# Opportunities and challenges of grid-based statistics

Tammilehto-Luode, Marja
*Statistics Finland*
*PO Box 4 A*
*FI-00022 Statistics Finland*
*marja.tammilehto-luode@stat.fi*

## Introduction

Grid-based statistics are regional statistics in which regional entities are defined by geographically referenced grid cells and statistical variables are calculated and displayed on a regular grid net.

Grid-based statistics have been used in many applications and contexts and their advantages are recognized. However, grid-based statistics are not yet considered as official statistics and an administrative area is still most often the target area for regional statistics.

The purpose of this paper is to highlight the opportunities and challenges of grid-based statistics in comparison with general statistics. Some examples of their potential strengths are described by analyses made by Statistics Finland or its customers.

## Advantages of grid-based statistics

Grid-based statistics could offer a good alternative to statistics by administrative areas. Grid cells do not change as administrative areas do. Their sizes are always comparable, which does not apply to administrative areas. Local or regional changes are easy to analyse with grid-based statistical time series without further processing of the data. Administrative changes always create extra work in a maintenance of regional time series. Grid-based statistics have great potential for comparable territorial statistics and statistical time series (Tammilehto-Luode et al. 2003). For example, combining statistics by grid cells in a consistent way would produce comparable statistics on the urban or rural areas of different countries (Backer et al. 2002).

A regional statistical system is usually based on an administrative hierarchy. Depending on the administration and its regeneration the hierarchy is used and changed differently in different countries. In addition, within countries, different administrative areas are used by different parts of administration for the compilation of statistics of their interest. If based on small building blocks, grid-based statistics can be compiled flexibly by small or large areas or by areas defined by natural boundaries, distances or other spatial factors. They are a good instrument for harmonising the datasets on different kinds of territorial units, when data by their location need to be combined.

Grid-based data are ready to use with many GIS analysing tools. Since the grid is an even-sized and usually relatively small statistical unit when compared to conventional statistical areas, it can describe the real spatial distribution of phenomena far better. Problems arising from the use of averages to describe regional differences can be partially avoided (Martin 1998).

Grid-data can be easily generated from point-based georeferenced data. Disaggregation methods for generating grid data from area-based source data have lately also been developed quite actively (e.g. Steinnocher et al. 2010).

Grid data have great potential for cross-border studies and especially for studies or indicators where the data are heavily dependent on the spatial entity they relate to.

Grid-based statistics are like small area statistics and data protection measures have to be applied carefully. However, solutions for disclosure control that take into account the spatial/regional characteristic

of phenomena can be chosen somewhat more easily for them than for general regional statistics.

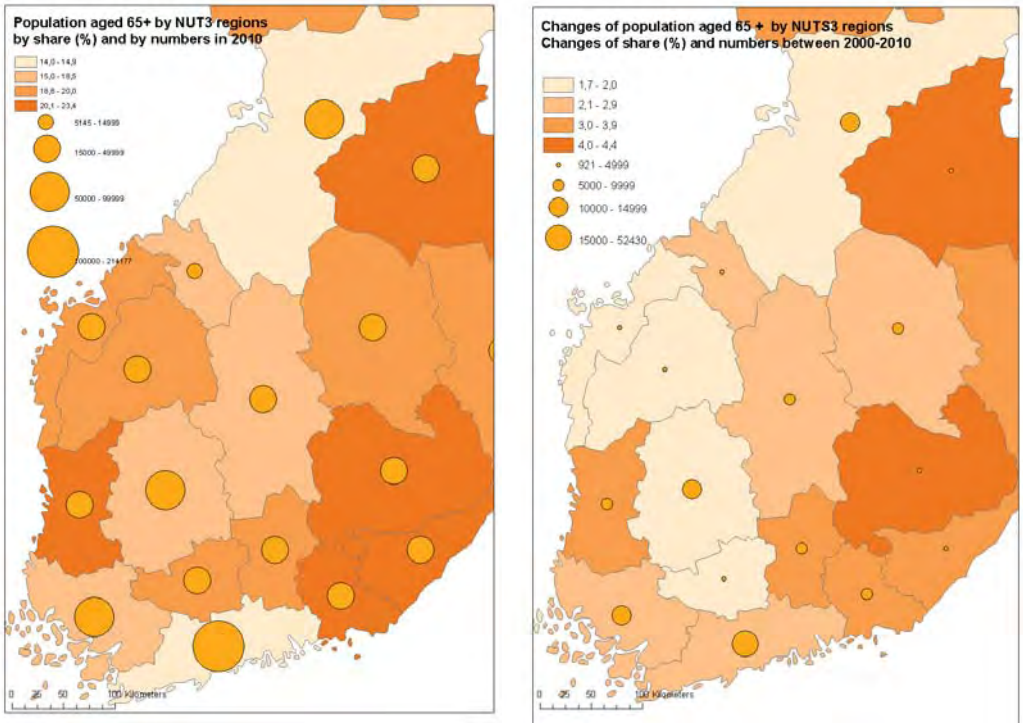
If grid data are already available they may serve as a good source of auxiliary data for small area estimations concerning variables that do not have primary data by detailed georeferences. Grid data also offer a good territorial framework for sampling when territorial representativeness in general or representativeness in e.g. clusters of certain kind of housing is needed.

