7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.
Edited by M. Caetano and M. Painho.

RENalyzer: A tool to facilitate the spatial accuracy assessment of digital cartography

Thomas Bartoschek[1,2], Marco Painho[1], Roberto Henriques[1], Miguel Peixoto[1], Ana Cristina Costa[1]

1 Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa
Campus de Campolide, 1070-312 Lisboa, Portugal
Tel.: + 351 21 387 04 13; Fax: + 351 21 387 21 40
tbarto@isegi.unl.pt, painho@isegi.unl.pt, roberto@isegi.unl.pt, mpeixoto@isegi.unl.pt, ccosta@isegi.unl.pt

2 Institute for Geoinformatics
University of Münster
Robert-Koch-Str. 26-28, 48149 Münster, Germany
Tel.: +49 (0) 251 83 33083; Fax: +49 (0) 251 83 39763
bartoschek@uni-muenster.de

Abstract

Managing spatial data in paper maps is quite different from managing digital spatial information. Sometimes manual vectorization of scanned maps is the only way to produce digital cartography, especially when the only source of spatial information is a paper map. Digital scanning and manual vectorization are two processes well known for adding error to data and accuracy is a particularly important issue for users of spatial information. The National Ecological Reserve (REN) established in the Portuguese national law protects areas with a diversified bio-physical structure and specific ecological characteristics. This information is often required to manage several human activities, such as mineral extraction, real estate, industry, tourism, etc. REN maps were originally produced in paper and were vectorized to produce digital cartography. The objective of this study is to measure the spatial accuracy and to assure the conformity with the original cartography. The accuracy of the REN digital cartography was assessed through a stratified sampling scheme, as described in Peixoto et al. (2006). The preparation of the digitized maps for the needs and the appliance of the sampling method as well as the data quality control were combined in the RENalyzer application. The application calculates areas of all polygons respective to their attribute classes using object oriented computation methods. As described in the sampling method, all class combinations (overlaying polygons) are considered. These areas and the global area are the base of the global sample size and the sample size per class and class combination. Considering these guidelines the application creates random points in randomly chosen polygons. The data quality control is done by facilitating the map reviewing process. After adding the scanned map and automatically setting map properties, the application zooms to each point and allows a fast and accurate quality control. Errors in digitizing are computed and classified in spatial and thematic errors. The attributes distance to original class and error description are added to the table.

Keywords: data quality, quality control, geocomputation, National Ecological Reserve

1 Introduction

The National Ecological Reserve (REN) established in the Portuguese national law protects areas with a diversified bio-physical structure and specific ecological characteristics. This information is often required to manage several human activities, such as mineral extraction,

real estate, industry, tourism, etc. REN covers 28 classes of ecological importance, examples are: islands, areas with erosion, areas with tendency of floods. REN maps were originally produced in paper, characterizing the classes with different shaded markers and were vectorized to produce digital cartography. This study started with the vectorization of 153 REN maps of the Alentejo-region in south Portugal, where each county produced its own maps with mostly slightly different legends. This fact, as well as a lot of cases of overlaying classes could add error to data and accuracy during the manual vectorization. And we know that, the most studied errors in GIS are those introduced during digitizing, particularly by inaccurate placement of the digitizer (Goodchild, 1989). The objective of this study is to measure the spatial accuracy and to assure the conformity with the original cartography.

The methodology in short form is to use a stratified random sampling scheme with proper global sample size and adapted stratum sizes and compare the REN classification of the digital database versus the reference paper cartography in the sampled points to calculate the positional and thematic errors proportions estimates and the sampling error. The exact sampling methodology is detailed by Peixoto et al. (2006).

This section covers preparing computations, the implemented sampling algorithm and the description of a user interface to facilitate visual comparing of analogue and digital cartography.

