes the metamodel closer to the reality of the cities.

It is hoped that this work can really be used to make citizens' lives better, and that the techniques presented here can be implemented so that managers can select better strategies for optimizing resources and offering public services. If this thesis helps to improve the life of some citizen, the work will have been well done, however, if it improves the life of a city, the work will have reached the initial desire of the author and his advisers…

REFERENCES

A. MAEDCHE AND S. STAAB, “Discovering Conceptual Relations from Text,” Proc. European Conf. Artificial Intelligence (ECAI-00), IOS Press, Amsterdam, 2000, pp. 321–325.

ALVES, A.; REVOREDO, K.; BAIÃO, F.. Alinhamento de ontologias baseado em instâncias utilizando técnicas de mineração de dados. WTDSI, 2012.

BEEBE K. R.; PELL, R. J.; SEASHOLTZ, M. B.; Chemometrics: a practical guide, John Wiley & Sons: New York, 1997

BIOLCHINI, J. et al. Systematic Review in Software Engineering: Relevance and Utlity.[Sl], 2005. Citado, v. 4, p. 47.

BOLLIER, D. How Smart Growth Can Stop Sprawl, Essential Books, Washington, DC, 1998.

BOWERMAN, B., BRAVERMAN, J., TAYLOR, J., TODOSOW, H., & VON WIMMERSPERG, U. The vision of a smart city. (2000, September) In 2nd International Life Extension Technology Workshop, Paris (Vol. 28).

BRAUNER, D., F.; E CASANOVA, M. A. (2008) Alinhamento de esquemas baseado em instâncias. Rio de Janeiro, 2008. 83p. Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

BRITO, K.S., et al. (2014). Using parliamentary Brazilian open data to improve transparency and public participation in Brazil. In Proceedings of the 15th Annual International Conference on Digital Government Research (dg.o '14). ACM, New York, NY, USA, 171-177. DOI=10.1145/2612733.2612769 http://doi.acm.org/10.1145/2612733.2612769

BRITTO, J., STALLIVIERI, F., CAMPOS, R. R., & VARGAS, M. (2007). Padrões de aprendizagem, cooperação e inovação em aglomerações produtivas no Brasil: uma analise multivariada exploratória. ENCONTRO NACIONAL DE ECONOMIA– ANPEC, Recife.

CABOT, J. et al. Integrating sustainability in decision-making processes: A modelling strategy. In: ICSE-Companion 2009. 31st International Conference, 2009. p. 207-210.

CARAGLIU, A., DEL BO, C., & NIJKAMP, P. (2011). Smart cities in Europe. Journal of urban technology, 18(2), 65-82.

CARDOSO, F.. Smart Cities PT. Primeira Norma Iso Para As Cidades. URL: http://www.smart-cities.pt/pt/noticia/primeira-norma-iso-para-as-cidades37120

CARROLL, S. R.; CARROLL, D. J. Statistics Made Simple for School Leaders (illustrated ed.). Rowman & Littlefield.

CCSPJP. Conselho Cidadão para a Segurança Publica e Justiça Penal.(Consejo Ciudadano para la Seguridad Pública y la Justicia Penal). 2014 Available at https://goo.gl/jEPrhr

CEF. Caixa Economica Federal. Programa Habitacional Popular “Minha Casa, Minha Vida”. 2015 Available at http://www.caixa.gov.br/habitacao/mcmv/index.asp

CHOURABI, Hafedh et al. Understanding smart cities: An integrative framework. In: System Science (HICSS), 2012 45th Hawaii International Conference on. IEEE, 2012. p. 2289-2297.

CHRISTIE, O. H. J.; Chemometr. Information Systems. Intell. Lab. 1995, 29, 177.

COCCHIA, A. "Smart and digital city: A systematic literature review." Smart City. Springer International Publishing, 2014. 13-43.

COHEN, B., 2012. What Exactly is a Smart City? In Co.Exist, Available at https://goo.gl/xGV1Z7 (accessed 27 February 2014).

