Identification of Urban Areas Using Raster Spatial Analysis

Miguel Valentim[*]

Instituto Superior Técnico, Universidade Técnica de Lisboa

Abstract.

The impacts on the environment and landscape from the adoption of spatial occupation models that promote and enable fuzzy urban sets represent a fundamental challenge to the development of a territory.

As a result, it is crucial to understand within planning the spatial distribution of built areas typologies and if possible to have methods and tools available for its automatized identification.

Nowadays, Geographic Information Systems (GIS) facilitate spatial analysis, not only in a rapid and objective manner, but also by being oriented to local improvement analysis and less dependent of disperse information which quality cannot be easily controlled.

In this paper a model for the identification of urban areas is proposed. The method is based on spatial statistics and uses discreet variation data. A spatial distribution analysis is made considering densities and applying a set of restrictions enabling solutions for the problem. The capacity to identify urban spaces, allows a better resource allocation and action taking within the planning process for the control of urban settlements and support of rural activities and uses in the remaining territory.

In addition to the analysis of similar methods, a practical application of the proposed method is also made, complemented by suggestions for procedures that allow the assessment and validation of its quality.

Keywords: Urban areas, GIS, Spatial analysis, Raster, Density, Quality measures

1. Urban Areas and Spatial Analysis

1.1 General description

The dichotomy between traditional mental conceptions of *urban* and *rural* landscapes is being threatened, not only by the growing fuzziness of its spatial borders, but also by technical and technological challenges faced by planners regarding the definition of distinctive criteria that allow the binary spatial classification between city and countryside. The importance of such task is increased by the need to contain urban sprawl and its undesirable impacts in terms of tensions and conflicts created in land tenure and use, access to services and other measures of social, economic and political integration, as well as environmental degradation. Another aspect is the need by the public sector of insuring the social function of land use. By defining which spaces are fit for urban uses the issue of equality can be adequately addressed by ensuring that investments made in qualifying urban spaces (namely in terms of infrastructure and social facilities) are not dispersed in a way that affects its effectiveness.

Urban settlements emerge in different typologies, scales and spatial distributions. The system's complexity is increased by the existence of numerous types of transition zones. As a result, the use of spatial analysis methods is an essential part of the planning process, allowing the identification of

[*] valentim.miguel@gmail.com

urban areas by the analysis of the settlement spatial distribution. This is possible because the continuity of buildings as well as the proximity to infrastructure networks are the basic elements in which the method's relations and interactions are set.

The present paper focus mainly on distribution of buildings as a tool for evaluating spatial interaction and organization. The distribution of buildings through the landscape allows the understanding of the patterns of spatial interactions at different scales and levels and the establishment of similarity relations. Variations of the built area density at the local level permit to classify it into two categories: *urban* / *non urban*. This is possible by the evidence that the proportion of space occupied by buildings increases as we approach city centres.

Spatial analysis can be understood as the process of spatial data manipulation, of which additional data regarding a particular geographical area is obtained. The use of GIS tools applied to spatial analysis allows an immediate visualization of the outcome of the analysis as well as the preservation of relevant information in a rigorous manner.

1.2 Typologies of urban areas

The evidence of the existence of different types of urban settlements goes beyond standard definitions. The rapid urban population growth due to migration's phenomena and the consequent enlargement of the built up area also contributes for the existent urban typologies beyond the conventional city centre. Numerous differentiations can be established, namely by nature of the growing process. A common aspect to all are the motives, particularly in the case of areas in expansion. These motives are closely related to the social and economic structure of the migrant population; physical barriers; dimension and organization of the original city; access to communication, transport and infrastructure networks; land value, tenure and taxation as well as political and administrative incoherencies within the area.

As a way of establishing a clear definition of *urban area* and normative parameters of edification so an area can be classified as *urban*, both urban Portuguese legislation and plans try to establish normative parameters of edification. In the table below some of these parameters are synthesized.

