Grid-Based Generalisation of Area Polygons for Automated Application in Small-Scale Statistical Maps

Marion D. Simon[1(✉)] and Hartmut Asche[2]

[1] Faculty of Science, University of Potsdam, Karl-Liebknecht-Street 24-25, 14476 Potsdam, Germany
marion.simon@uni-potsdam.de

[2] Computer Graphics System Group, Hasso Plattner Institute, University of Potsdam, Prof.-Dr.-Helmert-Street 2-3, 14482 Potsdam, Germany
hartmut.asche@hpi.de

Abstract. The added value of map representations of statistical mass data is now highly appreciated in big business. Most statistical maps are produced by non-experts using to GIS technologies or mapping functions embedded in statistical programmes. However, the range of semi-automated tools offered does not exploit the full potential of cartographic visualisation. Most map representations of statistical data colour administrative units resulting in area cartograms or choropleth maps. Visualisation of these area-related data need to focus on correct representation of the area rather than on the topographic elaboration boundary polygons. Hence the simplification or generalisation of the polygon lines becomes important for visual communication and perception. Up to date, no generally accepted rules exist for the generalisation of the polygon lines of area cartograms in which presentation scales. This paper deals with solution approaches in line generalisation using square grids. As well as an answer to the question, if square grids could be an option for fully automated map generalization in small scaled statistical maps.

Keywords: Geovisualisation · raster · generalisation · simplification · statistical maps · small-scale maps · business maps · cartographic generalisation · automated generalisation · raster grids

1 Description of the Problem

Most statistical map representations visualising quantitative enterprise (geo)data are areal representations, since the data provided usually refer to a defined area. Areas are differentiated by area colouring (e.g., choropleth) or map symbols (e.g., diagrams) placed in the respective areas (see Fig. 1). Administrative boundaries are most frequently used to delineate the areas in question. This line network of area boundaries is the basic element of this simple map type. Alternatively, quantitative data with area reference can also be plotted on artificial line networks in addition to the administrative reference area.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14107, pp. 371–383, 2023.
https://doi.org/10.1007/978-3-031-37114-1_25

Applying this so-called field principle results in the field cartogram map type. Administrative boundary lines and area units are freely available on the internet for defined scales. On its GISCO platform [1], Eurostat of the European Commission offers, among other data, linear and areal representations (lines and polygons) of administrative boundaries (worldwide) for representations at scales of 1:1 million, 1:3 million, 1:10 million, 1:20 million, 1:60 million. For small-scale statistical maps, the geometric precision of boundary polygons contained in these data sets needs to be reduced. That is because the unwanted level of detail of boundary line representations affects processing time, computing power, memory requirements. Moreover, simplification (or generalisation) is mandatory for map representation in a frequently reduced output scale to and the amount of generalisation appropriate to the output scale to ensure readability of the respective map.

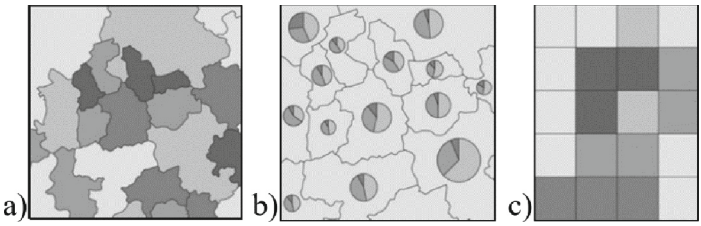


Fig. 1. Mapping principle of a) choropleth maps, b) area diagram maps and c) grid net maps (source: [2])

A range of software tools, such as Mapshaper, are available for the generalisation of boundary lines. These tools offer various mathematical algorithms with an individually adjustable degree of line generalisation [3]. The assessment whether the results generated meet the requirements of the base map lies exclusively with the user. Buttenfield [4] segments line elements for ordering into hierarchical tree structure with subsequent statistical examination of the geometries of the line segments. Lay and Weber [5] deal with the generalisation of mosaic maps (specifically forest areas) on a raster basis considering the automated generalisation of screen maps whose raster cells are used as a measure of the mesh size of the grid. However, no definition can be found how simple "highly simplified" line polygon generalisation is. The Eurostat national borders polygon datasets mentioned are provided for a scale of map provision. Specific recommendations for the usability in concrete application-related map types and the necessary generalisation requirements per scale are not given.