CONNECTEDSMARTCITIES. Cities of the Future of Brazil. 2016. Available at: https://goo.gl/4L1g74 accessed January 22, 2017.

CONTE, T.; CABRAL, R.; TRAVASSOS, GUILHERME, H. Aplicando Grounded Theory na Análise Qualitativa de um Estudo de Observação em Engenharia de Software–Um Relato de Experiência. In: V Workshop" Um Olhar Sociotécnico sobre a Engenharia de Software"(WOSES 2009). 2009. p. 26-37.

COOPER, H. M. Organizing knowledge syntheses: a taxonomy of literature review. Knowledge Society, 1988, 1, 104–126.

CORBIN, J. M.; STRAUSS, A. Grounded theory research: Procedures, canons, and evaluative criteria. Qualitative sociology, v. 13, n. 1, p. 3-21, 1990.

CORCHO, O., GÓMEZ-PÉREZ, A. A Roadmap to Ontology Specification Languages. Madrid, Espanha. EKAW'00. Springer-Verlag. 2000.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M.; RODRIGUES, A. Análise multivariada para os cursos de administração, ciências contábeis e economia. Fundação Instituto de Pesquisas Contábeis, Atuariais e Financeiras - FIPECAFI. São Paulo: Atlas, 2009

CORREIA, P.R.M; FERREIRA, M.M.C. Reconhecimento de padrões por métodos não supervisionados: explorando procedimentos quimiométricos para tratamento de dados analíticos. Química Nova, v. 30, n. 2, p. 481, 2007.

DEAN M, CONNOLLY D, HARMELEN F., HENDLER J, HORROCKS I., MCGUINNESS D., PATEL-SCHNEIDER P., E STEIN, L.A.. Web ontology language (OWL) reference version 1.0. W3C Working Draft, 2003.

DONATH, J.S. Inhabiting the virtual city: The design of social environments for electronic communities. 1997.

DYBA, T.; KITCHENHAM, B. A.; JORGENSEN, M. Evidence-based software engineering for practitioners. IEEE software, v. 22, n. 1, p. 58-65, 2005.

EISENBERG, A.; MELTON, J.. SQL: 1999, formerly known as SQL3. ACM Sigmod record, v.28, n. 1, p. 131-138, 1999.

ERGAZAKIS, K.; METAXIOTIS, K.; PSARRAS, J. Towards knowledge cities: conceptual analysis and success stories. Journal of knowledge management, v. 8, n. 5, p. 5-15, 2004.

EUROPE 2020 STRATEGY. URL: https://goo.gl/4etIIE

EUZENAT, J. AND SHVAIKO, P. Ontology Matching. Springer-Verlag, Berlin 2007

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. (1996) The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, v. 39, n. 11, p. 27-34.

FBS (2015). Fórum Brasileiro de Segurança. 2015 Available at https://goo.gl/tsyKFH

FENSEL, D., VAN HARMELEN, F., HORROCKS, I., MCGUINNESS, D. L., & PATEL-SCHNEIDER, P. F. (2001). OIL: An ontology infrastructure for the semantic web. IEEE intelligent systems, (2), 38-45.

FERNÁNDEZ-LÓPEZ et al. Building a chemical ontology using methontology and the ontology design environment. IEEE Intelligent Systems, v. 14, n. 1, p. 37-46, january/february 1999

FIELD, A. (2009). Descobrindo estatística utilizando o SPSS. Tradução Lorí Viali. 2. ed. Porto Alegre: Artmed, 2009. 688p.

FOLHA (2015). "Só 20% dos médicos fazem diagnóstico baseado em evidências, diz especialista." Seminários Folha – Fórum de Tecnologia e Acesso a Sáude. Jornal Folha de São Paulo. Disponivel em: https://goo.gl/sAFNh2 acessado em 03 de Fevereiro de 2017.