## Examples

### Ageing society

Ageing is defined as e.g. an increase over time of the percentage share of people aged 65 and over in the total population of a given area (Goll 2010). When comparing the aged population and its changes by different regions over years one may find confusing differences in the distribution of absolute and relative figures. There are also more variations in absolute figures than in relative figures. To describe this by thematic maps we need two different techniques: Choropleth maps and Proportional symbol maps (Figure 1).

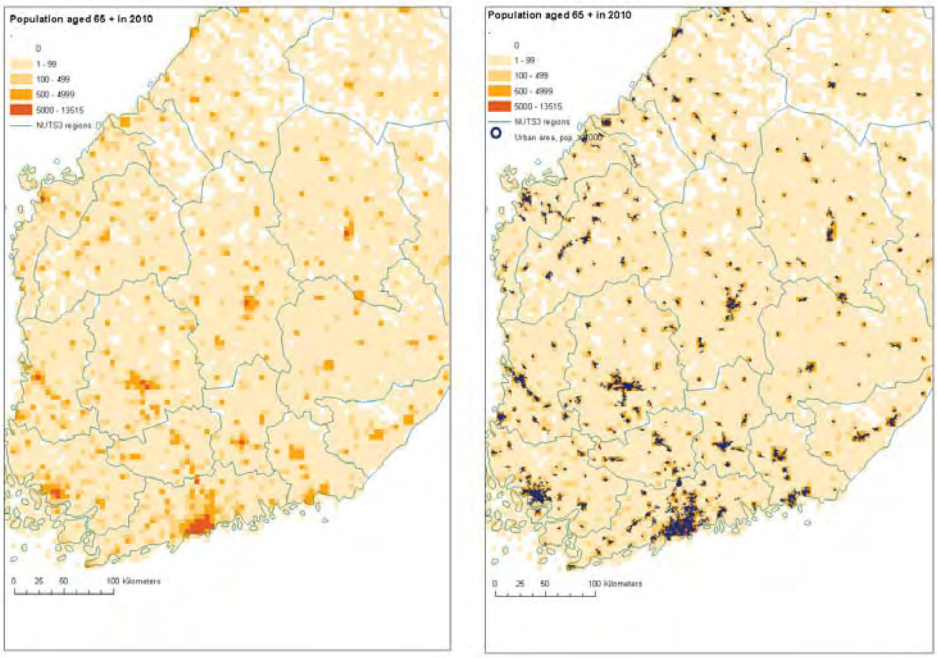
***Figure 1.*** *Share and number of people aged 65 and over in 2010 (on the left) and changes in the share and number of people aged 65 and over between 2000 and2010 (on the right) by NUTS3 regions in southern Finland*




If the regions are of different sizes, as they usually are, comparing them is not always easy neither feasible. The figures and maps only describe averages within the areas, which may hide interesting trends inside the areas.