This work is intended to contribute to the research of line generalisation by grids with a view to increase the degree of automation in the research field of automated map construction. The aim is to define scale ranges and to assign them degrees of generalisation for administrative units which enable a standardisation for thematic map representations. An acceptable generalisation solution for an island representation adapted to the output scale range is provided in this work for a specific scale range. Zooming in or out of the screen display is effected with the determined degree of generalisation within the scale

window to be defined. The objectives of the study are to assess the degree of automation, speed, correctness, and effectiveness of the resulting map representation. For this purpose, the following parameters are assessed:

Recognisability of chosen administrative unit (yes/no),
Ease of map exploration (high, medium, low),
Correctness of boundary depiction (yes/no),
Data maintenance effort (high, medium, low),
Performance assessment (good, medium, poor),
Usability rating (good, medium, poor),
Degree of automation (high, medium, low).

2 Solution Approach

The generalisation of a linear polygon delimiting an administrative area unit is tested in this work by using grids of various cell size in combination with different line generalisation approaches. First, approaches for base point reduction will be carried out on the basis of vector data. This is followed by a generalisation approach on a so-called raster basis. Based on the "raster", which originally refers to non-vector pictorial information, square grids will be used for line generalisation solutions.

1. The first vector-based approach (area-area) assumes a minimum overlap of grid cell area and underlying area unit (polygon), after which relevant contour components of the area ring are to be generalised (line generalisation). Further runs are to be made with a variation of the mesh size and the minimum coverage.
2. The second vector-based approach (line-line) considers only line intersections of boundary lines of the grid cells and polygon boundary lines (area perimeter, e.g., national boundary line).
3. The third line generalisation approach is the grid-based approach. Similar to the quadtree principle it pursues the division of a grid cell (grid cell as square) into exactly four daughter cells, if the grid field (areal) lies over a section of the linear area ring (area-line). After a defined number of iterations the smallest square is depicted by means of colour filling to delimit the area.

The results of these three line generalisation approaches will be compared with a generalised reference at a certain scale.

This conceptual approach results in following work packages:

Select reference map

- Definition of a generalised base map for statistical map representation consisting of administrative units.
- Adoption of the scale used in the reference map to derive a display scale range based on this base map.
- If necessary, digitalisation of the reference if available in analog form and determination of the number of base points (coordinate pairs) used.

Compile investigation material

- Geodata research and, if necessary, data extraction of the spatial section that occurs in the reference.

– Description of the data set elected regarding the map projection and accuracy (provision scale, number of support points used).

Develop implementation concept

– Set up the investigation procedures for each grid-based line generalisation approach.
– Requirements formulation for the evaluation of the results.

Application

– Implementation of the concepts developed for a defined scale range on extracted examination material.
– Presentation of results.
– Comparison of the graphical and numerical results with the reference.
– Validation based on at least two further administrative units for mapping in the same scale window.

The polygon generalised as an administrative unit is covered with a network of regular square grids of a certain mesh size. The reduction of polygon support points, each located with x and y coordinates, is to take place at those partial lines whose surrounding cuboids are "filled" to more than 60% by the polygon. A line segment is cut at the entry and exit point of a raster cell. Two new coordinate pairs are created at the intersection points. In order to avoid an increase in memory instead of reduction, the intersect conditions need to be described. To ensure reduction it is required that only cells that containing more than three grid points are allowed to intersect.

