Development of a new and open approach to dissolve polygons storing count data based on areal threshold

Marcello Schiavina
European Commission,
Joint Research Center
Via E. Fermi 2749
Ispra (VA), Italy
marcello.schiavina@ec.europa.eu

Sérgio Freire
European Commission,
Joint Research Center
Via E. Fermi 2749
Ispra (VA), Italy
sergio.freire@ec.europa.eu

Abstract

While many analyses require smart and flexible aggregation of adjacent and complex polygons storing quantitative data, currently available tools do not meet this need.

In this work we present the 'SmartDissolve v1.0' toolbox for ArcGIS (Esri) and MATLAB, a new approach and tool that handles minimum mapping unit, resolution mismatch between layers, or spatial uncertainty problems in GISc. This tool automatically dissolves polygons below a threshold area, updating fields' values. The toolbox allows to select the ordering of polygon analysis (i.e. from the smallest to the biggest area, vice versa, or order of IDs), different dissolve rules (i.e. with smallest, largest, or maximum-border-share adjacent polygon, minimum total perimeter or maximum compactness) and different field updating operations (i.e. sum, mean or text concatenation).

The approach is illustrated by combining raster-based population data from LandScan with vector administrative units to capture the 'ambient population' of a set of municipalities in Naples, Italy.

Keywords: dissolve; GIS generalization; smart aggregation; minimum mapping unit (MMU); raster-vector upscaling; MAUP.

1 Introduction

Within Geographic Information Science (GISc), the issues of resolution and spatial scale remain significant topics, especially in raster-based remote sensing and image processing. These aspects are particularly critical for spatial analysis and modelling, also in socioeconomic or environmental studies where vector-based data are frequently used, often in combination with raster layers (Mu and Wang, 2008). Many analyses involve combination of layers having quite different spatial scale or resolutions, requiring aggregation of the finer-scale data set to the resolution of the coarser one.

Aggregation of spatial data is typically performed through the 'dissolve' spatial operation, which is one of the more common and useful procedures applied in Geographic Information Systems (GIS) (Davis, 2001). In available dissolve tools, the aggregation is based on polygons sharing the same category or code (i.e. boundaries are removed between adjacent polygons that have the same value for a specified attribute). However, some analyses require more sophisticated aggregation by dissolution of adjacent polygons, such as reaching defined target areas (surface) correctly handling multi-part features and accurately summing quantitative count variables.

Several scientific studies (Bader and Weibel, 1997; Fraley and Raftery, 1998; Murrey and Shyy, 2000; Mu and Wang, 2008) proposed GIS-based generalization (*sensu* Johnston *et al.,* 1999: 'the appropriate representation of the two-dimensional polygon resolution') methods for polygons, especially for polygon clustering where the aggregation of features is subject to similarity in attributes or to balances of compactness and within-area homogeneity, while there is still need for ready-to-use simple algorithms to improve the GIS operation 'dissolve'.

Martinez-Llario *et al.* (2009) improved the standard GIS tool 'dissolve', but only enhancing computing performance. While some GIS vendors include enhanced 'dissolve' functionalities, such as 'ETGeoWizard Dissolve Polygon', such products are commercial and the way multi-part features are handled does not ensure preservation of the initial spatial distribution and the total volume, which are paramount requirements for count data such as population censuses.

However, between the standard dissolve tool, that is still limited to the aggregation by field, and the polygon clustering methods, there are many procedures that are usually only accomplished through several non-automated and computer intensive GIS tasks (Laurini *et al.*, 2016). These tasks include dealing with object attributes during the dissolve process and/or setting thresholds to dissolve only polygons with certain characteristics. Moreover, numerous studies may require that analysed units reach a certain minimum mapping unit (MMU) for results to be meaningful (e.g. when sampling or summarizing other data, i.e. points; or when upscaling from raster data to vector units in the case where some polygons are smaller than the resolution, i.e. cell size, of the raster layer),

while not losing the information stored in those small polygons. In addition, polygons having low geometric/positional accuracy but having high thematic/attribute reliability would benefit from a robust aggregation that preserves the attribute values while decreasing the positional uncertainty of overall quantities. This would mitigate the frequent mismatch between thematic and spatial resolution.