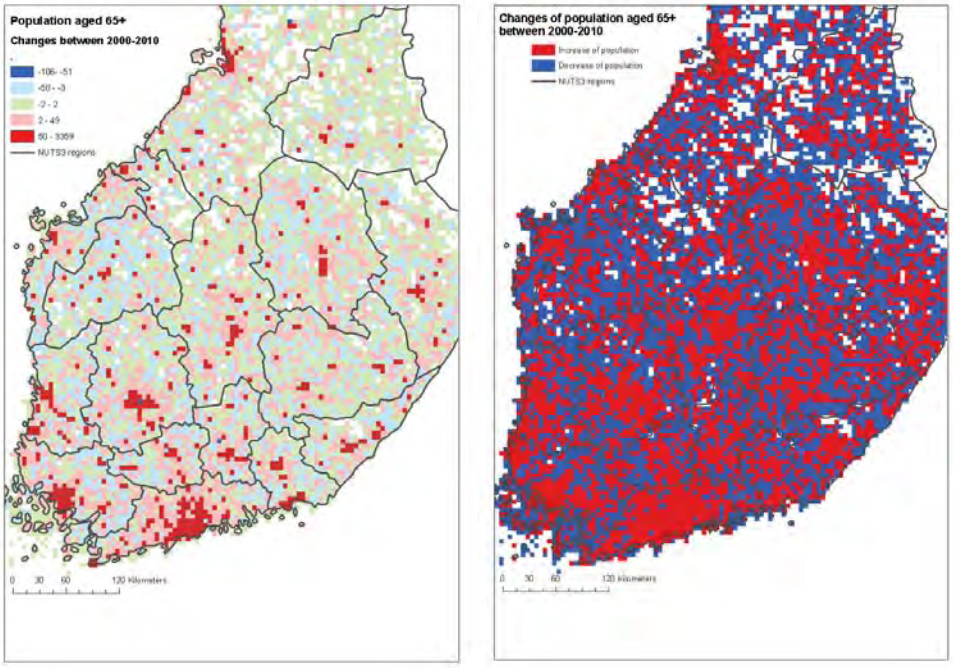
In Figure 2 the population aged 65+ is illustrated by 5 km x 5 km grid cells. Because the grid cells are even-sized the absolute numbers may be visualised and studied on choropleth maps. The maps show that the population distribution including aged people is very concentrated in Finland. If the urban delineation is added to the map (map on the right) hardly any clusters are seen outside urban areas.

***Figure 2.*** *Number of people aged 65 and over in 2010 (on the left) by 5 km x 5 km grid cells together with the urban delineation (on the right) in southern Finland*




Spatial changes can be studied easily by comparing grid cells of different years. Changes in the numbers of aged people in different parts of the regions or simply areas of increase or decrease of in the aged population can be illustrated (Figure 3).

***Figure 3.*** *Changes in the number of people aged 65 and over between 2000 and 2010. (on the left) and increase (red) or decrease (blue) in the number of people aged 65 and over from 2000 to 2010 (on the right) by 5 km x 5 km grid cells in southern Finland*



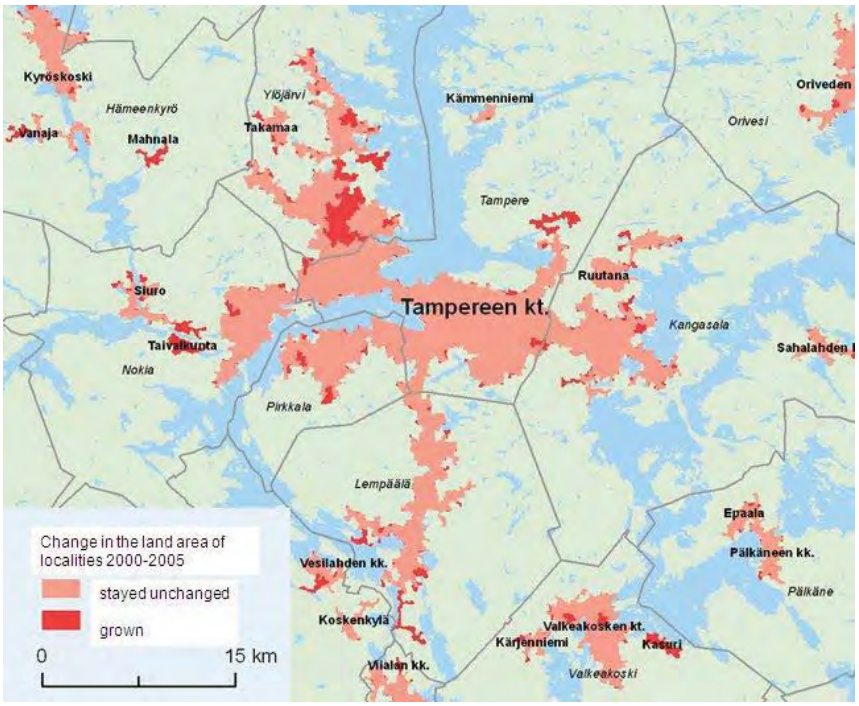
A study of depopulation can be made by identifying the grid cells from which all inhabitants have moved away after a certain year. The threat of depopulation can be identified by the grid cells where the youngest inhabitant is aged over 50 (e.g. Harala et. al. 1999).

