Decision Tools To Analyse Immigrant Territorial Distribution (The Porto Metropolitan Area)

Luís Tiago Paiva
Departamento de Matemática Aplicada
Faculdade de Ciências, Universidade do Porto, Portugal
email: ltpaiva@fc.up.pt

Emília Malcata Rebelo
Departamento de Engenharia Civil
Faculdade de Engenharia, Universidade do Porto, Portugal
email: emalcata@fe.up.pt

February 2006

Abstract

The main purpose of this study consists in the development of a set of mathematical tools - based on decision tree methodologies - that allows the systematised approach to the spatial immigrant population distribution in metropolitan areas.

The case study[1] reported is applied to Porto city (Portugal), and analyses the spatial distribution of immigrant dwellings (according to their country of origin), as well as the distribution of work places considering immigrant dwelling locations.

The developed tools, as well as their cartographic interface, also developed in this research study, are particularly important in demographic and population studies, because they allow the cross-sectional characterization of definite population groups, as well as the anticipation of probable dwelling and work location behaviours. They are also relevant in the definition of immigrants policies, namely in relation to employment and habitation, and in the monitoring of their respective evolution processes.

1 Immigration in Porto Metropolitan Area

The Porto Metropolitan Area[2] is located in the North of Portugal, and is set up by nine municipalities: Espinho, Gondomar, Maia, Matosinhos, Porto, Póvoa de Varzim, Valongo, Vila do Conde and Vila Nova de Gaia, as can be seen in the figure 1.

[1]This article reports part of the research project "Urban Planning for Immigrant Integration", financed by Fundação para a Ciência e a Tecnologia (Portugal)

[2]This article reports to 2001 census data

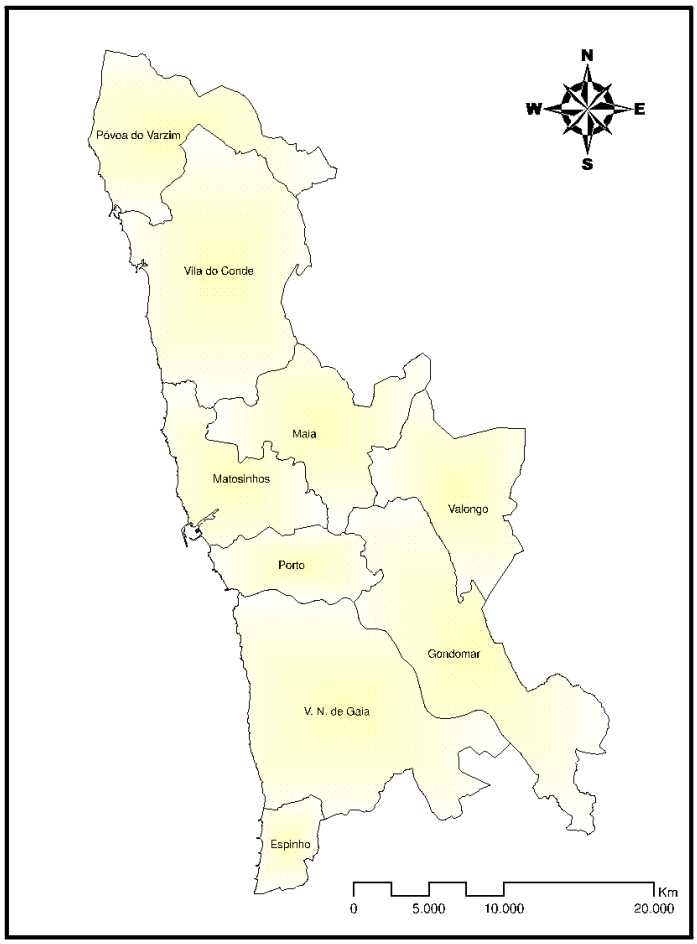

Figure 1: Municipalities in Porto Metropolitan Area

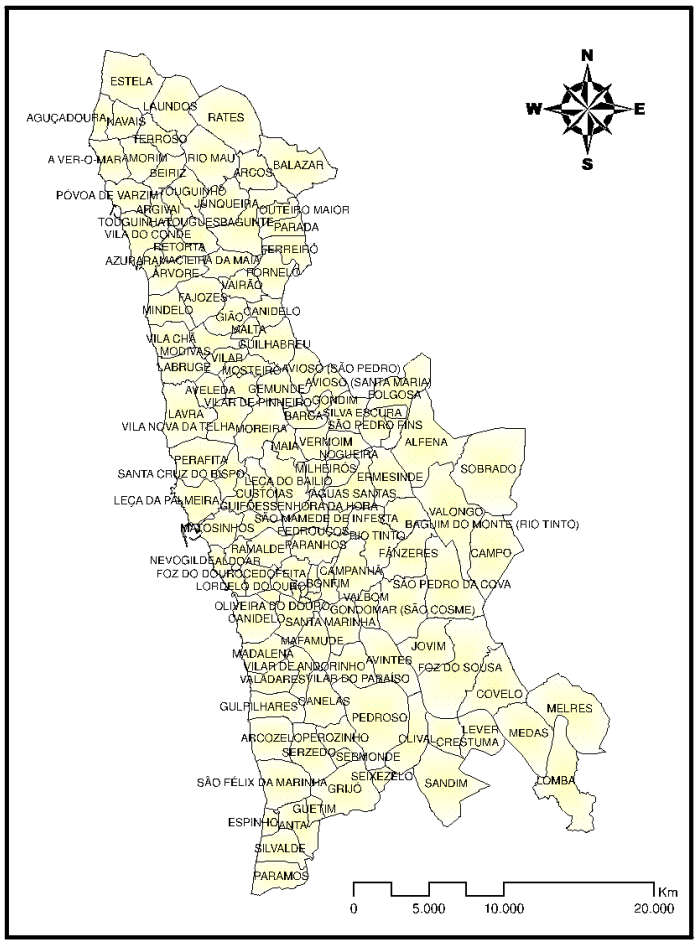

Figure 2: Parishes in Porto Metropolitan Area

The parishes that make up Porto Metropolitan Area are displayed in the figure 2.

In Porto Metropolitan Area inhabited in 2001 1.208.026 Portuguese people and 52.654 immigrants, that represented about 4.4% of total population. According to the countries of origin, foreigners from African countries with Portuguese official language were predominant (45.3%), whereas 23.4% came from European Union countries, 17.3% from other foreign countries, 12.4% from Brazil, and 1.6% from Eastern countries [4].

In order to develop the models reported in this article, a huge amount of data was collected from the population and the habitation census performed in 2001 by Instituto Nacional de Estatística [1]. The information collected concerned demographic and professional characteristics,

- Country of origin.
- Employment situation (with economic activity: employed or unemployed).
- Sector of economic activity (according to the economic activities' classification of instituto Nacional de Estatística).
- Professional assignment (enumeration of the professions according to the national professions classifications, of the same institute)
- Professional group (army; public administration, directors and firms upper staff; intellectual and scientific experts; intermediate level technicians and professionals; administrative staff and similar; service staff and sellers; farmers and qualified workers of agriculture and fishery; workmen, craftsmen and similar workers; operator of plants and engines and assembly workers; non-qualified workers)
- Professional situation (boss/employer; worker on his own account; worker on others' account; other situations).
- Dwelling location (municipality of residence and parish of residence).
- Work location relatively to dwelling location (located in the parish of residence, in another parish of the same municipality, in another municipality, or abroad).

