City Typologies: Classification and Characterization

Bruno Miguel Fonseca de Oliveira
Oliveira.bmf@gmail.com

Instituto Superior Técnico, Lisboa, Portugal

October 2014

Abstract

The aim of this thesis is to: create a typology of cities based on indicators from the economic sector; characterise it with several other indicators of consumption and emissions and predict the typology of a random city through a decision tree.

Available for this analysis were 298 global cities and 23 indicators that were processed through various clustering algorithms (K-Means, DBSCAN, EM and hierarchical) in order to obtain the typologies. These were then characterised in two ways. First using only the variables from the economical sector and second using the variables from population, economy, emissions and consumptions. This was done so one could construct typologies based on economic sectors and to avoid problems with the missing values in the second set of variables. With the typologies and characterisation done, the correlations between variables, with and without highlighting the typologies, were analysed and discussed. Finally the decision trees were constructed.

The results will show: the existence of five typologies using K-Means clustering; a characterisation of the typologies based on 8 economical sector and 15 other variables; several important correlations between indicators; the relevancy of the typologies between indicators correlations and two decision trees that predict the city's typology with good accuracy levels.
Keywords: Typology, City, Classification, K-Means, Indicator

1. Introduction

World population has been increasing at an incredible pace for a few decades now and, depending on the estimates, it may very well keep rising for a few more [1]. Accompanying this factor, is also the increasing migration of populations from rural areas to cities [2]. The combined factors are responsible from producing enormous cities without precedent and for increasing the urban population from 13% in 1900 to 67% in 2050 [1].

Cities are also the home to the majority of human activity and it is there that most of the mineral and human resources are concentrated [2]. This concentration of resources creates poles of creativity, wealth creation, invention and investment [3] but they also are responsible for much of the production of waste, greenhouse emissions and water contamination of the planet.

Therefore, cities play a key role in the issue of sustainability. Not only are they a major cause for the deterioration of the environment, but also in them relies the human capital to create solutions for those same problems.

If now one considers the constant increasing in urban population, it is possible to see how important this subject is today and will be in the future.

1.1. state of the art

Cities are the subject of many papers, however city typology has always remained elusive much due to the enormous lack of available data.

The concept of city typologies first began with city hierarchization. In this area one should notice the important work of Peter Hall in 1966 [4] with is book "The World Cities", Friedmann's in 1986 with is "World City Hypothesis" [5] and Sassen in 1991 with "The Global City" [6]. However these works create a hierarchy of cities and not a typology. This hierarchization shows the degrees of importance of each city on the world stage, based on the amount of connections between firms and services in different cities. This type of research is still very active today. The Globalisation and World Cities (GaWC) Research Network provides several data bases and research papers on different forms of hierarchization.

A different approach, called a Metabolic Profile, can be used if now one considers a city as a living organism that consumes resources, produces activ-

ity and expels waste. This approach was first used by Abel Wolman in 1965 [7]. After him, several other authors [8] [9] [10], used also the concept of urban metabolism to create metabolic profiles for the cities. These works did establish some differences between cities and shown that different types of activity are related to different types of consumption and emissions, thus different metabolic profiles.

Alongside with Hierarchization and Metabolic Profiles there are also City Rankings. There are countless number of reports involving city indicators: PWC - Cities of Opportunity, Siemens Green Index, Sustainability Cities Index 2010, Mercer - Cost of Living, UN-HABITAT - State of the World Cities, ATKearney - 2014 Global Cities Index and Emerging Cities Outlook, and many other.

Each of these reports reflects on a certain aspect of the city as the name properly indicates. The process is quite simple, for each report a certain amount of variables are selected and converted into one single index. This index will reflect the state of the city, the higher in the rank the better the city is. This does not divide the cities into groups, although it could be done in a very simplistic way by choosing good, medium and bad cities. These reports also differ widely among themselves, with regard to variables and methodologies used. This means that depending on the data and the methodologies, a city that is well placed in one rank may not be in another.

The approach used in the thesis is different from these three options, but has a little bit on every one.

With the development of the technology, statistical tool and data collection in general, there is a new option. One can construct and characterise city typologies based on several variables, with the help of data mining process. These typologies can then be used to search for geographical distributions and possible connections. Each typology will have certain unique characteristics, which is the same as saying profiles, and ultimately cities within typologies can be compared in search for a certain level of efficiency, this will show which cities are performing better and producing a kind of Ranking.

In all the literature that was consulted, the only attempt to create a typology of cities based on indicators of consumption and emissions was performed by [11]. There is another work [12] that also ties this but it is not so complete. The author uses country values of resource consumption and emission, to estimated values for the corresponding cities via Zipf Law []. Then it uses a decision tree model that classifies the cities. This work appears to be the first of its kind and therefore presents some conceptual issues. In particular, the improper use of the decision tree.