The results of this approach are generated for different square grid sizes. After these passes, a mean result is extracted from each scale range as a standard. This scale-dependent cascade can be used in the automated construction of thematic map representations.

3 Implementation

First, the scale range is defined starting from the initial data set for maps at a scale of 1:1 million. Since the next data set from the same data source is available for the scale 1:3 million, a scale range from 1:1 million to 1:3 million is defined for which the results are to be produced.

Various grid shapes (hexagon, triangle, rectangle, square, trapezoid) are available with their respective advantages and disadvantages for mesh formation. For this work, a grid shape is chosen and its origin superimposed on the origin of the map grid (bottom left). This facilitates localisation or geocoding, if necessary. Following this procedure, a square grid is preferable to a rectangular grid because sides are of equal length.

3.1 Initial Data Set and Reference Map

For users of the EU member states, the geographical data of the European Commission from the geographic information system (GISCO) [1] managed by Eurostat can be used. Selected country sections are extracted from the data set "NUTS_RG_01M_2021_3035" covering Europe. These data from the year 2021 can be mapped in high detail for a scale

of 1:1 million in Lambert's areal azimuthal projection (ETRS_1989_LAEA, EPSG-code 3035). For this purpose, the land area of Albania (27,398 km^2 – [6]) is extracted from the presented dataset. The land area ring of this original dataset comprises 785 points.

As a reference and for the evaluation of the results, the country polygon of Albania generalised for statistical maps from the Albanian National Atlas [7] is used. The map, which is available in analogue form, was digitised for this purpose at the working scale of 1:500,000.

This scale ensures the recognition of each line segment and the precise positioning of support points. This digitised national polygon is described by 181 support points. As has been mentioned, the projection used for the atlas is not known. Georeferencing is hence necessary for a direct comparison of the results. If the result is positive, the solution approach can be validated on countries with about 10,000 km^2 larger and smaller areas (Switzerland 41,285 km^2, Slovenia 20,151 km^2), [6], as these can be mapped in the same target scale range.

3.2 Vector-Based Grid Generalisation

The first vector-based generalisation approach (area-area) presented in Sect. 2 is aimed at reducing existing line support points. A square grid of different mesh size is placed on the area unit to be generalised. Subsequently, coverage percentages of the land area (x) are set in relation to the area of the grid squares (100%). All grid squares where the coverage by the land area is between 60% and 99.9% are generalised in the following. The following steps are carried out (see Fig. 2):

- Base point generation at grid edge and country ring intersections,
- Deletion of the intervening original support points

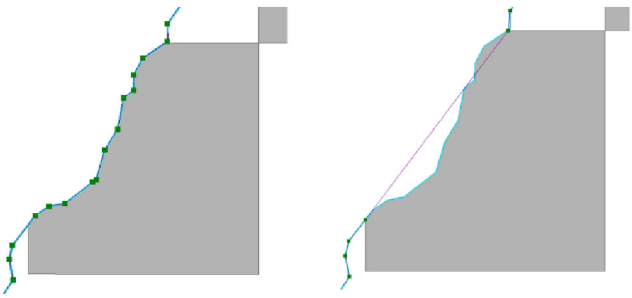


Fig. 2. Left: before, right: after line generalisation through grid-cell based generalisation with partly filled grid cell by underlying polygon (own source)

In this way, sections of the country's perimeter will remain in the original and other line sections will be represented as straight lines.

The second vector-based generalisation approach presented (line line), uses a square grid of a defined size placed on the polygon of the area unit to be generalised. This way a new linear feature is created. It consists of pairs of coordinates, defined at the intersections of the grid edges (or nodes) and the area boundary line. This is done by setting vertices and connecting them with a line. The original vertices of the area perimeter remain unaffected.