**Urban delineation - urban-rural areas**

The new urban-rural typology of the EU was developed by using grid data (Poelman 2010). The main reasons for the new methodology were distortions caused by large variations in the statistical areas (of LAU2 and NUTS3) in Europe. The distortions were large especially in Finland and Sweden due to the large sizes of the administrative areas by which the data for the former typology were collected.

In the Nordic countries, urban settlements have been delineated at five or ten year intervals since 1960 by using georeferenced data but using slightly different methods within the same framework (Tammilehto-Luode et al. 2000). In Finland, urban delineation is today based purely on data by (250m x 250m) grids. The automated delineation process takes into account clusters in the numbers of buildings, the floor areas of buildings and inhabitants. The method is objective and gives data that are comparable both in space and time. Urban delineation is used as such for the defining of urban and rural population and for other statistics by urban and rural areas. The statistical classification of municipalities is based on data by the urban delineations. The share of urban population of the total population of a municipality is the criterion on state subsidies. The delineation of urban areas is also an essential basis for the planning and steering of land use in Finland.

***Figure 5:*** *Delimitation of the Tampere urban settlement in two different years. The dark red areas represent growth areas. The grey lines are boundaries of municipalities*

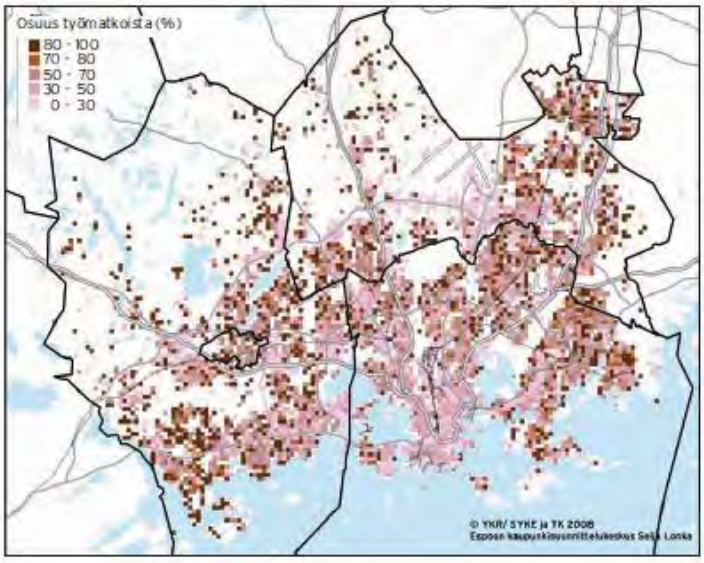


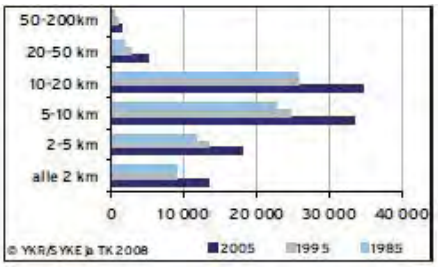

**Monitoring changes**

The Finnish Environment Institute (SYKE) has built a geographic information system, "Monitoring system for changes in urban structure" (YKR), which is widely used in urban and regional planning among public authorities. The YKR consists of nationwide grid data about the population, housing, workplaces and

travel-to-work information from the years 1980-2008. It is a web-based system with which the users can summarise grid data to areas defined by themselves, track changes between years and use ready-made thematic approaches for analyses (Oinonen 2007). An example of an YKR analysis is one of travel-to-work distances and their changes between 1985and 2005 in the metropolitan Helsinki area.