	DL n.º 400/84	**L 26/2003**	**PROT OVT**	**PDM Tomar**
	relative to urban alottments	**relative to municipal taxation**	**regional master plan**	**Tomar municipal master plan**
Minimum number of buildings	–	10	80	10
Maximum separation radius	10 meters	10 meters	25 meters	15 meters

1.3 Density calculation

The fuzziness evidenced by edification patterns in urban margins make the necessary task of delimiting urban sets very difficult. At this point, several spatial analysis approaches can be used. However, by classifying built areas according to the patterns of the buildings spatial distribution, a spatial statistics approach is applied. Since the density of the buildings is the main distinct feature of the built up area, the selected approach in this paper is constructed upon this aspect, which is a discreet variation data. Therefore, the use of a raster dataset was considered the most appropriate when dealing with this type of data.

The number of buildings per unit area – density (*d*) – is used so a continuous surface is created, where each location on the surface is a measure of the existing density. The density match the magnitude per unit of area from the building centroids that fall within a circular distance around a cell, called Neighbourhood Radius (*r*).

$$d = \frac{n.^{\text{o}}\,buildings}{\pi . r^2}$$

Equation 1 – Density

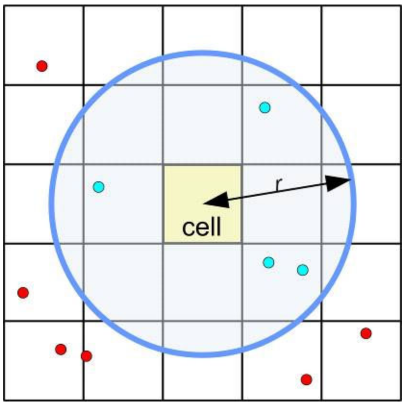

Figure 1 – Density calculation

2. Methods for the Identification of Urban Area

Spatial statistical-based methods developed through GIS tools are not a novelty. The data models in which they are based mark the biggest difference between the two and define its strengths and weaknesses. As an example, three of these approaches are briefly described and analysed.

The first one is a vector data based method, proposed by Gonçalves *et al.* (2007), where the modelling is accomplish by a contiguity analysis of regions formed by *atractivity* buffers applied to the buildings, forming homogeneous regions in terms of the existence of a minimum number of buildings. The *atractivity* defines how far and along the roads the influence of one particular urban set is felt, depending of its dimension and moderated by an exponential function. Since this function is standardized by the extreme values of localities dimension, its choice will affect the sensitivity of the method to local density phenomena. Also, being a vector data based model, it is also highly susceptible to topographic errors and the management of high quantities of data from different origins and scales can be difficult and very demanding in terms of resources.

A raster data based method was proposed by Borruso (2003), where the urban areas were defined by a spatial density estimator based on a Kernel Density Estimation, applied to urban road network junctions. It's a simple and direct method, based in a single modelling variable and very sensitive to the identification of local building phenomena. However, the direct relationship between road network and urban structure can only be evidenced in consolidated non-linear built areas. This vector data information is also very susceptible to topographic errors.

Finally, a mixed vector and raster method was proposed by the *Geographic Information Team of Statistics Finland* (2001), where the buildings are clustered according to a minimum of resident population and distance between them. This method requires a raster to vector data conversion, which besides having a significative impact on the outcomes, the way the software makes the interpolations is also very difficult to control leading to the generation of errors. Another particularity of this method is the attempt of including an administrative definition of locality, which can be particularly inadequate in some cases.

3. Developed methodology

3.1 Generic description

The method used is supported by spatial statistics, based on a set of discrete variation data. A spatial distribution analysis is then made over this dataset, superimposing a number of restrictions at the object aggregation and classification level around contiguous regions with homogeneous density combined with a limitation on the number of building occurences. The method consists in converting the buildings into points and creating a raster surface based on their spatial density, where the value of the cell matches, at least, the minimum value taken as relevant to be considered as *urban area*. These cells are then clustered by a contiguity criterion and finally secluded by pertinent number of buildings.

