AUTOMATIC ITERATIVE GENERALIZATION FOR LAND COVER DATA

Olli Jaakkola
Finnish Geodetic Institute
Geodeetinrinne 2 (PL 15), FIN-02431 Masala, Finland
E-mail: olli.jaakkola@fgi.fi

ABSTRACT

The problem of generalizing spatial data is studied in the context of land cover data. The theory behind automatic generalization is reviewed and a raster based deterministic iterative generalization method is proposed. The methodology is based on the Map Algebra. The multivariate statistical testing methods are extended to deal also with generalized land cover data.

The case study is related to the production of a European CORINE Land Cover map from Finland. It presents an implementation of a new methodological concept for land cover data production, supervised classification and automatic generalization.In the case study, the existing supervised classification is supported with digital maps and attribute databases. All input data is combined to a detailed land cover with a very small minimum feature size, and automatically generalized to the standard European Land Cover. According to the quality tests performed, the automatically generalized method used here meets the present quality specifications. The method is fast, and it gives an opportunity for multiple data products in various scales, optimized for different purposes.

INTRODUCTION

The study includes both a theoretical examination of generalization concepts and methodologies as well as a practical implementation of these theories to the generalization of CORINE (Coordination of the information on the environment) land cover data.

The theoretical part describes the automatic generalization as it is reviewed in present scientific papers and how much we could apply these theories to the generalization of land cover data. The theories are partly modified in order to make them fit better with the used raster-based methods. Also the operations are redefined and the guidelines for executing operations is given. The methodology for handling both spatial and semantic knowledge in generalization operations is presented. Finally, some proposals for testing the quality of generalized land cover data are given. The accuracy testing methods are based on extensions of discrete multivariate testing, and on graphical overlays of generalized and original data.

The case study is related to the conversion of an existing land use database of Finland to a standard CORINE land cover database. At present, most of European countries still use the visual photo-interpretation of satellite images for the production of land cover features (see CORINE land cover 1992 and CORINE land cover - Guide technique 1993). In the Finnish version the conversion includes multiple input data in different spatial data forms, combination of different input data into one coverage, and a automatic generalization of the coverage to the CORINE land cover. This concept has also been tested in other European countries with common test data in Spain. The input data consisted of supervised classification, which was automatically generalized to CORINE land cover and coarser outputs. Some unlabeled areas were generalized manually. The paper shows that it is possible to automatically generalize a more detailed land cover to the coarser CORINE land cover database and fulfill the requirements specified for the database. The methodology provides a possible for multi-scale land cover databases generalized at different levels.

There is a report on the feasibility study for the Finnish CORINE land cover, with greater consideration on the theory of generalization and quality of generalized land cover data, as well as a detailed description of the contents of the case study (Jaakkola, 1994). The improved methods with heterogeneous classes are presented in Jaakkola (1995a, 1995b). The theoretical consideration is widened in Jaakkola 1996.

THE AUTOMATIC GENERALIZATION

Why, when and how to generalize

Cartographic generalization is a comprehensive process comprised of several interrelated processes. The generalization can be done for producing either graphical outputs as maps or for producing spatial data for geographical or statistical analysis. In both cases we may use similar operations, although we may not optimize the generalized data for all purposes. Therefore, it is good to keep in mind that good generalized maps are not always good spatial data for analysis. The cartographic generalization has, until so far, resisted the attempts for automatization.

Automatic generalization can be defined as the process of deriving, from a detailed (large scale) geographic data source, a less detailed (small scale) data set through the application of spatial and attribute operations. In attribute operations the spatial contiguity is not taken into account, whereas in spatial operations the value of the class at one point depends on spatial associations with the area surrounding it.

Objectives of a generalization process are: to reduce the amount, type, or cartographic portrayal of the mapped data consistent with the chosen purpose and intended audience, or to maintain clarity of presentation at the target scale. Clearly, the objectives state that the selection of generalization operations and parameters is dependent on the intended purpose of the generalized data.