FROST & SULLIVAN. Strategic Opportunity Analysis of the Global Smart City Market. Available at: https://goo.gl/W7GccM accessed January 22, 2017

GAMA, K.; ALVARO, A.; PEIXOTO, E. Em direção a um modelo de maturidade tecnológica para cidades inteligentes. Simpósio Brasileiro de Sistemas de Informação, VIII, 2012.

GIFFINGER, R., FERTNER, C., KRAMAR, H., KALASEK, R., PICHLER-MILANOVIÜ, N., & MEIJERS, E. (2007). Smart Cities: Ranking of European Medium-Sized Cities. Vienna, Austria: Centre of Regional Science (SRF), Vienna University of Technology. Available at https://goo.gl/lN7cXC

GIFFINGER, R.; GUDRUN, H. Smart cities ranking: an effective instrument for the positioning of the cities?. ACE: Architecture, City and Environment, v. 4, n. 12, p. 7-26, 2010.

HAIR JR., J.F., ANDERSON, R.E., TATHAM, R.L. E BLACK, W.C. (2005). Análise Multivariada de Dados, Tradução, 5ª ed., Bookman, Porto Alegre.

HANDSCHUH, S., MAEDCHE, A., STOJANOVIC, L., VOLZ, R., KAON – The Karlsruhe Ontology and Semantic Web Infrastructure, URL: https://goo.gl/iqODfs

HENNINGER, S.; PADMAPRIYA A. "An ontology-based metamodel for software patterns." (2006).

HERNÁNDEZ-MUÑOZ, J. M. et al. Smart cities at the forefront of the future internet. In: The Future Internet Assembly. Springer Berlin Heidelberg, 2011. p. 447-462.

HOLLANDS, R. G. Will the real smart city please stand up? Intelligent, progressive or entrepreneurial?. City, v. 12, n. 3, p. 303-320, 2008.

HORROCKS, I., & SATTLER, U. (2001, August). Ontology reasoning in the SHOQ (D) description logic. In IJCAI (Vol. 1, No. 3, pp. 199-204).

IBGE. Censo Populacional 2010 Censo Populacional 2010 Instituto Brasileiro de Geografia e Estatística (IBGE) (29 de novembro de 2010).

IBGE. IBGE apresenta nova área territorial brasileira. URL: https://goo.gl/5g9kCs

IDEB. Índice da Educação Básica do Brasil. 2015 Available at https://goo.gl/Ql9fik

IDEC. Available at Available at https://goo.gl/pSQOIT

INABA, A., OHKUBO, R., IKEDA, M., MIZOGUCHI, R., TOYODA, J.. Design and Analysis of Learners' Interaction based on Collaborative Learning Ontology . In: Proc. of EuroCSCL01, pp.308-315, 2001

INSTITUTO INTERNACIONAL DE ESTATÍSTICA. ISE (2015) Available at https://goo.gl/w2gq7I

IPEA. Instituto de Pesquisa Econômica Aplicada. 2015 Available at https://goo.gl/Y1rf0T

ISSO. “ISO 37120: Sustainable Development of Communities – Indicators for City Services and Quality of Life”, International Organization for Standardization, First Edition, 2014-05-15, ISO37120:2014(E).

JEAN, S.; AÏT-AMEUR, Y.; PIERRA, G.. A language for ontology-based metamodeling systems. In: East European Conference on Advances in Databases and Information Systems. Springer Berlin Heidelberg, 2010. p. 247-261.

JEAN, S.; AÏT-AMEUR, Y.; PIERRA, G.. Querying ontology based database using ontoql (an ontology query language). In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer Berlin Heidelberg, 2006. p. 704-721.

JOHNSON, R.A. E WICHERN, D.W. (1992). Applied Multivariate Statistical Analysis, Prentice-Hall, New Jersey.