2 Independence Test

Municipality of work versus Municipality of residence

The analysis of the comparative location of immigrants' dwellings and work places suggest a relation between these two indicators, as can be observed in the figures 3, 4 and 5.

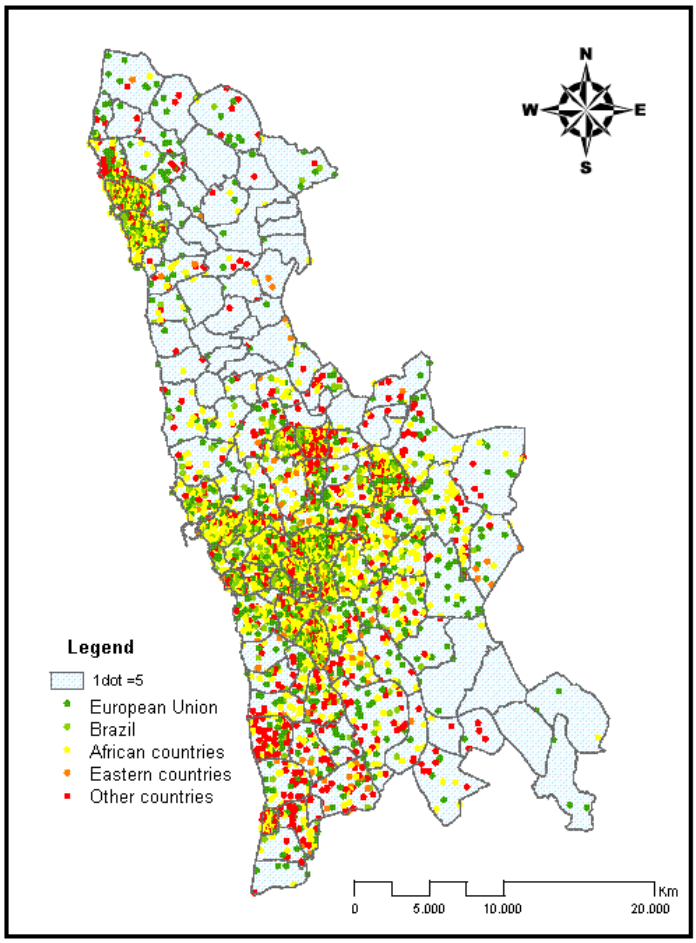

Figure 3: Spatial distribution of immigrants that work in the parish of residence, in Porto Metropolitan Area

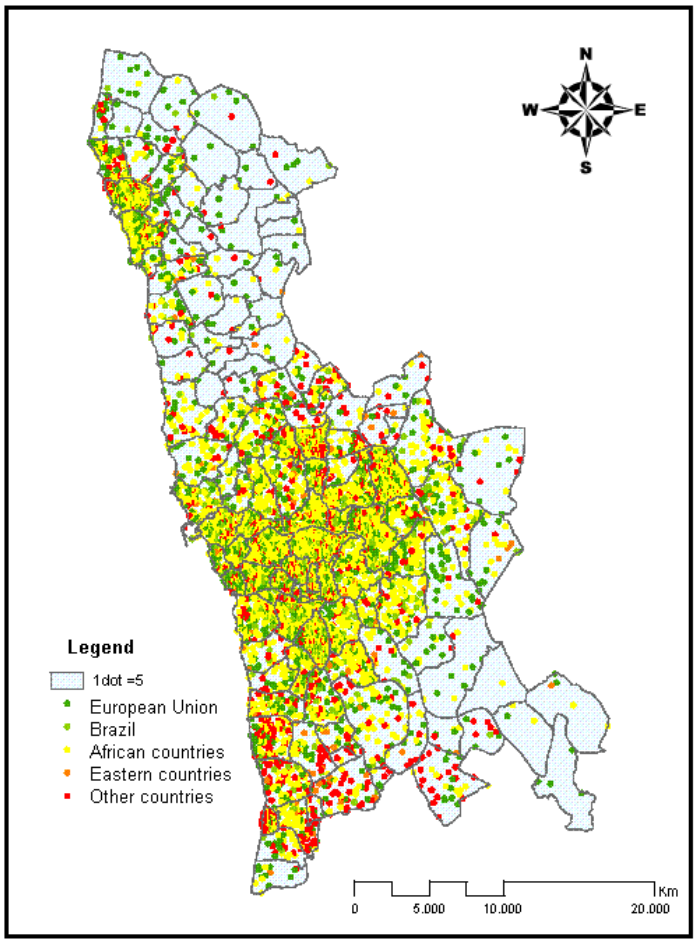

Figure 4: Spatial distribution of immigrants that work in another parish of the dwelling municipality, in Porto Metropolitan Area

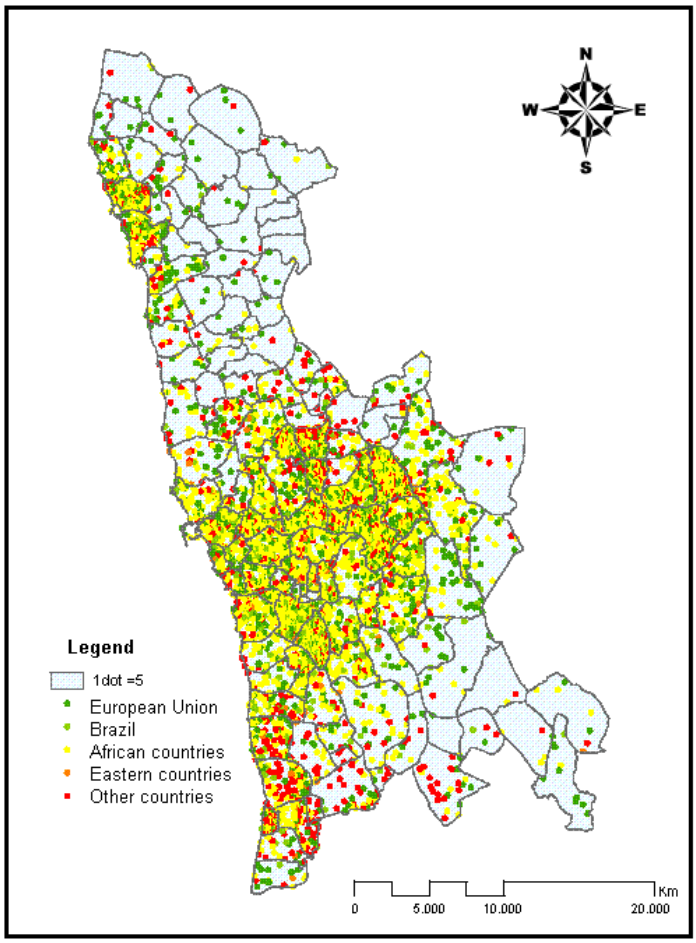

Figure 5: Spatial distribution of immigrants that work on other municipality, in Porto Metropolitan Area

Thus, a study was carried through on the correlation between the variables:

X: Municipality of work

Y: Municipality of residence

and another one considering the variables:

X': Municipality of work in Porto Metropolitan Area

Y': Municipality of residence in Porto Metropolitan Area

The variables Y and Y' are identical since in this study were considered only the individuals resident in the Municipalities of Porto Metropolitan Area. However X and X' are different, since X' excludes the individuals that, despite inhabiting in the Porto Metropolitan Area, work outside of this area and, therefore, $X' \in X$.

