# Land use/cover maps by RS and ancillary data integration in a GIS environment

J. Rocha & P.M. Sousa
*Centro de Estudos Geográficos, Lisbon University, Portugal*

J.A. Tenedório & S. Encarnação
*e-Geo, Centro de Estudos de Geografia e Planeamento Regional, New University of Lisboa, Portugal*

Keywords: remote sensing, ancillary data, Geographic Information Systems (GIS)

ABSTRACT: The main purpose of the research presented in this paper is the development and validation, through the application to a case study, of an efficient form of satellite image classification that integrates ancillary information (Census data; the Municipal Master Plan; the Road Network) and remote sensing data in a Geographic Information System. The developed procedure follows a layered classification approach, being composed by three main stages: (1) Pre-classification stratification; (2) Application of Bayesian and *Maximum-likelihood* classifiers; (3) Post-classification sorting. Common approaches incorporate the ancillary data before, during or after classification. In the proposed method, all the steps take the auxiliary information into account. The proposed method achieves, globally, much better classification results than the classical, one layer, *Minimum Distance* and *Maximum-likelihood* classifiers. Also, it greatly improves the accuracy of those classes where the classification process uses the ancillary data.

## 1 INTRODUCTION

Since the generalization of satellite images, a huge effort has been done to identify, delineate, and measure, in an automatic or semi-automatic form, the different features of the urban space. If for studies over agricultural areas digital image-processing tools have shown their value, when transposed for constructed areas the results seem to be poorer (B.C. Foster 1985, Sadler & Barnsley 1990). The spatial resolution of the images functioned initially as justification for all the difficulties, and was considered as the main limiting factor in various studies (B.C. Foster 1980, R. Welch 1982).

Some authors (B.C. Foster 1985, Haack *et al*. 1987, Martin *et al*. 1988) have noticed, however, that, paradoxically, the increase of the spatial resolution may also increase the problems related with the urban area analysis, due to the greater spectral heterogeneities of the urban environment, that leads implicitly to an increase of the variability – as the spatial resolution of an image increases, the details of the image (e.g. roads, houses) start taking form, promoting an erroneous and confused image handling, compromising the extraction of global information, and becoming problematic for a coherent and homogeneous image classification. In fact, urban areas involve spectrally heterogeneous land use classes, making impracticable a correct classification based solely in spectral information, as in traditional classification algorithms.

Geographical Information Systems (GIS) allow an easy integration of multi-source information. This fact can be exploited to produce image classification methods where information other than that collected from remote sensing, known as ancillary information, is also used. The main purpose of the research presented in this paper is the development and validation, through the application to a case study, of an efficient form of satellite image classification that integrates non-spectral data

(Census data; the Municipal Master Plan; the Road Network) and remote sensing data, in a Geographic Information System.

Till present, only a few works integrated Census data in satellite images classification. One of the most recent and innovative work, was held in Portugal, by the *Portuguese Geographic Institute* (IGP former *National Center for Geographical Information* – CNIG) in partnership with the *National Statistic Institute* (INE). This study (M. Caetano 1997) aimed to evaluate the growth dynamics of the Great Lisbon Area, using spectral (SPOT images) and ancillary data (Road Network-RN, Digital Terrain Models – DTM and Census). The followed methodology used the DTM for the correction of topographic effects on the images, the RN in a pre-classification stratification (urban/non urban mask) and the Census data in the creation of classification rules for post-classification sorting. In 1998, another important work was presented (V. Mesev 1998), which used the Census data to produce '*a priori*' occurrence probabilities of land-use classes and then, using a Bayesian classifier, integrated those probabilities in satellite images (Landsat TM) classification. The methodology presented in this paper can be considered as a step forward on these two works.

As already mentioned, the hierarchical classification methodology studied in this paper incorporates spectral and auxiliary (non-spectral) information. Both are described in section 2. In the pre-processing stage, which is detailed in section 3, this information is made compatible after correction of acquisition errors. Also, using the auxiliary information alone, a binary mask is produced that subdivides the area on study into two clusters – "land" and "water" – and '*a priori*' occurrence probabilities for some urban classes are computed. The classification procedure, presented in section 4, is composed by three main stages: Pre-classification stratification; Application of Bayesian and *Maximum-likelihood* (ML) classifiers; Post-classification sorting. Common approaches incorporate the ancillary data before, during or after classification. In the proposed method, all the three steps take the auxiliary information (or data extracted from it) into account. Experimental classification results are presented in section 5, together with common classification error measures. Also, a comparison with classic, one layer, classification strategies (*Minimum Distance* and *Maximum-likelihood*) is provided.