One of the major breaktrough was the report from Brooking Institute that standardizes the values for the economical sector for several global cities. This allows for a general view of how the economic sectors of each city are distributed and construct a typology of cities based on such divisions.

1.2. Objectives

The main objective is to divide the available 298 cities into different typologies based on their economical sectors. This division will be made trough several clustering algorithms (K-Means, DBSCAN, EM and Hierarchical) in order to identify the most adequate.

Furthermore, one shall proceed to the characterisation of each cluster, first with the variables from the economical sector and secondly with variables from population, economy, consumptions and emissions. Thus, creating a detailed profile for each typology.

Another objective, is to observe the correlations between indicators, with and without the typologies highlighted, so to infer about the importance of the typologies for the global correlation.

Lastly, the construction of two decision trees, will enable the prediction of the typology of a city, using economical or consumption and emissions variables, without the need to reuse the clustering algorithm .

1.3. Limitation

This thesis provides a photograph of what can be perceived as a typology today. It does not give, or try, to produce a general definition of a typologies for cities that can evolve with time. The characterisations presented here do not reflect the typologies for cities in future nor do they reflect what they have been in the past.

2. Data

To accurately construct a city typology, one must be supported by reliable data. As was mentioned, the use of national data to estimate values for cities is not enough, because cities have a unique speed and state of development when compared to the country [2]. This emphasises the need for real data collected from cities. In order to obtain the best representation of cities and indicators, several sources were consulted resulting in a database of 298 cities and 23 indicators.

2.1. Variables and cities

The 298 selected cities are mostly country capitals, state capitals, cities with large populations or simply, cities that generate a great deal of influence or importance at national level. The geographical distribution is as homogeneous as possible, although there is a significant discrepancy between industrialised and developing countries. The majority of

cities come from Europe, North America, Japan and Chine, while South America and Africa are the continents with least representation.

Considering accessibility and relevance to the work, a total of 23 variables were chosen for the analysis of the cities. These variables were divided into 2 groups. The first with 15 variables, deals with the specificity of the city and contains information about city area, population, density, Gross Domestic Product, Waste Production, Recycling Rate, Water consumption, non-revenue water, PM10 annual emissions, CO2 emissions (total, energy and transport). For simplicity this group shall be known as PECE. The second group, has 8 variables that contain the percentages of the economical sectors (ES) of the city: Commodities, Local, Manufacturing, Finance, Construction, Tourism, Transport and Utilities.

Since the purpose of this thesis is to create a typology based on the activity of the city, and not on the intakes and outtakes, the second group will be the one responsible for creating the city typologies and determine their differences from a productive stand point. While the first group will be responsible for identifying and characterising the differences between typologies.

2.2. Sources

Information is a tradable good, so it was not surprising that the best databases were only accessible at a fee. Among these, one can mention City Benchmarking Data, Oxford Economics, Mercer or Mckinsey's Cityscope 2.0, just to name a few.

Thus, it was necessary to choose from sources that were open to all, like municipal sites, statistic offices, specialized sites, reports, articles, some open databases and others in order to obtain the necessary information.

Some of the most important sources were: Brooking Institution, Eurostat (New Cronos, Urban Audit, Eurostat Yearbook) Canada and USA online government sites, China (China Yearbook), ODCE (Metropolitan Explorer), UN-HABITAT (Urbaninfo v.2, State of the World's Cities 2012/2013), World Health Organization (PM10 report), UNDATA, IIASA, Siemens (Siemens Green Index), CDP (CDP Cities 2012 Global Report), PWC (Cities of Oportunities), KPMG, C40 (C40 cities), Global City Indicators .

2.3. Problems and Limitations

Although there were several sources for the collection of data, some problems were encountered regarding the collection, accessibility and standardisation of data and the limits to which it can be applied [13].

Most of the available sources and data are for developed countries, mostly Europe, USA and Japan. This makes any analysis a little biased towards these economies and dismisses the reality in developing countries.

The number of missing data was also an issue. Although, from the 23 variables, the 8 from the ES had no missing values, the remaining 15 PECE, were in average only 50% complete. The reason for this average value, lies not only with the lack of available data but also with the lack of time for collecting it. This process is sometimes quite strenuous, as it is often necessary to look individually, each variable for each city.

Another issue had to do with the definition of a city, since currently there is no convention in how to define its area. This implies that the data recovered from the cities, may vary substantially from country to country depending on their definition. Some reports try to provide a standardised model for these areas [14] [15] [16], However they are far from being the norm.

City area is not the only indicator suffering from lack of definition. Some variables change theirs from country to country or even from city to city. Water consumption may or may not include the consumption by the industrial and commercial sectors, and the rate of recycling is confusing because in some places, burned waste is considered energy recycling.