The conditions for generalization include both object based and holistic measures, which have to be specified. The object based spatial measures specify the land cover features itself, such as its size, shape or distance to other features. They can be given as parameters to the generalization operators, e.g. the minimum feature size, the minimum feature width, the minimum distance between the features, the minimum inlet/outlet width etc. The parameters used specify the content of the generalized data set. The holistic measures specify e.g. the distribution or density of the land cover features. So far, the holistic measures have not been used in the developed automatic generalization procedures.

The forms of spatial and attribute operations possibly needed for the generalization of the land cover may vary for different data sets. Presented here are the ones needed for the CORINE land cover. As the terms originally have been defined for vector transformations, in the raster domain these terms may be somewhat fuzzy (McMaster & Shea 1992). The generalization operations are combined from the standard Map Algebra (Tomlin 1990) operations. The approach given here is similar to that described as object-based raster generalization (Schylberg 1993:109).

In the generalization, we have used one attribute operation, namely reclassification, which is the regrouping of features into classes sharing identical or similar attribution. Five spatial operations have been used. Aggregation is the lassoing of a group of individual point or areal features in close proximity and representing this group as one continuous area. Merging is the joining of contiguous features together and representing it as one area. Amalgamation is the joining of contiguous feature to an another one, either by attaching the feature to the semantically closest one, or by dividing it between the neighbouring features. Smoothing is the relocating or shifting of a boundary line to plane away small perturbations and capturing only the most significant trends. Simplification is the reducing of a boundary line complexity by removing changes smaller than a certain threshold.

The operations must handle both the spatial and semantic aspects of the data. The spatial aspect is handled with the object-based spatial measures. The distance and shape measures are given directly as a single parameter for the aggregation and amalgamation. The size is iteratively increased in geometrical series until we have amalgamated the features to the specified minimum feature size. The nature of the amalgamation procedure enables to stop the iteration at any time, and thus produce maps at all feature sizes.

The semantic information is stored in a hierarchic division of land cover classes, in a CORINE land cover nomenclature based on three levels of classes (see CORINE land cover 1992 and CORINE land cover - Guide technique 1993). In the procedures the semantic information is handled by using the groups of classes at these levels. Mainly we select groups of classes under level 1, although in merging we select classes under level 2. In the amalgamation we enlarge the possible neighbour classes from the level 1 to all land areas and to all areas, and amalgamate the features smaller than minimum feature size to nearest permitted neighbour class. In overall the generalization use hierarchy of classes specified for certain theme, at this case the theme is land cover. The method can be extended to other classifications as well by building the hierarchic groupings for the specified theme, and thus getting the priorities of classes during amalgamations.

The quality of generalized data

Most of the spatial data quality measures are scale dependent. Actually, none of the present quality measures is suitable, as such, to describe the quality of the generalized data. The generalization reduces the complexity of the data structure and adds error to the database, therefore the quality is always deteriorated in favor of simplicity and legibility. We have to understand that the error rate in the generalized database includes both the degree of generalization and the real error, the bias in summary measures and unintended positional and attribute errors produced by generalization. Also, the different generalization operations have different effects on the quality of the result (Jaakkola 1994:16-17), and the generalization operations are dependent on one another. Thus, we need to test the quality of different generalization operations combined with each other.

For testing the output data, we have modified the normal discrete multivariate analysis, based on the error or confusion matrix (see e.g. Sotkas et al. 1992), to take into account generalized land cover data. From the error matrix we have derived quality measures for the feature based attribute accuracy. Also, from the error matrix we have derived measures (Jaakkola 1994:20-21), for analysing the quality of land cover class shares after different generalization methods, namely the residual differences between the areas of the generalized and the original land cover. Visual inspection have been used for detecting the major positional and attribute errors.