## 2 DATA SET

The multi-source information used in this paper can be divided in two main groups: spectral data (raster format) and ancillary data (vector and ASCII formats). The former corresponds to SPOT 5 –P and HVIR (2004) and Landsat 7 ETM+ (2000) satellite images; the latter corresponds to the results of the Portuguese Population General Census statistics (2001), referenced to the subsection vector basis (equivalent to the UK enumeration district, known as BGRI), the Municipal Master Plan (1994) of the area under study and the Roads Network (2000). The Portuguese Population General Census statistics comprises a large set of data, divided into four subgroups: "Families", "Dwellings", "Individuals" and "Buildings". Taking into account previous works experience (Weber & Hirsch 1992, V. Mesev 1996), a first selection of census data was made, leaving out the families data.

The Road Network was selected because it already exists in vector format and so could be inserted in the final classification with a 100% precision. Moreover, the inclusion of the Road Network allows removing from the classification a class – "roads" – that has great similarities (identical digital values) with the multi-family housing area, allowing calculating with better precision this last one. In what concerns the Municipal Master Plan (MMP), although the represented classes are predominantly an indication of what is or not, allowed to construct in the future, some classes represent the existing land uses at the plan elaboration date. These were used in post-classification sorting rules, with the objective to improve the classification results.

Considering the spectral data, it was decided to use all the SPOT bands (1 panchromatic and 4 multispectral) and Landsat ETM+ (with exception of band 6 that was excluded due to its weak spatial resolution). This option allows to simultaneously exploiting the better spatial resolution of SPOT images, and the higher spectral resolution of Landsat images. The methodology proposed in

this paper was tested over a geographical area with 2.3 km height by 4 km width, belonging to the Lisbon Metropolitan Area (LMA). The classes considered are those desired for the LMA land-use cover map.

## 3 ANCILLARY DATA PROCESSING

In this stage, it was possible to fulfill all the operations related with the ancillary data handling, namely: creation of a water/land binary mask; production of a first binary map representing the urban and non-urban uses; calculation of the '*a priori*' occurrence probabilities for the urban classes; creation of a contextual urban band – all to be used in the satellite image classification process.

In order to obtain the water/land mask, the region corresponding to Tagus River was detected on the area on study. This was accomplished using the BGRI, that presents the coast limits perfectly (statistical subsections do not exist inside the water), and transforming the result in a binary image ('1'– water; '0'– land). It should be noted that even many of the traditional classifiers obtain almost always 100% accuracy when classifying the pixels representing the "water" class, they usually fail in areas presenting a mixture of uses (mainly near the coast), often referred to as mixels. In these situations, it is common that elements of the shoreline be classified as water and vice-versa. The statistical data (Census data and BGRI) also allowed to identify the subsections without buildings and considered as non-urban areas. These sections were unified with the Road Network, giving origin to the first urban/non-urban binary mask ('urb_1').

Concerning the '*a priori*' occurrence probabilities for the urban classes, from the first selection of data, presented in Table 1, one second choice was made, pointing only to the information related with the sub-group 'buildings'. Several reasons supported this option: (i) buildings data contain information that can be efficiently correlated with reflectance (in contrast to what happens to the dwellings, that are "invisible" from satellite images); (ii) buildings data provide the sufficient information for the '*a priori*' probabilities computation, for some of the urban classes seek for the LMA land use/cover map (see Figure 1), namely: 'Commerce and Services', 'Ancient Urban Nucleus', 'Multi-family Houses', 'Single-family Houses'.

Once selected the Census indexes, it was necessary to establish formulas to calculate the occurrence probabilities of the four above-mentioned urban classes. The buildings exclusively or mainly destined for residential purposes, with more than two floors, have been considered as Multi-family Housing. The buildings exclusively or mainly destined for residential purposes, with one or two floors have been considered as Single-family Housing. The Ancient Urban Nucleus was established as the set of buildings constructed up to 1945, inclusively. In fact, in the 40's occurred an alteration in the type of materials used in construction, fact that is also perceivable on the satellite images, which allows a significant correlation between the two types of data (spectral and statistical). The Commerce and Service buildings have been considered those buildings mainly destined for non-residential purposes, independently of the construction date or number of floors.