3.3 Raster-Based Grid Generalisation

The grid-based solution approach is suitable for scales ranging from 1:1 million to 1:2.5 million. The country area to be generalised is covered with a square grid dimensioned according to the map grid design used. Starting with a mesh size of 7,500 m, the overlap of a grid field with the line segment is checked. In case of overlap of a line segment, the grid cells are further divided by a number of iterations following to the quadtree approach. An iteration step means a further subdivision of a grid field into exactly four equal squares. The final step consists of filling the smallest squares. In this application, the colouring is used to represent the polygon ring. With this approach all area polygons are scanned first and thus no island of relevant size is deleted from the map display.

4 Application

4.1 Vector-Based Grid Generalization

The bay of Vlora with the island of Sazanit is strongly distorted when applying a mesh size of 15,000 m (see Fig. 3, b). When using a smaller mesh size of 10,000 m the topographic features are retained (see Fig. 3, a, left circle). The recognition of the southern tip of the country is also more likely with the smaller mesh size than the result with 15,000 m mesh size allows. The grid calculation refers to a bounding box surrounding the land area with the grid origin at the bottom left. Variant a) (Fig. 3) appears to be graphically more correct, since the differences in level of detail of generalised versus ungeneralised line segments are visually more apparent.

Despite larger mesh size and, as a consequence, a greater number of interpolation point reductions, the opposite occurred. The number of grid points generated and retained is 376 in a) and 424 in b). Thus, doubling the area of the grid meshes is not an option for halving the result from a). Another option for reducing the number of grid points is to halve the coverage of a grid field (10,000 m). This procedure leads to the fact that the entire national boundary polygon becomes relevant for generalisation (see Fig. 4) and would have to be generalised completely according to the approach described in Sect. 3.2.

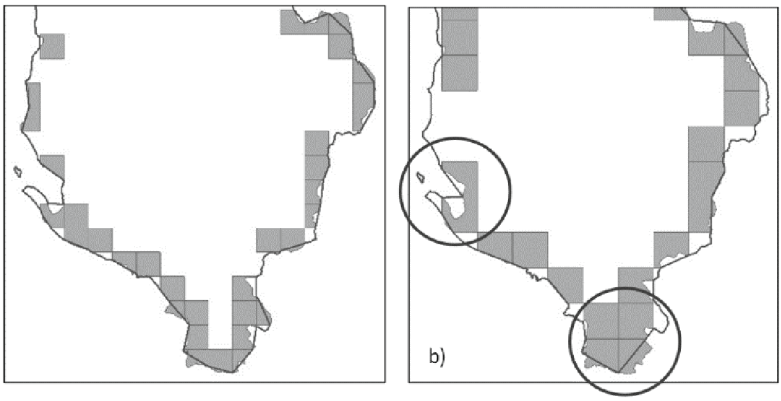


Fig. 3. Comparison of the border polygon of South Albania with grid-based line generalisation using mesh sizes of a) 10,000 m, b) 15,000 m and grid fillings of 60–99.9%, respectively. (own source)

As a consequence, the coverage parameter in the approach is replaced to generate a new line element (country polygon). The modified procedure is to first place a square grid a with mesh size of 7,500 m on the country polygon of the initial data set for Albania. In a second step a support point is placed at each intersection of the grid and the country polygon and subsequently connected them by open polygon lines. The result is shown in Fig. 5 in combination with the underlying reference data set. The reference data set is described by 181 vertices, the generalised original data set of the output data with 176 interpolation points. A visually pleasing appearance is also guaranteed.

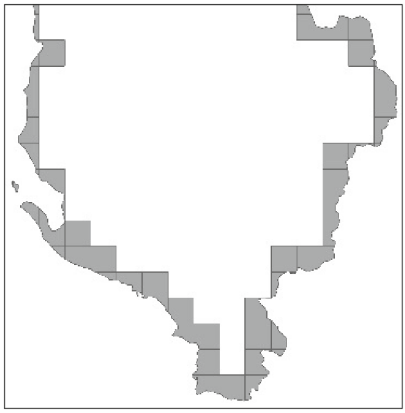


Fig. 4. Reduction of vertices halving the area overlap 30–99.9% (own source)

