

CONFERENCE OF EUROPEAN STATISTICIANS

**Joint UNECE/EUROSTAT Work Session on Methodological Issues Involving the Integration of
Statistics and Geography**

(Tallinn, Estonia, 25-28 September 2001)

Topic (i): New opportunities created by cooperation and partnership

**IN SEARCH OF A COMMON GEOGRAPHICAL BASE TO COMPARE STATISTICS ACROSS
THE EU: THE TANDEM GIS PROJECT**

Submitted by The Tandem Consortium¹

Invited paper

ABSTRACT

This paper summarises the objectives, results and ideas produced by the Tandem GIS consortium within the framework of the Tandem GIS project. The aim of the project is to explore the result of a combined Grid- and a Region-based approach needed to tackle the limitations inherent in the NUTS system. Two methodologies are presented, discussed and compared. A case study applied to the delimitation of rural/urban areas is presented, and the feasibility for wider use over Europe is discussed. This project is not concerned with any changes in the NUTS or other hierarchies of administrative areas generally used for decision support.

The "Tandem consortium", which consists of representatives from GIS groups at Statistics Finland, Statistics Sweden and the Office of National Statistics (UK), is actively pursuing this goal in a project funded by a grant from the Commission (85%) and the participating NSI's (15%). Final results and recommendations will be presented at the annual "Workshop for Geographical Information and Statistics" (to be organised by Eurostat / Gisco in October 2001).

The purpose of this paper is to provide the attendees of the ECE conference and other interested parties with a preliminary report. The definite results will be available for publication just in time for the conference in Tallinn in September 2001.

I. PRESENTATION OF THE TANDEM PROJECT

1. Eurostat/GISCO has for some time argued the need for "Basic statistical area" that could be used to increase the comparability of European "area-based" statistics. On the other hand, developments both in the field of remote sensing and efforts on the part of many NSI's to collect information with point-based strategies have led to a growing need to agree on a common system of grids and grid methods to increase the comparability of both types of spatial statistics.²

¹ Prepared by Philippe Gublin, Office of National Statistics, United Kingdom; Marja Tammilehto-Luode, Statistics Finland; and Lars H. Backer for Statistics Sweden.

² See (Rase 2000) *Technical Specifications; Improving comparability of statistical data across EU Member States*.

2. In response to papers submitted by United Kingdom³, Sweden and Finland⁴ at the meeting of the Working Party on “Geographical Information Systems for Statistics” held in Luxembourg on 20 and 21 October 1999, it was suggested that a combined Grid- and a Region-based approach would be needed to tackle the limitations inherent in the NUTS system. As a result of these and other developments, the Tandem consortium, consisting of GIS groups from the Office of National Statistics (UK), Statistics Finland and Statistics Sweden, was invited to apply for a commission grant to study these questions further.

II. INTRODUCTION AND PROBLEM STATEMENT

3. According to an analysis of “User needs” it seems that there are two main uses of statistics and statistical systems:

- The use of statistics aggregated onto hierarchies of administrative areas (such as the NUTS system) to be used for “accountant” type decision support;
- The use of statistics aggregated on to systems of small statistical areas (if necessary for disclosure reasons) for the analysis of spatial distribution patterns etc.

This project is not concerned with any changes in the NUTS or other hierarchies of administrative areas. It is concerned solely with the problem of studying the feasibility of developing a system of small statistical areas across the EU.⁵

4. The core argument here is, first of all, that there is a need for two systems of small statistical areas, one consisting of regular and the other consisting of irregular tessellations. Both systems should, when complete

- include comparable area units;
- include the smallest possible area units;
- eventually cover the whole EU territory;
- have a reasonable set of statistical variables connected to them.

5. In this paper we are looking for two systems of small statistical regions that can be used in parallel (in combination). Both of these systems have their own special problems and advantages. For the regular tessellations we face the problem of the demand for equal size area units. The problem here lies in the differences of the available input data, the definition of optimum size of the grids and the projection system used for harmonising geo-statistical analysis on grids.

6. In the case of the irregular tessellations, the demand for a system of regions with roughly the same “population” poses some major problems, since the population varies according to the variables studied. Thus a set of regions with an equal distributions of “natural” persons will show a very unequal distribution of “juridical” persons. (The solution to this problem is probably to make these regions as small as possible so that they may later be clustered into new systems of statistical regions depending on the variables studied.)

III. METHODS

III.1 Preliminary questions

7. The ambition of the project was to provide answers to these preliminary questions:
Technical feasibility: is it possible to produce a twin system of statistical areas that would make it possible to compare analysis of spatial statistics across the EU?

³ See (Wagget 1999, (20-21 October)) *Towards improved statistical comparability across Member states – A better geographical framework.*

⁴ See (Backer, Tammilehto-Luode and Rogstad 1999) *The use of grids to improve the comparability of statistical data.*

⁵ The need for such a system as this has become apparent in the current need to monitor and analyse information in connection with the “foot and mouth” disease in Europe.

Feasibility for Europe: is it possible to compile a system of statistical areas from a variety of small-area systems? Is it possible to build a harmonised system of statistics on areas that are smaller than the smallest administrative unit? Is this system desirable for Europe?

8. This paper will present a methodology to create homogeneous building blocks and propose a case study focused on the delimitation of urban/rural areas. Test areas, representative of the whole of the EU have been selected from Finland and United Kingdom(see “Practical aspects and Case Study” section).

III.2 Theoretical aspects

9. The two methodologies described below correspond to the recent developments carried out in Finland (grid square approach) and United Kingdom (construction of ‘blobs’).

III.2.1 Grid squares – regular areas

10. The theoretical assessment of the production of grid-based data includes the choice of applications to be used for spatial transformation of different types of source data, discussions about a possible optimisation of the size of the grid cells and about the choice of a proper projection system.

11. If the system of grids needs to cover all Europe, it is obvious that the source data for grids will differ from country to country. The methods which should be used for spatial transformation depend primarily on:

- characteristics of the source data (polygon, line or point geo references);
- the nature of output data required (raster, vector, ascii);
- spatial character of phenomena being analysed (gradual or disjunct change over space - intensive, extensive or categorical variables) (Eurostat 1999).

12. The most precise results will be achieved if individual statistical units (input areas/points) are allocated to grid cells. This is the case in Nordic countries, where established links from statistical units to geocoded centroids of buildings (Finland), centroids of real estate (Sweden) or standardised addresses (Norway) are available. The gridded data are simply aggregated micro data performed using point in polygon method, or even more simply obtained by calculating frequencies within grid cells using information of map coordinates of each unit point (Tammilehto-Luode, Backer 1999). The practical way to compile statistics by grids from georeferenced, point-based micro data is described in the final report of the Tandem project.