JUDD, W.S., CAMPBELL, C.S., KELLOG, E.A., STEVENS, P.F., DONOGHUE, M.J. (2007) Taxonomy. In Plant Systematics - A Phylogenetic Approach, Third Edition. Sinauer Associates, Sunderland.

KITCHENHAM, B. AND CHARTERS, S.: Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report (2007)

KITCHENHAM, B., BRERETON, O. P., BUDGEN, D., TURNER, M., BAILEY, J., & LINKMAN, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. Information and software technology, 51(1), 7-15.

KITCHENHAM, B.A., CHARTERS, S., 2007. Guidelines for Performing Systematic Literature Reviews in Software Engineering. In: Technical Report EBSE 2007-001. Keele University and Durham University Joint Report.

KITCHENHAM, B; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. Technical Report, School of Computer Science and Mathematics, Keele University, 2007. 65 p.

KITCHENHAM, B. A.; DYBA, T.; JORGENSEN, M.. Evidence-based software engineering. In: Proceedings of the 26th international conference on software engineering. IEEE Computer Society, 2004. p. 273-281.

KOMNINOS, N. The Architecture of Intelligent Cities, Conference Proceedings Intelligent Environments 06, Institution of Engineering and Technology, pp. 53-61.

KOMNINOS, N. et al. Smart city ontologies: Improving the effectiveness of smart city applications. Journal of Smart Cities, v. 1, n. 1, 2016.

KREYSZIG, E. Advanced Engineering Mathematics (Fourth ed.). Wiley

LASSILA, O.; SWICK, R. Resource Description Framework (RDF) model and syntax specification.1.0, 22 Feb 1999. (Recomendação do W3C). Disponível em: <http://www.w3c.org/TR/REC-rdf-syntax>. Acesso em março de 2008.

LEVINE, D. M. et al. (2008) Estatística: Teoria e aplicações. Tradução Teresa Cristina Padilha de Souza. Rio de Janeiro: LTC, 2008. 752p

LI, ZHENG ET AL. On evaluating commercial Cloud services: A systematic review. Journal of Systems and Software, v. 86, n. 9, p. 2371-2393, 2013.

LI, ZHENG, ET AL. "Towards a taxonomy of performance evaluation of commercial Cloud services." Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on. IEEE, 2012.

LISBOA, L. B., GARCIA, V. C., LUCRÉDIO, D., DE ALMEIDA, E. S., DE LEMOS MEIRA, S. R., & DE MATTOS FORTES, R. P. (2010). A systematic review of domain analysis tools. Information and Software Technology, 52(1), 1-13.

LÓPEZ, F. Overview of methodologies for building ontologies. In Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends. CEUR Publications, 1999. Intelligent Systems, 16(1):26-- 34, 2001.

M. AL-HADER, A. RODZI, A. R. SHARIF AND N. AHMAD, "Smart city components architicture", Proc. Int. Conf. Comput. Intell. Modelling Simulation, pp. 93-97. (2009)

M. KLEIN, "XML, RDF, and Relatives," IEEE Intelligent Systems, 15(2), pp. 26-28, 2001.

MACADAR, M. A.; LHEUREUX-DE-FREITAS, J. (2013). Porto Alegre: a Brazilian city searching to be smarter. In Proceedings of the 14th Annual International Conference on Digital Government Research (dg.o '13). ACM, New York, NY, USA, 56-64. DOI=10.1145/2479724.2479736 http://doi.acm.org/10.1145/2479724.2479736

MAFRA, S. N.; BARCELOS, R. F.; TRAVASSOS, G. H.. Aplicando uma metodologia baseada em evidência na definiçao de novas tecnologias de software. In: Proceedings of the 20th Brazilian Symposium on Software Engineering (SBES 2006). 2006. p. 239-254.

MAFRA, S. N.; TRAVASSOS, G. H. Estudos Primários e Secundários apoiando a busca por Evidência em Engenharia de Software. Relatório Técnico, RT-ES, v. 687, n. 06, 2006.