For both cases, SPSS 13.0 software was used to carry through the χ^2 Test for the essay of hypotheses H_0 versus H_1 [3], being considered:

H_0: The variables are independent

H_1: The variables are not independent

For both tests, taking the significance level $\alpha = 0.05$, the results have been considered as being statistically significant and have resulted in the rejection of H_0, what means that H_1 should be accepted, that is, the variables X and Y are dependent, and the variables X' and Y' are dependent as well.

The second test, considering the variables X' and Y', was carried out because, even though the first one is statistically significant, it did not verify all the applicability conditions. An imposed condition is the fact that all the modalities of X and Y must have an expected value greater than 5, $E(X_i) > 5$ and $E(Y_j) > 5$. However, as some municipalities out of Porto Metropolitan Area were considered, the majority of these present an insignificant amount of individuals, thus the expected value of some modalities is greater than 5. For variables X' and Y', however, the imposed condition is verified.

Finally, an analysis of the correlation coefficient of Pearson was performed. Once more, in both cases, the results have been considered statistically significant by SPSS 13.0: the correlation coefficient between X and Y was 0.146, and the correlation coefficient between X' and Y' amounted to 0.729. Considering only Porto Metropolitan Area, the correlation between the municipality of work and the municipality of residence is high, what shows the strong dependence between these variables, even though there is evidence of a lower correlation between the municipalities of residence and the ones of work wherever their location is.

3 Tree-structured Classifiers

The terminology of trees is graphic, although conventionally trees are shown growing down the page. The *rout* is the top node, and examples are passed down the tree, with decisions being made at each *node* until a terminal node or *leaf* is reached. Each non-terminal node contains a question on which a split is based. Each leaf contains the label of a classification [5].

A classification tree is a sequence of questions that can be answered as yes or no, plus a set of fitted response values. Each question asks whether a predictor satisfies a given condition. Depending on the answers to one question, you either proceed to another question or arrive at a fitted response value.

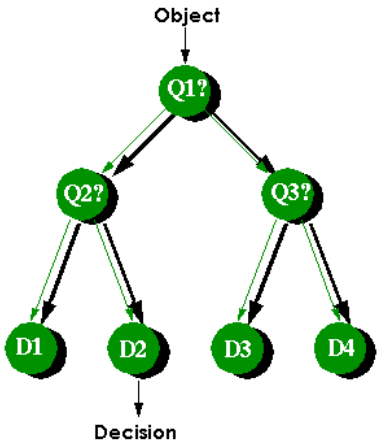

Figure 6: Structure of a decision tree

A classification tree partitions the space $\mathfrak{X}$ of possible observations into subregions corresponding to the leaves, since each element of the data set will be classified by the label of the leaf it reaches. Thus decision trees can be seen as a hierarchical way to describe a partition of $\mathfrak{X}$.

3.1 Theoretical introduction

A hierarchic method does not produce an only partition of the data set $\mathfrak{x}$ but a family of partitions indexada by a parameter T that it is assigned by scale. The family of partitions must possess the following property: *if two patterns x and y belong to the same aggregate for a scale T, then they belong to same aggregate for all the scales T.*

This property allows us to represent the process of grouping by means of a tree whose we are associates the values of the scale T. The leves of this tree represent patterns of $\mathfrak{x}$, and the nodes represent sets of patterns. Two nodes are related in the same scale T if the respective patterns belong to one same aggregate for all the scales $T' > T$.

In a hierarchic method, the final partition depends on the value chosen for T. If it is sufficiently small, each pattern forms an aggregate. When T tends for infinity, each pattern forms an aggregate. As T controls the scale which the data is analyzed, it is necessary to determinate the most appropriate T-scale that fits the problem.

The methods of hierarchic classification divide in *agglomerative methods* and *partitive methods*. The agglomerative methods start with a large number of aggregates and then proceed to a process of fusing aggregates, controlled by the parameter T that increases along the classification process. This paper reports the application of a partitive method of hierarchical classification. It means that it begins by a unique aggregate that is successively divided as the T parameter diminishes, thus the graphic is constructed growing up.

A decision tree consists in constructing a classifier from a data set. Thus, it is usual to grow the tree successively in the nodes. The construction of the tree is easier when there is an exact partition of X (a partition that classifies

all data correctly). In this situation, the tree develops continuously until each element of the data set is correctly classified. However, to do so would be too complex and in these cases two possible strategies arise: to stop the growth of the tree prematurely or to prune the tree after constructing it.

How to at each node

The type of partition used in each node can be decisive on the performance of the construction of the decision tree. The treatment of discrete characteristics is distinct of the continuous case.

A branch for each attribute

The most usual test is the one that attributes to each characteristic a specific branch. This type of partition, even if allows the extraction of all the informative content of each characteristic, has a main disadvantage that consists in the creation of a great number of branches, sometimes completely unnecessary, which implies the construction of trees of exaggerated dimensions. On the other hand, the evaluation of the quality of the partitions is influenced by the number of subgroups that these lead to, becoming difficult the comparison of partitions based on characteristics of different sizes.

The Hunt solution

The solution presented by Hunt in 1966 suggested for this problem the creation of binary nodes attributing to one branch the value of a characteristic and to the other all branches the other values. This solution is obviously limited, because doesn't use all the information of each characteristic. However, it presents a great simplicity, what is especially important because a human user will interpret results.

Ordinate characteristics

This methodology is more efficient than the previous one, and it was adopted in the research reported in this study. A commanded characteristic defines a relation of order between the values it can take. When handling these types of values, it is possible to define binary tests of the type $x_n \leq C$ - where x_n represents the variables and C is known as a *cut-off value* - making possible the construction of binary trees. For a characteristic of size N, it will be possible to create N-1 different partitions. It becomes, thus, necessary to make some tests in order to choose the best one. This method has the inconvenience of not using all the information of each characteristic, providing, however, one tree sufficiently intelligible for the human observer.

3.2 Minimum-cost tree

When a leaf of the decision tree is reached, it is necessary to determine which class should be associated to it. There are two distinct approaches that can be adopted: the attribution of the most likely class that minimizes the error of classification, or the attribution of the class that allows the minimization of the classification costs.

The cost of an error can be faced as the penalties imposed to the system when it produces a certain error. If the main goal of the system consists in the minimization of the costs, in place of minimizing the error of classification, penalties must be defined. Suppose costs are attached to different misclassifications, say the costs C_{ij} of misclassifying examples of class i as class j. One approach is to consider that the tree construction consists in the mere modelling of the probabilities $p(k \mid x)$, and the costs should be used to choose the classification at each node. In this case, distinct costs suggest that a more accurate model should be preferred for some classes than for others.