Let consider the following symbols, referred to a given sub-section:

- MFB_t: number of multi-family buildings;
- SFB_t: number of single-family buildings;
- CS: number of buildings mainly destined for non residential purposes;
- TB: total number of buildings;
- MFB: number of multi-family buildings, constructed after 1945;
- SFB: number of single-family buildings, constructed after 1945;
- AUN: number of (exclusively or mainly) residential buildings, constructed before 1945.

The '*a priori*' occurrence probabilities will be given by:

p('Multi-family Houses') ≡ p(MFB) = MFB/TB
p('Single-family Houses') ≡ p(SFB) = SFB/TB

p('Ancient Urban Nucleus') ≡ p(AUN) = AUN/TB
p('Commerce and Services') ≡ p(CS) = CS/TB

Of course, the four probabilities summation must equal one. However, some of the required indexes (namely MFB, SFB and AUN) are not directly available from census data. For instance, these data tell us the number of multi-family houses and the number of buildings constructed after 1945, but not the 'number of multi-family buildings, constructed after 1945', as required. Taking these facts into account, the following strategy for the 'a priori' probabilities computation was adopted:

1. Compute p(CS) = CS/TB
2. If p(CS) > 0.8 then
   p(CS) = 1, p(MFB) = p(SFB) = p(AUN) = 0
   else
   p(CS) = 0
3. If p(CS) = 0 then
   p(AUN) = AUN/TE
   p(MFB) = (MFB_t – $\alpha$ AUN)/TE
   p(SFB) = (SHB_t – $\beta$ AUN)/TE

The parameters $\alpha$ and $\beta$ ($\alpha + \beta = 1$) account for the fraction of ancient buildings that are multi-family and single-family, respectively. In the experimental results we used $\alpha = \beta = 0.5$.

The final step of the ancillary data processing block is the creation of both a spectral and an ancillary contextual urban band. The first one was achieved calculating the features fractal dimension of the SPOT panchromatic band (2.5 m pixel) and the second establishing three great groups of subsections – subsections strongly urbanized, subsections low urbanized and subsections with little probability of having buildings. This was accomplished with a simple clustering technique using, as parameters, the buildings density (number of buildings/km$^2$) and the population density (number of individuals/km$^2$), in each subsection.

## 4 CLASSIFICATION METHODOLOGY

The developed classification procedure follows a layered classification approach, being composed by three main stages: 1 – Pre-classification stratification; 2 – Application of Bayesian and *Maximum-likelihood* classifiers; 3 – Pos-classification sorting. Common approaches incorporate the ancillary data before, during or after classification, through the aforementioned steps. In the proposed method, all the three steps take the auxiliary data (or data extracted from it) into account.

### 4.1 *Pre-classification stratification*

The objective of this stage is to have, at the classifier input, three main stratums – "water", "urban areas", "non-urban areas" – that will be processed individually. The pre-processing stage already produced a "water/land" binary mask, that allows isolating in all the bands (satellite images and the contextual urban bands) the use "water", and a first approximation of the "urban/non-urban" binary mask (urb_1). In this stage, a more accurate version of the "urban/non-urban" binary mask (urb_fin) is produced.

The pre-classification procedure consists of an ISODATA algorithm having, as inputs: the spectral bands, the contextual urban band, and two TVI (Transformed Vegetation Index) bands obtained from each group (SPOT and Landsat) of satellite images (the purpose of the water mask is merely to exclude, from all these inputs, the water region). These indexes represent the surface biomass content, constituting a valid aid in the urban/non-urban differentiation. The reason for choosing a ISODATA algorithm is related with two factors: (i) the intention to not define land use/cover classes at this time; (ii) the knowledge of previous works (Gong & Howarth 1990), stating that the conjugation of a ISODATA classifier with a contextual spectral band (similar to the one

used in this work), may improve the classification results in 10% (in our case the improvement was only of 3%). The TVI was applied due to its strong correlation with the construction density (S.B. Jayamanna 1996) and its capability to improve the discrimination of urban areas (M. Achen 1992).

