

Work Session on Geographical Information Systems
(Ottawa, Canada, 5-7 October 1998)

Item 9 of the provisional agenda

**POINT-BASED AND AREA-BASED GEOGRAPHIC ANALYSIS
OF STATISTICAL DATA**

Submitted by Statistical Office of Estonia ¹

CONTRIBUTED PAPER

Revised 29. Sept 1998

¹ Prepared by Teet Jagomägi (Regio Ltd.) and Inge Nael.

I. RE-CALCULATION OF POPULATION COUNTS AFTER CHANGES OF HAMLET BORDERS. DIFFERENCES OF AREA-BASED AND POINT-BASED SOLUTION

I.1 Background

1. The administrative division of Estonia has undergone many changes during the last century. There were 9 major administrative reforms between 1881 to 1997, and Estonia is now planning to decrease the number of local governments by 1/3. In addition to reforms (campaigns), smaller changes routinely take place – renaming, splitting or joining of administrative units.

2. As a rule, statistical data has been (and still is, in most cases) collected by administrative units. If borders of units change, special care has to be taken in order to ensure the compatibility of time-series analysis.

3. Adjusting statistical time-series over changed borders is fairly simple as long as changes are limited to joining adjacent units. In this case data values of old units must be summed up for calculating values for new units.

4. If new administrative borders are formed by splitting former units and new borders follow lower-level units, re-calculation is also easy where statistical data on lower-level units is available. For example, data on splits of a municipality can be found provided the division line follows village borders and there is statistical data available on villages.

5. The situation is much more complex, if new borders do not follow any previously known border or if there is no data available on smaller units. This may be the case if data grouped by parishes needs to be represented by local governments (in Estonia, parish boundaries do not follow administrative boundaries). In this case, only GIS offers tools for re-calculations.

6. Probably the best method is to overlay old borders with new borders, forming “shiver” polygons, to estimate the (population) count for each shiver polygon, and then join shivers together to form new borders. For vector data (polygons are usually represented as vectors) the task is computing-intensive.

7. A more detailed approach could be used if statistical data was available not as a sum for the whole administrative unit, but on points described by the data (population number by buildings, contamination data by wells, etc.). In this case, a simple point-in-polygon algorithm gives us sums for any polygon overlaid with scattered points.

I.2 The situation

8. In Estonia, population counts (and other statistical data) are released by counties (15, average size 2900 sq.km, population 98400), local governments (198 rural, average 215 sq.km, population 2200, and 57 urban municipalities, average size 13 sq.km, population 18400) and hamlets (subdivisions of rural local governments).

9. There have been two major changes in the nomenclature of hamlets during the last 20 years, plus a number of smaller ones. The next change will take place at the end of 1998. After this, all changes will be frozen until the Population and Housing Census 2000.

10. Traditionally, buildings in Estonian rural areas have been addressed by building (farm) name + local government name. Hamlet names were also in use, but they have been considered rather as placenames than areal features with clearly delineated borders. Estonian rural population is fairly scattered – buildings closer than 200 meters from each other are considered neighbors, closely built-up villages are relatively rare. In hilly Southern-Estonian areas, a couple of farmhouses may form a village. There was extensive no-man's-land between hamlets (usually massive forests or bogs) which did not belong to any hamlet.

11. Up to 1978, about 10 000 official placenames (towns, villages and hamlets) were recognized in Estonia, and this nomenclature was also used for Censuses. During the wave of centralization, there was a substantial decrease in the number of hamlets, and as a result ca. 3500 placenames remained. There exists an old 1:50 000 map showing the changes. From a geographical point of view the importance is that no-mans-land no longer existed. In many cases, the local population did not accept the changes, road signs continued to use the old nomenclature, and precise maps preferred to show the old “unofficial” names.

12. The change of nomenclature of placenames was initiated in 1997. The job was governed by the Ministry of Internal Affairs and carried out by local governments. Preliminary Census maps prepared by Statistical Office of Estonia, were used for this task. The number of official placenames grew from ca. 3500 to ca. 4500. Another wave of changes are underway in 1998, to be established legally on 01.01.1999.

13. Although the Estonian Council of Placenames has a certain control over placename creation, local governments took varying approaches to the changes. Some of them did little renaming, others re-designed the whole system. The latter usually created new borders, which do not follow previous ones.

14. Currently, there is no database of population counts available for new hamlets. Statistical time series are continued accordingly to previous nomenclature, but fresh data is created by new nomenclature. There is an urgent need to provide both current and old statistical data grouped by 1) the old nomenclature and 2) the new one. The bridge between 1997 statistical data (which has been increasingly built based on 1989 Census data) and 1998 data has to be created.