CASE STUDY: THE FINNISH LAND COVER

The purpose of the Finnish Land Cover -project is to use existing land cover classification with auxiliary data and automatic generalization to produce a standard CORINE land cover database. The purpose of the CORINE land cover is twofold: to provide quantitative data on land cover for statistics, and to provide maps of different scales for European environmental policy. The land cover database in the nominal scale of 1:100000 is considered accurate enough, but not too large in volume. The spatial measures, when to generalize, are given with a minimum mapping feature size of 25 hectares and a minimum feature width of 100 metres. The quality specifications in the original CORINE procedure include a positional accuracy limit of 75 metres, and a feature based attribute accuracy of 85 percent for a total classification (CORINE land cover - Guide technique 1993). There is no specification for the areal shares of the classes, since the original method does not include a separate generalization task.

The procedure

As an input data we propose to use existing supervised satellite data classification (forest and semi-natural areas), map masks (fields and wetlands) , statistical records with position (buildings), and statistical records with digitizing (FIGURE 1). The digitizing was not included into the feasibility study.

Firstly, we aggregated point features, namely the individual buildings to the areal feature, built-up area. Thus we expanded the ground areas of certain buildings in artificial surfaces to also cover the surrounding concrete areas. Secondly, we combined the SLAM, and the building register to an ungeneralized land cover database (FIGURE 1). At the same time with combination we reclassified 64 SLAM and some other classes to CORINE classes. Thirdly, we merged the heterogeneous classes (CORINE classes 242, 243 and 313), reclassified and statistically compared the merged features (method see Fuller & Brown 1994:10-11). Fourthly, we amalgamated small areal features to larger ones. We performed iterative amalgamations in the different hierarchic levels of the CORINE nomenclature. With the hierarchic grouping we actually defined the priorities of amalgamation, first grouping and amalgamations for each class group in the CORINE class level 1, then for all land classes in level 1, and finally for all the classes. The amalgamations were done in several iterative size levels, geometrically increasing the minimum feature size from 0,0625 hectares to 25 hectares by each time doubling the minimum feature size (FIGURE 1). Fifthly, we smoothed the border lines, and thus reduced the narrow inlets and outlets in all features with a iterative majority filtering (method see e.g. Goffredo 1995). After raster-vector conversion we simplified arcs with too many points by giving a weed-tolerance of 25 metres, and obtained the final standard CORINE land cover database from the test area.

The quality of products

The generalization level for the CORINE land cover is the area of generalized features divided by the whole area, and is about 25 percent. It could be interpreted that generalized map features have on average about one fourth of other classes included, although one third of the test area is sea, and thus in land areas the generalization level is higher.