In this work we present the development of a new approach and tool, the 'SmartDissolve v1.0' toolbox for ArcGIS (Esri) and MATLAB (MathWorks), to automate the dissolution of polygons in a smarter and more flexible way, by handling minimum mapping unit requirements, resolution mismatch between layers, or spatial uncertainty problems in GISc. This tool automatically dissolves polygons that are below a user-defined threshold area, updating their fields' values. The toolbox allows to select polygon analysis ordering (i.e. from the smallest to the biggest area, vice versa or order of IDs), different dissolve rules (i.e. with smallest, largest or maximum-border-share adjacent polygon, minimum total perimeter or maximum compactness) and different field updating operations (i.e. sum, mean or text concatenation).

We illustrate the approach by aggregating vector administrative units to a MMU that is more suitable for combining with raster-based LandScan population data (Dobson *et al.*, 2000), to better capture the 'ambient population' of a set of municipalities in Naples, Italy.

2 Methodology and tool development

The 'SmartDissolve' tool can be imported into ArcGIS 10.X toolbox or used directly as a function in MATLAB. 'SmartDissolve' requires seven inputs to run (Figure 1): (i) the feature layer to work on; (ii) the area threshold and its unit of measure (MATLAB version requires only a value in squared meters); (iii) the polygon areal-analysis ordering (iv) the polygons dissolve rule; (v) the list of fields to be updated, (vi) the updating rule for each field selected and (vii) the number of cores to be used.

Figure 2 shows the 'SmartDissolve' algorithm flowchart. The input feature layer ('Input Feature') could be a shapefile or, only in the ArcGIS toolbox, a vector layer within a Geodatabase (i.e. gdb file). Three different polygon ordering methods are possible: (i) 'ID' order; (ii) 'MinMax' or (iii) 'MaxMin'. With 'ID' order, the algorithm will cycle polygons following the original order (i.e. how the polygons are stored in the feature layer). Using the other rules (i.e. 'MinMax' or 'MaxMin') the algorithm sorts polygons before cycling through them: from the smallest to the largest and vice versa by selecting 'MaxMin' or 'MinMax', respectively. Polygon's area and perimeter are evaluated using 'polygongeom' function that differentiates the cases of projected and unprojected maps. In the first case the area and perimeter are easily calculated with

Figure 1 – 'SmartDissolve' GUI for ArcGIS 10.X.

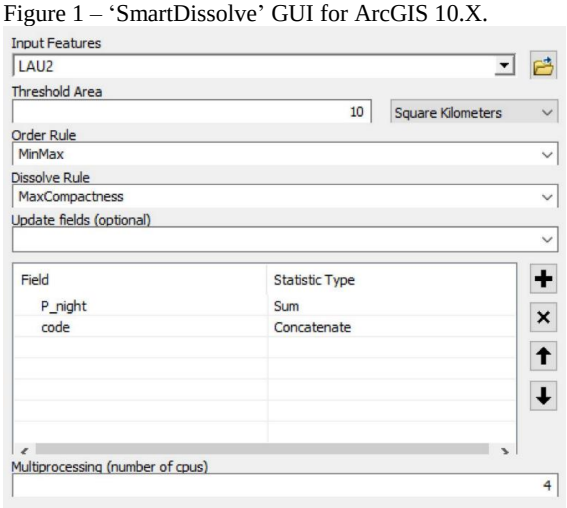
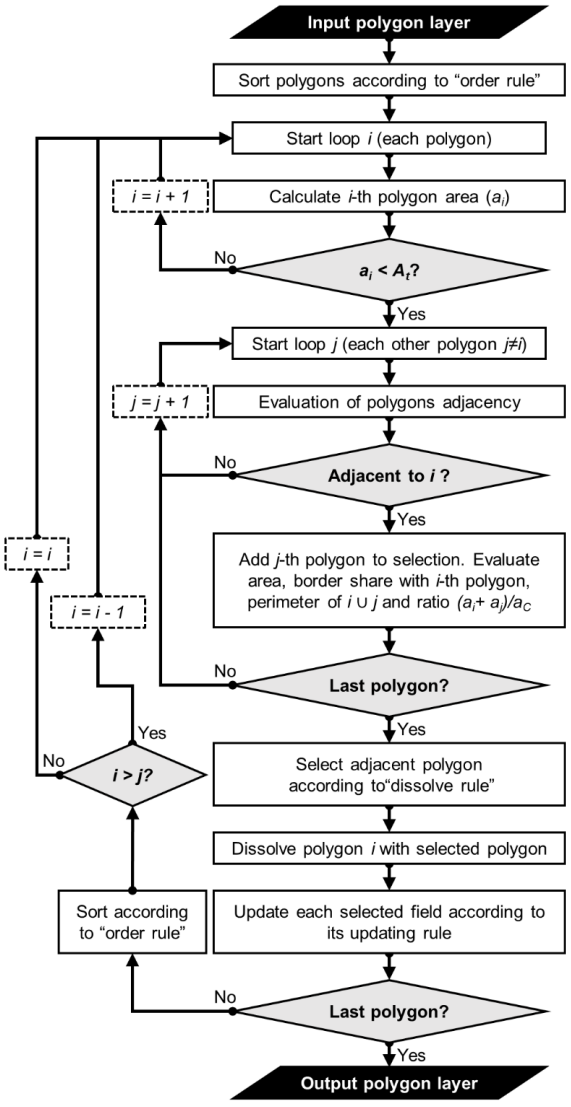