The best-known procedure for tree pruning is as follows. Let $R(T)$ be a measure of a tree formed by adding the contributions from the leaves. An obvious candidate is the number of misclassifications on the training set. Let the size of a tree be the number of leaves. A *subtree* of T is a tree with root at node of T; and it is called a rooted *subtree* if its root is the root of T. The objective consists, thus, in finding a rooted subtree T_0 of the full tree T that minimizes

$$R_\alpha(T_0) = R(T_0) + \alpha size(T_0)$$

where $size(T_0)$ represents the size of the tree. T_0 assigns for *minimum cost tree*.

4 Development of the algorithm

The steps taken towards the construction of the classification tree are described along this section.

Using the instruction `xlsread` of Matlab, this algorithm loads the information of a file and creates the matrix *data* of values that will generate the classifier, and a vector *class* with the information related to the classes.

Then, using the function `treefit` with the parameter `classification`, a classification tree is built, that adjusts to the input data. In this stage of the process special care is required, since variables are categorical and the procedure is different from the one adopted with quantitative variables.

Thus, all the columns must be specified as categorical variables and, for this purpose, the parameter `catidx` is used [2].

After the tree construction, having as many branches as required by the classifier, there is a danger that this tree fits the current data set well but would not do so at predicting new values. Some of its lower branches might be strongly affected by outliers and other inaccuracies of the current data set. If possible, a simpler tree that avoids this problem of *overfitting* is preferred.

The best tree size can be estimated through *cross-validation*. First, a *resubstitution* estimate of the error variance for this tree can be computed, what leads to a sequence of simpler trees. This can be done using the `treetest` instruction.

This reckoning probably under-estimates the true error variance. Then, a cross-validation estimate of the same quantity must be computed. The cross-validation method also gives an estimate of the best level of *pruning*, which is necessary to reach the ideal size of the tree. In this procedure the function `treeprune` is used, and the pruning level is defined using the argument `level`.

In order to see the classification process, the structure of the minimum cost decision tree is displayed graphically, using the instruction `treedisp`.

Finally, the algorithm asks for the information related to a new immigrant, treats it and gives the probable answer to intended issues.

The first part of the algorithm builds a decision tree that classifies the new immigrant, showing the place of residence that it considers more adequate in accordance with his personal information. The second part, analogous to the first, gives the decision tree that classifies the professional characteristics, showing the most likely place of work, according to his personal information and to the place of residence given in the previous step.

In a sequential perspective, the first phase of the process decides which is the most adjusted place of residence for new immigrants, and uses this decision in the second phase, where it points out the place of work, already knowing the residence place. This procedure makes sense, since, as shown before, these variables are correlated.

5 Application of the algorithm

In the first phase of this algorithm the variables *Municipality of Residence* and *Parish of Residence* were modelled according to the *Country of Origin*, *Professional Group*, *Professional Assignment*, *Sector of Economic Activity*, *Employment Situation* and *Professional Situation*.

It is necessary to consider that in this problem all the variables are discrete or categorical, thus all the elements of the data set were attributed a number that identifies the different values that variable can take. The described algorithm will be used to determine the most likely *Municipality of Residence* for a new foreigner, according to its characteristics.

The first output of the program relates to the size of the minimum cost tree. Figure 7 presents the leaf number of a succession of trees and respective costs. The result of the cross-validation method is also displayed in hatched line.

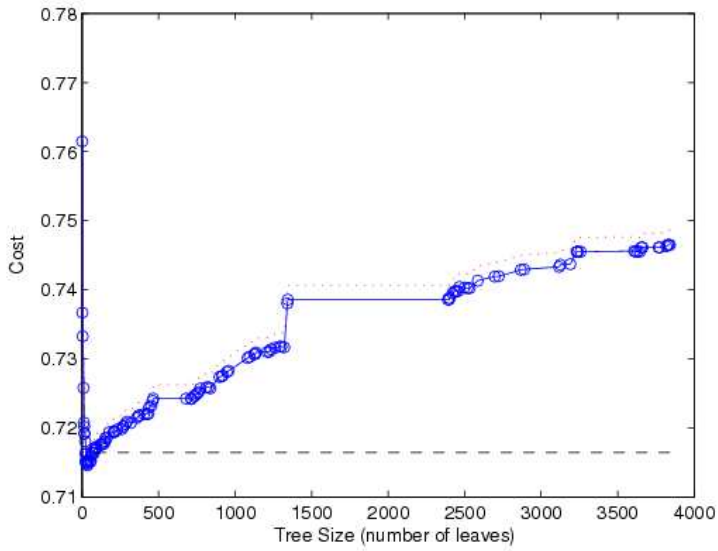

Figure 7: Search for the minimum cost tree
that classifies the municipality of residence