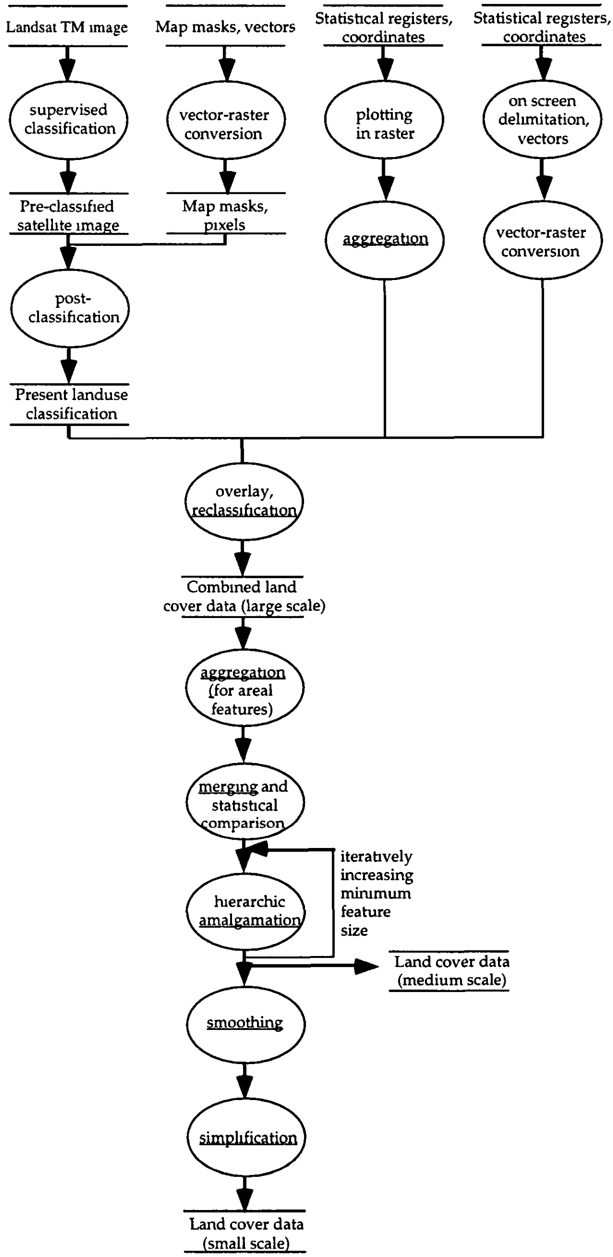


Figure 1. The classification, combination and generalization procedure.

The positional accuracy was verified by overlaying the original raster map with vector borderlines of the produced CORINE land cover. The borderlines of the main classes in level 1, e.g. water bodies, forests, and agricultural areas, are preserved quite nicely [Jaakkola 1994, appendices B5,B6]. The feature based attribute accuracy was tested with a comparison to the original combined data. In the derived measures we have the real errors produced by generalization and the level of generalization together (Jaakkola 1994, appendices A5,B5,B6]. The generalized pixels are graphically presented as a difference map between the original and generalized land cover map (Jaakkola 1994, appendix B7]. The quality of areal shares of different classes is important for the statistical use of the results. The areal shares changed systematically during the generalization, and the quality of shares were tested with the residual differences. In overall, the classes covering small areas with small average feature size tend to decrease in area or disappear totally, and the classes covering large areas with large average feature size tend to increase in area. Nevertheless, we may compensate these changes by redefining different parameters for different classes, which has not been tested yet. Note, that the residual differences (Jaakkola 1994, appendix A6) include both the degree of generalization and the real systematic errors, and that the heterogeneous classes should be reduced from the differences.

CONCLUSIONS

The presented procedures for the automatic generalization are fully operational. The processing time for the present generalization procedure for producing the standard CORINE land cover from the test area of 1200 km^2 is only a couple of hours. The automatic generalization can flexibly produce multiple products in different scales. Actually, the present procedure can produce different databases: an original combination map, maps of 1, 5, 10 and 25 hectare , or even maps of 50 or 100 hectare minimum feature size (Appendix 1). We can get all maps as side-products of the standard CORINE land cover. Maps are directly in digital form and can be stored on a database.

The accuracy tests confirm that the optimal solution fulfills the CORINE land cover quality specifications, and provides a fast, consistent, homogeneous, and accurate enough generalized land cover database. It produces a database of good positional accuracy, as well as of reasonably controlled attribute accuracy, for the class features and for the areal statistics of classes. The tests have been done in several sites in the Finland and elsewhere in the Europe. Also, the method has been tested with different minimum feature sizes for different land cover classes, and with topological constraints such as roads and other linear features. Tests has included maps as large as 40 Mega byte in size, and the method works well. The automatic generalization is considerably faster than the manual digitizing.

The size and shape parameters can be modified for different classes, e.g.

we may give a smaller minimum feature size for some important classes. Also, we can test the influence of different operations and parameters on the results, and select the most optimized ones. The procedure is areally homogeneous and objective if the supervised classification is controlled. In addition, the method is open for topological constraints, i.e. the land cover features can be kept topologically correct with certain point or linear features. The generalization method could also be used for homogenizing different data classifications in different countries, and for producing a truly consistent database in smaller scale over large areas, over the whole Europe.