MAN. (2015). Media Awareness Network. Knowing what´s what and what´s not: The 5Ws (and 1 "H") of Cyberspace. Media Awareness Network, October, 2015. URL: https://goo.gl/g8yGjJ

MARCONI, M.; LAKATOS, E. Metodologia científica. 4. ed. São Paulo: Atlas, 2004. 312 p.

MARTINS, C. R., DE ALBUQUERQUE, F. J. B., GOUVEIA, C. N. N. A., RODRIGUES, C. F. F., & DE SOUZA NEVES, M. T. (2007). Avaliação da qualidade de vida subjetiva dos idosos: uma comparação entre os residentes em cidades rurais e urbanas. Estudos Interdisciplinares sobre o Envelhecimento, 11.

MCGUINNESS ,D. L.; HARMELEN , F. VAN. "OWL Web Ontology Language Overview," W3 Consortium, https://goo.gl/PQn44t /, Accessed January 31, 2017.

MEC. Ministério da Educação. 2015 Available at http://www.mec.gov.br

MILLER E., SWICK R., BRICKLEY D., MCBRIDE B., HENDLER J., SCHREIBER G., WOOD D., CONNOLLY D., "W3C Semantic Web," World-Wide Web Consortium, http://www.w3.org/2001/sw/, Accessed January 31, 2017.

MILONE, G. Estatística geral e aplicada. São Paulo: Centage Learning, 2009.

MMA.. Ministério do Meio Ambiente. 2015 Available at www.mma.gov.br

MME. Ministério de Minas e Energia. 2015 Available at http://www.mme.gov.br/mme

MOHER D., SCHULZ K.F., ALTMAN D.G. The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomised Trials, The CONSORT Group, Annals of Internal Medicine, 134(8):657-662, April 2001

MT. Ministério dos Transportes. 2015 Available at http://www.transportes.gov.br

NAJERA, K. et al. An Ontology-Based Methodology for Integrating i* Variants. In: iStar. 2013. p.1-6.

NAM, T., & PARDO, T. A. (2011). Conceptualizing smart city with dimensions of technology, people, and institutions. Digital Government Innovation in Challenging Times (pp. 282-291). ACM.

NEIROTTI, P., DE MARCO, A., CAGLIANO, A. C., MANGANO, G., & SCORRANO, F. (2014). Current trends in Smart City initiatives: Some stylised facts. Cities, 38, 25-36.

NOY, N. F. E MCGUINNESS, D. L. Ontology Development 101: A Guide to Creating Your First Ontology, Knowledge Systems Laboratory, Stanford University, mar. 2001.

NOY, N. F., & MUSEN, M. A. (2000, August). Algorithm and tool for automated ontology merging and alignment. In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831.

ODM. Portal ODM, Acompanhamento municipal dos objetivos do milênio. 2015 Available at http://www.portalodm.com.br/

OMG, "MetaObjectFacility(MOF) Specification, v1.4," Object Management Group, https://goo.gl/JbLPDT, accessed January 31, 2017.

OMS. Organização Mundial da Saúde. 2015 Available at http://new.paho.org/bra

OPEN KNOWLEDGE FOUNDATION. OKF (2015). “Empowering Though Open Knowledge”. Available at http://index.okfn.org/place/brazil/ accessed January 31, 2017.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. ONU. (2015) Available at http://www.un.org/en

PHIPPS, J. B. (1971). Dendrogram topology. Systematic Biology, 20(3), 306-308.

RACINE, J. S. RStudio: A Platform-Independent IDE for R and Sweave. Journal of Applied Econometrics, v. 27, n. 1, p. 167-172, 2012.

REGOCZEI, S. AND G. HIRST, Knowledge and Knowledge Acquisition in the Computational Context. 1994.

REM-F. Ranking de Eficiência dos Municípios. Available at: https://goo.gl/yoIDTe, accessed January 22, 2017.