The result presented has been verified by an adaptation of Salistščev's nomogram, which systematically arranges recommendations for scales or scale ranges for mapping defined countries or continents onto a given mapping surface [8]. This procedure raises



Fig. 5. Georeferenced national boundary polygon of Albania 1:1,750,000: thin line shows borderline generalised for statistical map (Albania atlas) (181 control points), thick line shows georeferenced result with 6 control points according to vector-based grid generalisation approach with grid size of 7,500 m (own source)

the question whether the underlying methodology is generic and can be applied to other countries in a similar scale range (1:1,000,000 to 1:2,500,000) as well. For this purpose, countries with a land area of approximately 10,000 km^2 smaller and larger than Albania were used as an example (Slovenia < Albania < Switzerland) (Figs. 6 and 7).

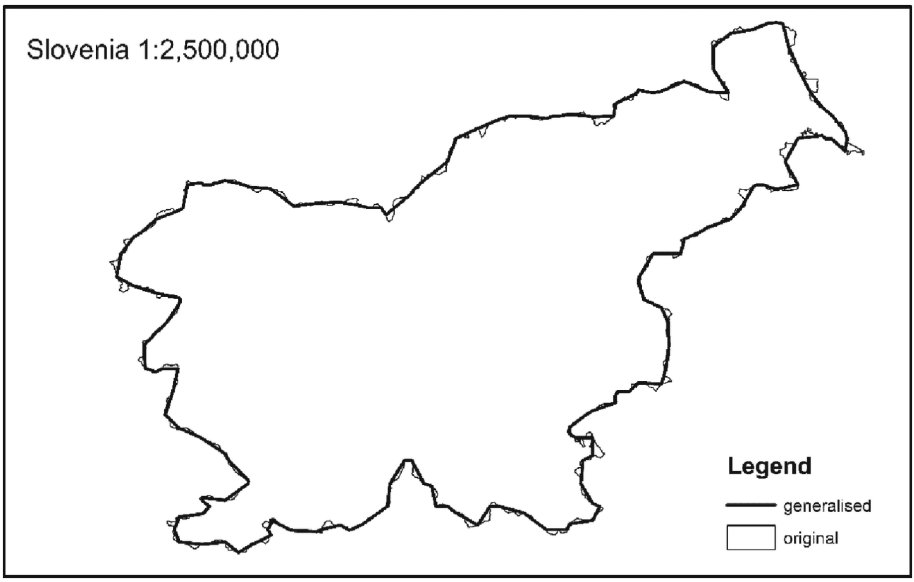


Fig. 6. Slovenia 1:2,500,000 generalised by pure grid-based line generalisation (own source)