2 Computed working processes

The manner in which spatial data is represented in an information system is key to the efficiency of the computational processes that will act upon it (Worboys and Duckham, 2004). The application works with spatial data given in shape or coverage format, which represents a simple geodatabase. Upon this database the calculations and randomizing computations can be done very efficiently and interoperability is granted by using the same output formats. Figure 1 gives a brief overview about the computed working processes, their input and output.

2.1 Preparing Calculations

To assure a sufficient global sample size and probable testing conditions we first need to know about the classes and class combinations in the REN data in each county. The input data contains overlaying polygons with each having one single class as attribute. As described below we use an object-oriented algorithm to get all existing class combinations and to calculate their respective area. As part of the equation for the global and stratum sample size we need these areas and their relation to the stratum size.

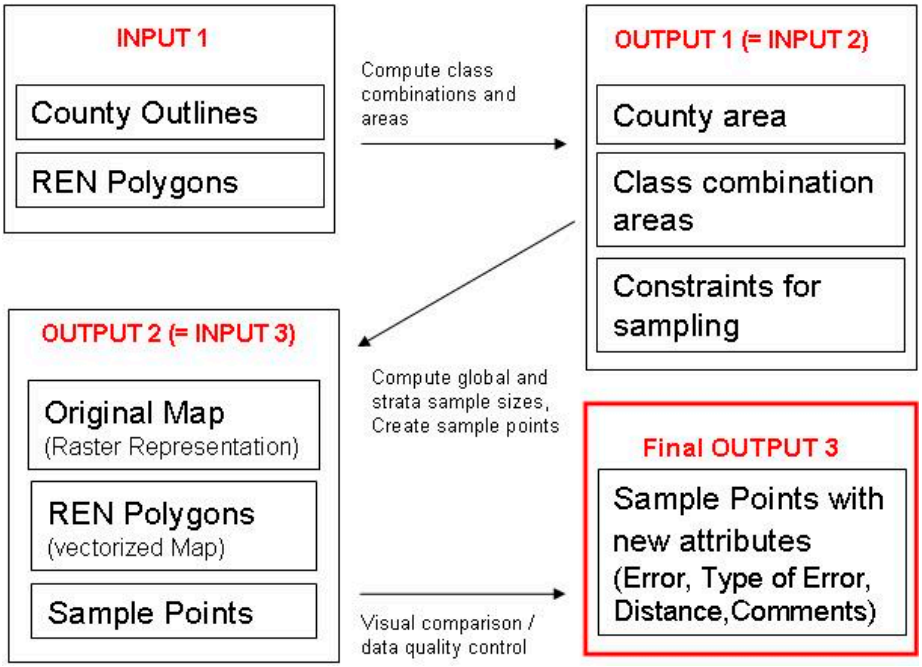

Figure 1 Input/Output diagram for computational processes.

2.1.1 Algorithm

The algorithm uses two classes representing a REN-class combination and a collection of REN-class combinations. It can be applied to all counties at the same time, if they are added as different layers to the map. First a collection of REN-class names is set. For each layer in the map (county) we set a new object of the combinations-collection-class. Then for each record we encode a binary string with 1 representing a contained class and 0 a not contained, respective to the sorting of the collection of REN-class names. If the encoded string exists in the combinations-object, its area is added to the area saved before, if not, a new combination object is created and its name and area is saved (Figure 2). After the computation the results can be printed in an ascii file or shown on the screen.

The result of this algorithm is a list of all single classes and class combinations with the respective area for each stratum, in our case county.

With the information of the number of classes and class combinations, the area of each class and class combination, the area of non-REN polygons, the total county area and constraints to assure probable testing of spatial accuracy and data quality assessment the sample size per class combination is calculated, following the equations described in Peixoto *et al.* (2006).

```
Set collection classnames
For each map layer (county)
   Combinations = New Classes Object
   For each record (polygon)
      For i = 0 to classnames.count -1
         Create string 'where 1 encodes class, 0 not
         If combination exists in Combinations Then
            Add area
         Else
            Create new combination object in Combinations
            Save area and name
         End If
      Next i
   End for
End for
```

Figure 2 Pseudo-code of preparing calculation algorithm.