SACKETT, D.L., STRAUS, S.E., RICHARDSON, W.S., ROSENBERG, W., AND HAYNES, R.B. Evidence-Based Medicine: How to Practice and Teach. EBM, Second Edition, Churchill Livingstone: Edinburgh, 2000.

SANTOS, M. A Natureza do Espaço: Técnica e Tempo, Razão e Emoção. (4 ed. Vol. 2. reimpr). São Paulo: Editora da Universidade de São Paulo.

SEI. SEI - Software Engineering Institute. Smart Grid Maturity Model (SGMM). 2015 Available at http://www.sei.cmu.edu/smartgrid/

SHARAF, Muhammad A.; ILLMAN, Deborah L.; KOWALSKI, Bruce R. Chemometrics. John Wiley & Sons, 1986.

SHVAIKO, P. AND EUZENAT, J. A survey of schema-based matching approaches. Journal on Data Semantics, IV, 2005, 146-171.

SIMPSON, M. G.. Plant Systematics. 2nd ed. [S.l.]: Academic Press, 2010.

SNEATH, P.H.A.; SOKAL, R.R. Numeric taxonomy: the principles and practice of numerical classification. 1973. San Francisco: W.H. Freeman, 1973. 573p. SOKAL, R.R.; ROHLF, F.J. The comparison of dendrograms by objective methods. Taxon, Berlin, v.11, p.30-40, 1962

SOUZA, J. F. et al. Uma abordagem estrutural para calcular similaridade entre conceitos de ontologias. Revista de Informática Teórica e Aplicada, v. 17, n. 2, p. 249-269, 2010.

STRAUSS, A.; CORBIN, J.. Basics of qualitative research: Techniques and procedures for developing grounded theory . Sage Publications, Inc, 1998.

STRAZZA, L.; AZEVEDO, R. S.; CARVALHO, H. B. Prevenção do HIV/Aids em uma Penitenciária-modelo feminina de São Paulo–SP, Brasil. DST–J Bras Doenças Sex Transm [Internet], v. 18, n. 4, p. 235-40, 2006.

STUDER R., BENJAMINS V. R., FENSEL D., Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering, 25, pp. 161-197, 1998.

THE ECONOMIST. Bright lights, big cities. Urbanisation and the rise of the megacity. URL: https://goo.gl/FKkPkv

TOLÓN-BECERRA, A.; BIENVENIDO, F.. Conceptual modeling in a meta-model of sustainability indicators. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. Springer Berlin Heidelberg, 2008. p. 716-723.

TRANSPARENCY BRAZIL. TB. Portal da Transparencia no Brasil. Available at https://goo.gl/bg0IlB

UNICEF. Água é vida, saneamento é dignidade. 2015 Available at https://goo.gl/jrXw6e

VAN DEN BESSELAAR, P.; BECKERS, D. Demographics and sociographics of the Digital City. In: Community Computing and Support Systems. Springer Berlin Heidelberg, 1998. p. 108-124.

VENABLES, W. N.; SMITH, D. M. The R development core team. An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics, 2005.

VOM BROCKE, J., SIMONS, A., NIEHAVES, B., PLATTFAUT, R., & CLEVEN, A. Reconstructing the giant: on the importance of rigour in documenting the literature search process. ECIS 17th European Conference on Information Systems (pp. 2–13).

WEISS, M. C., BERNARDES, R. C., & CONSONI, F. L. Cidades inteligentes: casos e perspectivas para as cidades brasileiras. URL: https://goo.gl/gJV0az

XI-JUAN L., YING-LIN W. E JIE W. Towards a Semi-Automatic Ontology Mapping - An Approach Using Instance Based Learning and Logic Relation Mining. pp.269-280, Fifth Mexican International Conference on Artificial Intelligence (MICAI'06)

ZEISE, N.; LINK, M.; ORTNER, E.. Controlling of dynamic enterprises by indicators–A foundational approach. In: International Conference on Business Process Management. Springer Berlin Heidelberg, 2010. p. 521-530.