13. If the input data are already aggregated and geo referenced by points or polygons (or lines) a conversion of input data into grid data is an estimation process. There are plenty of different methods which can be applied. Many of them are described by Flowerdew et al. (Eurostat 1997), Eurostat (1999) and Briggs (2000). Some of those methods can be implemented using standard GIS overlay functions. Because of drawbacks in standard methods, Eurostat has also developed special programs to transform data for zones to a coverage of grids squares of desired size (Eurostat 1997). Three candidate methods were chosen for further testing in this study (Appendix 1). The test results are used for choosing the best methods for constructing a prototype of grid-based statistics.

14. The grid cells can be encoded in raster or vector format. Often gridded data are required to be in raster-format for GIS–tools, which use this data format for different kinds of spatial analysis (for example, ArcView Spatial analyst or ArcInfo GRID). It is fairly easy to convert data from vector to raster and vice versa. The main advantage of using vector based grid cells compared to raster based cells is the possibility to assign multiple attributes to vector based grid cells. The most important disadvantage of the vector based grid cells is the lack of any compression techniques.

15. Aggregating point-based information to grid cells is a similar procedure to that of aggregating points to any kind of polygons. However, as regular tessellation, the grids include information that simplifies the process considerably. Producing numeric data by grid cell only requires an addition

procedure. To produce categorical data, more complicated procedures may be needed. If a cell contains points representing different categories there can be several solutions to the problem: one can choose a certain dominating value, e.g. a major value, an area-weighted value, and so on.

16. If there is a need to convert polygon-based information to grid-based information, it is important to differentiate between intensive and extensive variables (Goodchild, Lam 1980). An intensive variable is expected to have the same value in each part of a polygon as it has in the whole polygon. Ratios, like population density, and categorical data, like land use, are examples of intensive variables. A value is considered to be extensive if a larger region is expected to be the sum of the values for its component parts. Population, for example, is an extensive variable.

17. Intensive variables are easier to handle, because the original polygons are derived from the spatial variation itself (NERI Technical Report 2001). The classic method for converting extensive data is based on the combining of source zone values, weighted according to the area of the target zone they make up. This area-weighted overlay method assumes that the variable of interest is evenly distributed within the source polygons. This seems to be unlikely in most cases. However, if no information is available about the distribution of values within the source polygons, this procedure may give good results.

18. It is obvious that the optimum size of the grid cells in a grid-based statistical system is dependent on the quality of the input data and on the scope of the analyses in which the data are used. However, the optimum size of the grid cell is often also its minimum size, which is determined not only by quality but also by confidentiality reasons. Flowerdew et al. (Eurostat 1997) suggests that, if the input data are area-based data, the ideal cell size should be equivalent to the larger regions.

19. The size of the cell should also be considered from the point of view of processing speed and disk space. If disk space is unlimited and processing speed irrelevant, the analysis should determine the cell size. For example, it is not practical to have a 100m resolution for data that are to be used to examine the population structure of the whole of Europe.

20. The production of grid-based data is also sensitive to map coordinate references, their location and projection system. If the origin is different to same size grid-nets, the configuration of the data in grid cells are different (Grasland 2000). The construction, visualisation and analysis of grid-based data require a rectangular coordinate system that does not distort regular polygons, or distorts them as little as possible.

21. The projection system has to be a map projection with distance units as meters or kilometres. The idea of grid squares as effective analytical tools is based on usage of distance units within data. A common reference system for geographic information is needed to ensure that the data are compatible across Europe. The reference system has to be universal, and mathematical transformations from national systems need to be available.

22. A hypothetical projection system for the whole of Europe is the Universal Transverse Mercator (UTM) system, which is widely used in grid-based systems for nature and land use information (EEA 2001). UTM coordinates define two-dimensional, horizontal positions. The UTM system includes sixty zones, which keeps the distortions reasonably small even on medium- and large-scale maps. The UTM projection is used in Nordic studies with grid-based data (Tammilehto-Luode et al. 2000).

III.2.2 Blobs –irregular areas

23. The theoretical assessment of the region-based approach consists of the application of ‘automated zoning procedures’ (AZP) thanks to Openshaw (1977) who originally developed the basic algorithm. This AZP algorithm is used to improve the consistency of displaying data at different geographical levels. The AZP algorithm can be defined as a local boundary optimiser

24. The key points in automating design:

- **The basic algorithm** was developed by Openshaw (1977) to explore modifiable areal unit effects. This AZP algorithm is used in order to improve the consistency of displaying data at geographical levels.
- **The AZP (automated Design Procedure) algorithm:** a local boundary optimiser
- **Optimising an objective function:** the aim is to optimise a function of the data generated by a zoning system defining an aggregation of N original zones into M regions or output zones or building blocks ($M < N$).

The optimisation is driven by three principles:

- given a study of N zones to start with, a set of M larger but fewer “pseudo-regions” could be derived for which an objective function is optimised;
- each new pseudo-region should be internally connected;
- the resultant set of pseudo-regions (regionalisation) provides a new sets of output areas (OA) at a higher level of geographical definition.

The objective function is a mathematical expression $F(Z)$ where Z is not a simple set of linear or non linear parameters but defines an aggregation of N initial zones into M output zones. A value of Z is then associated with each possible output geography. A graph of this function can then be drawn giving the values of F for the successive iterations of the algorithm. Minimising (maximising) $F(Z)$ consists of finding the overall minimum (maximum) value of $F(Z)$. This absolute optimum defines the “best output” geography.

- **Optimisation under constraints:** For the purpose of the study, design criteria are preferred to others by the user. Those conditions guided by the study and mathematically defined by the user are called “constraints”. There are implicit constraints on Z so that each of the original N zones has to be assigned to exactly one output zone and all the members of the same output zone have to be connected so that when the internal boundaries are dissolved they form a single polygon. Numerous constraint terms are then defined as mathematical expressions and added to the terms of the already defined objective function in order to form a bigger objective function to be optimised. Wise et al (1997) suggest that the zones should also be of equal size. Martin (1997) uses population size, shape, and homogeneity, and defined the objective function as the weighted sum of up to three different design functions.

The function to minimise is :

$F(Z)$ = Equal population zoning constraint function + shape design function + homogeneity design Function A description of these functions is given in Appendix 2.