Figure 2: 'SmartDissolve' flowchart. i is the external loop cursor (loop for areal analysis); j is the inner loop cursor (loop for adjacency test); A_t is the user-defined 'threshold area'; a_i is the area of polygon i; a_j is the area of polygon j and a_C is the area of the circle having the same perimeter of $i \cup j$.



planar coordinates, while in the second case the area is estimated using a line integral, based on Green's Theorem (Kaplan, 1991), and the perimeter is calculated as arc-distances (the algorithm uses the MATLAB 'areaint' and 'distance' internal functions), both under the assumption of a spherical Earth.

Cycling the polygons according to the selected 'order rule', the algorithm tests if polygon's area is below the threshold. Once it finds a polygon *i* that satisfies the areal constraint, it looks for all adjacent polygons (i.e. not overlapping polygons that shares segment(s) on the boundary) using 'isadjacent' function (developed within the 'SmartDissolve' package). All polygons with no overlapping bounding box are excluded, while, temporarily breaking up all multi-parts polygons into single-parts, the algorithm checks for border share (i.e. borders that are overlying and not crossing) calculating its planar length (unprojected map) or the length on a sphere (projected map). Given $\boldsymbol{x_c}$ and $\boldsymbol{y_c}$ the coordinates of intersection points between polygon A and polygon B, and $\boldsymbol{l_c}$ the two-column line segments indices (i.e. the *k*-th row indicates which polyline segments

Figure 3: Selected municipalities of Naples' Province (a). Comparison between the original polygon layer (a) and the dissolved layers using different 'dissolve rules' (b-f) with a minimum mapping unit of 6 km^2 and 'MinMax' 'order rule': (b) 'MinArea'; (c) 'MaxArea'; (d) 'MaxBorder'; (e) 'MinPerimeter' and (f) 'MaxCompactness'. Highlighted polygons (grey) are the results of dissolve procedure. The dashed lines represent the original borders of the dissolved polygons.

give rise to the intersection point $x_c(k)$ and $y_c(k)$) as generated by the internal 'polyxpoly' MATLAB function, and $\boldsymbol{p_A}$, $\boldsymbol{p_B}$ two vectors indicating the membership of the intersection points to the list of vertex, respectively of polygon A and polygon B, adjacency between polygons (single-parts) occurs when:

$$\Big((\exists k: l_c(k,1) - l_c(k+1,1) + l_c(k,2) - l_c(k+1,2) = 0) \wedge \big(\exists n: p_A(n)\big) \wedge \big(\exists n: p_B(m)\big)\Big) \vee \Big(\exists h: \Big((|l_c(h,1) - l_c(h+1,1) + l_c(h,2) - l_c(h+1,2)| \leq 1) \wedge \big((l_c(h,1) - l_c(h+1,1) = 0) \vee (l_c(h,2) - l_c(h+1,2) = 0)\big)\Big)\Big) \quad (1)$$