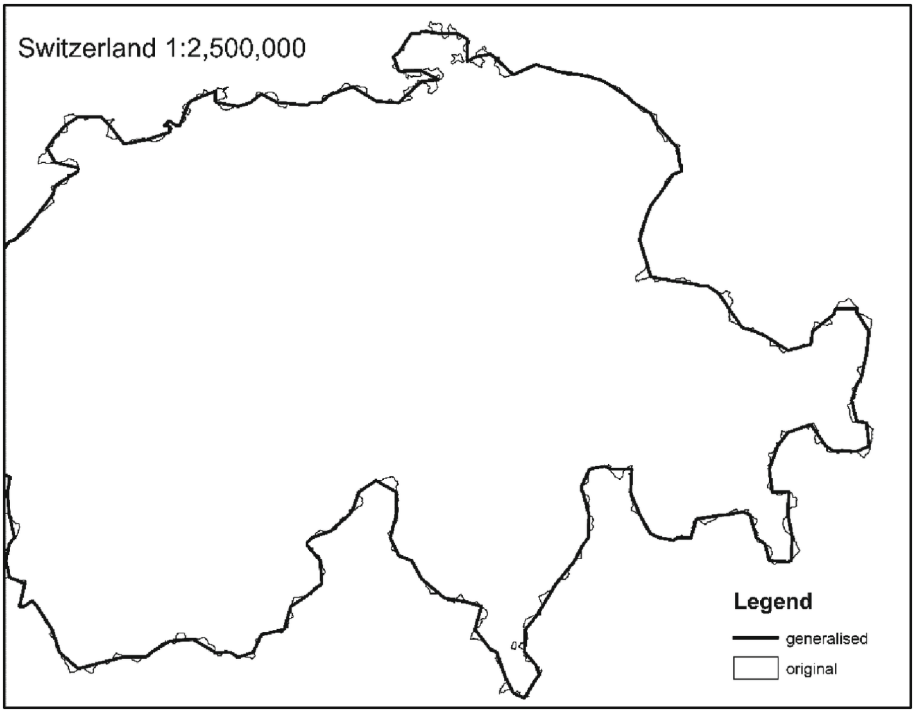


Fig. 7. Switzerland 1:2,500,000 generalised by pure grid-based line generalisation (own source)

4.2 Raster-Based Grid Generalisation

To start the process, the initial mesh size was set to 7,500 m. This figure is based both on the positive results from the second vector-based line generalisation and considering the relationship with the reference map scale. Koch and Stams [9] emphasise the "close interdependence" of "field size and scale". Based on this initial mesh size, the grid-based line generalisation approach presented in Sect. 3.2 is then iteratively processed using a square grid. The grid field division is a halving of the grid edge lengths and a re-check of the overlap of new grid patches and the line from the area outline (Fig. 8).

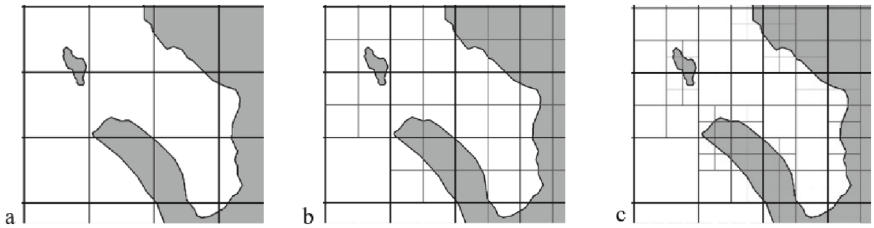


Fig. 8. Quadtree approach of grid field reduction when covered with polygonal ring; mesh sizes: a) 7,500 m, b) 3,750 m, c) 1,875 m (own source)

The number of iterations is performed up to a defined mesh size. This depends on the application or the type of map to be created (see Fig. 9). For an area cartogram based on administrative units, a linear representation and thus a smaller mesh size is required than a cartogram produced following the field principle. The latter is characteristically created using semantically coloured/filled raster cells. The result is a map of the field cartogram type. This automatically generated map representation is optimally suited to provide an easy-to-perceive overview of the geographical distribution of the statistical data in small scales. It is apparent that the defined cell size controls the number of iterations required to generate the delineation of the country's area.

Code is necessary for automation. A possible code could be explained as follows:

/* Function to load the shapefile CNTR.shp, project it in EPSGcode 3035, overlays a rectangular gridnet with cellsize 7,500 m, divide each cell in 4 rectangles, if cell overlies polygon line of the loaded polygon, the iteration of cell splitting continues until cell size reaches 468.75 m. */

Using the example of the grid-based approach, the content of a suitable code, here based on the programming language C, could have the following structure:

```
Int main (void);
int loadShapefileAndDivideCells(…) {
    // Initialize GDAL
    // Open the shapefile
    // Get the layer
    // Get the spatial reference
    // Set the EPSG code
    // Create a rectangular grid
    // Iterate over the cells
            // Check if the cell overlaps the border of the polygon
            // Divide the cell into 4 rectangles
                  // Check if the rectangles overlap the border of the
                     polygon
                       // Decrease the cell size
            // Destroy the feature
    // Destroy the rectangle
    // Close the shapefile
    return 0;
}
```