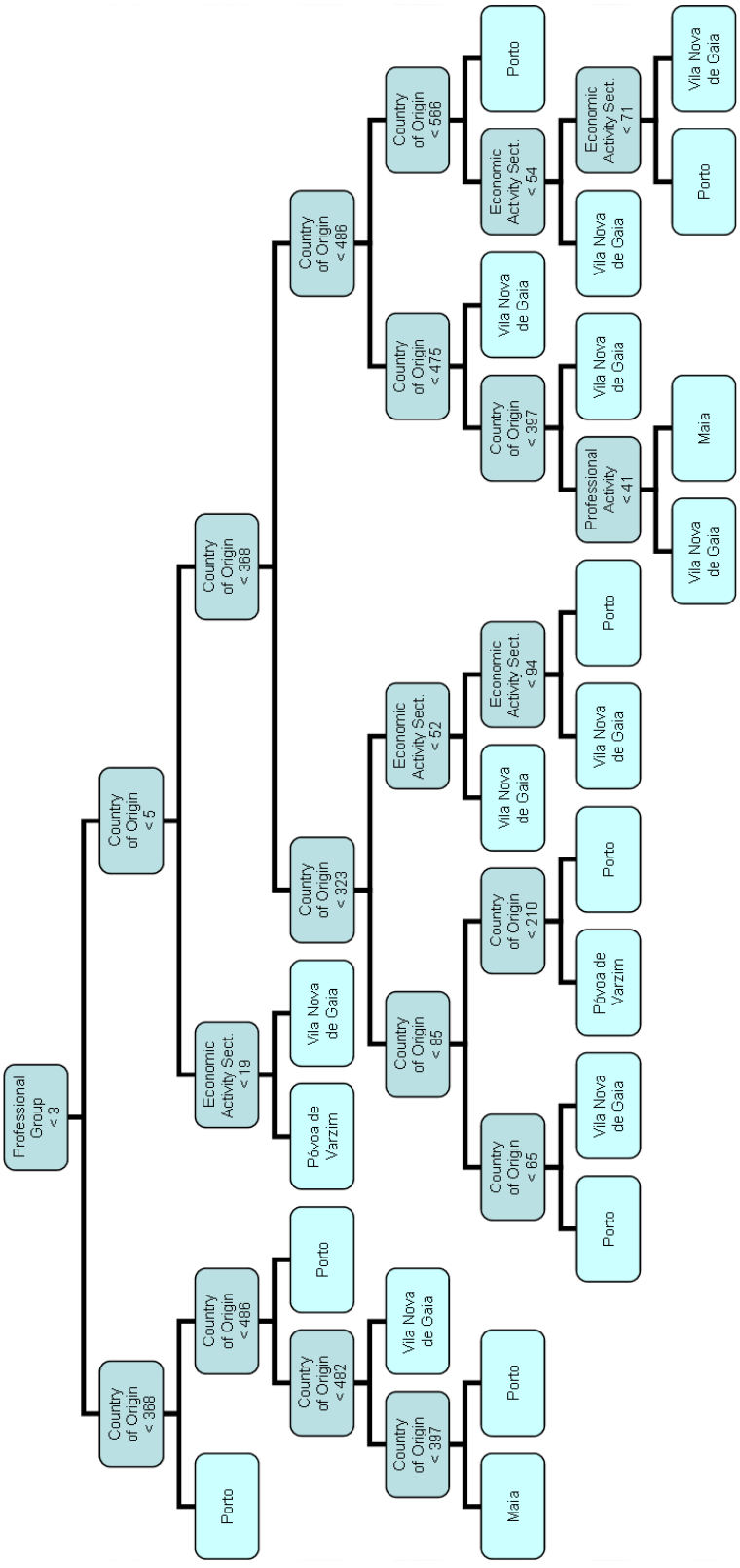

Figure 8: Decision tree for the municipality of residence

As can be seen in figure 8, one of the most important characteristics is the *Professional Group* since the first division of the data depends on its value. Another important characteristic is the *Country of Origin* since it appears sometimes as a division criterion. This phenomenon was already expected since the immigrants of the same country join themselves together intuitively because they have the same language and cultural customs. The *Sector of Economic Activity* also exerts a significant influence in the decision process.

Once the tree is built up, new immigrant dwelling and working locations can be anticipated. Suppose the arrival to Porto Metropolitan Area of an individual with the following characteristics: Brazilian (Country of Origin: 508), non-qualified worker (Professional Group: 9), cleaner (Professional Assignment: 913), usually works for families with house servants (Sector of Economic Activity: 95), employed (Employment Situation: 1) and worker on others' account (Professional Situation: 3). The program considers the Municipality of Porto the most probable one for his dwelling.

Figure 8 displays the way the decision was taken. Starting from the top node, the number that identifies the *Professional Group* is less than the *cut-off* value, 3, and then the analysis takes the left branch. The value that corresponds to the *Country of Origin* is greater than 368, so now it takes the right branch. Following in a descending trajectory along the tree, a terminal node is reached that identifies the most appropriate grouping procedure to study the individual. In this case, the *Country of Origin* is again greater than 486, thus it follows for the right branches until it arrives at a leaf: the immigrant will probably find dwelling in the municipality of *Vila Nova de Gaia*.

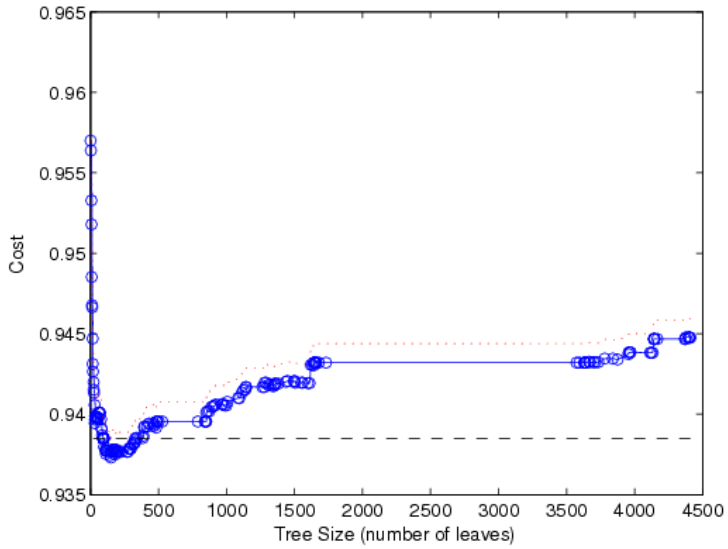

Figure 9: Search for the minimum cost tree
that classifies the parish of residence

Another interesting study consists in the construction of an analogous tree that is more specific than the previous one, and that classifies an individual according to its Parish of Residence. This study was carried through, however

this decision tree is rather more complex, as figures 10, 11, 12 and 13 show.

Following the same methodology, but now using the information concerning the place of work, a decision tree that classifies the *Municipality of Work* considering the *Municipality* or *Parish of Residence*, Country of Origin and *Employment Situation* was built.

In order to classify the place of work of a new immigrant, it is necessary to know his place of residence. This information was already supplied to the algorithm issued to classify the dwelling location, and now is used to find the place of work.

Figure 14 shows, once again, a succession of trees and their respective costs.

Figure 15 displays the minimum cost tree to classify the municipality of work, considering the municipality of residence. As can be noticed, the municipality of residence is the most significant characteristic in this tree. This makes sense because the variables are dependent (as has already been proved).

Considering, once again, the Brazilian citizen (Country of Origin: 508), resident in Vila Nova de Gaia (Municipality of Residence: 1317) and employed (Employment Situation: 1) the decision tree presented in figure 15 can be used to anticipate the most probable municipality for him to work. Starting from the root, the municipality of residence is equal to 1317, thus the immigrant will obtain a job in *Vila Nova de Gaia* municipality.

A similar study was developed considering, instead, the parish of residence. A decision minimum cost tree was built to classify the municipality of work, considering the parish of residence. A succession of trees and respective costs were computed once more (figure 16).

Finally, figure 17 shows the minimum cost tree of the work parish, and it can be noticed that it is rather more complex than the previous tree. The most significant characteristic of this tree is, again, the parish of residence, what makes sense because these variables are dependent.

Taking once again the Brazilian citizen example, following the ramifications of the tree lead to the conclusion that this individual - that will probably choose *Mafamude* parish to live in - also presents high probability of finding a job in *Vila Nova de Gaia* municipality (what is, of course, coherent with the previous classification).

6 Conclusions

The set of decision trees developed adds to the geographical understanding of immigrant residential and work place distribution in Porto Metropolitan Area, according to their country of origin. Additionally, it allows the prediction of the desired place of residence as well as the probable place of work of a new set of immigrants that potentially arrive to this Metropolitan Area.

Thus, this article makes an important contribution to municipal power decision agents, because it allows the systematisation of information (as well as their cartographic display) that supports better understanding of population choices and opportunities, as well as the definition and implementation of policies in order to enhance better integration of certain population groups, thus promoting urban and metropolitan quality of life.