2.2 Sampling

The inputs in this stage of processing are the areas of all REN classes and class combinations, for each county and the county area. This data is used to create some constraints for the sampling like global sample size, stratum sample size (for each county), assurance of considering each class and class combination with at least one point. These constraints are discussed in Peixoto *et a.l* (2006). After following the constraints the sampling is done by the use of a special randomizing algorithm explained as followed.

2.2.1 Randomizing Algorithm

The input for the algorithm is a list of REN classes and class combinations with the respective number of sample points. Out of the string, providing the REN class and class combinations, we create a SQL expression which is passed to a randomize-function together with the number of points for this class or class combination. The SQL expression is used to query the layer and create a selection containing the proper classes. This recordset is mostly composed by many polygons. The randomize function chooses first randomly one of these polygons, gets its extent, creates randomly a point using the polygons extent coordinates, checking if the point is inside the polygon (the extent coordinates are rectangular – the polygons are not), if the point is inside it continues until the number of points per class is reached. If the point is outside the polygon, it creates new points randomly until one is inside the polygon (Figure 3). Finally a shapefile with the sample points and their class as an attribute is created and added as a layer to the map.

```
get recordset(REN-class combination, SQLexpression)
for i = 0 to samplesize
      randomPolygon = random * numberOfRecords + 0.5
      get extent of randomPolygon
      notEnoughPointsBool = True
      while notenoughPoints = True
            randomPoint.X = random * extent
            randomPoint.Y = random * extent
            notEnoughPointsBool = True
            If randomPoint is inside randomPolygon Then
                  notEnoughPointsBool = True
                  ExportToShape randomPoint
            End If
      While End
End for
```

Figure 3 Pseudo-code of randomizing algorithm.

3 User interface

An application's interface is of extreme importance. It is inextricably related to the usefulness of a geographic information system. An interface should be easy to learn, appear natural, and be independent of implementation complexities such as data structures and algorithms (Egenhofer and Frank, 1988). The applications User Interface design is kept simple (Figure 4); it contains basic GIS-Viewer capacities like zooming, panning, measuring and an identification button in the toolbar, to provide a deeper map analysis and interaction. A legend and a map control are centred and buttons for visual comparison are placed in the bottom. An extra form containing a zoom on the raster maps legend can be opened too facilitate the data quality control (Figure 4, upper right).

3.1 Functionalities

To facilitate the comparison of the raster map versus the vectorized representation the User Interface takes some work usually done on the user side. After creating the random points some map properties are changed automatically. The REN-layer is shown in a bright transparent colour, the names of classes/class combinations are centred in each polygon. The user adds the proper raster image to the project and can open an extra form with a zoom on the raster legend. This helps comparing the raster map without scrolling to the legend or printing it. After adding the map the user sees the raster map in the background, above it are the REN polygon layer and the sample points layer, with point attributes (class and class combinations names) are shown. Beside that the user has two arrows to navigate through the actual dataset. By clicking analyze points, the application makes a zoom to the first record, centering at the point and allowing the user to compare directly the class or class combination on the raster map with the class or class combination in the point sample, which was read out from the REN layer. If the class / class combination is equal the user continues comparing the points. If the user finds an error, he has to classify it (thematic or spatial error) and can save a comment as attribute into the shape file. The newly changed class or class combination is also saved in the shapefile and in the case of spatial errors the distance to the probable polygon with the correct class is calculated.

3.2 Implementation

The tool was implemented with Visual Basic 6.0 due to its capacities in progress effectiveness, using the component software technology Map Objects® 2.2. from ESRI.