I.3 The solution

15. The best method is described in paragraph 7 of this paper. New population counts can be calculated with GIS tools.

16. The data is available for both methods in Estonia: i) 1:50 000 digital map with old borders and population data of old placenames; 1:50 000 digital map with new borders and population data of old placenames; and ii) 1:50 000 digital map of 1989 building centroids, database of 1989 Census data with links to building centroid numbers; 1:50 000 to 1:10 000 digital map of 1997-1998 building centroids, 1:50 000 digital map with new borders.

17. The polygon-based approach overlays two 1:50 000 maps. New borders split old border polygons into “shivers”. The percentage of area of each “shiver” will be calculated.

Example 1: old hamlet X could be divided into three new ones as follows: 10% to new hamlet A, 50% to new hamlet B and 40% to new hamlet C. Old hamlet Y could be divided between new hamlet A 70% and new hamlet C 30%.

Next, the population of each “shiver” polygon is calculated using its relative size as an old hamlet.

Old hamlet X with population 100 inhabitants and old hamlet Y with population 50 inhabitants will give us shivers YA 10, XB 50, XC 40, YA 35, and YC 15.

Finally, the population of new villages can be calculated by grouping shivers into new hamlets.

New hamlet A has population $XA+YA=45$, B has population $XB=50$ and C has population $XC+YC=55$.

18. The result is naturally not precise, as the population distribution in the hamlet is never uniform. The whole territory of local government is divided between hamlets, including forests and bogs. Thus, some “shivers” may represent 50% of the territory of the hamlet, but include 90% of the population. In addition, as hamlet borders are considered more as “placenamesheds” than as legal dividers, they are delineated with relatively low positional accuracy, which introduces errors in area calculation and creates sliver polygons on the overlay.

19. For certain applications, such a polygon-based recalculation is fairly acceptable, as long as it carries a disclaimer regarding possible deviations from reality due to approximation by the algorithm.

20. The point-based algorithm overlays new borders and old building centroids. Old centroids must have an attribute describing in which old hamlet this particular building belonged (if this attribute is not available, it can be created with point-in-polygon algorithm as well). A less precise approach is to count all centroids falling into new borders, sum them by old hamlet number and use the percentage of each sum for calculating the population of new hamlets.

*Example 2: new hamlet A may consist of 15 buildings from old hamlet X and 10 buildings from old hamlet Y; new hamlet B 10 buildings from old hamlet X and 10 buildings from old hamlet Y; and new hamlet C 5 buildings from old hamlet X. Given the number of buildings in old hamlets $X=30$ and $Y=20$, shiver AX represents 50% of hamlet X buildingwise, AY 50% of Y, BX 33% of X, BY 50% of Y and CX 17% of X. If the population of old hamlets $X=100$ and $Y=50$, the population of new hamlets $A=AX*100+AY*50=75$, $B= BX*100+BY*50=58$, and $C=CX*100=17$.*

21. A more precise approach is to use the actual number of inhabitants in each building instead of percentage of buildings in given area.

Example 3: new hamlet A may have 35 inhabitants in buildings from old hamlet X and 20 inhabitants in buildings from old hamlet Y; new hamlet B 15 inhabitants in buildings from old hamlet X and 30 inhabitants in buildings from old hamlet Y; and new hamlet C 50 inhabitants in buildings from old hamlet X. The population of new hamlets $A=AX+BX=55$, $B=BX+BY=45$ and $C=CX=50$ inhabitants.

22. The result is certainly more precise than that obtained with a polygon-based algorithm. As the population distribution in the buildings is never equal the actual number of inhabitants should be used. Unfortunately, this is fairly difficult to achieve, especially in situations where registration of population is not enforced. Locations of buildings can be mapped without intervention into peoples’ private lives; estimation of current population based on the type of building and last count inhabitants in the building works fairly well. The Statistical Office of Estonia has compiled the population counts on an annual basis from the previous Census, but this data is presented

according to placenames rather than buildings. This is also a factor leading to possible deviation from today's real population.

I.4 Test area

23. The theoretical difference in two abovedescribed methods was illustrated in North-West Estonia, Toila Commune (160 sq.km, 2400 inh.), refer to Figure 1. Population of the commune is distributed unevenly, occupying mostly the area between the shore and transit railroad. Southern part is covered by bogs, oil-shale quarries and forests (Figure 2).

24. During 1998 placename reform (refer to paragraph 12) the number of settlements grew from 8 to 11. Majority of changes was made in area of densest population, virtually no-man-area was left intact. New hamlets were formed by changing borders rather than dividing or aggregating old ones (Figure 3).