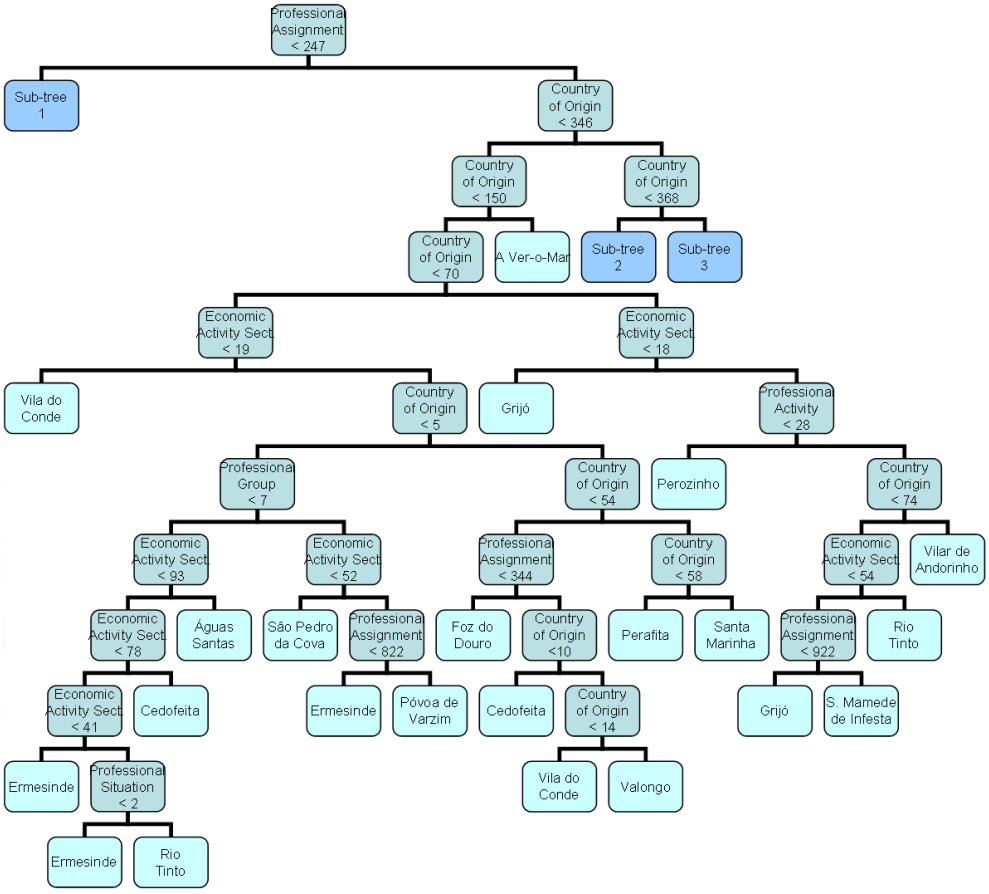

Figure 10: Decision tree for the parish of residence

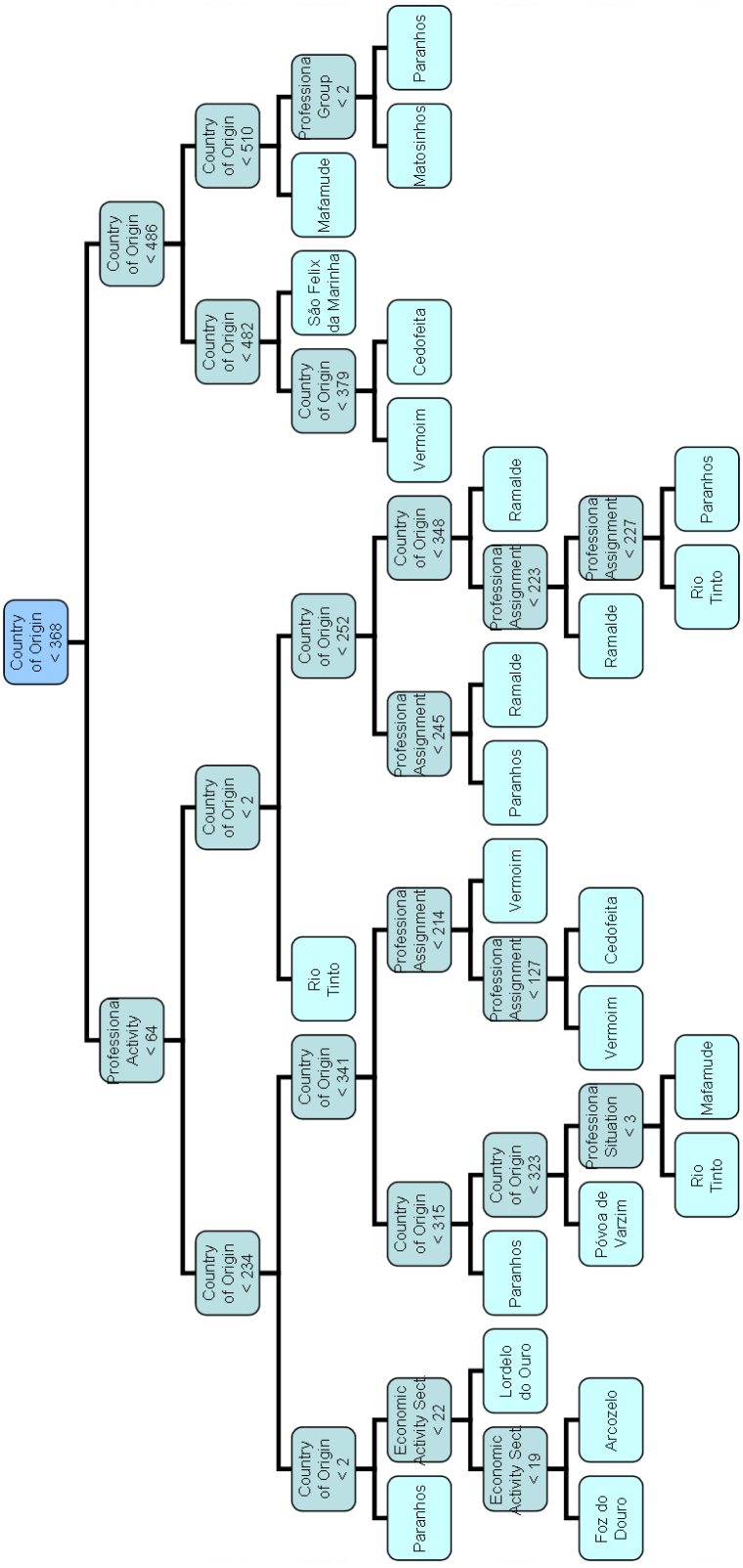

Figure 11: Sub-tree 1

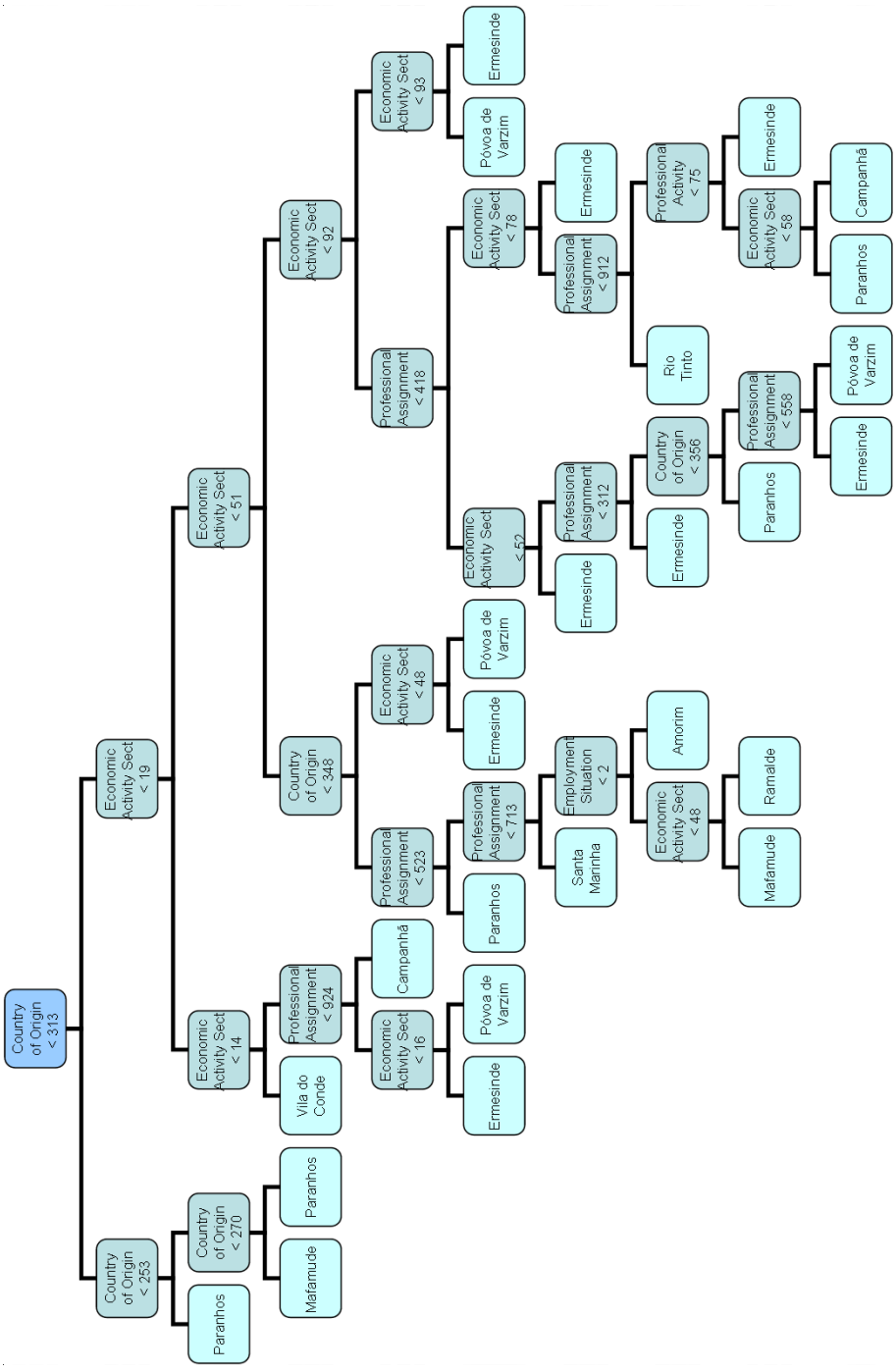