3. Algorithms

The process of Data Mining allows to categorise and summarise large quantities of data into useful information, in order to observe correlations that otherwise would not be possible. This process is composed of several steps like data preparation, cleaning, clustering algorithms, selection and interpretations of the results [17].

In the case of the clustering algorithms, there are several options available depending on the size and type of data or even on the speed and accuracy desired [18].

3.1. Hierarchical VS Partitional clustering

Clustering algorithms can be divided into two main groups: Hierarchical and Partitional [19].

Hierarchical algorithms repeat, at each cycle, the process of either merging smaller clusters into larger ones or dividing larger clusters to smaller ones. Either way, it produces a hierarchy of clusters called a dendrogram that is a very clear visual representation of the merging process.

Partitional algorithms creates several partitions that are then evaluated and then rearranged, optimising the initial partitions until a stopping criteria is archived.

Hierarchical and Partitional Clustering have key differences in running time, assumptions, input parameters and resultant clusters. Partitional clustering is faster and more accurate than hierarchical

clustering but requires stronger assumptions such as number of clusters and the initial centers. Hierarchical clustering requires only a similarity measure, does not require any input parameters, the dendrogram returns a much more meaningful and subjective division of clusters, and is more suitable for categorical data as long as a similarity measure can be defined accordingly [20].

3.2. K-Means

K-Means is a partitional clustering method, where data is divided into K clusters and where each object is assigned to precisely one cluster. The algorithm works in this manner: first select K points as initial centroids; Assign all points to the closest centroid; Recompute the centroid of each cluster; Repeat 2 and 3 until the centroids don't change [19].

What this really does, by moving the centers, is minimise the sum of the squared distance between points and the cluster "center" [20][38] or in mathematical form to minimise the expression:

$$\sum_{i=1}^{k} \sum_{\overrightarrow{x} \in c_i} \|(\overrightarrow{x_j} - \overrightarrow{c_{ij}})\|^2 = \sum_{i=1}^{k} \sum_{\overrightarrow{x} \in c_i} \sum_{j=1}^{d} (x_j - c_{ij})^2 \quad (1)$$

where $\overrightarrow{x}$ is a point, $\overrightarrow{c_i}$ is the "center" of the cluster, d is the dimension of $\overrightarrow{x}$ and $\overrightarrow{c_i}$, and, x_j and c_{ij} are the components of $\overrightarrow{x}$ and $\overrightarrow{c_i}$.

In the absence of numerical problems, this procedure always converges to a solution, which is typically a local minimum due to the dependency of the initial values. [](Selim and Ismail, 1984). Thus, this dependency becomes a key step of the basic K-Means procedure because it greatly influences the end results. Since, in most cases, the best option is unknown, choosing the initial points may require trying a few options.

K-Means is easy to interpret, simple to implement, has a good speed of convergence and good adaptability to sparse data. In the case of a sample with m instances, N attributes, T iterations and K clusters the complexity is given by: $O(T*K*m*N)$. Its linearity makes it more advantageous in comparison to other clustering methods which have non-linear complexity like hierarchical. However this is an algorithm that is very sensitive tho the initial values; works best with data that have isotropic clusters and can be somewhat slower due to the recalculations of all distances in each step.

3.3. DBSCAN

DBSCAN [21], is a density based hierarchical algorithm that creates clusters by connecting regions of high density of points, separated by regions of low density. This algorithm is particularly efficient in cases were the distribution of data is not circular or has strange shapes and is also very good in dealing with noise and outliers. However, if the density of points in each clusters is not constant, the algorithm starts to perform very poorly.

3.4. Expectation Maximisation (EM)

The EM algorithm [20] [22] is similar to K-Means, but instead of unequivocally assigning each point to one cluster it computes the probability of cluster membership based on one or more probability distributions. In other words, each data point belongs to all clusters with a certain probability in each one. The process by which the algorithm is capable of assigning probabilities is somewhat long, and is further described in [23].

This algorithm applies especially for those cases where clusters might not be completely separated, implying a certain level of overlap. In those cases, it is necessary to determine to which extend each point belongs to each cluster.

3.5. Hierarchic

The hierarchical algorithm [19] [20] produces a series of nested clusters, ranging from clusters of individual points at the bottom to an all-inclusive cluster at the top. This process is best described by a diagram, called a dendrogram, which displays graphically the order in which points are merged. In order to agglomerate points or clusters, one must first define what is the proximity between them. There are several options for this depending on the objective, being the most common the Single linkage, Complete linkage, Group Average Linkage and Ward's Method [20]

The algorithm is characterised by its versatility. It maintains good performance on data sets containing non-isotropic clusters if the clusters are well separated. Another advantage is that it does not assume any particular number of clusters. Instead, any desired number of clusters can be obtained by cutting the dendrogram horizontally at the desired level.