3.2 Requirements and bases for the proposed method

The baseline data includes a topologic error free vector data representative of all buildings, with exception for the small ones and therefore considered irrelevant to the urban network. The method was developed recurring to *ESRI@ArcMap*^{TM} *9.2*. software and it revealed to be simple and efficient in terms of cost and resources. The sequence of operations of this method is illustrated in Figure 3:

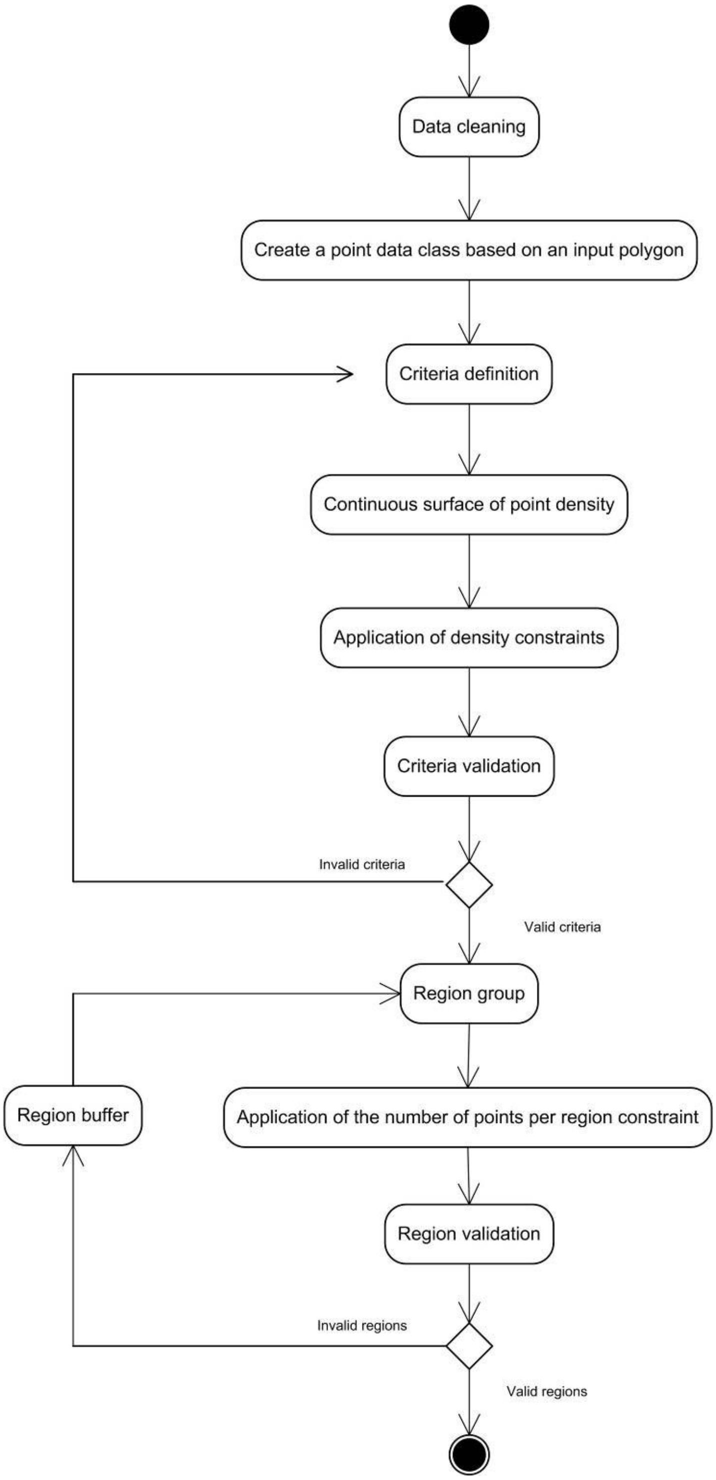
Data cleaning
Create a point data class based on an input polygon
Criteria definition
Continuous surface of point density
Application of density constraints
Criteria validation
Invalid criteria
Valid criteria
Region group
Region buffer
Application of the number of points per region constraint
Region validation
Invalid regions
Valid regions

Figure 2 – Proposed method flowchart

Each step of the diagram above is explained in more detail below:

1. **Data Cleaning** – Correction of topologic errors and elimination of polygons that represent buildings with no evident urban function, such as sheds and similar. This step aims to the elimination of elements that can compromise the method's accuracy.