Figure 12: Sub-tree 2

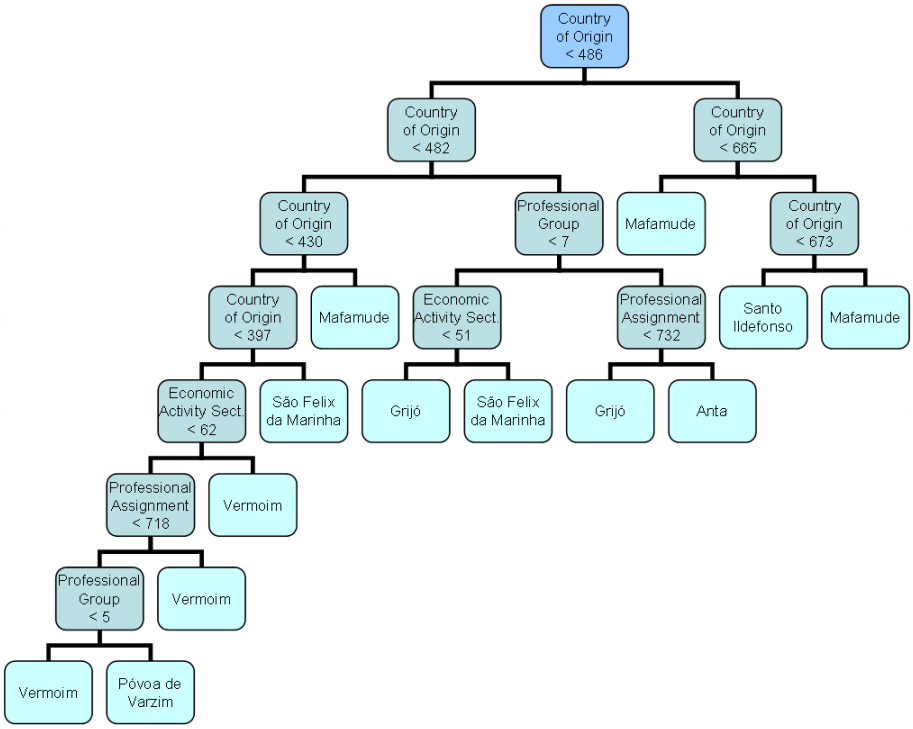

Figure 13: Sub-tree 3

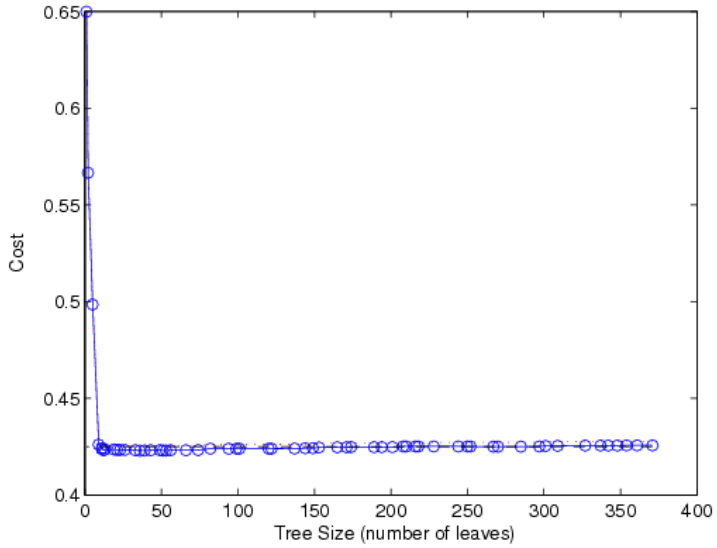

Figure 14: Search for the minimum cost tree
that classifies the municipality of work