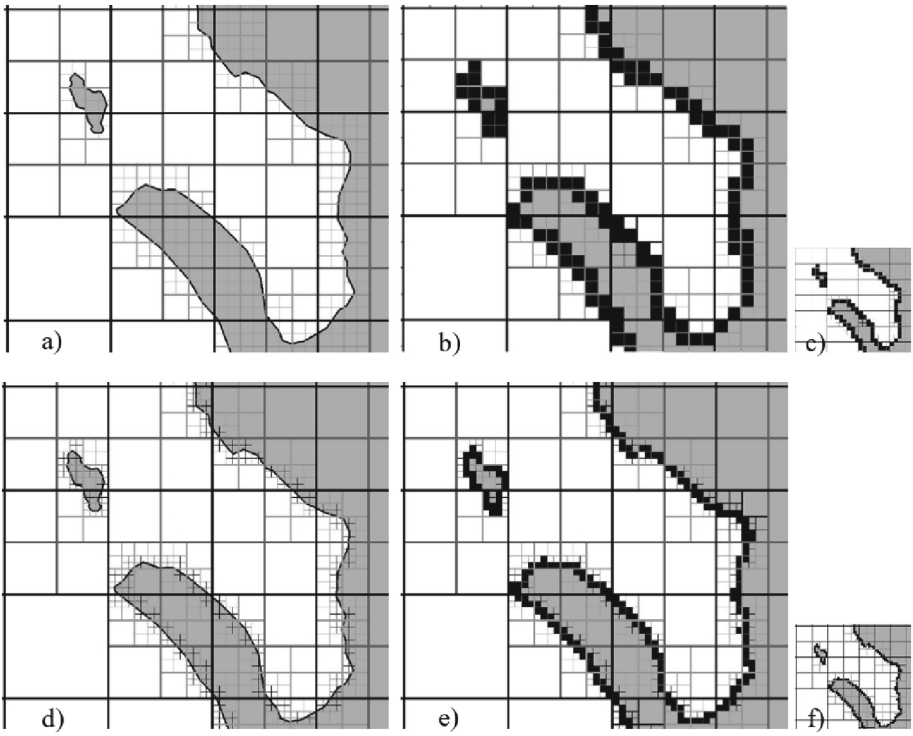


Fig. 9. Quadtree approach of grid field reduction with closed polygon coverage; mesh sizes: a-c 937.5 m; d-f 468.75 m; a), b), d), e) at working scale 1:150,000; c) and f) at target scale 1:1,750,000 (own source)

5 Concluding Remarks

The quality of the various grid-based line generalisation approaches are assessed against the objectives listed at the beginning (Table 1). In detail, the following results have been found: Both vector-based approaches (area-area; line-line) include island polygons off the sea-land border polygon only if a vector or grid cell intersects with or covers the topographic form of an island (polygon or line). Otherwise it is excluded from generalisation and subsequent map representation. The vector-based generalisation approach for area boundary lines described in Sect. 3.2 can be neglected in the form of the area coverage-dependent generalisation approach because of the high data maintenance effort (compared with the original data set) and the comparatively high computational effort. It also emerged that a correct and speedy perception of the administrative units and a visually pleasing appearance of the latter is strongly dependent on the selected mesh size and the original data set.

Table 1. Evaluation of the investigation objectives per grid-based line generalisation method

	Vector-based area – area (meshes: 10,000 x 10,000 m, coverage: 60%)	Vector-based line - line (meshes: 7,500 m x 7,500 m)	Raster-based area - line (initial mesh width: 7,500 m)
Recognisability	yes	yes	yes
Interpretation speed	medium	fast	very fast
Data storage effort	high	low	medium
Performance assessment	poor	good	good
Usability	good	good	good
Degree of automation	medium	high	high

The variant of the pure line intersection-based solution derived from the results of the above, but neglecting the degree of area coverage, is computationally less intensive. This is due to a high degree of data reduction while preserving the recognisability of the country polygons at the same time. The solution is optimal for the scale range mentioned (1:1 million to 1:2.5 million) and area sizes from about 20,000 km^2 to 40,000 km^2 because it keeps the usual vector-based procedure of base point reduction and shifting in line generalisation.