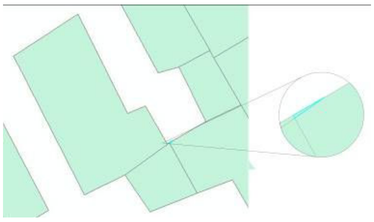

Figure 3 – Delimitation error

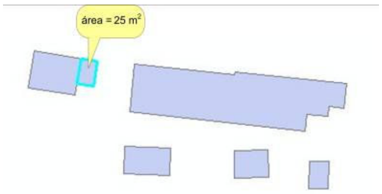

Figure 4 – Building with no urban relevance

2. **Conversion of polygons into a point data feature** – Creates a point feature class based on the input feature class polygons that represent the buildings. This step intends the substitution of the polygons by points, located in the respective centroids.

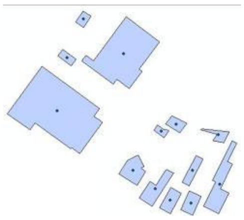

Figure 5 – Polygon centroids

3. **Construction of a continuous surface of point density** – Calculates the number of events (number of buildings) per unit of area, within a circular neighbourhood around each cell. For the application of this step a previous and preliminary definition of criteria is required, namely the output cell size and the neighbourhood radius.

4. **Application of density constraints** – Segregates the cells, defining which ones are considered *urban* and which ones are *non urban* by relevant intervals of density.

5. **Region group** – It records for each cell in the output, the identity of the connected region that is formed by spatial continuous cells. An identity number is assigned to each region within the analysis data.

6. **Region buffer** – The previous step can be – according to the flowchart in Figure 2 – complemented by an expansion in a considered number of cells of the output regions. In this way, is possible to force the aggregation of cells that are notoriously part of a particular region.

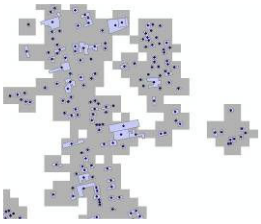

Figure 6 – Homogeneous region

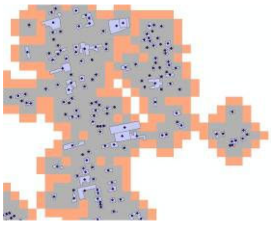

Figure 7 – Homogeneous region 1 cell buffer

7. **Application of the dimension constraint** – Segregates the regions created by number of events. In this way, only those with a sufficient number of buildings are taken into account, which allows its identification as an *urban area*.

8. **Result validation** – Since the accuracy of the method is yet to be proved in different types of landscapes, spatial parameters to fulfil the definition of *urban area* and its sensitive adjustment are recommended.

4. Assessment in a case study

4.1 Description

To test this method and finding the relevant parameters to be used as input, the described methodology was applied to a case study. The case study was Tomar municipality and a model was constructed using building vector data for a particular area of the municipality. This area was large enough to include the city centre, the countryside and the transition zone between the two. Since recent data was available – a dataset of the delimitation of urban areas in Tomar's Master Plan – a set of quality measures was considered as a way of assessing the method's adequacy, taking that dataset as reference. Being the goal of the method to provide an indicative identification of urban areas, the comparison with a not yet definitive delimitation was considered an adequate choice.

Given the binary nature of the classification – urban versus non urban – the quality assessment was conducted recurring to a *Weighted Assessment* (WA) between *Sensitivity* (*S*) and *Specificity* (*E*):

Sensitivity – measures the proportion of area identified by the method as *urban area* which is correctly identified as such by the reference data:

$$S = \frac{TP}{FN + TP}$$

Equation 2 – Sensitivity

with:

$TP =$ number of cells correctly classified as *urban area*

$FN =$ number of cells wrongly classified as *non-urban*

Specificity – measures the proportion of area correctly identified by the method in areas classified as *non urban* in the reference data.

$$E = \frac{TN}{FP + TN}$$

Equation 3 – Specificity

with:

$TN =$ number of cells correctly classified as *non-urban*

$FP =$ number of cells wrongly classified as *urban area*

Weighted Assessment – measures the global performance of the method

$$Weighted\_Assessment = \alpha \times Sensibility + (1 - \alpha) \times Specificity$$

Equation 4 – Weighted Assessment

The need for a weighted assessment arises from the spatial segmentation introduced by the grid, limiting the number of *TN* cells. However, since the data extent is always the same, this effect is mitigated. In addition, the evidence that the correct classification should be valued made that in this case 2/3 was the chosen value for α .