If no adjacent polygons are found (i.e. the polygon is isolated and external, *sensu* Johnston *et al.*, 1999), the algorithm proceeds to the next polygon (i.e. $i = i + 1$). If the algorithm finds adjacent polygons, it selects among them the polygon j to be dissolved with polygon i, according to the defined 'dissolve rule'. Five possible 'dissolve rules' are implemented: (i) 'MinArea'; (ii) 'MaxArea'; (iii) 'MaxBorder'; (iv) 'MinPerimeter'; (v) 'MaxCompactness'. 'MinArea' selects the adjacent polygon with the smallest area; vice versa 'MaxArea' selects the adjacent polygon with the biggest area; 'MaxBorder' selects the adjacent polygon that shares the longest border with polygon *i*; 'MinPerimeter' selects the polygon that gives the minimum total perimeter (i.e. the perimeter of the dissolved polygon) and 'MaxCompactness' selects the polygon that gives the highest isoperimetric quotient (i.e. the ratio of the total area over the area of the circle having the same perimeter, Croft *et al.*, 1991). Figure 3 shows the comparison among the results of the different 'dissolve rules' applied to a set of local municipalities within the province of Naples (Italy). Polygon attributes are updated according to the updating rules, by summing or averaging values, selecting the minimum or the maximum, or concatenating text fields. For all those fields for which no rule has been defined, the new polygon will inherit the polygon i values. Polygon i is then substituted with the newly created one and polygon j is removed from the list. The algorithm continues sorting the polygons according to the 'order rule' by placing the newly created polygon at the right position (i.e. according to the 'order rule'); it leaves i cursor to the same value, if $i < j$, or updates the cursor to the previous value (i.e. $i = i - 1$), if $i > j$. The resulting layer is saved as a shapefile or exported into the GeoDatabase of the input feature layer.

3 Example application

The estimates of population distribution of LandScan product (Dobson *et al.*, 2000) can be a useful for those places where population data are missing or incomplete, or as indicator of activities and population dynamics. We captured the LandScan 'ambient population' at the Naples' (Italy) municipality level by upscaling the LandScan population grid. We used the Italian Statistical Institute (ISTAT) census 2011 (ISTAT, 2011) data to compare the residential population of each municipality in the Naples surroundings with the obtained results.

3.1 Materials and Methods

LandScan 2012 is a global high-resolution population distribution raster dataset. It is the result of a multi-dimensional dasymetric modeling approach that downscaled population count to 30 arc second (~800 m) grid cells and estimated the 'ambient population' in each cell.

We used the 2011 census of Italian resident population in a set of municipalities in the province of Naples, Italy (37 communes), produced by ISTAT and available online, and the respective polygonal layer of municipal boundaries. We 'SmartDissolved' the local municipalities shapefile in order to match a MMU of 9 times the LandScan cell size (~5.76 km^2), in order to avoid abusive upscaling of LandScan population in too small polygons. 'SmartDissolve' was performed by selecting the 'MinMax' ordering rule and the 'MaxCompacteness' dissolve rule to obtain more 'regular' polygons (i.e. more compact according to the isoperimetric quotient). We conducted a 'zonal analysis' to upscale and aggregate LandScan values to the 'SmartDissolved' municipalities layer to compare resident population values with the LandScan estimate of 'ambient population'.

3.2 Results

Among Naples 37 municipalities, eighteen show an area below the defined MMU. The 'SmartDissolve' process produced a layer with 23 polygons summing the resident population of the dissolved polygons (Figure 4a). Using this new polygon layer, LandScan 'ambient population' estimate was upscaled by respecting a threshold of 9 cells per polygon. Figure 4b shows that 'ambient population' varies between 10,000 individuals (Commune of Procida) and 1,000,000 individuals (Commune of Napoli). This upscaling exercise shows that the estimate of the 'ambient population' from LandScan 2012 upscaled at the dissolved municipality level displays little variation (3% on average) from residential population as assessed by ISTAT in census 2011 (Figure 4b).