Depending on the chosen map type (field cartogram, choropleth/area diagram map), the new, raster-based line generalisation approach is extremely well suited for line generalisation for area delineation for the scale range being processed. A significant advantage of this generalisation approach is that several cartographic steps can be combined in a

single operation. The production of statistical maps and their cartographic generalisation of the base map are processed together using square grids.

When used to create a field cartogram, the map content can be perceived and analysed easily and quickly. This map type with its boundary generalisation proves to be particularly efficient in communication area-related statistical information to a wide audience. One reason is that the focus of attention is directed a to the map topic, and not to topographic details of the delimitation of the respective administrative units. Furthermore, in contrast to vector-based approaches, a fast combination with raster-based information from satellite data is possible.

In summary, grid-based line generalisation can be a useful option for the production of statistical maps. It simplifies the initial input data, reduces file size and processing time, and allows the production of maps at different scales. For these reasons, an extension of the grid-based line generalisation approach to other scale ranges and area sizes is conceivable. The results provide an efficient, easily automated generalisation component for area delineation in the base map. In a broad sense, grid-based line generalisation can be considered one of the most effective statistical map representation. It thus significantly contributes to the full automation of the map construction process for statistical representations.

References

1. Gisco, Europäische Kommission: Geobasisdaten. https://ec.europa.eu/eurostat/de/web/gisco/geodata/reference-data. Accessed 17 April 2023
2. Bollmann J., Koch, W. G. (Hrsg.): Lexikon der Kartographie und Geomatik. In two volumes, vol. 1, p. 118, 248, 250. Spektrum Akademischer Verlag. Heidelberg, Berlin (2001)
3. Mapshaper: Liniengeneralisierungswerkzeug. https://mapshaper.org/. Accessed 17 Apr 2023
4. Buttenfield, B. P.: Digital Definitions of Scale-Dependent Line Structure. 1986. https://www.researchgate.net/publication/245928529_Digital_Definitions_of_Scale-Dependent_Line_Structure
5. Lay, H.-G.; Weber, W.: Waldgeneralisierung durch digitale rasterdatenverarbeitung. In: Institut für Angewandte Geodäsie. Bundesamt für Kartographie und Geodäsie (Hrsg.): Nachrichten aus dem Karten- und Vermessungswesen. Reihe 1: Originalbeiträge. Heft Nr. 91., p.61–71. Verlag des Instituts für Angewandte Geodäsie. Frankfurt am Main (1983). ISSN 0469–4236
6. CIA: Worldfactbook. https://www.cia.gov/the-world-factbook/countries/albania/. Accessed 17 Apr 2023
7. Bërxholi, A.: (Hrsg.). Stability Pact for South Eastern Europe (Hrsg. Organ): Atlas of Albania. Demographic Atlas of Albania. Tiranë: Shtypshkronja Ilar (2003)
8. Sališčev, K.A.: Nationalatlanten. Vorschläge zu ihrer Vervollkommnung. In: Neef, E. (Hrsg.): Petermanns Geographische Mitteilungen. 104. Jahrg. pp.77–88. VEB Hermann Haack. Geographisch-Kartographische Anstalt. Gotha (1960)
9. Koch, W.G., Stams, W.: Felderprinzip. In: Bollmann J., Koch, W.G. (Hrsg.): Lexikon der Kartographie und Geomatik. In two volumes, vol. 1, p. 241. Spektrum Akademischer Verlag. Heidelberg, Berlin (2001)