25. The principal criticisms can be summarised as follows (Openshaw and al.).

- a) There is no assurance that either any or all of the design functions will meet whatever minimum or maximum limits are placed upon them;
- b) The quality of the solutions depends on the respective weighting given to the design functions and how each of them are scaled or standardised;
- c) Using multiple objective functions in zone design increases the difficulty of finding a good optimum. Alternative objective functions have to be used in order to find a stable solution. This process might be complex and involve interactive trade-offs between different design functions.

IV. A CASE STUDY

IV.1 Test areas and data sets

26. A set of test areas was selected. For the test a realistic set of test areas, described by adequate statistics, is needed. It seemed clear within the project framework that the test data should be representative of the whole of EU. For practical reasons we have decided to use data from two different countries constructed with two different types of source data.

IV.1.1 The Helsinki Region (Finland)

27. The Finnish data contain population data of the year 1998 covering 56 municipalities (NUTS5 areas) in the surroundings of the city of Helsinki. The data are compiled from point-based information of georeferenced buildings.

28. Two levels of resolution are considered (one for each data set):

- Population at post-code level (168 postal code areas);
- Population on 1km x 1km grids (13 003 grid cells).

IV.1.2. Cardiff region (United Kingdom)

29. The British data set covers the county of South Glamorgan (NUTS3 containing the City of Cardiff and the Vale of South Glamorgan) and contains the 91 census population counts at Enumeration District (ED) level (818 Eds).

30. **Remarks**

- a) Each data set was converted into UTM coordinate system. (30 for Cardiff, 35 for Helsinki);
- b) Tests for transformation of point or polygon based data to grids and tests for displaying and analysis of grids are performed by standard and applied tools of ArcView and ArcInfo;
- c) The AZP algorithm was performed using Fortran and Visual Basic routines developed by David Martin. It required the creation of contiguity files by ArcInfo. No homogeneity function will be used in this case study.

IV.2 Grid squares – regular areas

31. Concerning the critical points of the grid-based approach, four different kinds of tests were chosen to be made with empirical data:

- tests of the candidate algorithms for converting polygon-based data to grids;
- tests of an optimum size of the grid cell and an optimum coordinate system;
- tests of usability of the prototype of grid-based statistics for visualising population structure;
- tests of usability of the prototype of grid-based statistics for delineating urban areas.

A case study for delineating urban areas with the prototypes of grid-based statistics is thus a final test. The purpose is to compare results not only with current practices (a delineation with NUTS 5-based data) but with results of parallel study with optimised irregular tessellation.

IV.2.1 Building up a system of grids

32. The point-based data (the Finnish microdata) were converted to grid-based statistics with methods which have been developed in Nordic countries (Tammilehto-Luode, Backer 1999, Tammilehto-Luode et al. 2000). The Finnish gridded data of the test area of the Helsinki region represented a prototype of grid-based statistics compiled from point-based information. However, methods to convert polygon-based data to gridded data were tested before final selection of the method for the English prototype of gridded data.

33. The quality of the results was studied by comparisons of the Finnish data; weighing real grids (compiled from point-based, accurately georeferenced data) against the estimated grids (converted from polygon-based postal code area data) produced using three different algorithms. The test conversions were made using two different kinds of variables, population count (extensive) and population density (intensive). Two different sizes of grid cells were tested, 1km x 1km and 10km x 10km. The statistics on the “estimated grids” from each conversion were compared against statistics on the “real grids” (Appendix 3).

34. According to the tests, the RegionGrid algorithm (extensive) gives the best results in converting population counts from polygon-based to grid-based statistics (Table 1). It maintains the original statistical data structure by producing fairly similar statistics to those on the “real grids”. Large (10km x 10km) grids seem to preserve the structure even better than small (1km x 1km) ones. This supports the earlier comment that the grid size should be chosen to match large regions. About 6% of the postal code areas are even larger than 10km x 10km. Both the PolyGrid and the PointGrid algorithm overestimate the data for small grids and underestimate them for large grids (Table 1).

35. Using the other variable, i.e. population density, the variation in the results is smaller (Table 2). All conversions seem to overestimate the data for small grids and underestimate them for large grids. However, the RegionGrid algorithm retains the statistical structure of the data. The PolyGrid algorithm converts data to small grids as well as the RegionGrid algorithm does, but it badly underestimates data to large grids. The PolyGrid algorithm takes the data of the polygon with the greatest area in the cell concerned. At least in Finland, large areas do not directly indicate highest population densities. The PolyGrid and PointGrid algorithms can be used with weighted values, which could have made the results better. No weighted values were used in this study.

36. With the PointGrid algorithm, the postal code area data were assigned to the nodal points of the postal code areas. When no points fall within a cell, it is assigned the code NODATA. This is why the “counts” of PointGrid statistics differ from the others.

IV.2.2 Delineation of urban areas

37. With reference to Eurostat studies, the hypothesis of this study was that the new building blocks, i.e. grid squares created with more detailed source data, are better for this task than NUTS5 areas. The definition of an urban area in the application is: **Urban area is a densely-populated area**, a contiguous set of local areas, each of which has a density superior to 500 inhabitants per square kilometre, where the total population for the set is at least 50,000 inhabitants (Eurostat 1998).

38. The method of delineation by grid-based statistics is an application accomplished with the ArcInfo 8.2 software (ESRI 2000). The basic idea is to create zones from selected grid-polygons with population densities equal to, or above, 500/km² and then join the population data (summarised by zone) to these zones. The zones whose populations equal or exceed 50,000 are selected.

39. The visualisations of the results are contained in Appendices 4: Figures 1-4. Figure 1 includes the “points of departure” of both the test areas and of the delineation of urban areas using NUTS5-based statistics. By comparing the test results (Figures 2-4) to these reference maps, one can detect a big difference. Grid-based applications should be more realistic, because they do not support ecological fallacy, which may be a problem with large NUTS5 areas (Martin 2000).

40. The basic statistics in the table attached (Appendix 3: Table 3) show that the urban area defined using the Finnish test data relating to 1km x 1km grids is 45% smaller than that defined using the data on NUTS5 areas. On the other hand, the urban area of the UK test area, South Glamorgan, is a little bit larger by 1km x 1km building blocks than that defined using the data on NUTS5 areas. This shows a big difference in areas of NUTS5 regions in the two different countries. The South Glamorgan test area is somewhat small to be using a grid size of 1 km x 1km to describe continuous phenomena – the whole

region of Wales seems to be better suited to this size of grid cells. This size seems appropriate for the Helsinki region as well.