4 Discussion and conclusions

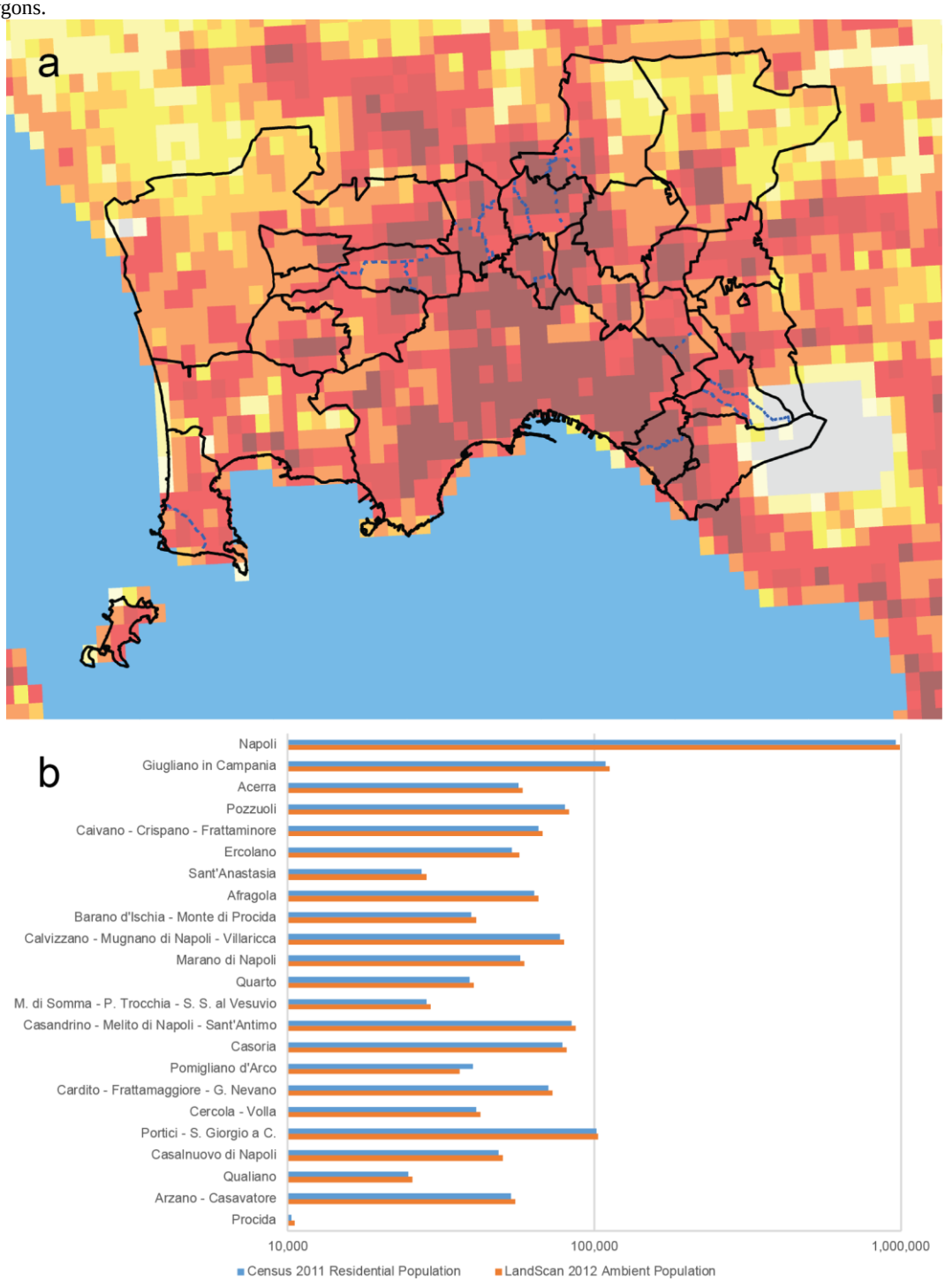
Dissolving polygons storing count data is not a trivial procedure, especially when a user requires more control over the dissolution process. The process is especially challenging when the application requires reaching a specific MMU or multi-part features are involved. We developed an automated and flexible approach and built an open and free tool that meets the needs of spatial analysis and required aggregation of vector features while preserving the native integrity of count values, both in output aggregated feature and in the study area (volume). Contrary to cartographic aggregation procedures, this approach fully preserves the input geometry of the area of interest without generalizing outlines or changing the total surface of analysis. In the case of population counts, the aggregation ensures that population is not transferred to units originally unpopulated.

This approach can mitigate the famous MAUP problem (Openshaw, 1983), by allowing the analysis to be conducted at

a different scale (aggregation) and geometry from the one used to report the phenomenon.