The best result (in terms of maximizing the discrimination between constructed and non-constructed areas) was achieved with six classes on the ISODATA algorithm. Reclassifying the resulting image, allowed to create a new urban/non-urban mask (urb_2), which was intersected with the previous one (urb_1) to obtain the final mask (urb_fin). The three urban masks allow a progressive improvement in the urban-rural differentiation, as shown by the increase of the resulting *overall accuracies*: 48% for urb_1, 80% for urb_2 and 93% for urb_fin.

### 4.2 *Application of Bayesian and ML classifiers*

In the second stage of the classification procedure the urban (urb_fin) and water (water_mask) binary masks, obtained in previous steps, are crossed with all spectral bands, producing two new images for each original image: one with the non-urban uses ('Forest', 'Shrubs', 'Bare Soil', 'Agricultural', 'Roads', 'Beach', 'Public Equipment') and another with the urban uses ('Commerce and Services', 'Ancient Urban Nucleus', 'Multi-family Houses', 'Single-family Houses', 'Industry', 'Public Equipment').

The 'Forest', 'Shrubs', 'Bare Soil' and 'Agricultural' uses, belonging to the non-urban images, were extracted using a *Maximum-likelihood* classifier. As mentioned before, the 'roads' use was already available through the Road Network ancillary data. The 'Beach' and non-urban 'Public Equipment' are obtained in the next classification step, using post-classification rules, together with the Municipal Master Plan (MMP).

The urban images and the four '*a priori*' probabilities images (previously obtained) constitute the base for the Bayesian classifier application. These probabilities allow discriminating four urban classes ('Commerce and Services', 'Ancient Urban Nucleus', 'Multi-family Houses', 'Single-family Houses') that, due to their similar spectral responses, would be misclassified if a ML classifier was applied. The 'Industry' and urban 'Public equipment' uses are extracted in the next classification step, using both post-classification rules and the MMP.

The training areas (necessary in the first step of both Bayesian and ML classifier) were delimited over a color composition offering an adequate visual discrimination of the different uses (see Figure 2-a)). This composition was obtained through the following steps: 1 – production of a first composition using the 7, 4 and 1 TM bands assigned, respectively, to the R, G and B channels; 2 – transformation of the color space from RGB to IHS; 3 – substitution of channel I by the SPOT-pan band; 4 – transformation of the color space from IHS to RGB.

### 4.3 *Post-classification sorting*

The last stage of the classification procedure re-classifies pixels attributed to the wrong classes in the previous steps, through the application of contextual rules and the use of information directly available from the ancillary data (Road Network and MMP).

For the area on study, four different classes of use are considered in this step: 'Roads', 'Beach', 'Industry' and 'Sports Equipment' (a sub-class of the 'public equipment'), like football fields and tennis courts. The 'roads' class is directly available from the Road Network. The remaining classes result from the information available in the MMP or, for the missing cases, from the following contextual rules:

- 'Beach' – pixels that, although belonging to the urban areas (value '1' in the urb_2 mask), were not classified in any of the urban classes, and at a geographical position not far from 200 m of the line coast;
- 'Industry' – pixels classified in urban classes (value '1' in the urb_fin mask), but signalized as 'industry' in the MMP;
- Sports Equipment – pixels classified as 'Shrubs' or 'Bare soil', comprising a region with an area and perimeter (compactness ratio) plus half pixel tolerance, typical of football fields or tennis courts, and at a geographical position not far from 500 m of a road.

- Other Equipments than sport – pixels classified in urban classes (value '1' in the urb_fin mask), where the ratio between resident population and dwellings is very high, indicating the presence of hospitals, prisons, and so on.

All the extracted classes are then integrated in the final classification map, according to the following algorithm:

- The replenishment of the final map is performed from the highest to the lowest confidence classes, namely:
  1. 'Roads' (extracted from the Roads Network)
  2. 'Commerce and Services'; 'Ancient Urban Nucleus' (obtained from the Bayes classifier)
  3. 'Sports Equipments'; 'Beach' (resulting from the MMP and/or contextual rules)
  4. 'Industry' (resulting from the urb_fin mask and MMP)
  5. Other urban and non-urban classes (resulting, respectively, from the Bayes and ML classifiers)
- Once a label (class) has been attributed to a pixel, it cannot be overwritten by another label.