***Figure 6.*** *Analysis of travel-to work distances and their changes between 1985 and 2005. The map illustrates the share of commuters of 5 km distance or less in each grid. The graph illustrates the number of commuters in different distance categories in different years. The table shows the share of commuters by different distance categories and years*




Taulukko. Espoolaisten työmatkojen jakautuminen

| © YKR/SYKE ja TK 2008 | | | |
|---|---|---|---|
| | 1985 | 1995 | 2005 |
| alle 2 km | 12,5 % | 11,7 % | 12,6 % |
| 2-5 km | 16,4 % | 17,5 % | 17,0 % |
| 5-10 km | 32,2 % | 32,3 % | 31,6 % |
| 10-20 km | 36,1 % | 33,9 % | 32,6 % |
| 20-50 km | 2,3 % | 3,3 % | 4,8 % |
| 50-200km | 0,5 % | 1,3 % | 1,4 % |

## Challenges

Public institutions still have only little understanding of the importance of spatial information. Digital maps and geographic information systems are relatively new. The compilation of core datasets with consistent georeferences is not yet part of the production of official statistics.

The production of grid-based statistics is dependent on the available data and the used methods. The so-called bottom-up method by direct aggregation with the help of point-based, detailed georeferenced data produces the best quality results. However, disaggregation from area-based data together with relevant auxiliary data usually generates good estimates as well. Disaggregation methods may vary according to the nature of the variable to be estimated. The availability and quality of the auxiliary data to be used in an estimation process seems to be a very crucial factor. There is a great need for projects to develop and test disaggregation methods further. At least among national statistical institutes disaggregation methods are still only on an initial stage.

In small area statistics, such as grid-based statistics, disclosure control is seen as one of the major challenges recognised by spatial data users in many countries. A detailed map as background information for grid data may increase the need for data protection. In sparsely populated countries like Finland, grid-based statistics face confidentiality problems especially in rural areas. Increasing the size of the grid cell does not always solve the problem. On the other hand, confidential data that have been suppressed may cause crucial effects on the results of a spatial analysis. The user of the data must therefore be aware of how confidential data have been processed in order to understand the potential impacts on his/her analysis.

The data protection measures usually follow the general guidelines on disclosure control with the help of simple data suppressing. Hardly any methods are used which take into account the special features of

geographic information. One example of methods that may suit grid data is the Local Restricted Imputation method (LRI), where data protection is made locally so that the data will always be accurate at a hierarchically higher area level. (Markkula 2003).

Grid-based statistics are heavily dependent on the co-ordinate sytem by which they are produced and displayed. An obvious target is that the grid cells stay rectangular and that the data can be analysed and displayed together with other map data. One may also want to merge different grid datasets together. There is often a need to change from one co-ordinate system to another. The change of co-ordinate system has to be made in theory in the original data before their compilation into grids, which makes it relatively laborious. How the change of co-ordinate system with already aggregated grid-based statistics should be properly done is not so clear at a moment.

A major European terrestial reference system (ETRS89), which will gradually be implemented in the European countries standardises the framework for grid-based statistics and other georeferenced data. The grid, proposed as the multipurpose Pan-European standard, is based on the ETRS89 Lambert Azimuthal Equal Area (ETRS89-LAEA) co-ordinate reference system with fixed centre of projection. However in practice national implementations vary and different projections are needed for different applications and for different part of the world.

Another challenge to the harmonisation of grid data is their coding system. Inspire specifications (INSPIRE 2010), which require the ETRS89-LAEA framework say that the cell code should be composed of the size of the cell and the coordinates of the left cell corner. How this should be done in detail is still an open question. Some of the controversial issues concern whether the coding should take into account scale intervals suitable for spatial analysis and whether the quadtree solution should be included in the coding system.

Errors are difficult to find and correct in the grid data. They are essentially due to the quality of the raw data, such as accuracy and types of georeferences in the input data. A quality specification is necessary for the data, but often difficult to produce. The critical question is the size of the grid cell to be used, which should be considered in the light of quality as well as in the light of the confidentiality aspects.