25. New population counts were calculated using two techniques - point-based (refer to paragraphs 20, 21) and area-based (paragraph 17). The difference is given in the following table:

Old settlement (until 31.12.1997)	Number of inhabitants in 1997	Territory sq. km.
Uikala	9	8.8
Martsa	159	14.2
Toila	753	2.3
Pühajõe	152	25.8
Voka	1176	2.3
Konju	136	78.0
Vaivina	22	20.2
Päite	34	8.2
SUM:	2441	159.8

New settlement (from 01.01.1998)	Number of inhabitants in 1997		Difference between two methods		Territory sq. km.
	Calculated from village territory (i.e. area-based)	Calculated from building centroids (i.e. point-based)	Absolute numbers	Percent	
Uikala	8	12	4	33%	8.2
Martsa	26	21	-5	24%	3.3
Mäetaguse	9	11	2	18%	0.8
Toila alevik	780	495	285	37%	1.8
Altküla	107	61	-46	75%	9.6
Pühajõe	399	104	-295	284%	22.1
Voka	29	84	55	65%	5.0
Voka alevik	1176	1176	n/a	n/a	2.4
Konju	133	133	-1	1%	77.0
Vaivina	26	25	-1	4%	21.3
Päite	32	34	2	6%	7.8
SUM:	2441	2441			159.3
		Average:	69.6	55%	

26. As expected (due to uneven distribution of the population) the difference between two

methods was significant (55%). One of the hamlets has the vast difference, and should be treated exceptional. But even after ignoring the exceptional 284%, the average difference remains **29%**.

I.5 The conclusion

27. The methodology has been worked out for re-calculation of population counts of the new 1998 administrative division of Estonia. A study has been carried out in the Northwest part of the Estonia and procedures have been worked out for calculations for the whole country. The job will be most likely postponed to 1999 to incorporate the latest changes before the Population and Housing Census 2000.

28. The difference between point-based statistics and area-based statistics can be clearly seen from this example. The population behavior, availability of spatial data and presence of GIS knowledge make the point-based approach the most suitable for Statistical Offices of Estonia and Sweden. However, the same approach is not necessarily the best for all countries and all cases.

II. ISARITHMIC MAPS - EMERGING PRESENTATION TECHNIQUE FOR POINT-BASED STATISTICS

II.1 Choropleth maps – the most used thematic maps

29. The most widely used thematic map type is the choropleth map – a map representing certain types of areas (counties, EAs), colored by certain variables. The choropleth map can be combined with pie charts.

30. There are many reasons for using choropleth maps. First, data is usually tabulated by administrative units and it is very natural to present the data on a contour map with borders of these administrative units. There is also a traditional technological reason – choropleth maps were the easiest way to create thematic maps manually; the areas were simply coloured in on a contour (template) map. Subsequently, users of maps became accustomed to choropleth maps (and in today's market-driven world one must always consider customer needs). And, finally, to change anything is more inconvenient than to keep going in the old manner.

31. The biggest problem of the choropleth map is its quality in terms of cartographic representation. The most carefully calculated statistical results can be illustrated on the choropleth map in such a way as to give a totally wrong impression of spatial distribution. Putting data onto the map is made so simple with modern desktop mapping tools that even professionals may slip up.

32. The most widespread misuse of the choropleth map is showing absolute value rather than rates or density. For example, in Nordic countries counties in northern areas are tens of times bigger than those in the south. If number of higher educated people was shown on the choropleth map, then relatively few people would result in vast areas being coloured and would give the impression of high-tech arid areas. The density of higher-educated people would reveal a more realistic picture. The same phenomena is illustrated on the Figure 4.

33. Another problem of the choropleth map is the trend to make “jumps” of value along borders, but illustrated phenomena rarely follow administrative borders. For example, air pollution crosses all borders; percentage of arable land may follow state borders, but usually neglects local government boundaries. Quite often areas available for illustration are too big - they exceed the size of clusters of phenomena several times. This problem is illustrated on the Figure 5.

34. Misuse of choropleth maps is difficult to pinpoint. Users are accustomed to relying on the illustration and an impression is made within the first 1-2 seconds of the glance. Even if an educated and careful mapreader finds probable errors, the first impression biases the final understanding.

35. As a matter of fact, most choropleth maps have something “wrong”. Mr. Robert Cromley, chief editor of Journal of Cartography and GIS, even has a special term - a choroplethical map - choropleth map, where everything has gone wrong.

II.2 Isarithmic maps

36. Isarithmic maps were originally developed to describe three-dimensional features (such as surface of landscape or height of water table) on a planar map. In the course of seeking for impressive communication means, many phenomena can be imagined to be a three-dimensional surface. For example, population density can be quite easily illustrated as “relief”.

37. Isarithmic maps have two subtypes:

- i) Isometric maps – depicting features which are located at a certain point, e.g. height;
- ii) Isoplethic maps – depicting features spread over a certain area, e.g. density.