Further improvement in quality can be introduced with more precise objectives. It is also good to bear in mind, that visual interpretation of the Finnish landscape is very difficult (Ahokas et al. 1992). The advantages of the new method compared to the old method are that it is fast, it provides multiple products, it provides consistent and homogeneous data, the positional accuracy is good, and the attribute accuracy is controllable since it produces systematic errors. The disadvantages is that it is not an easy task to add more detailed rules to procedure, and therefore only part of the human intelligence of interpreting is included into procedure.

Finally, we should understand that the aims of generalization can be multiple, and some procedures produce high quality maps, some high quality statistics for certain specified purpose, and these two aims can be conflicting.

REFERENCES

Ahokas, E., J. Jaakkola & P. Sotkas (1990). Interpretability of SPOT data for general mapping. European Organization for Experimental Photogrammetric Research (OEEPE) Publ. off. No. 24. 61 p.

CORINE land cover (1992). Brochure by European Environment Agency Task-Force. 22 p.

CORINE land cover - Guide technique (1993). CECA-CEE-CEEA, Bruxelles. 144 p.

Fuller, R.M. & N.J. Brown (1994). A CORINE map of Great Britain by automated means: a feasibility study. Institute of Terrestrial Ecology (Natural Environment Research Council). ITE project T02072J1. 37 p.

Goffredo, S. (1995). Knowledge-based system for automatic generalization of satellite-derived thematic maps. Proceedings of the 17th International Cartographic Conference, vol 1, pp. 108-117.

Jaakkola, O. (1994). Finnish CORINE land cover - a feasibility study of automatic generalization and data quality assessment. Reports of the Finnish Geodetic Institute, 94:4, 42 p.

Jaakkola, O. (1995a). Automatic generalization of land cover data. Proceedings of the 17th International Cartographic Conference, vol 2, pp. 1984-1989.

Jaakkola, O. (1995b). Theme-based automatic hierarchic generalization for the European Land Cover data. EUROCARTO XIII proceedings, Workshop on Scale and Extent. JRC, Ispra, Italy. pp. 80-93.

Jaakkola, O. (1996). Quality and automatic generalization of land cover data. Publications of the Finnish Geodetic Institute, n:o 122, 39 p.