The grid cell is an abstract and artificial spatial unit which is sometimes difficult to explain to those unfamiliar with the system. It is often necessary to visualise it by identifying the location of the grid on a map. Although grid-based statistics can be further analysed without specific software, GIS tools clearly extend the possibilities to process the data. However, we need to know how to use the tools and how to interpret the results correctly. It is a well-known fact that with a cartographic presentation grid data can be misleadingly manipulated.

## Future of grid-based statistics

According to the “Data provider survey” made by the GEOSTAT project (Geostat 1A 2011), at least 15 national statistical institutes are going to compile grid-based statistics from the 20101/2011 census data. Most of them are going to use the bottom-up approach. The results of the survey indicate that disaggregation methods are not widely known among national statistical Institutes. In order to extend the possibilities to widen grid-based statistics to cover the whole of Europe there is a need to adopt disaggregation methods as well. The development of disaggregation methods would also extend the possibilities to offer different kinds of statistics or longer time series by grid structure. Co-operation with researchers is necessary here.

The GEOSTAT -project promotes the production of grid data by collecting experiences from national statistical institutes and international organisations about methods and good practices for making grid data. The first goal of the project is to compile guidelines for making grid data, especially for those who have not even started yet. A harmonised data delivery, such as grid data compatible from country to country and common dissemination practices are objectives of the project in the longer run. The GEOSTAT project also facilitates the activities of the European Forum for Geostatistics, which has national contact persons in 29 European states and territories who hold annual conferences and meetings (EFGS website).

The European Forum for Geostatistics together with the GEOSTAT -project forms a solid foundation for the future of grid-based statistics. However, the user needs of international organisations and national actors will in the end assure further development and even production of grid-based statistics. Good examples of relevant analyses together with constructive discussion about new challenges that are faced when disseminating small area statistics are necessary to enable more extensive launching of new type of regional statistics.

## REFERENCES

Backer, L., M. Tammilehto-Luode & P. Gublin (2002). Tandem GIS_I. A Feasibility study towards a common geographical base for statistics across European Union. Eurostat. Working papers.

EFGS -website. www.efgs.info.

Geostat 1A (2011). Intermediate report. ESSnet -project GEOSTAT – Representing Census data in a European population grid. Unpublished.

Goll, M. (2010). Statistics in focus 26/2010. Eurostat. ISSN 1977.

Harala, R. & M. Tammilehto-Luode (1999). GIS and Register-based Population Census. Statistics, Registers and Science. Edited by J. Alho. Statistics Finland. Helsinki.

INSPIRE. 2010. INSPIRE Specification on Geographical Grid Systems - Guidelines (GGS_v3.0.1). In *INSPIRE Infrastructure for Spatial Information in Europe*: European Commission.

Markkula, J. (2003). Geographic Personal Data, their Privacy Protection and Prospects in a Location-based Service Environment. Jyväskylä Studies in Computing 30. University of Jyväskylä. Jyväskylä

Martin, D. (1998). Census output areas. From concept to prototype. *Population Trends* 94, 19-24.

Oinonen, K. (2007). Monitoring system for changes in urban structure. Nordic Forum for Geo-Statistics Seminar 2007.Helsinki. http://www.stat.fi/geostatistics2007/session2_oinonen_presentation.pdf.

Poelman, H. (2010). A new urban-rural typology, developed using grid data. e-Proceedings of European Forum for Geostatistics Conference, Tallinn, Estonia. 5-7 October, 2010.

Steinnocher, K. , I. Kaminger, M. Köstl and J. Weichselbaum (2010). Gridded Population – new data sets for an improved disaggregation approach. e-Proceedings of European Forum for Geostatistics Conference, Tallinn, Estonia. 5-7 October, 2010.

Tammilehto-Luode, M., L.Backer & L.Rogstat (2000) Grid data and area delimitation by definition. Towards a better European territorial statistical system. Statistical Journal of the United Nations. ECE. 17,109-117.

Tammilehto-Luode, M., M. Ralphs & L. Backer (2003). Tandem II: Towards a common geographical base for statistics across Europe. The Final Report. Unpublished. Eurostat. Luxembourg