38. Relief maps are typical isometric maps – they represent the height of a particular point in the mapped area. Use of isometric maps for statistical analysis requires an explanation for why such a presentation technique may be used. There is no meaning for “population density (count) at this point”. But by defining the “height” of each point by “number of people residing within 100 meters from this point”, population density can be shown on an isarithmic map (there are certain nuances for defining isometric or isoplethic maps; this difference is not crucial for the current paper).

39. Isarithmic maps can be represented with various options:

- i) isolines (contour lines). Isolines do not overload the map image, but they are relatively difficult to follow if the feature has many local minimum and maximum extremes (Figure 6).
- ii) isosurfaces (color-filled contour lines). In cases where the number isosurfaces is big and contours are omitted, continuous-tone graded map is created (Figure 7).
- iii) perspective views of DEM surfaces. Three-dimensional views are usually visually attractive (Figure 8). For experienced users, they may be very useful analysis tools, especially with modern software, which enables real-time DEM rendering and navigation.

40. Data shown on isarithmic maps is usually scattered, or has a sparse regular pattern. High quality maps require much more valuable; they can be computed with interpolation techniques. If choropleth maps are too sensitive to classification method, isarithmic maps are very sensitive to selection of interpolation method and its parameters. There are 3-4 most widely known interpolation methods, each of them with unique characteristics. The most important feature of the interpolation method is, whether it creates under- and overshoots during the interpolation.

Example: We might have chosen to use Inverse Distance Weighted Interpolation method. One of the issues is choosing correct Search Radius and Exponent, among other

parameters. Exponent is the variable that defines the exponential rate of decay of influence of neighbouring points the farther they lie from the grid node. Increasing the exponent will decrease the relative influence of more distant neighbours.

The question is, does the crime site make particular corner of the street or block or the whole city region criminal? Or does a house full of kids make this building, or street, or neighbourhood youthful?

41. The major disadvantage of using isarithmic maps is the relative difficulty of estimation of the value of features shown on the map at any given location (but it can be made user-friendlier in certain extent, refer to Figure 9). In addition, desktop mapping software does not support isarithmic maps or requires special add-ons.

42. Isarithmic maps are not entirely an alternative to choropleth maps, but are a valuable replacement for illustrating complex spatial distribution, especially if:

i) there is a threat that the distribution of natural phenomena does not follow given geographies (e.g. administrative boundaries), refer to Figure 10;

ii) if there is a need to illustrate many features, which cannot be combined on choropleth and dot density maps (Figure 11);

iii) thematic map is created as intermediate step in the analysis (figure 12).

II.3. Grid maps – on a way from choropleth map to isarithmic map

43. A grid map (often used in illustrating statistical features) can be considered as a choropleth map or as an isarithmic map. A unique feature of the grid map is **the same size of all areas on the map**. This automatically removes the possibility of making the most frequent mistake of choropleth maps (showing absolute values on the map), as the absolute value grid maps and density grid maps give exactly the same distribution. At the same time, the grid map is a reduced-resolution color-graded isarithmic map (Figure 13).

44. As grid maps use much smaller areas than “ordinary” choropleth maps, it is easier to follow the distribution of features, which do not necessarily honour administrative boundaries (water quality, agricultural yield, education).

45. Grid maps help to increase the resolution of disseminated statistical maps while preserving privacy. If data is released by many geographies, there is a threat that polygon overlay discloses information on too small areas. Grids with fixed point of origin reduce the risk significantly.

46. Statistics Finland and Sweden extensively use grid maps; the population density map presented at UN/ECE Work Session on Geographical Information Systems (Brighton, UK, 22 - 25 September 1997) was an excellent product. On the map, the distribution of population along rivers in distant northern areas was very clearly shown; this could have been lost on the traditional choropleth map. In addition, the administrative units in Finland and Sweden differ in size, which could have distorted the cross-country choropleth map.

II.4 Point-based data and isarithmic maps

47. Isarithmic maps should be generated from data, the spatial extent of which is much smaller than the grid cell. Point-based statistical data is particularly suitable for such an application, because points are also smaller than any grid cell size.

48. Certainly, grids can be generated from enumerator areas (or any type of choropleth

representation), but in this case the grid cell has to be fairly large, otherwise grid cells involve too big amount of interpolation.

II.5 Animation

49. The most suitable structure for storing base data for isarithmic maps is raster. Contour maps or 3D perspective views are already vector forms, but the data engine in the background is usually a regular grid. Grid-like structure is very convenient for creating animation, which is one of the most interesting areas of study in computer cartography. All other kinds of maps can be done with traditional means as well; the animation is a new dimension computers may introduce into cartography. For statisticians, animated maps are the perfect tool for illustrating time-series data.