However, since no global objective function is being optimised, this means that merging decisions are final. This means that even though decisions at a local level may be good, the final global result may be poor.

3.6. Decision Tree

Decision tree model, is a tree based graphic that allows for the classification of previously unclassified data points. For this purpose it uses dependable variables (inputs) in correlation with an independent one (target or classification), to allow the creation of paths that classify the points [20].

Decision Trees are generated by recursive partitioning. Depending on the splitting method, in each step the method divides one variable creating two or

more groups. In general, the recursion stops when all instances have the same label value, i.e. the subset is pure or if a certain stopping criteria is reached.

Several level of complexity can be obtained by this method. However, to avoid the problem of overfitting one should consider the tool called pruning. This option removes paths that do not add to the discriminative power of the decision tree, making the resulting nodes more simple and easy to understand at the cost of some accuracy.

It is important to mention that this method requires a previous classification to work properly. Thus, in this thesis, the K-Means was first used to create the classification, and then the decision tree gave the model based on that information. This, of course, involves some drawbacks.

This model is important to avoid the recalculation of the K-Means algorithm every time one new city is added to the database and to provide a simple model that can be easily interpreted.

4. Results

Each algorithm listed above produced a set of results. These were analysed individually and then compared, so one could choose the best algorithm, taking into account the characteristics of the data and the quality of the results.

4.1. K-Means

This method has two main difficulties: its great dependency on the initialisation of points and the impossibility to determine " priori" the optimal number of clusters.

To minimise these issues, one explored two different ways of initialising the K points: Maximising the Initial Distance (MD) and Constant Intervals (CI) [24]. For each initialisation the algorithm was run several times, varying the value of clusters (K) from 2 to 20. In each run, two values were collected: total distance and distance within clusters. These values were then used in the determination of the best initialisation and value of K.

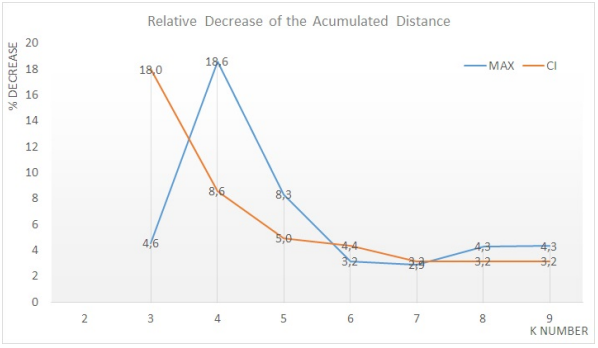

Figure 1: Relative distance decrease in K clusters

The results showed that although both initialisations produced very close results, CI produced a more specialised set of clusters, i.e, cities that had their predominant characteristic more evident. After all considerations, the best result was obtained for 5 clusters with CI.

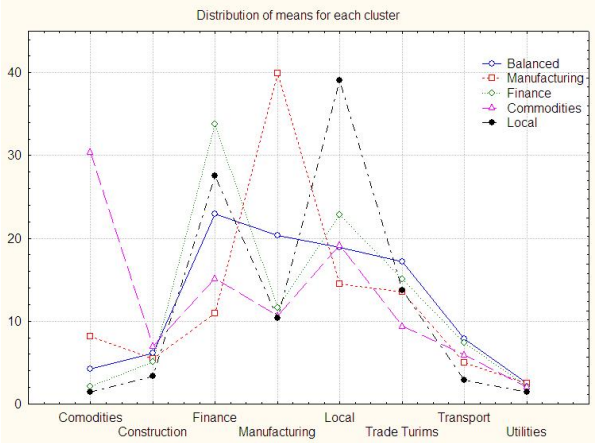

Figure 2: Average values for the economical sectors of the typologies

Each cluster was analysed for its characteristics and given the name of its predominant variable, thus obtaining the names: Finance, Local, Manufacturing, Commodities and Balanced. The cluster Balanced earned its name by not presenting any predominant variable.

Furthermore, analysing both the average values and the variables distribution for each cluster, enabled the construction of a characterisation for each typology.

4.2. DBSCAN

DBSCAN has used in hopes of improving the clustering process, since the natural shape of the clusters was unknown and this algorithm is more flexible in the presence of non circular shapes.

DBSCAN uses only two parameters, distance (ε) and core points ($MinPts$), to produce the clusters. In order to obtain the best combinations of both, they were chosen in a grid pattern, varying ε between 4 and 7,5 in intervals of 0,5 and $MinPts$ from 3 to 7 in intervals of 1. Since this algorithm does not depend on initial points, to each combination of parameters there is only one possible result.