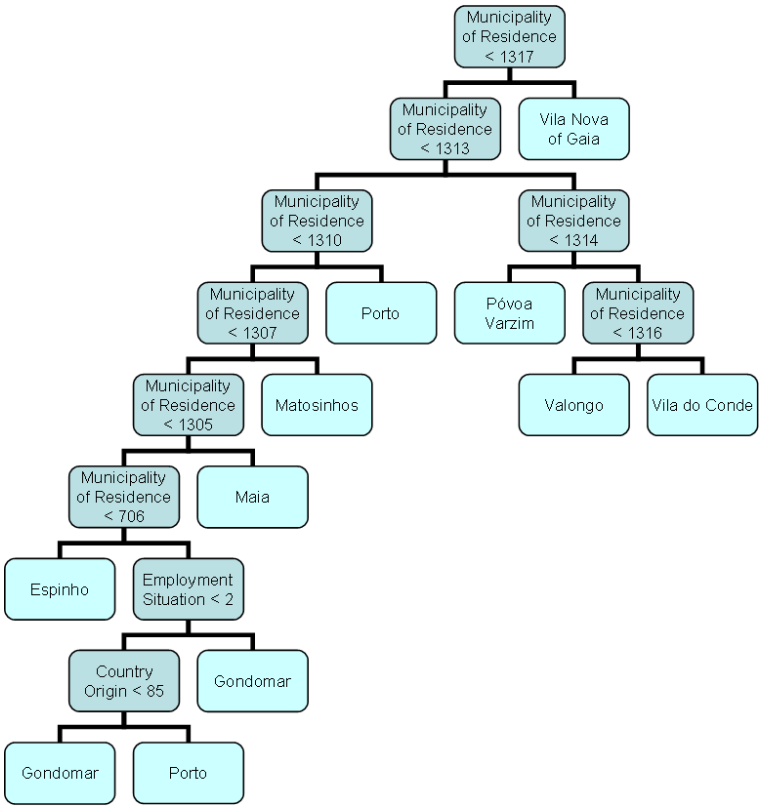

Figure 15: Decision tree for the municipality of work

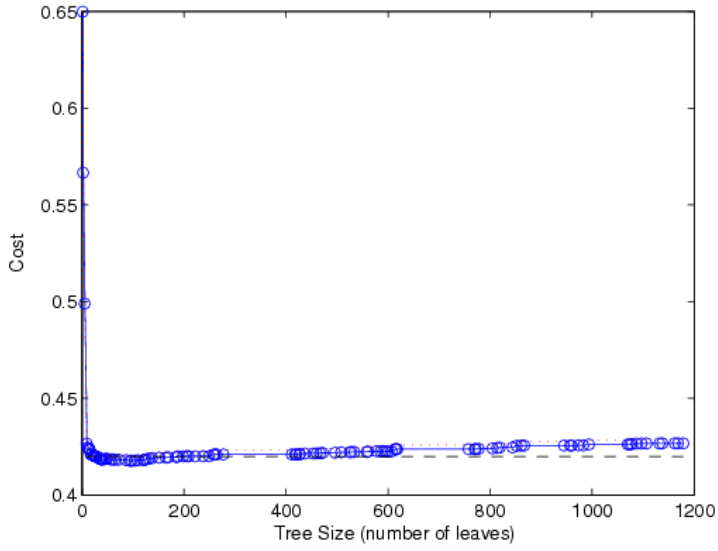

Figure 16: Search for the minimum cost tree
that classifies the parish of work

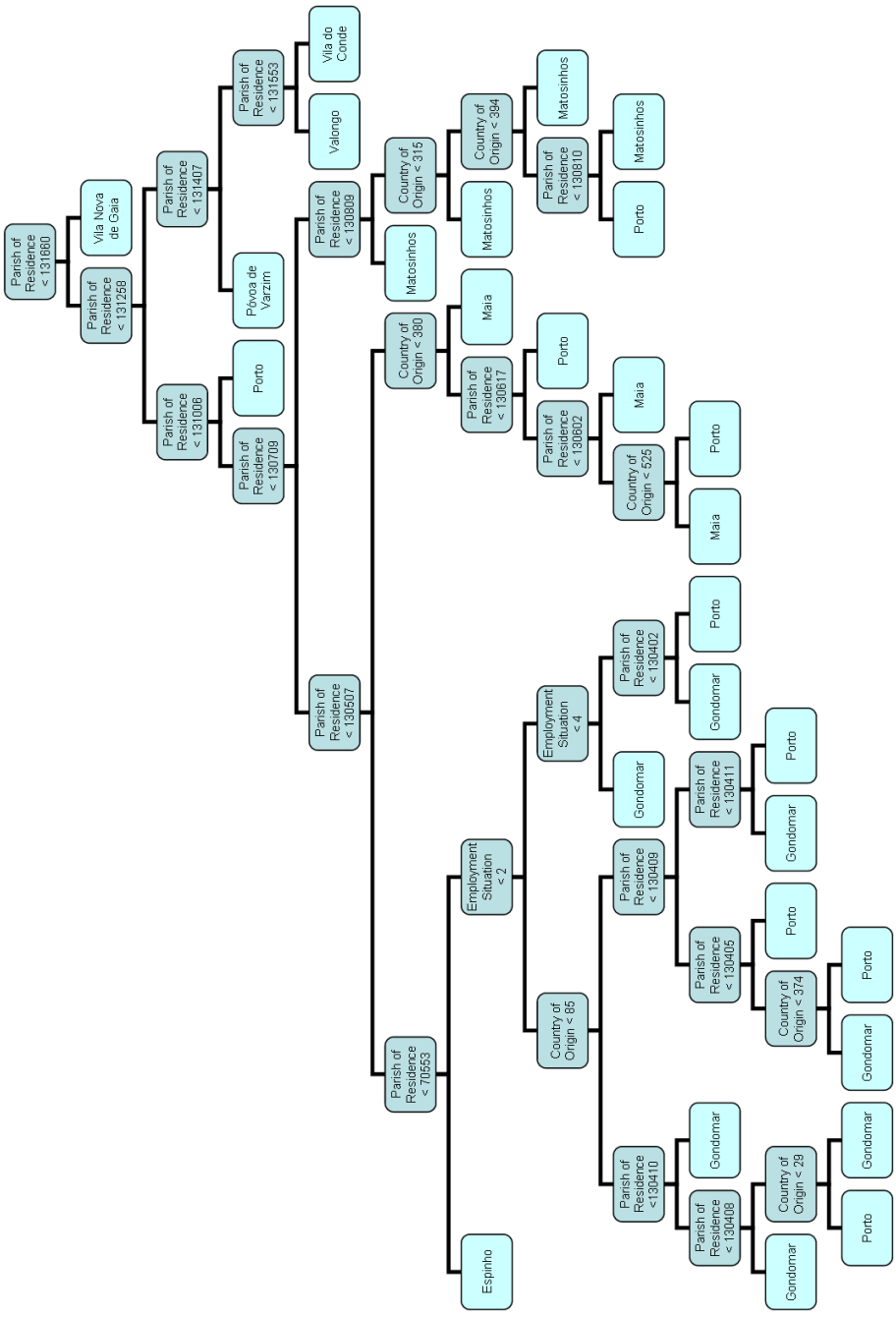

Figure 17: Decision tree for the parish of work

Ackowledgements

The authors acknowledge the Portuguese Science and Technology Foundation (FCT) and the Portuguese National Statistics Institute (INE) for all the given support.

References

[1] Instituto Nacional de Estatística. *XIV Recenseamento Geral da População IV Recenseamento Geral da Habitação*. Lisboa: Imprensa Nacional - Casa da Moeda, 2001.

[2] Howard Demuth and Mark Beale. *Neural Network Toolbox User's Guide*. The MathWors, Inc., Natick, MA, 4th edition, September 2000.

[3] Edward J. Dudewicz and Satya N. Michra. *Modern Mathematical Statistics*. Jonh Wiley & Sons, Inc., 1988.

[4] L. T. Paiva and E. M. Rebelo. Planeamento Urbano para a Integração de Imigrantes. Report of the project IME/AUR/49901/2003, financed by Fundação para a Ciência e a Tecnologia, 2005.

[5] B. D. Ripley. *Pattern Recognition and Neural Networks*. Press Syndicate of the University of Cambridge, Cambridge, United Kingdom, 1996.