4.2 Choice of the Spatial Parameters

The first step consisted in finding an appropriate value for both Density (*d*) and Neighbourhood Radius (*r*). A first set of tests was made using neighbourhood radius of 30 m, 40 m, 50 m, 60 m and 70 m. For these values, a one cell wide region buffer was also tested. Since feasible solutions for *d* take the form of discrete intervals, only those that allowed a minimum classification of half the total area were tested. A single cell buffer was also applied and tested to the homogeneous regions. The best results for each one are presented in the table below:

Table 1 – Best results of the first neighborhood radius and density sensitive variation

r ***[m]***	***d*** ***[buildings/m²]***	**buffer** **[# cells]**	***S*** ***Sensitivity***	***E*** ***Specificity***	***WA***
30	≥ 0,000353	1	0,96	0,84	0,92
40	≥ 0,000198	-	0,98	0,79	0,91
50	**≥ 0,000127**	**-**	**0,95**	**0,85**	**0,92**
60	≥ 0,000088	-	0,98	0,80	0,92
70	≥ 0,000064	-	0,99	0,75	0,91

A criterion was set to choose amongst the results that presented the same WA value: the selected option was to choose the parameter combination which provided the least extensive classification of cells. This imposition was made because, being the result of the method application a broad classification, the number of TP cells would increase, and, consequently, both *S* and *WA* values would rise without being able to guarantee the method's accuracy. As a result, the finest approximation was accomplished using the 50 m radius and a non-restrictive interval of density.

Using this result, a second set of combinations was tested, trying to establish which was the relevant number of buildings located in an homogeneous region so this could be considered as an *urban area*. Testing multiple of five building combinations helped to detect that the best result was being provided by a minimum of 35 buildings per region.

Table 2 - Best results of the minimum number of building per region sensitive variation

r ***[m]***	***d*** ***[buildings/m²]***	**# buildings**	***S*** ***Sensitivity***	***E*** ***Specificity***	***WA***
50	≥ 0,000127	20	0,96	0,84	0,92
50	≥ 0,000127	25	0,96	0,85	0,92
50	≥ 0,000127	30	0,95	0,85	0,92
50	**≥ 0,000127**	**35**	**0,95**	**0,85**	**0,92**
50	≥ 0,000127	40	0,95	0,85	0,91

This procedure allowed the establishment of the base for a third set of tests where, with the previous result in mind, the focus would be again on radius and relevant density intervals. All significant intervals of density for 55 m and 45 m radius were tested.

Table 3 - Best results of the second neighborhood radius and density sensitive variation

r	*d*	# buildings	*S*	*E*	*WA*
[m]	*[buildings/m^2]*		*Sensitivity*	*Specificity*	
45	**≥ 0,000157**	**35**	**0,93**	**0,88**	**0,92**
45	≥ 0,000314	35	0,80	0,95	0,85
45	≥ 0,000472	35	0,50	0,99	0,67
45	≥ 0,000629	35	0,37	1,00	0,58
55	≥ 0,000105	35	0,97	0,82	0,92
55	≥ 0,000210	35	0,90	0,90	0,90
55	≥ 0,000315	35	0,80	0,94	0,85
55	≥ 0,000420	35	0,69	0,97	0,79

As a result it was now possible to understand that the sensitive variation of the modelling parameters and for the case study used the best result (Figure 8) was obtained by using a 45 meters neighbourhood radius, a minimum of 35 building per homogeneous region and without setting any density constraints. A fourth set of tests was yet made recurring to a further reduction of the radius, but it only came as a confirmation of the previous results.