The best results were obtained for ε=5 with MinPts= 4 , 5 and 6 and for ε=5.5 with MinPts=6. For the remaining cases, either there were too many clusters with few points or too few clusters with one exceedingly large.

DBSCAN presented some difficulties. Most of the clusters were too small and in every run, of the algorithm, there were always a considerable amount of unclustered points. Most of these difficulties come in part due to the different densities of points and

because the clusters were not compact and spaced out. Ultimately, this algorithm only served to get a sense of the shape and distribution of data.

4.3. Expectation Maximisation (EM)

Due to the presence of a Balanced cluster, that represents a group of non specialised cities, the possibility of overlapping clusters arose.

These overlapping areas could be the intersection of the other four typologies or any combination of intersections between two or three clusters. The objective was, check to which extent clusters shared points between each other and if could be possible to obtain only four clusters, where Balanced cluster would correspond to the overlapping area of the other four.

Two programs were used to produce the results, Statistica and KNIME. Expectation Maximisation produced only slightly different results depending on the program that was used and the chosen parameters. Most of these results remained consistent with the characterisation of the clusters produced with K-Means.

The cluster of Commodities was the most affected by this algorithm, getting completely mischaracterised if too much overlapping was forced, and disappearing completely if only 4 clusters were considered. This is partly explained due to the fact that the cluster is very small and very specialised. Meanwhile, the cluster Balanced suffered very few changes, when the value of K changed, reinforcing the premise that it exists by its own merit and not as a result of overlapping clusters.

The presence of several cities with significant probability of existing in more than one cluster, evinced in both results from both programs, indicate the possibility of a transitional phase between clusters, and shown the possibility of a continuous evolution path, from one typology to another.

However, the algorithm proved to be somewhat volatile, due to its dependency on the initial values. Because of that, the most important information that was retrieved from the algorithm, was the fact that for small values of overlapping, the result from K-Means is almost recover with the added information about the presence of several cities with mix probabilities.

4.4. Hierarchical

In the hierarchical algorithm, several options were available to choose from. However, since the objective was to obtain clusters with close characteristic and small deviations, the Complete Linkage was the best choice.

However, since this algorithm suffers from the inability to adjust or optimise the results once the points are clustered, which can cause severe errors, there were little expectations for this algorithm.

Despite of this, one found that the results were actually very good. In fact, for the case with 7 clusters, in reality 5 clusters plus 2 clusters of isolated points, the values were remarkably similar to the K-Means case. The overall characteristics of the produced clusters were the same as in K-Means and EM with only small deviations.

4.5. Best Result

In sum, one should notice that the results from K-Means, EM and Complete Linkage, although not absolutely equal, all agree among themselves about the characteristic variables and about the average values of the economic variables in each cluster. Nonetheless K-Means was chosen as the best option because it was the one that proved to be the best fit and the one who best performed in relation to the other algorithms. Although there is not an absolute certainty about the results obtained by this algorithm, when comparing these results with other algorithms it is possible to observe some consistency between them, thus solidifying the results.

4.6. Decision Tree

The results from K-Means opened the possibility to assign a label to each city (Finance, Manufacturing, Commodities, Local and Balanced), enabling the construction of a decision tree that identifies the typology of a city through a simple diagram. This will provide an approximation model that can be easily used each time a new city is added. This avoids the need for countless recalculations of the K-Means algorithm each time a new city need a classification. It also provides a simple way for a common person to use and understand these results without requiring computational skills or extensive knowledge about algorithms.

It should be mentioned that this process is somewhat flawed due to its dependency on the K-Means classification. Two tree models were constructed. The first model was completely expanded, showing all the possible paths necessary to identify unequivocally the typology of a city (overfitting), while the second was reduced in size sacrificing some accuracy but making the model much simpler (pruning).

5. Results 2

With the typologies created in the previous section, one may now proceed to identify differences among typologies based on the indicators of the PECE variables. This was done by comparing average values and variable distribution for each cluster with the new variables. After this, one analysed the correlations between variables, with and without the typologies highlighted in search of different correlations. Lastly, the creation of a new decision tree, allows to predict the typology, now based in the PECE variables.

5.1. Characterisation of typologies
Considering the average values of the PECE variables for each cluster, several important differences between the typologies have appeared.