'SmartDissolve' includes different algorithms to meet specific user needs. Its main strengths are the use of area threshold to guide dissolution, dissolution by polygon

Figure 4: LandScan upscaling to Naples municipalities. (a) LandScan 2012 (grid) and SmartDissolved Naples local municipalities (black polygons) using a minimum mapping unit of 9 times the LandScan cell size (~5.25 km^2) and "MaxCompactness" dissolve rule. Blue dotted lines represent the dissolved polygon borders. (b) Comparison of resident population from Census 2011 with ambient population from LandScan 2012 within the obtained 'SmartDissolved' polygons.



adjacency alone, possibility of different dissolve rules, different polygon areal analysis ordering, and correct handling of multi-part polygons, while taking advantage of multiprocessing computing.

The tool automates tasks that previously should be performed manually, such as an area-threshold based approach for dissolving polygons updating selected count fields. In addition, differently from other commercial software, 'SmartDissolve' is an open source toolbox available as a toolbox for ArcGIS (Esri) and for MATLAB (MathWorks), providing a widely commented code, which can be easily customized, as well as further developed for including new dissolve rules and formulas for updating fields. This feature also facilitates the conversion of the code to other programming languages, such as Octave, Python and C++, which may reduce the computational requirements of the implemented MATLAB code. Not fully coding the tool in Python allows 64-bit and multiprocessing computing, as Python for ArcGIS is still a 32-bit version. On the other sided this requires the installation of the MATLAB Compiler Runtime (MCR).

'SmartDissolve' is freely available for non-commercial, research and educational uses through the 'Tools' page on the Global Human Settlement Layer (GHSL) website (http://ghsl.jrc.ec.europa.eu/). We intend to maintain and further develop the 'SmartDissolve' toolbox for responding to users' needs and suggestions. Beside extending some functionalities already implemented, potential updates would include (i) the dissolution rule 'affinity' to aggregate only similar polygons, according to a priority table (defined by the user), and the 'strict affinity', that would avoid aggregating dissimilar polygons, allowing for some polygons not to respect the area threshold imposed. Future releases will be available in the same repository.

Acknowledgements

This work was performed in the frame of the JRC institutional research project *GHS-CORE*. We thank all colleagues that supported this work.

References

Bader, M., Weibel, R., 1997. Detecting and resolving size and proximity conflicts in the generalization of polygonal maps. Proc. 18th Int. Cartogr. Conf. Stockholm, Sweden

Croft, H.T., Falconer, K.J., Guy, R.K., 1991. Unsolved Problems in Geometry. Springer-Verlag, p. 23, New York

Davis, B., 2001. GIS: a visual approach. 2nd ed. OnWord Press, New York.

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., Worley, B.A., 2000. LandScan: A global population database for estimating populations at risk. Photogramm. Eng. Remote Sens. 66 (7), 849-857.

Fraley, C., Raftery, A.E., 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J. 41 (8), 578-588.

ISTAT (Istituto Nazionale di Statistica) (2011) Dati provvisori del 15° Censimento popolazione e abitazioni 2011 (available at http://www.istat.it/it/, last accessed 16/12/2016).

Johnston, M.R., Scott, C.D., Gibb, R.G., 1999. Problems arising from a simple GIS generalization algorithm. Proc. 11th Annu. Colloq. Spat. Inform. Res. Cent. Dunedin, New Zealand.

Kaplan, W., 1991. Green's Theorem. Chapter 5.5 in Advanced Calculus. 4th ed. Addison-Wesley, Reading.

Laurini, R. Servigne, S., Favetta, F., 2016. An Introduction to Geographic Rule Semantics. Proc. 22nd Int. Conf. on Distrib. Multim. Sys., Italy

Martinez-Llario, J. Weber-Jahnke, J.H., Coll, E., 2009. Improving dissolve spatial operations in a simple feature model. Adv. Eng. Softw. 40, 170-175.

Murrey, A.T., Shyy, T.K., 2000. Integrating attribute and space characteristics in choropleth display and spatial data mining. Int. J. Geogr. Inform. Sci. 14, 649-667.

Mu, L., Wang, F., 2008. A scale-space clustering method: mitigating the effect of scale in the analysis of zone-based data. Ann. Ass. Am. Geogr.98 (1), 85-10

Openshaw S. 1984. The Modifiable Areal Unit Problem. Geobooks, Norwich, England.