41. There are big differences in population figures, too. In Finland, the population in urban areas is 10% smaller when defined by 1km x 1km building blocks than when defined by NUTS 5 building blocks. On the other hand in the UK test area population is greater when defined by 1km x 1km building blocks than when defined by NUTS 5 building blocks.

42. Compared to the urban areas defined by NUTS5 building blocks, grid-based delineation gives much more comparable results. There are big differences in surface areas, in particular, which in turn may mislead interpretation of population densities in NUTS5 areas. When population density is the critical component in the defining of urban areas, the building blocks should be standardised by area. Grid-based statistics provide a good alternative for this purpose.

IV.3 Blobs – irregular areas: building up a system of blobs: running the AZP algorithm

43. In area-based systems, the geo-reference relates to some small census area, such as collection area, output area or census tract. In the United Kingdom the smallest units for which Census data are published are the Enumeration Districts (EDs) in England, Wales and Northern Ireland, and Output Areas (OAs) in Scotland.

44. The problematic considered here is slightly different from the grid-approach. The result expected is the production of a very low level of data resolution, a very flexible geography able to provide more homogeneous aggregated data at a higher level.

45. The practical assessment consists of the application of the AZP algorithm written by S. Openshaw (1977) and developed in a Fortran / Visual Basic environment by D. Martin. Previous tests have been carried out by Martin (1998).

46. The work carried out within the Tandem GIS project was performed within the Office for National Statistics on Census 1991 population data at Enumeration District level for the County of South Glamorgan, covering the city of Cardiff and on population data at postcode level for the area of Helsinki. The algorithm will also be tested on gridded data covering the region of Helsinki.

IV.3.1 AZP on ED data: South Glamorgan area

IV.3.1.1 Creation of a new boundary set of building blocks

47. The data used to perform this case study are 1991 British census population data at ED level for the county (NUTS3) of South Glamorgan (Wales). The set of ED boundary is presented in Figure 1 (see Appendix 5) and covers the 2 districts of the city of Cardiff and the Vale of South Glamorgan. Initially, attention is focused on population counts for the 818 EDs to illustrate the AZP outputs. The objective is to create a new set of optimised boundaries of building blocks. The second part of the study is to define rural/urban zones using criteria defined previously. This case is relatively simple, only 11 EDs have a total population equal to zero. We also precise that in order to simplify the study no physical features (e.g. rivers, islands) have been taken into account. There is no easy solution to solve this question (see Openshaw and Rao, 1994).

48. In order to perform the rezoning of the County of Glamorgan, we have used the Visual Basic program, AZM, developed by David Martin. This exercise has used the standard AZP algorithm although the use of the simulated annealing algorithm is also possible. The result is shown in Appendix 5 – Fig. 2. The results, basics statistics and comparison with the input data set are presented in Table 1 (below). In order to achieve the rezoning, the program requires 2 input parameters: the threshold population value defining a minimum population value for the building blocks and the target population value. In these

examples, we have entered a minimum value (threshold) close to the median to ensure a substantial ratio of aggregated areas and a target close to the mean.

Input Areas						
Number of EDs	Minimum Population / non zero	Maximum Population	Median/ Average population	Minimum Area (m ²)	Maximum Area (m ²)	Average Area (m ²)
818	0 / 64	1030	492 / 477.384	10,285.31	13,575,610	510,904.34
Set of optimised Building Blocks:						
Population Threshold: 500; target population: 500; shape constraint: on; homogeneity constraint: off.						
Number of Building blocks	Minimum Population / non zeros	Maximum Population	Median / Average population	Minimum Area (m ²)	Maximum Area (m ²)	Average Area (m ²)
527	500 / 500	1909	648 / 740	34,313.47	19,724,55	792,580

Table 1: Comparison of basic statistics for population and area sizes for the Input and AZP-output set of areas

IV.3.1.2 Delimitation of rural/urban areas

49. The use of the AZP algorithm seemed to us an interesting approach able to provide an alternative way to the grid-based approach for the definition of rural/urban boundaries.

Since the density population is the criterion used to define the rural/urban zones, the density has been calculated as the ratio population/area. Maps of population densities are presented on the Figure 1 and 2. They were produced using the GIS software MapInfo. The map on Appendix 5- Figure 1 shows the input areas (Eds for Cardiff and the Vale of South Glamorgan) three different colours are used to represent 3 classes of populated areas:

- Densely-populated area: population of 500 inhabitants per kilometre squared;
- Intermediate area: population between 100 and 500 inhabitants per kilometre squared;
- Thinly-populated areas: population of less than 100 inhabitants per kilometre squared.

50. Only equal population zoning and a shape constraints (perimeter squared/area) were used.

51. The interesting points are:

- a) the three delimitation zones look fairly more distinguishable on the 'optimised' map (Figure 2), as if the data were smoothed. The transition from densely populated zones to thinly populated zones seems better designed;
- b) within each of the three group of zones the areas are of nearly equal size, 'they could more readily be ranked or subjected to other forms of spatial analysis and modelling' (as suggested by Openshaw and Rao, 1994).
- c) The amount of population within the urban areas is 2% smaller after rezoning.

52. In this simple example an equal weight was given to both equal population constraint and the shape constraint. It also has to be mentioned that no other source of heterogeneity was considered such as geography and/or social class. These two issues need further investigation.

Remark

53. The criterion used for the delimitation of urban/rural areas is the density of population (count/area). Using an equal density zoning constraint added to the equal population zoning constraint seems relevant but need more investigation.

54. A final map presenting the delineation between urban zones (area containing population of more than 50,000 people) is presented on Figure 3.

IV.3.2. AZP on the Helsinki postcode data

IV.3.2.1. Creation of a new boundary set

55. The data used in that case study are Census population data collected at post-code level for the whole region of Helsinki (44 municipalities, 437 post-codes). Unlike the South Glamorgan area, the particularity of the Helsinki area (see Figure 4):

- The existence of several Islands which cause problems in the aggregation process. Since islands are not contiguous to any of the other post-codes it has been decided to leave them and not to integrate them in the aggregation process;
- The existence of very large empty zones or zones with a very low density (see table below):

Remark

56. The criterion used for the delimitation of urban/rural areas is the density of population (count/area). Using an equal density zoning constraint added to the equal population zoning constraint seems relevant but need more investigation.