## 5 EXPERIMENTAL RESULTS AND CONCLUSIONS

Figure 1 presents the land use/cover map for the working area obtained through *Minimum Distance* classifier photo-interpretation and by the proposed method. The whole photo-interpretation map was used as 'ground-truth' for validation.

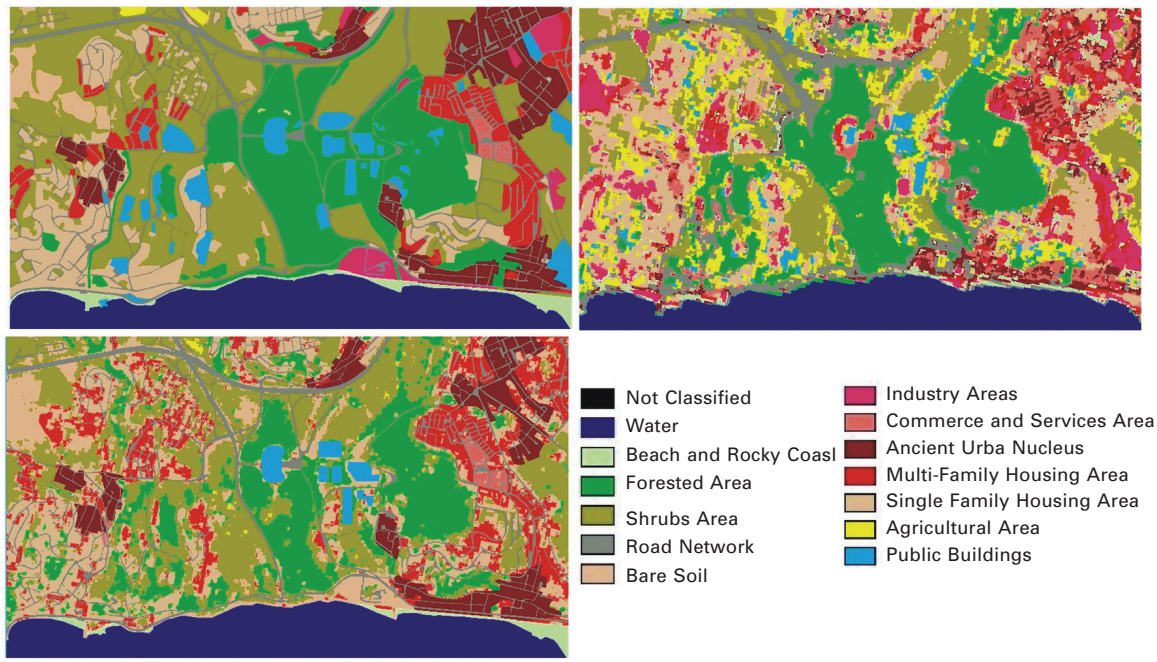


Figure 1. Study area Land Use/Cover map by photo-interpretation (top-left), by minimum distance classifier (top-right) and by the proposed method (bottom-left).

To assess the accuracy of the proposed method, and compare it with classical classifiers, several descriptive measures were also produced. The *Overall Accuracy* and *Kappa index*, reveal a better performance of the proposed method (70.5% and 0.67) comparatively to *Minimum Distance* (37.8% and 0.36) and *Maximum Likelihood* (36.5% and 0.35). Table 1 shows accuracy measures on a category-by–category basis, namely: *Producers Accuracy* and *Consumers Accuracy*.

Table 1. Both methods producers accuracy and consumers accuracy for each land-use class.

| Classes | Minimum Distance | | Proposed Method | |
|---|---|---|---|---|
| | Producers Accuracy (%) | Consumers Accuracy (%) | Producers Accuracy (%) | Consumers Accuracy (%) |
| 1. Water | 94.9 | 98.6 | 99.9 | 99.9 |
| 2. Beach and Rocky Coast | 21.6 | 12.8 | 98.9 | 99.0 |
| 3. Forested Area | 70.3 | 70.0 | 70.2 | 75.0 |
| 4. Shrubs Area | 38.6 | 84.6 | 65.1 | 75.9 |
| 5. Road Network | 23.6 | 26.8 | 100.0 | 100.0 |
| 6. Bare Soil | 25.1 | 14.8 | 54.3 | 21.7 |
| 7. Industry Areas | 11.9 | 3.6 | 0.0 | 100.0 |
| 8. Commerce and Services Area | 21.5 | 2.0 | 99.1 | 98.5 |
| 9. Ancient Urban Nucleus | 17.0 | 25.7 | 98.7 | 98.9 |
| 10. Multi-Family Housing Area | 18.0 | 23.0 | 57.6 | 39.3 |
| 11. Single-Family Housing Area | 33.1 | 35.4 | 53.1 | 54.8 |
| 12. Agricultural Area | 22.3 | 0.5 | 26.9 | 34.3 |
| 13. Public Equipment | 14.6 | 24.6 | 31.8 | 97.3 |