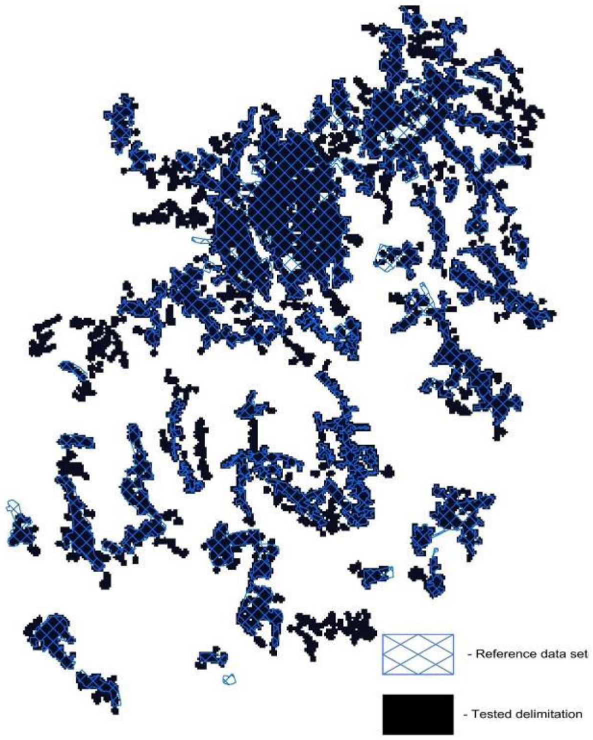

Figure 8 – Method application to a case study

5. Conclusions

This paper discussed on the possibility of engaging an operational solution in the process of identifying urban areas using raster spatial analysis. The quality of the results showed that this possibility is real and permits to consider this proposal as an indicative urban identification method able to operate at the municipal master plan scale. It also revealed the possibility of applying sensitive assessment measures so that the quality and consistency of the method is appraised. The relative small number of modeling parameters requires a more careful assignment when choosing the appropriate values, due to the largest influence in final results. Having only one layer of input data requires also that the information must be reliable and error free. At the operational level, the results showed that the one cell buffer applied to homogeneous regions does not improve the global quality of the method. This effect is particularly visible when combining parameters with the best results, although for others this procedure was advantageous. The quality of the test did not improve with the imposition of density containment, which showed that since the density values correspond to discrete intervals imposed by the number of buildings in the neighborhood radius, the segregation of some of these values implies a perceptible decrease of the classified area with negative consequences to the capacity of classifying *TP* cells. When this happened, the value of *S* and *WA* was also affected, resulting in a poor final score. Therefore, the tuning of the method should be made mostly according to both values of the search radius and of the minimum number of buildings in the homogeneous region. Notice that an “over extensive” method would result in the opposite effect, without a respective correspondence in accuracy.

According to the results obtained with the best performance combination of parameters, an area can be considered as *urban* if it has homogeneous regions with at least 35 buildings, separated for a maximum distance of 45 m and establishing a minimum density pattern of 1,57 buildings/m^2. Since these results are not in accordance with the majority of the parameters defined by the urban Portuguese legislation and Plans, it can be asserted that the introduction of normative rules would only decrease the method quality. The exception would be the minimum number of building per region proposed both by Law 26/2003 and Tomar’s municipal master plan. Nonetheless, only with the application of this testing procedures to other sample areas disseminated throughout different types of landscapes in the territory would allow the confirmation of the universality of these parameters and its potential as a cost-effective process.

6. References

BORRUSO, G., Network density and delimitation of urban areas, Transactions in GIS 2003, Blackwell Publishing Lda, Reino Unido, 2003.

CCDRLVT, Plano Regional de Ordenamento do Território do Oeste e Vale do Tejo -: Versão para Discussão Pública, CCDRLVT, Portugal, 2008

Decreto-Lei n.º400/84. D.R. n.º 301, Série I de 1984-12-31

GONÇALVES, T., PAULINO, L., VALENTIM, M., O Periurbano de Tomar – Identificação do Fenómeno e Proposta de Ordenamento, Projecto de Final de Curso em Engenharia do Território, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal, 2007.

Lei 26/2003 de 30 de Julho, D.R. n.º 174, Série I-A de 2003-07-30.

Resolução do Conselho de Ministros n.º 100/94. D.R. n.º 233, Série I-B de 1994-10-08

STATISTICS FINLAND, The Development of Delineation Methods of Urban Areas for the Census 2000, Statistical Commission and Economic Commission for Europe – Conference of Europeans Statisticians, Finlândia, 2001.