57. A final map presenting the delineation between urban zones (area containing population of more than 50,000 people) is presented on Figure 3.

IV.3.2. AZP on the Helsinki postcode data

IV.3.2.1. Creation of a new boundary set

58. The data used in that case study are Census population data collected at post-code level for the whole region of Helsinki (44 municipalities, 437 post-codes). Unlike the South Glamorgan area, the particularity of the Helsinki area (see Figure 4):

- The existence of several Islands which cause problems in the aggregation process. Since islands are not contiguous to any of the other post-codes it has been decided to leave them and not to integrate them in the aggregation process;
- The existence of very large empty zones or zones with a very low density (see table below):

“Empty” Zones	
Density (inhabitants / km²)	Number of post-codes
0	6
< 5	27
< 10	74
< 50	180
< 100	207

Table 2: Distribution of “empty “ zones

59. The case presented here take the total population as a target variable, results are presented below. Results are summarised on the table 3.

Input Areas						
Number of postcodes	Minimum Population / non zero	Maximum Population	Median / Average population	Minimum Area (m ²)	Maximum Area (m ²)	Average Area (m ²)
436	0 / 3	24,334	1759.5 / 3,526	60,888.88	337,757,389.64	28,005,140.36
Set of optimised Building Blocks:						
Population Threshold 1500; target population: 4000; shape constraint: on; homogeneity constraint: off.						
Number of building blocks	Minimum Population / non zero	Maximum Population	Median / Average population	Minimum Area (m ²)	Maximum Area (m ²)	Average Area (m ²)
266	0 / 1551	24,829	4302 / 5,780.29	120,930.86	665,677,122.66	45,903,162.4

Table 3: Comparison of basic statistics for population and area sizes for the Input and AZP-output set of areas

IV.3.2.2. Delimitation of urban/rural areas

60. Here again only equal population zoning and a shape constraints (perimeter squared/area) were considered and the threshold was chosen close to the median.

Looking at the maps, it seems that the main difference between the input system and the building blocks system concerns the rural zones. The intermediate (transition) zones seem also slightly more defined and distinguishable.

61. A final map presenting the delineation between urban zones (area containing population of more than 50,000 people) is presented on Appendix 5. – Fig. 5.

Remark:

62. The population within urban areas is 12% bigger after rezoning.

IV.3.3. AZP on gridded data: Helsinki Region, 1km x 1km grids

IV.3.3.1. Creation of a new boundary set

63. Applying the AZP algorithm on gridded data involved adaptations in the Visual Basic program and to reconsider the data set on which the program should be applied on. The data provided for the study and analysed in parallel within the grid-based approach cover the region of Helsinki. They consist of population counts on 1 km x 1 km grids. Due to the amount of grid cells (18,625) and the limitation of the computer memory involved in the analyses it was necessary to reduce the amount of data. Only a 'chunk' of the Helsinki grid data were used for the study covering the city of Helsinki plus the fringe. A box of 1074 grid cells was defined. The study area is shown on two maps presenting the gridded area within the Helsinki region studied in previous analyses.

64. Since the grid size is 1 km x 1 km the density of each grid cell is equal to the population. The cells being all contiguous the problem caused by the islands in the previous analyses do not persist anymore. As for the previous analyses a case study considering a target population of 4000 people per building blocks with a threshold of 1500 inhabitants was performed. The result is summarise in table below:

Input Areas						
Number of grids	Minimum Population / non zero	Maximum Population	Median / Average population	Minimum Area (m ²)	Maximum Area (m ²)	Average Area (m ²)
1074	0 / 1	22,028	87 / 986	1,000,000	1,000,000	1,000,000
Set of optimised Building Blocks:						
Population Threshold 1500; target population: 4000; shape constraint: on; homogeneity Constraint: off.						
Number of building blocks	Minimum Population / non zero	Maximum Population	Median / Average population	Minimum Area (m ²)	Maximum Area (m ²)	Average Area (m ²)
243	0 / 1551	22,039	3752 / 4358	1,000,000	119,000,000	44,427,984

Table 4: Comparison of basic statistics for population and area sizes for the Input and AZP-output set of areas

65. A figure showing the result of the aggregation process is presented on the Appendix 5 - Fig. 6.
66. The only building block corresponding to a minimum equal to 0 is the only island considered in the study and not aggregated to any other building block.
67. Empty spaces here have been considered as any other cell. They play an important role in the aggregation process in terms of calculation duration. Keeping the empty cells as such in the process leads to the creation of low populated zones in the middle of densely populated areas.

IV.3.3.2. Delimitation of urban/rural areas

68. A final map presenting the delineation between urban zones (area containing population of more than 50,000 people) is presented on Appendix 5– Fig.7. The use of AZP lead to the creation of zones with intermediate populated density in the middle of densely populated zones. This is due to the presence of empty grids in the input gridded data sets. For the creation of the Urban zones those arbitrarily less dense zones have been aggregated to the densely populated zones.

V. Conclusion of the analyses

69. It was a part of the project to organise the technical implementation of the both grid-based and region-based systems. Both methods propose their own way to harmonise data production.
- To perform well both approaches:
 - need the provision of a ‘good’ initial set of areal units. Both methods need to incorporate confidentiality restrictions (for the dissemination of statistics). The AZP method incorporates it within the internal process of calculation, the grid-based approach incorporate it in a step-by-step process performed by the user;
 - Face the definition of optimality (e.g. optimal grid-size, optimisation of an objective function);
 - Did not manage to take into account several specific problems. For example how the data near outer boundaries of the study area is taken account (grid-based) or data of islands or other special physical structures (both);
 - Need to deal with limitations due to processing speed and disk space;
 - A good GIS software environment.
 - A system of grid-based statistics has a lot of advantages, which has been discussed in the final report. However the production of grid-based statistics can be complicated depending on input data. The European wide system of grid-based statistics needs harmonised production methods;

- The performance of the AZP algorithm is subject to the MAUP (Modifiable Areal Unit Problem) and to the uncertainty due to the optimisation process.
- The AZP algorithm is able to incorporate social constraints. This aspect has not been considered within the framework of this project. Hopefully it will provide an interest for further practical studies.
- This study is hopefully one step towards a common geographical base to compare statistics across the Europe. Before the final recommendations for construction of the European wide statistical system can be given there is need for further studies. There is a need for methodological developments. There is also a need for more case studies with different kind of input data, with different scale of analyses and with different type of spatial analyses.