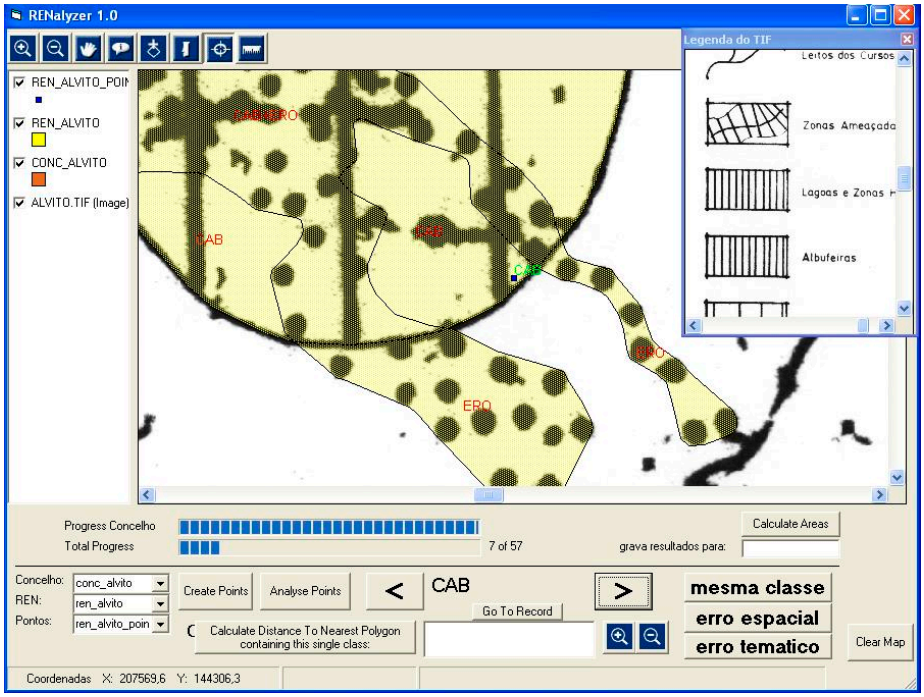

Figure 4 RENalyzer during visual comparison process.

4 Results and discussion

The result of this study is a stand-alone application, providing help and facilitation in data quality assessment by implementing the statistical framework and automating a lot of work processes. The class names and sample sizes are variable and easily changeable so the tool is applicable to different cases in different frameworks. The output of the tool is a shapefile for each county containing the sample points, their original attributes from the vectorized map. Iin case of thematic error the corrected attributes are saved and in case of spatial error the distance to the nearest correct feature is also saved. The shape format allows interoperability with other GIS tools for deeper data analysis and accuracy assessment. The data quality control is done by visual comparison through the user, so the source of probable errors could produce new errors during comparison. But, controlling the points out of the global sample size is comparatively few work and errors are little too. The results of the particular study and resulting statistical indicators are discussed in Peixoto *et al.* (2006).

5 Conclusion

In terms of these results and the positive study results presented in Peixoto *et al*. (2006), we conclude that RENalyzer is a tool providing helpful functions for spatial accuracy assessment in digitalized cartography. The work process is much shorter than the analysis in a full GIS, which costs a lot of manual operations.

References

Egenhofer, M. and Frank, A., 1988, Designing Object-Oriented Query Languages for GIS: Human Interface Aspects, Proceedings of the Third International Symposium on Spatial Data Handling, International Geographical Union Commission and Mapping; Falls Church, VA.

Goodchild, M., Gopal, S., (1989), Accuracy of Spatial Databases, London: Taylor & Francis.

Peixoto, M., Costa, A. C., Painho, M., Bartoschek, T., 2006(accepted), A stratified sampling approach to the spatial accuracy assessment of digital cartography: an application to the Portuguese Ecological Reserve, In: Spatial Accuracy Research Group (Eds.), Proceedings of the Seventh International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.

Worboys, M. and Duckham, M., 2004, GIS A Computing Perspective – 2nd ed., CRC Press LLC.