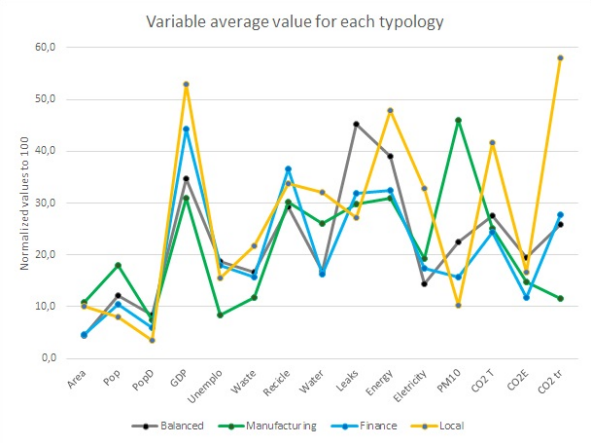

Figure 3: Average values for the PECE group of indicators, per typology

It was found that the typology Local tops in 7 out of 15 variables. These cities have the most GDP per capita, the biggest consumption of water and energy, produce the most waste and emit more CO2 Total and CO2 Transport. By opposition the Manufacturing typology appears on the bottom of 5 variables, and in 3 that are very close of being last. This typology is the one with the least GDP per capita, waste, energy consumption, unemployment and general CO2 emissions. Balanced and Financed typologies, almost coincide in 9 variables, and the only variable where they differ significantly is in the percentage of Non-Revenue Water. In most cases they are located between Local (above) and Manufacturing (bellow).

Analysing now the distributions of the variables, it was seen that in the majority of cases, the results reflected the same reality as the average values, confirming that the typologies with most and least production of waste, water consumption or CO2 emissions remained the same.

5.2. Correlations with typologies
The purpose was to identify the most relevant global correlations and observe if the introducing of typologies, would produce the same type of correlation.

Analysing first the global correlations, one found that most of them follow a power law or a logarithm law. This may indicate that some variables, may be influence by a scale economy like outlined in [3].

The introduction of the typologies showed a clear subdivision of the global correlation. Although most of the typologies follow the same trend line as the global correlation, they clearly differ between

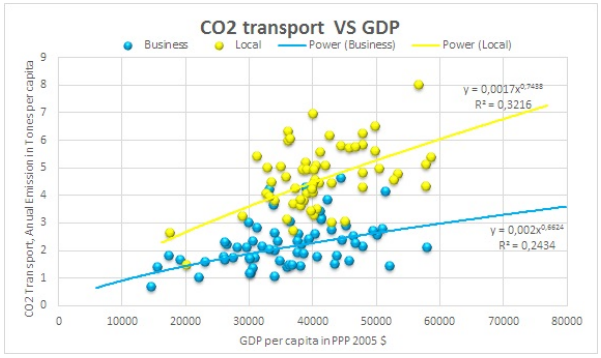

Figure 4: Correlation between CO2 and GDP, including typologies

them, like the cases of GDP Vs Area or CO2 Transport Vs PM10. In some cases the sub-correlations did not even follow the expected trend line, like PM10 Vs GDP or Water consumption Vs Density. These cases show that a global correlation, is frequently composed of several sub correlations that do not necessarily share the same trend line.

Also very remarkable, were the cases where the introduction of the typologies did not produce any effect, as Unemployment Vs GDP, Non-Revenue Water Vs GDP and Non-Revenue Water Vs Recycling are such cases. This clearly indicates that some correlations are immune to the typology of the city.

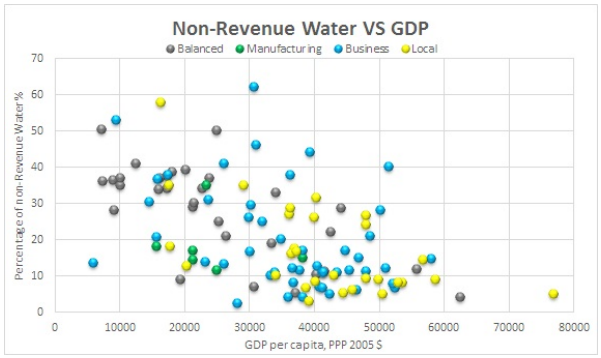

Figure 5: Correlation between Non-Revenue Water and GDP, including typologies

5.3. Decision Tree
The program used to construct the decision tree was the IBM SPSS Statistics 20, because it offers a wide range of choices and parameters to optimise the results. The program was run for three growing methods: CRT, CHAID and exhaustive CHAID. The method that produced the best results, with high accuracy and low number of nodes, was the CHAID [25].

The advantage of this method is that it includes ramifications for the cases of missing data, mak-

ing the tree much more understandable and usable, since without this option, it would be very tricky to select a path, if a random variable had no value available.

The method had full use of the 15 PECE variables. Even though it only used PM10, CO2 Transport, CO2 Energy, Energy, Unemployment Rate and GDP it obtained a 73,5% of correct prediction with 26 nodes. There is a curious fact about this model. It is possible to verify that the cities of the Manufacturing typology are the ones with the least amount of indicators available. Thus, it is not difficult to understand how the model assumes that cities with no data are from this type. This shows one of the weaknesses of this decision tree and reinforces the need to work with databases that are as complete as possible.

5.4. Conclusion

In conclusion, with the typologies that were previously created, it was possible to define characteristics for each typology and observe differences between them based on the PECE variables. This approach was necessary, not only because it made more sense to first use the economical sector to establish the typologies, but also because the high level of missing values in the PECE variables, would severely impair the creation of the typologies.

The first results, compare the average values of each typologies withing each variable. They showed several differences between the typologies, especially in variables like the PM10 and CO2 Transport emissions and Energy and water consumption. These results were further confirmed by comparing the variable average values with their distributions. The majority of cases showed that the mean values were in agreement with the distributions and only in the cases of GDP and Energy were there found some small deviations.

The correlation between variables, with and without the typologies inserted, showed some very interesting cases. For the most part, the introduction of typologies showed that a global correlation is composed of several sub correlations belonging to each of the typologies that may or may not follow the same distribution as the global correlation. The most evident cases of differentiation between typologies were found in CO2 Transport VS Area and PM10 VS GDP per capita. There were also cases where the introduction of typologies did not produced any kind of differentiation. This may indicate that some correlations are immune to different typologies and exist above them. These were the cases of Unemployment rate VS GDP per capita and Non-Revenue Water and Recycling Rate.

The use of the tree model here is debatable, given the conditions that lead to the creation of typologies in the first place. However, since there is no other way to circumvent these issues, and despite only working in a qualitative way, the decision tree that was constructed, is the best tool that one has, currently available to classify a city based on their population, economy, emissions and consumptions. The level of precision of the decision tree, depends on the objectives of the work, however the model that was produced, gave a reasonable balance between accuracy and complexity.

The number of missing values is of the most importance. One should expect that a more complete data base will produce a more accurate set of results and possibly enable the construction of the typologies based in more than just the variables.

6. Conclusions

In this thesis, there were two main objectives: first and foremost, the creation of a typology for cities and second the characterisation of these typologies so they could be differentiated.

The first objective was achieved by combining the variables from the economic sector with several clustering algorithms. This was a slow process that required several attempts, but that produced a solid result, where cities were divided into five groups (Commodities, Balanced, Manufacturing, Local and Balanced), now called typologies.

These several algorithms had also served to identify some characteristics of the data, that are relevant for the clustering process. Most interestingly was to verify that the clusters were not equally dense, which causes problems for density based clustering algorithms, and that they overlapped in some points. Nonetheless, from all the clustering algorithms, it was the K-Means that proved to be the most reliable.

With the typologies defined, by comparison between them, one found that each one had certain characteristics that made them easily identifiable. The most important was that each typology had a characteristic variable with a very high value that did not repeat in any other typology. Thus, one proved that it is possible to separate the cities into several groups, based on their economical activities. Doing it, then, allowed for the characterisation of each cluster based on those same variables. Next, the first decision tree model was created and helped to classify each city based on the economical variables. Despite some conceptual problems, this model can predict the typology of a city with over 90% accuracy in a tree with only 10 nodes, thus avoiding new recalculations of the clustering algorithm.

In the second main objective of this thesis, the remaining variables related to GDP, population, consumptions and emissions were used to characterise

even further each typology. The two derived results, average values and variable distributions, showed significant differences between typologies making it possible to determine some patterns of consumption and emissions for each typology. For example it, was found that cities from the Local typology had, in average, more GDP per capita but they also had more waste production, water and energy consumption and CO2 emissions, while on the opposite side of the spectrum were the manufacturing cities with low values of these variables.

Looking at the correlation between variables, they presented several interesting cases. Alone, global correlation showed strong interaction between several variables like PM10 and GDP per capita or CO2 Transport and PM10. These were enough to identify some very important characteristics of the cities but the introduction of the typologies enriched much more this analysis. It was seen that although many typologies follow the global correlation, they display different rates between them and in some cases, the typologies do not even follow the global correlation. A few special cases even showed no differentiation when typologies were involved showing that those variables were transverse to all typologies. Thus, showing that beneath the global correlations between variables, there is specific reality for each typology that normally is hidden.

The second decision tree help to improve even further the classification of cities. Using now the consumption and emissions variables, the model predicts with little above 70% accuracy the typology of the city. This model, even though suffers from the same issues as the first decision tree, becomes very important due to the existence of missing data. Currently, this model is the only tool available to predict the typology of a city though emission and consumptions variables, that even allows for some missing data.

At the end of this thesis both main objectives were completed. It is now possible to classify a city, using one of two options and there is a characterisation of each typology based on their economic sectors and in their consumptions and emissions.

6.1. Future Work

There were several considerations and paths that were not explored in this thesis, some because of data availability, most because of lack of time and some due to technical difficulties.

It was seen with Expectation Maximisation that some clusters overlap, and this may in fact be a important stage of transition for cities, making this a subject that should be deeper investigated. Also, the number of algorithms should be expanded to include more recent one like CLIQUE, CLARANS or BIRCH in hopes of add more information or better results.

One remarks that an important connection should be done with the work produced by George West. It was found that certain variables follow a power law when the city size scales up or down. It would be very interesting, since this thesis show that not all typologies follow global correlations in the same way, to see if these power laws would be also applied for all typologies or if they only apply when considering cities in general. A creation of a level of efficiency for the cities should be investigated. With the typologies it is now possible to predict the values for consumptions and emissions taking into account several other indicators. It should be possible to calculate the efficiency of a city by comparing its values with the ones expected for a city like that.

Moreover, it would be very interesting to observe these cities for a considerable period of time, not only to understand their evolution but also to see how the typologies would evolve.

Another very important improvement would be the introduction of more indicators like, number of inhabitants per house, average utilisation of motorised vehicles, number of PhDs, and others that could be analysed in blocks so not to overburden the process of classification. Associated with the number of variables, the reduction of the number of missing values should also be a priority in order to obtain more accurate results..

Finally, it would be interesting to use the remaining 150 cities, for which only the values for the variables of consumption and emissions were available, to determine their typologies.

Acknowledgements

I would like to thank my girlfriend, for her incredible support and patience, Andre Pina for his inputs and help in crucial moments, professor Paulo Ferrao For giving me the opportunity to work in this amazing field and Carlos Silva for his help in the later stages of the thesis.

References

[1] United Nations. World urbanization prospects 2014, 2014.

[2] UN-Habitat. State af the world's cities. report HS/080/12E, UN-Habitat, https://portal.gs.com, 2012.

[3] Dirk Helbing Christian Khnert Lus M. A. Bettencourt, Jos Lobo and Geoffrey B. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306, Mar 2007.

[4] Peter Hall. *The world cities*. McGraw-Hill, 1966.

[5] Jonh Friedmann. *The World City Hypothesis*, volume 17, pages 69–83. 1986.

[6] Saskia Sassen. *The Global City: New York, London, Tokyo*. Princeton University Press, 1991.

[7] Abel Wolman. The metabolism of cities. *Scientific American*, 213(3):179–190, Sep 1965.

[8] John Cuddihy Christopher Kennedy and Joshua Engel-Yan. The changing metabolism of cities. *Journal of Industrial Ecology*, 11(2), 2007.

[9] Marina Fischer-Kowalski. Societys metabolism - the intellectual history of materials flow analysis, part i, 1860-1970. *Journal of Industrial Ecology*, 2(1):61–78, Jan 1998.

[10] Heinz Schandl Krausmann Fridolin, Marina Fischer-Kowalski and Nina Eisenmenger. The global socio-metabolic transition: past and present metabolic profiles and their future trajectories. *Journal of Industrial Ecology*, 12(5-6):637–656, 2009.

[11] Artessa Nicola D. Saldivar-Sali. Master's thesis.

[12] KPMG. City typology as the basis for policy. report, KPMG, 2010.

[13] Christa Anderson. City sustainability indicators. World Bank.

[14] OCDE. *Redifining "Urban" - A new way to mesure metropolitan areas*. OCDEpublishing, 2012. 10.1787/9789264174108-en.

[15] M. Piacentini and K. Rosina. Measuring the environmental performance of metropolitan areas with geographic information sources. *OECD Regional Development Working Papers*, 2012/05, 2012. http://dx.doi.org/10.1787/5k9b9ltv87jf-en.

[16] Eurostat. *Eurostat regional yearbook 2013*. Eurostat, 2013. ISSN 1830-9674.

[17] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 1996.

[18] Vladimir Estivill-Castro. Why so many clustering algorithms: a position paper. In *ACM SIGKDD Explorations Newsletter*, volume 4, pages 65–75. ACM New York, NY, USA, June 2002. ISSN: 1931-0145.

[19] Eui-Hong Han. An introduction to cluster analysis for data mining, February 2000.

[20] Lior Rokach and Oded Maimon. *Clustering Methods*, chapter 15, pages 321–352. Springer US, 2005.

[21] Martin Ester, Hans peter Kriegel, Jrg S, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[22] A. P. Dempster; N. M. Laird; D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1977.

[23] Ted Pedersen. The em algorithm: Selected readings, June 2001. Good starting point in EM.

[24] Statistica. *STATISTICA Electronic Manual*. StatSoft, 2005.

[25] IBM. *CHAID and Exhaustive CHAID Algorithms*. IBM. ftp://ftp.software.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/TREE-CHAID.pdf.