Bibliography

Alvanides, S., 1995, The investigation of a Zone Design System for reconstructing census geographies Dissertation 3031, School of Geography, University of Leeds

Backer, Lars H., Marja Tammilehto-Luode, and Lars Rogstad. 1999. The use of grids to improve the comparability of statistical data. Paper read at Meeting of the Working Party "Geographical Information Systems for Statistics". Joint meeting with the National Statistical Offices and National Mapping Agencies., 20-21 October, at Luxembourg.

Briggs, David. 2000, (10-12 April). *Spatial Transformation Methods for the Analysis of Geographic Data*. Edited by U. K. Imperial College of Science, *UN/ECE Work Session on Methodological issues Involving the integration of Statistics and Geography*. Neuchatel, Switzerland: Statistical Commission and Economic Commission for Europe.

Cliff, A.D., Haggett, P., Ord, K, Bassett, K., Davies, R., 1975, Elements of Spatial Structure CUP, Cambridge

Eurostat 1997. Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15 . Project Team: Flowerdew, Geddes and , Gatrell from Lancaster University, Diggle and Rowlingson from Lancaster University, Collins from Sheffield University and Briggs from Nene Collage Northampton. March 1997.

Eurostat 1998. Urban database. Working document for the Meeting of the Working Party “European infra-regional information system and urban statistics”. DOC. E/LOC/77.

Eurostat 1999. GIS Application Development. Final Report. SUP-COM 1997 – LOT 3. HTS Consultants in Association with Nene Centre for Research. May 1999.

Fletcher R, 1987, Practical Methods of Optimisation. Chichester, Wiley.

Geddes, A. and Fowerdew R., Geographical considerations in designing policy-relevant regions. 3rd AGILE Conference on Geographic Information Science – Helsinki/Espoo, Finland, May 25th-27th, 2000.

Grasland, Claude 2000. Spatial Homogeneity and Territorial Discontinuities. Paper read at UN/ECE Work Session on Methodological issues Involving the integration of Statistics and Geography, 10-12 April 2000, at Neuchatel, Switzerland.

Harala, Riitta, Tammilehto-Luode, Marja (1999). GIS and Register-based Population Census. Statistics, Registers and Science. Edited by Juha Alho. pp.55-72. Statistics Finland.

Martin, David. 1991. Understanding socioeconomic geography from the analysis of surface form. Paper read at Conference on Geographical information Systems, at Belgium.

- , 1997, Implementing an automated census output geography design procedure. Department of Geography, University of Southampton, Southampton. Draft 20/01/97 (copies obtained by the author)
- , 1998. Optimizing census geography; the separation of collection and output geographies. *Int. J. Geographical information Science* 12 (7):673-685.
- , 1998b. 2001 Census output areas: from concept to prototype. *Population Trends*: 94, 19-
- Openshaw, S. 1991. Developing appropriate spatial analysis methods for GIS. In *Geographical Information Systems: Principles*, edited by D. J. Maguire, M. F. Goodchild and D. Rindh. New York: Longman Scientific & Technical with John Wiley & Son Inc.
- Openshaw, S., and L. Rao. 1995. Algorithms for reengineering 1991 Census geography. *Environment and Planning A* 27:425-446.
- Rase, Daniel. 2000. Technical Specifications; Improving comparability of statistical data across EU Member States, April 2000.
- .Tammilehto-Luode, Marja, Backer, Lars 1999. GIS and Grid Squares in the Use of Register-based Socio-economic Data. Bulletin of the International Statistical Institute. ISI'99. 52nd Session. Proceedings. Book 1. Helsinki 1999.
- Tammilehto-Luode, Marja, Backer, Lars, Rogstad, Lars 2000. Grid data and area delimitation. *Statistical Journal of the United Nations ECE* 17 . pp 109-117. IOS Press.
- Tomlin, C D 1983. A Map Algebra. Proceedings, Harvard Computer Graphics Conference. Cambridge, Massachusetts. Usa.
- United nations 2000. Handbook on geographic information systems and digital mapping. Studies in methods ST/ESA/STAT/SER.F/79. New York.
- Wagget, Margaret. 1999, (20-21 October). Towards improved statistical comparability across Member states- A better geographical framework. Paper read at Meeting of the Working Party "Geographical Information Systems for Statistics". Joint meeting with the National Statistical Offices and National Mapping Agencies., at Luxembourg.

Appendix 1: The candidate methods to convert polygon-based data to grid-based data

In this study three methods were chosen for testing. All of them can be made with the ArcInfo software. Two of them are ArcInfo's standard algorithms; PointGrid and PolyGrid (ESRI 2000). The third method is developed by Eurostat and written in the ArcInfo macrolanguage, AML (Eurostat 1997).

The **PointGrid** algorithm converts data associated with point features to GRID cell format (which is a special ArcInfo format). Each cell in the grid is assigned a code according to the point(s) it overlays. If a cell has more than one possible code, the code with most occurrences in the cell is used. If no points fall within a cell it is assigned the code NODATA.

The **PolyGrid** algorithm converts data associated with polygon features to GRID cell format. Each cell in the grid is assigned a code according to the polygon(s) it overlays. If a cell has more than one possible code, the code of the polygon with the greatest area in the cell is used.

The third method consists of two different algorithms: the **REGION_GRID** command splits irregular polygons with regular grid squares to make new zones, termed "intersect zones". The **DATA_GRID** command is then applied to interpolate data from irregular polygons to the intersect zones and then to the grid squares. The interpolation is made differently for extensive and intensive variables. For both extensive and intensive data types, the new data values for the grid squares equal the sum total of the apportioned or weighted data for all the intersect zones that fall within their boundaries. The data are interpolated to a new grid square using the following equations:

$$\text{Extensive: } X_j = \sum_{i=1}^n X_i \frac{A_{ij}}{A_i}$$

$$\text{Intensive: } X_j = \sum_{i=1}^n X_i \frac{A_{ij}}{A_j}$$

where

X_j = data for grid square j

X_i = data for region i

A_{ji} = area for intersect of grid square j and region i

A_i = area for region i (extensive)

A_j = area for grid square j (intensive)

n = the number of regions

Appendix 2 :The functions used in the automatic zoning procedure

Equality population zoning is a basic and common function used in many census applications by geographers. Regions are devised that are equal or near equal in value in terms of a selected variable (e.g. population size or numbers of economically active people).