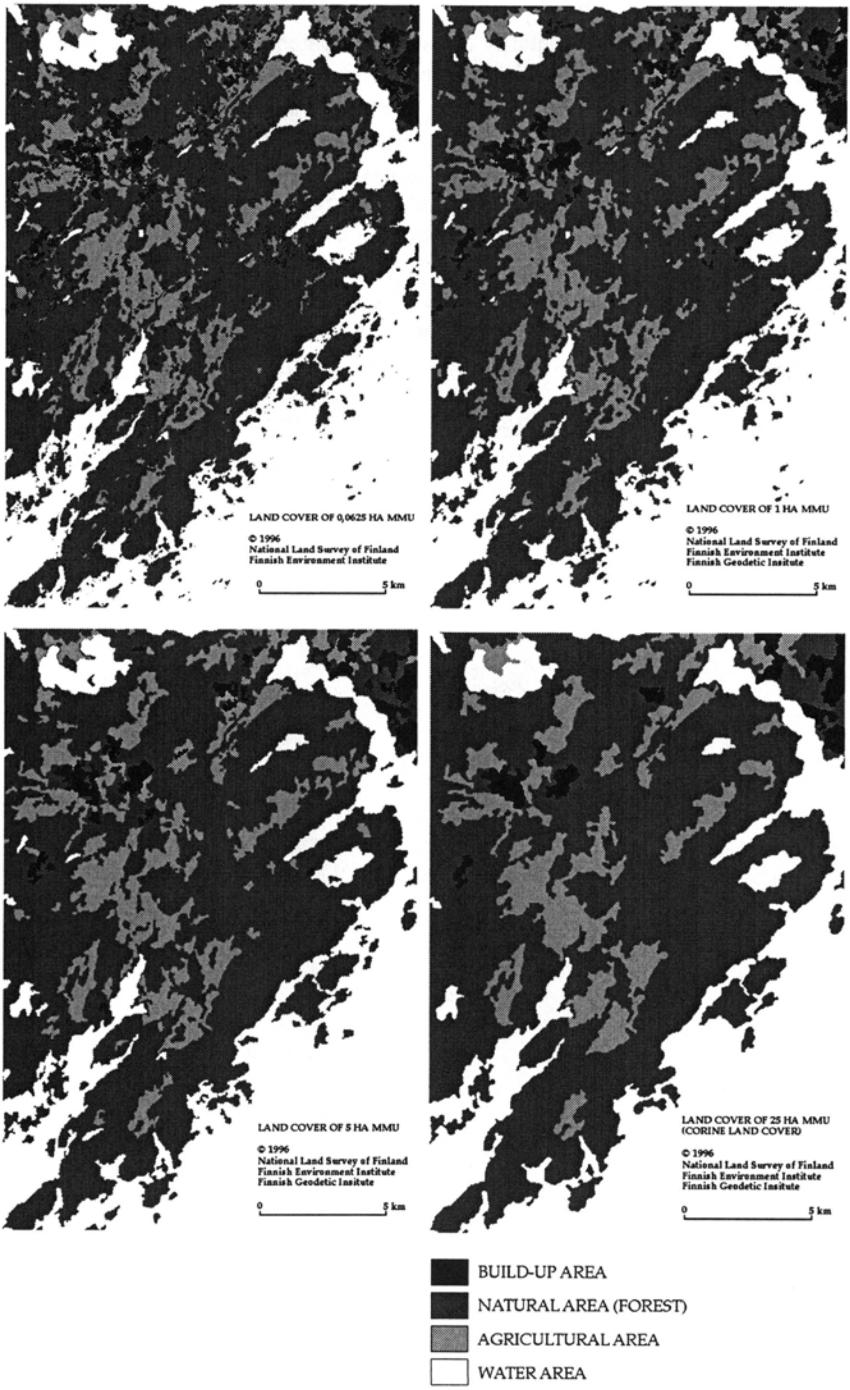
McMaster, R. & K. Shea (1992). Generalization in digital cartography. Association of American Geographers, Washington. U.S.A. 134 p.

Schylberg, L. (1993). Computational methods for generalization of cartographic data in a raster environment. Royal Institute of Technology, Department of Geodesy and Photogrammetry. Doctoral thesis. Stockholm, Sweden. 137 p.

Sotkas, P., J. Laaksonen & R. Kuittinen (1992). Satelliittikuvan tulkinta tarkkuuden määrittäminen (Determination of the accuracy of satellite image interpretations). Geodeettinen laitos, tiedote 5, 46 p. (In Finnish)

Tomlin, C.D. (1990). Geographic information systems and cartographic modeling. Prentice Hall, Englewood Cliffs, N.J., U.S.A. 249 p.

APPENDIX 1



LAND COVER OF 0,0625 HA MMU
© 1996
National Land Survey of Finland
Finnish Environment Institute
0
5 km
LAND COVER OF 1 HA MMU
© 1996
National Land Survey of Finland
Finnish Environment Institute
Finnish Geodetic Insitute
0
5 km
LAND COVER OF 5 HA MMU
© 1996
National Land Survey of Finland
Finnish Environment Institute
Finnish Geodetic Insitute
0
5 km
LAND COVER OF 25 HA MMU
(CORINE LAND COVER)
© 1996
National Land Survey of Finland
Finnish Environment Institute
Finnish Geodetic Insitute
0
5 km
BUILD-UP AREA
NATURAL AREA (FOREST)
AGRICULTURAL AREA
WATER AREA