From these results, some conclusions can be drawn:

- The proposed method achieved, in a global perspective, better classification results than the classical, one layer, *Minimum Distance* and *Maximum-likelihood* classifiers;
- In a category-by-category analysis, the proposed method has a higher accuracy for all classes except *shrubs* (4);
- The proposed method greatly improved the accuracy of those classes where the classification process uses the ancillary information, mainly in classes 1, 2, 5, 8, 9 and 13 (CA higher than 97%). Class 5 (´Road Network´) already exists in digital format, which justifies the 100% accuracy;
- The proposed classifier failed the analysis of industrial areas, because they were not signalized in the MMP (note that, as mentioned previously, this particular case is totally dependent on a post-classification rule that uses the MMP);
- The proposed method allows the identification of classes (e.g., 'Commerce and Services'), "invisible" on the satellite images.

## REFERENCES


Achen, M. 1992. Landsat TM Data for Municipal Environment Planning? Studies of Vegetation Indices on Urban Areas", *Proceedings of the XVII ISPRS congress*, Washington, USA, vol. 19.

Caetano, M., Santos, J.P. & Navarro, A. 1997. Uma Metodologia Integrada para Produção de Cartas de Uso do Solo Utilizando Imagens de Satélite e Informação Geo-referenciada não Espectral, *Cartografia e Cadastro*, 6, 71-78.

Forster, B.C. 1980. Urban Residential Ground Cover Using Landsat Digital Data, *Photogrammetric Engineering & Remote Sensing*, 46, 547-558.

Forster, B.C. 1985. An Examination of Some Problems and Solutions in Monitoring Urban Areas from Satellite Platforms, *International Journal of Remote Sensing*, 6, 139-151.

Gong, P. & Howarth, P.J. 1990. The Use of Structural Information for Improving Land-Cover Classification Accuracies at the Rural-Urban Fringe", *Photogrammetric Engineering & Remote Sensing*, 56, 67-73.

Haack, B., Bryant, N. & Adams, S. 1987. An Assessment of Landsat MSS and Tm Data for Urban and Near-Urban Land-Cover Digital Classification, *Remote Sensing of Environment*, 21, 201-213.

Jayamanna, S.B. 1996. Relation Between Social and Environmental Conditions in Colombo, Sri Lanka and the Urban Index Estimated by Satellite Remote Sensing Data, *Proceedings of the XVIII ISPRS congress*, Austria.

Martin, L.R.G., Howarth, P.J. & Holder, G. 1988. Multispectral Classification of Land Use at the Rural-Urban Fringe Using SPOT Data, *Canadian Journal of Remote Sensing*, 14, pp. 72-79.
Mesev, V. 1998. The Use of Census Data in Urban Classification, *Photogrammetric Engineering & Remote Sensing*, 64, 431-438.
Mesev, V., Longley, P. & Batty, M. 1996. RS/GIS and the Morphology of Urban Settlements, *Spatial Analysis: Modelling in a GIS Environment*, P. Longley and M. Batty, pp. 123-148, John Wiley & Sons.
Sadler, G.J. & Barnsley, M.J. 1990. Use of Population Density Data to Improve Classification Accuracies in Remotely-Sensed Images of Urban Areas, *Proceedings of the First European Conference on Geographical Information Systems (EGIS'90)*, Amsterdam, The Netherlands, 10-13 April, 968-977.
Weber, C. & Hirsch, J. 1992. Some Urban Measurements from SPOT Data: Urban Life Quality Indices, International *Journal of Remote Sensing*, 13, 3251-3261.
Welch, R. 1982. Spatial Resolution Requirements for Urban Studies, *International Journal of Remote Sensing*, 2, 139-146.