The function is:

$$F_{pop}(Z) = \sum_j^m \left(\sum_i^n (\delta_{ij} P_i - T_j)^2 \right)$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if zone } i \text{ is in zone } j \\ 0 & \text{otherwise} \end{cases},$$

P_i is the number of people in zone i ,

T_j is the target size for region j .

The design functions below are designed to ensure:

a limited size (“geometry”) for output areas. The area score or squared boundary length is used to characterise the shape of an area: the **shape design function** (Martin, 1998);

a limited distance from a completely uniform social composition (measured in terms of the proportion of households falling into various tenure classes): the **homogeneity design function** (Martin, 1998).

Shape design function

The function is :

$$F_S(Z) = \sum_j^m (OA_{scj} - H_{scj})^2$$

OA_{ascj} is the shape score for output area j ,

H_{scj} is the hexagone score j .

The shape score is the total squared boundary length of the output area (or hexagon surface).

Homogeneity design function

It measures the squared difference from a completely uniform social composition (measured in terms of the proportion of households falling into various tenure classes):

The function is :

$$F_H(Z) = \sum_i^n (1 - OA_{dhpj})^2$$

OA_{dhpj} is the dominant tenure homogeneity proportion of area j ,

If $OA_{dhpj} = 1$, only one tenure class is present in the OA j , the OA j is at its highest level of homogeneity.

The further to 1 OA_{dhpj} , the less homogeneous the OA j .

Appendix 3 : Grid approach – test results

Table 1 : The Finnish test data
Comparison of real grid squares with estimated grid squares

Real grid squares = Data aggregated from point-based data, accurate geo references

Estimated grid squares = Data converted from polygon-based data by different methods

Population 1998						
10 km x 10 km	sum	count	mean	max	min	variance
Real grid squares		1 541 507		168	9176	290 445 0 935 742 020
Estimated grid squares						
Polygrid	582 546 168		3 468	20 796 98		19 976 866
Pointgrid	412 732 128		3 224	20 796 0		18 413 823
Regiongrid (extensive)	1 537 092		168	9 149 298 399 0		942 603 994
Population 1998						
1 km x 1 km						
Real grid squares		1 418 793		12 988 109	19 478 0	329 220
Estimated grid squares						
Polygrid	37 908 582		12 969 2 923	24 334 0		15 049 853
Pointgrid	1 512 797		429 3 526	24 334 0		17 289 288
Regiongrid (extensive)	1 528 773		12 969 118	13 541 0		246 332

Table 2: The Finnish test data
Comparison of real grid squares with estimated grid squares

Real grid squares = Data aggregated from point-based data, accurate geo references

Estimated grid squares = Data converted from polygon-based data by different methods

Population density 1998					
10 km x 10 km	count	mean	max	min	variance
Real grid squares		168	918	29 045 0	9 357 428
Estimated grid squares					
Polygrid	168	85	2 287 2		65 562
Pointgrid	128	31	208 0		1 787
Regiongrid (intensive)	168	232	12 987 3		1 227 872
Population density					
1 km x 1 km					
Real grid squares		12 988 109		19 478 0	329 220
Estimated grid squares					
Polygrid	12 968 150		31 363 0		735 312
Pointgrid	429 3521		24 334 0		17 288 687
Regiongrid (intensive)	12 969 151		31 319 0		709 963

Table 3: Statistics of urban areas defined by different kind of building blocks

Finland	Total population	Pop density	Area (km2)
Helsinki region			
1 x 1km	903 116	2010,32	449,24
10 x 10km	904 499	1558,41	580,4
NUTS5	998 716	1282	811,4
UK			
Region of Wales			
1 x 1km	2 921 905	2259,14	1293,37
10 x 10km	3 455 239	1110,05	3112,69
NUTS5			
County of South Glamorgan			
1 x 1 km	586 786	3 810,2	134,70
10 x 10 km	953 383	1 440,53	661,83
NUTS5	340 947	2 749,35	124,01

Appendix 4: Grid approach,

Delineation of urban areas, Figures 1- 4

Figure 1:

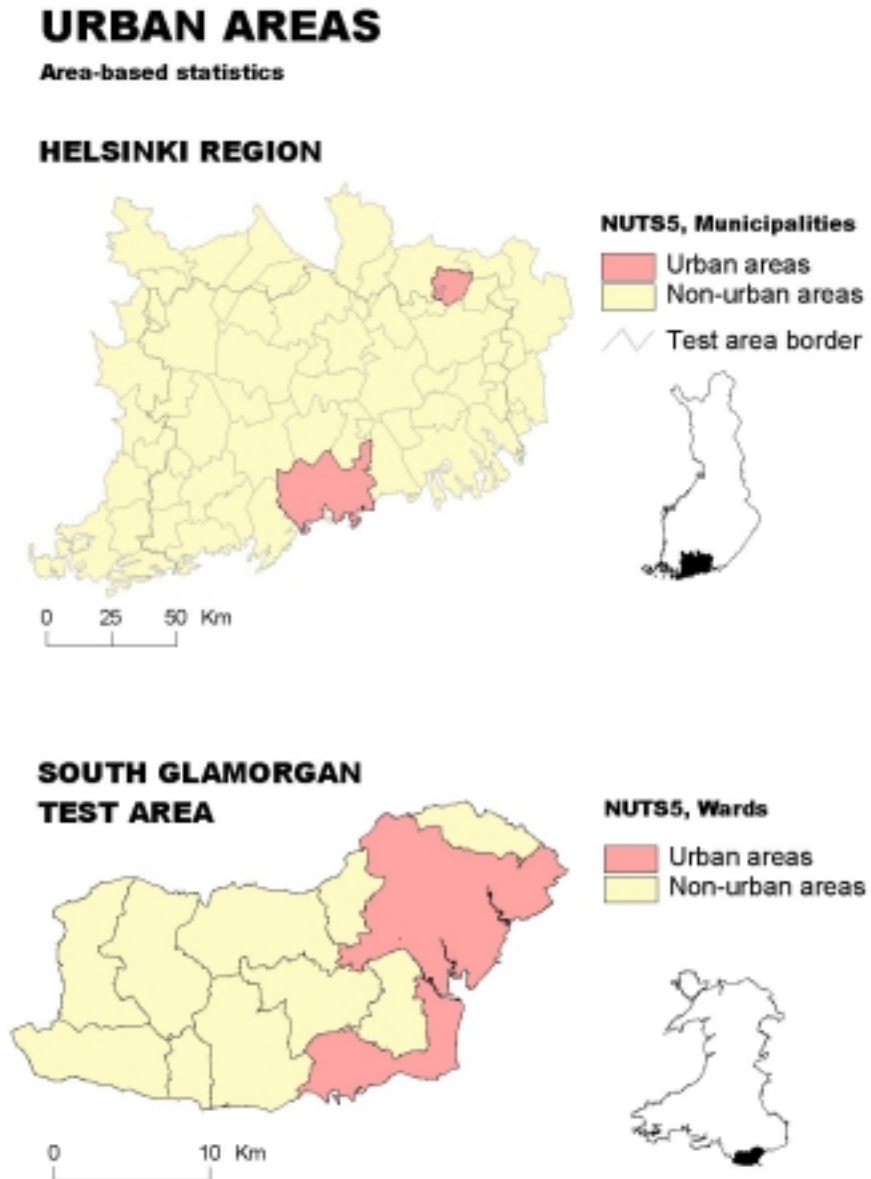


Figure 2:

URBAN AREAS - HELSINKI REGION

Grid-based statistics

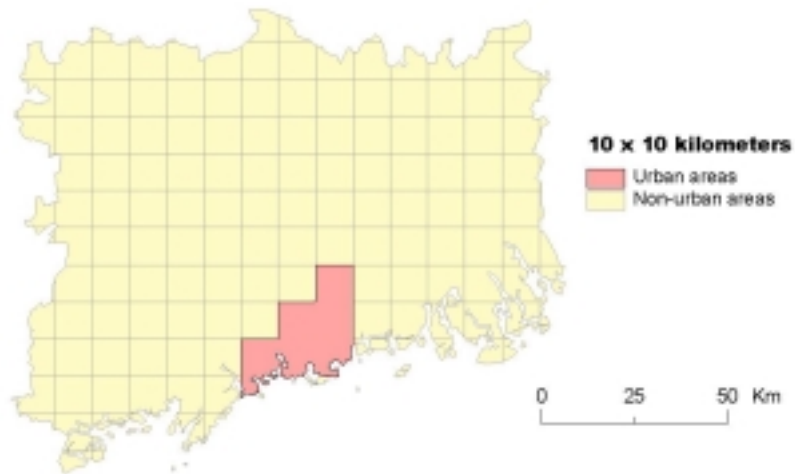
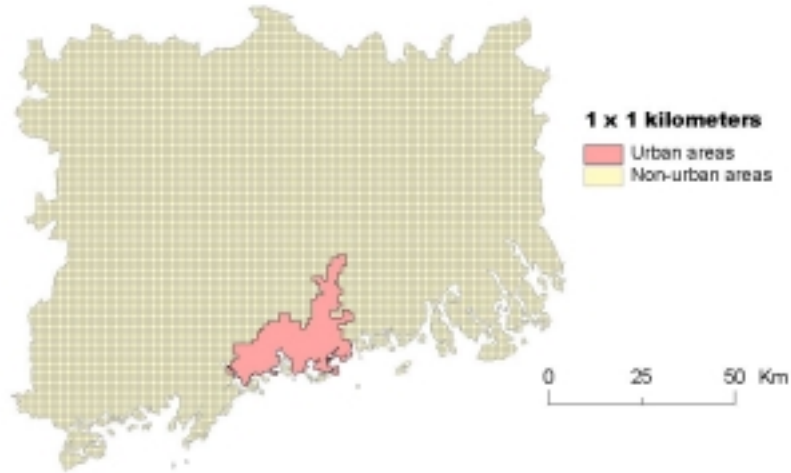


Figure 3:

URBAN AREAS - THE REGION OF WALES

Grid-based statistics

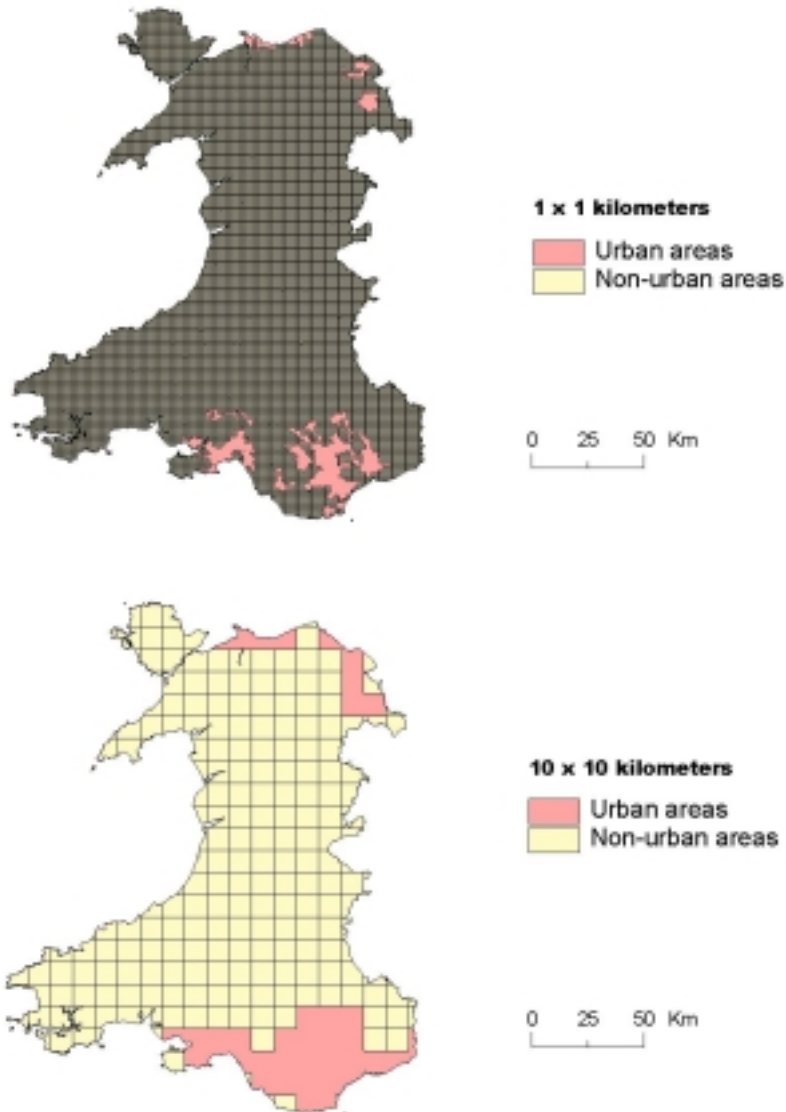


Figure 4:

