United Nations ECE/CES/2010/11

Economic and Social Council

Distr.: General
30 March 2010

Original: English

Economic Commission for Europe

Conference of European Statisticians

Fifty-eighth plenary session
Paris, 8–10 June 2010
Item 6 of the provisional agenda
Spatial statistics

Production and dissemination of grid data since the 1970 Census in Finland

Note by the Statistics Finland

Summary

Statistics Finland has produced and disseminated census data by 1 km^2 grids since the 1970 Census. The production process of statistical grid data as well as the dissemination of grid-based statistics have changed somewhat over time. The use of statistical grid data has been increasing gradually and their advantages are understood in many organisations outside national statistical institutes.

The purpose of this document is to describe the background of the Finnish statistical grid system, how it is implemented today and what kinds of potentials and challenges it brings out in comparison with general statistics.

I. Introduction

1. The statistical grid data are statistics by direct geographical references (georeferences). Georeferences of grid data are map coordinates of a rectangular system of coordinates. Each square grid cell constitutes a geographic region, a territorial unit and a tool for the territorial classification of data.

2. Administrative areas have conventionally been the most important territorial classification for statistics. The area code systems for the administrative areas have been the cornerstones of regional statistics. The use of administrative areas has turned out to be somewhat problematic. Administrative areas change constantly, their sizes differ within a country and between countries, they are usually large enough to comprise different kinds of areas inside them, and their comparability is weak from region to region and country to country, and over time.

3. The adoption of Geographic Information System (GIS) technologies at national statistical institutes has increased alternative ways of developing and compiling spatial statistics. Addresses and map coordinates have been introduced in addition to the codes of enumeration areas or administrative areas. According to an enquiry made by Eurostat in November 2009, 17 out of the 27 national statistical institutes which responded to the inquiry had established or had a strong commitment to develop detailed georeferencing for their 2010/2011 Census data (Eurostat 2010).

4. The GEOSTAT project which was launched at the beginning of 2010 by Eurostat is a consortium of nine European Union (EU) member states. It promotes the building and harmonisation of grid-based statistics by drawing guidelines and collecting examples with the help of the existing best practices (GEOSTAT 2010-2011). Statistics Finland is a member of the consortium because of its long history of producing grid-based statistics and the experiences it has gained of the growing demand for them.

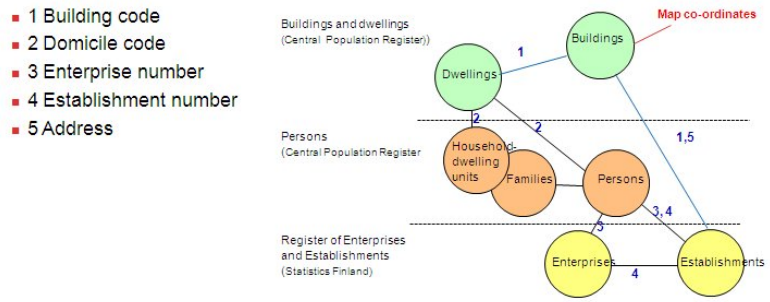
II. Preconditions of statistical grid data in Finland

5. The starting point for grid-based statistics in Finland was the decision to add map coordinates for the centroids of all buildings during the 1970 Census. After 1970, all data which can be linked to buildings can be summed up to grid squares.

6. Before the 1970 Census, there had been efforts to compile grid-based statistics but the results were based more or less on manual work on maps and estimations of population distribution. An example of such statistics is statistics by urban settlement in the years between 1960 and 1980 (General Census of Population 1960).

7. From 1970 onwards, the Finnish register-based census system was developed gradually so that by 1987 all units could be linked with each other with the help of identification codes of persons, enterprises and buildings and dwellings; individual persons could be linked into families and household-dwelling units and into the dwellings or buildings in which they lived, as well as into the employers for whom they worked. This also means that from 1987 onwards all census data can be located on the map with map coordinates via the centroids of buildings.

Figure 1
The basic units and their links in the register-based statistical system in Finland



8. Thanks to the availability of geocoded register-based data, the opportunities for exploiting socio-economic geographical data are exceptionally good in Finland. On the other hand, the quality of grid-based statistics has been very much dependent on the accuracy and quality of map coordinates of buildings (in the Register of Buildings and Dwellings). During the last decade, their quality and coverage have increased considerably. This would not be possible without close cooperation between register authorities and the National Land Survey of Finland together with the fast development of good quality digital maps in Finland.

9. Traditionally, the grid data were considered to form part of special statistical services that were not intended for the general public. The grid data were initially only produced by commissioned assignments. Standardised products have also been available since 1995.

10. Today, Statistics Finland has an objective to increase visualisation by maps to all kinds of regional statistics on the Internet. Interest in spatial statistics in general is clearly growing as well. One reason for this may be the implementation of the INSPIRE directive[1] together with the development of national spatial data infrastructures. The directive and its national decrees address 34 spatial data themes needed for environmental applications. The themes include the geographical grid system and demography. There are currently discussions about the role of statistical grid data or their applications in the national spatial data infrastructures. This may mean that in the future Statistics Finland, with other European national statistical institutes, should provide among other things standardised metadata, data specifications and network services concerning population grid data.

III. Production and dissemination of grid-based statistics

A. Production model

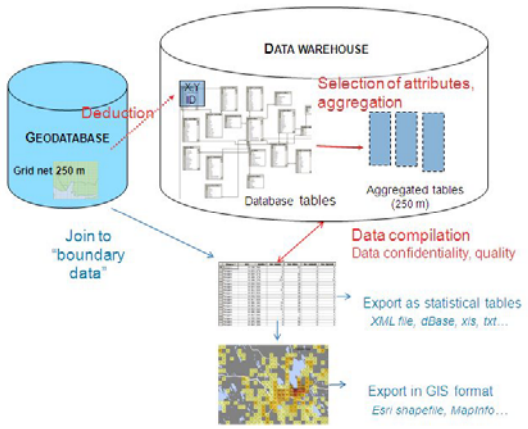
11. Most of the grid data are produced, like other types of census data, with spatial dimensions. In the current production model, which was implemented during 2005-2008, all census type data are stored in the single SQL database. It is a relational database where

[1] INSPIRE Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community http://inspire.jrc.ec.europa.eu/

territorial classifications are managed by different kinds of geographic information attached to buildings. This means that all buildings include all the geocodes that are used in statistical production; from those of provinces to those of municipalities and from postal codes to grid references and map coordinates. All attribute data in this database can be linked to the "building location table" from which the necessary geocodes are retrieved for the compilation of statistics by different domains and by different territorial units.

12. The boundaries of statistical areas are maintained in a central "Geodatabase" containing the grid nets with two different cell sizes. Statistical data can be linked to the boundary data in question by the geocodes. On the other hand, the geocodes of different territorial units for buildings are deducted spatially, e.g. all the buildings inside a certain grid cell or inside a certain postal code area get the geocode in question. This is not necessary for administrative geocodes, which are usually attached to the primary register data.

Figure 2
Production of grid data similar to other territorial statistics with spatial characteristics



13. Statistics Finland maintains national grid nets, one with the grid cell size of 1 km^2 and another with the grid cell size of 250 m^2. The grid nets are vector files covering the whole of Finland with a fixed origin. The grid cell georeference is the map coordinate of the lower left-hand corner of each grid cell. The individual cells of the grid nets have another identification code, the running number of each grid cell. The grid nets also contain information about the municipality to which each grid cell belongs. Any grid at the border between two municipalities is given the code of the municipality with the larger surface area in the grid.

14. At present the Finnish grid-based statistics are constructed on the grid reference system based on the national unified coordinate system (GRID27E). Finland is about to replace its national coordinate system with the European Terrestrial Reference System (ETRS) 89[2], which is to be a major European terrestrial reference system and which will also be the main reference system for grid-based statistics in Finland in the future. The change of the coordinate system has to be made in the original data before the compilation into grids, which makes it relatively laborious.

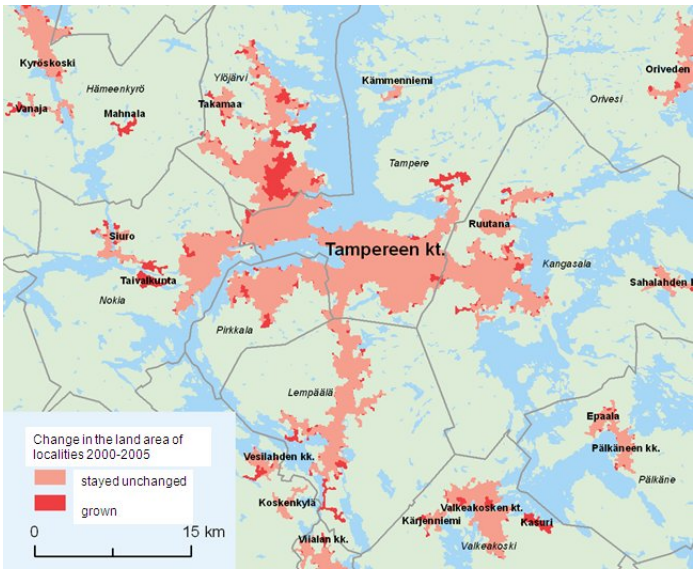
[2] http://etrs89.ensg.ign.fr/

B. Products and services

15. Statistics Finland uses grid data once in five years for the delimitation of localities and urban settlements. The delimitations are then used as statistical areas for the general production of statistics but also for the classification of data and municipalities to rural and urban municipalities[3].

Figure 3

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



16. Most of the grid data are disseminated in two ways: grid-based data are produced and maintained for ready-made products or produced for custom-made services ordered by individual customers.

17. A ready-made product, the Grid Database, includes all grids containing aggregated data of eight data groups[4]. It is delivered as complete or as a module extracted from it. After signing a contract, the customer can download the data via the web by using his/her identification code. The pricing of the Grid Database depends on the volume of data and on the number of users.

18. The Grid Database is delivered in MapInfo or the Environmental Systems Research Institute (ESRI) format. The data can be used with GIS software directly, even without a separate grid net. A grid net is available if needed. The Grid Database and the grid nets have been updated annually since 2003.

19. The contents and dissemination formats of the custom-made grid data vary according to the availability of data and disaggregation level of data. The data quality and the need for any measures to protect statistical confidentiality must always been taken into account, too. According to the Statistics Act, confidential grid data, like any other statistical data, can be delivered for statistical and research purposes with certain conditions and subject to a written and signed agreement about their restricted use.

[3] http://www.stat.fi/meta/luokitukset/kuntaryhmitys/001-2009/index_en.html

[4] http://tilastokeskus.fi/tup/ruututietokanta/index_en.html

20. In addition, "raw statistical grid data" are occasionally refined in-house and only the results of spatial analyses are delivered to customers. Examples of these kinds of analyses are population statistics on sparsely populated areas, or statistics by a certain distance from a power plant, main highway or coastline.

C. Data security

21. 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 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.

22. The Local Restricted Imputation method (LRI) was developed to take account of the special features of geographic information (Markkula 2003). With the LRI, data protection is done locally so that the data will always be accurate at a hierarchically higher area level. This fairly promising method is maybe best suited for research data based on geographic information.

23. Today, the LRI has not been implemented for grid-based statistics in general. The data protection measures follow the general guidelines of disclosure control for demographic statistics with the help of simple data suppression. These measures are normally dependent on the group to which the variables of the statistical data belong. The risk of disclosing individual data is thus mainly considered to be dependent on the sensitivity of the variables and the number of cases in the different cells (Tammilehto-Luode 2001).

D. Data quality

24. The spatial accuracy of source data (the data linked to buildings and their map coordinates) plays an important role when compiling small area statistics such as grid-based statistics. Today, when data on the population and their dwellings are linked to buildings, the coverage of map coordinates is 99 per cent. When workplaces are linked to buildings, the coverage of map coordinates is 92 per cent. Cooperation between the register authority (the Register of Buildings and Dwellings), Statistics Finland and data users is very important in tracing holes in the data. A feedback process towards the central register is still under development.

25. At Statistics Finland, validation checks are made annually to improve the quality of the spatial data. All buildings without coordinates or with locations in conflict with their municipality code are further examined and corrected. Cross-checks are made between the Business Register's establishments and the buildings in the Register of Buildings and Dwellings. Addresses that do not match between these two sources are examined in more detail. Geocoding of buildings with road network address data is a complementary source for quality checking. A GIS application tool for updating address links to the map co-ordinates of buildings is also used.

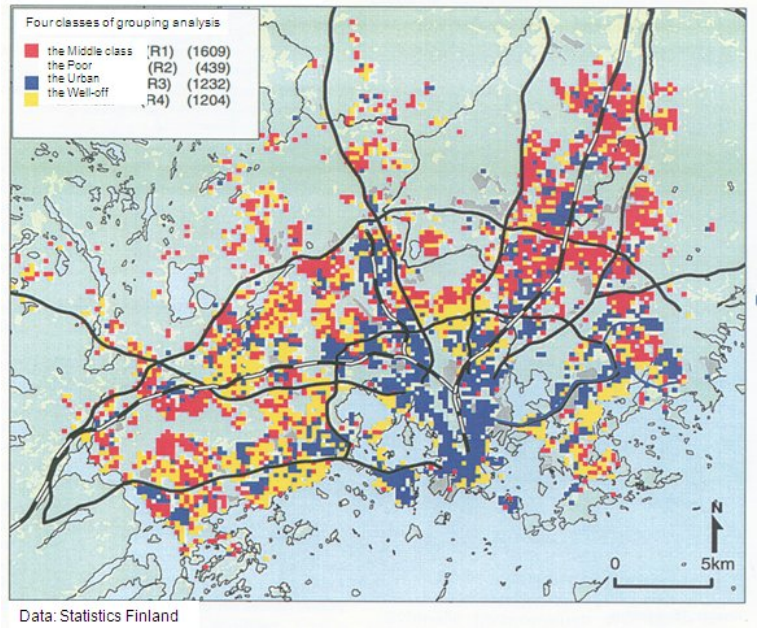
IV. Examples of end use and users

26. For a long time, grid data users were mainly researchers. Grid-based statistics were regarded as a practical calculating tool for the modelling and research of spatial changes that are not dependent on administrative boundaries. Researchers at the Department of

Geography of the University of Oulu are the pioneers in using socio-economic statistical grid data. “Dynamics of regional structure” is their nearly 20-year-old umbrella research programme aiming to analyse different aspects of Finnish regional structures and different phenomena of human actions at all regional levels (e.g. Rusanen 2009). Several doctoral theses have been written at the University of Oulu with the help of statistical grid data (e.g. Vaattovaara 1998). Since the early 1990s, other universities and research institutes have occasionally used grid-based data in the fields of residential differentiation, environmental impact studies, risk management, consumer research, urban and rural studies, and health geography. Many departments of geography of Finnish universities use statistical grid data as teaching material for their students.

Figure 4

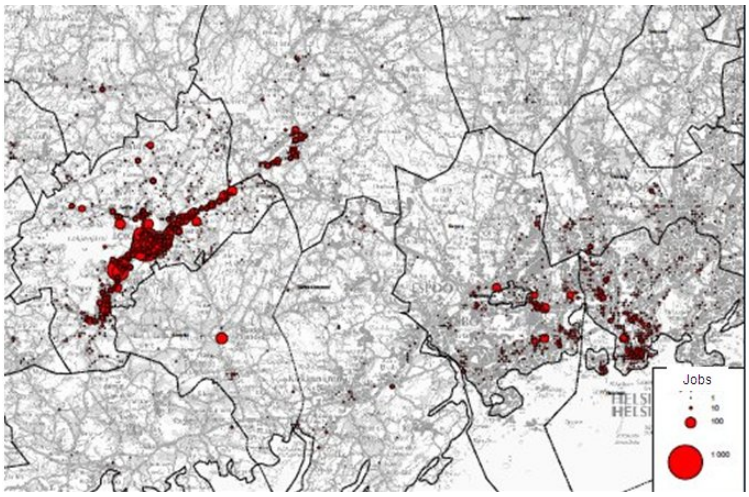
Residential differentiation within the metropolitan area of Helsinki. Visualisation of areas of four major classes, which were defined by factor analysis of grid data by 20 different variables (Vaattovaara 1998).



27. One of the major users of statistical grid data in the public sector is the Finnish Environmental Institute (SYKE). 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 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).

Figure 5

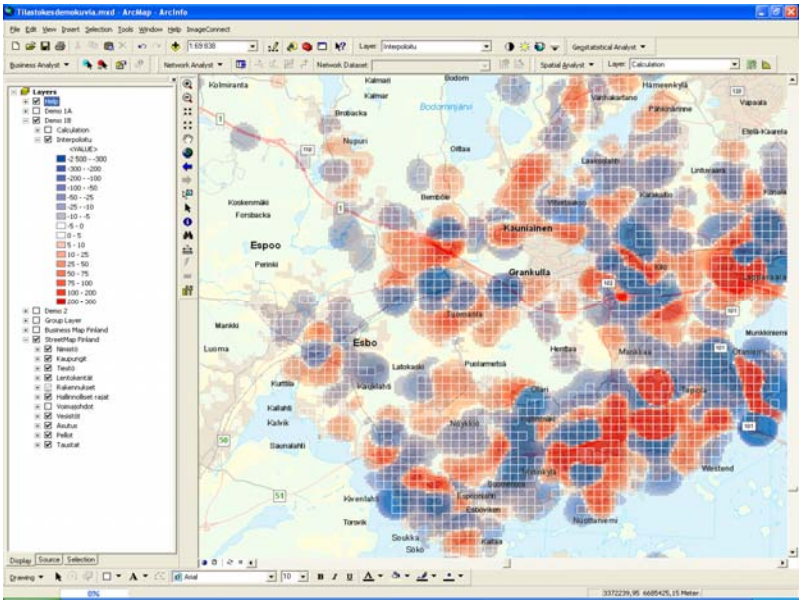
The distribution of the workplaces of employees living in the city of Lohja (Oinonen 2007)



28. The number of private enterprises which use statistical grid data for their businesses has been clearly growing from one year to the next. During the past three years, almost 70 per cent of the revenue from the Grid Database has come from the private sector. Typical users come from communication, banking, retail, insurance and consultancy businesses. Their use of grid-based statistics ranges from applications for marketing, site planning, risk management and planning to the design of all kinds of services. There are also some resellers of the statistical Grid Database, mainly GIS software vendors, but at the moment their input is quite minor.

Figure 6

An analysis (made by a private company) of the change of the number of workplaces between 2003-2005 in the area of southern Finland. The areas of decreased workplaces are blue and increased workplaces are red. This analysis was made for planning new housing for a growth area.



29. Occasionally there are inquiries about grid-based statistics from abroad. They mainly come from either international enterprises planning operations in Finland or a researcher with special interest in cross-border studies.

V. Potentials and challenges

30. In Finland, grid data are compiled mainly from census type data, which are regularly updated with the data from administrative registers. There are other potential sources of grid data, e.g. the Business Register, but for the time being they have not been exploited as much as census data.

31. The Finnish administrative data sources have accurate georeferences (or they can be linked to buildings with map coordinates), which allow a flexible spatial approach to many statistical topics. However, the regular, free of charge statistics production only covers data by major administrative areas. On the other hand, the maintenance of good quality spatial information and protection of data confidentiality in small area statistics require resources which would not be available today if there were no revenue from the services in question.

32. Today there are increasing demand and needs concerning grid-based statistics and all kinds of services by geographic information in statistics. The INSPIRE directive and its implementation rules put additional stress on publishing spatial statistics. It is obvious that there is a lot of potential, but implementing it requires extra funding.

33. Discussions about the role of national statistical institutes in producing regularly new kinds of indicators by using grid-based data are ongoing. At the moment there are 11 countries in Europe that have produced grid data for the whole of their country (GEOSTAT 2010). There are examples in the Scandinavian countries of new kinds of interesting statistics facilitated by grid data: the accessibility of residents to major services, green areas and public transport, population in urban and rural areas and sparsely populated areas, population above the Arctic Circle, concentration of certain workplaces, etc.

34. In general, grid-based statistics have a great potential for comparable territorial statistics and statistical time series (Tammilehto-Luode et al. 2003). They can be compiled flexibly by large areas or by areas defined by natural boundaries, distances or other spatial factors. They give an easy and reliable base for tracking spatial changes in different phenomena. They are a good instrument for harmonising the datasets on different kinds of territorial units, when combining data by their location is needed.

35. Stable grid-based statistics could offer a good alternative for statistics by administrative areas, because their areas never change and they are comparable from region to region. For example, combining statistics by grid cells in a consistent way would produce comparable statistics on urban or rural areas of different countries, which do not exist at the moment (Backer et al. 2002).

36. Grid data have a great potential for cross-border studies and for studies or indicators where the data are heavily dependent on the spatial entity to which they relate. Since the grid is usually a relatively small statistical unit when compared to conventional statistical areas, it can far better describe the real spatial distribution of phenomena. Problems arising from the use of averages to describe regional differences can be partially avoided by using small, evenly-sized area units like grids (Martin 1998).

37. An advantage of grid-based statistics is their ability to present “unpopulated” land or other non-existent variables in space. It would be difficult to describe depopulation in the Nordic countries if statistical figures were available by administrative units only (Harala - Tammilehto-Luode 1999).

38. Today, grid data are mainly compiled for national purposes and hardly any international standards are applied. However, there is a need for standards at least for cross-border and international studies. Standards are about to be introduced in connection with the Infrastructure for Spatial Information in the European Community (INSPIRE) implementation rules.

39. The grid cell is an abstract and artificial spatial unit which is sometimes difficult to explain to those unfamiliar with the system. Often it is necessary to visualise the results by identifying the location of the grids on a map. For statistical tables, grid cells need to be aggregated to an administrative or other well-known regional level for making final interpretation easier.

40. Statistical grid data are very effective with the map algebra applied in many GIS software packages. However, one needs to know how to use them and to interpret the results correctly. A responsible attitude towards cartographic presentation is essential because grid data are easy to manipulate.

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

42. Strict disclosure control has to be exerted if the grid cells only contain very small numbers of residents or employees, for example. Often the researcher need not know the exact location of each cell. However, there are applications, such as site planning and market analysis, which require all grids to be identified. In small area statistics, such as grid-based statistics, disclosure control is a major challenge recognised by spatial data users in many countries.

VI. 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.

Business Register. http://stat.fi/tup/yritysrekisteri/index_en.html

Eurostat (2010). Geo-referencing of Data from the 2011 Censuses of Population and Housing. Consultation November 2009. Document: E/GIS/103/EN. Meeting of Working Party “Geographic Information Systems for Statistics”. Luxembourg, March 08-09,2010.

General Census of Population 1960. Non-administrative urban settlements and their boundaries, Official Statistics of Finland VI C: 103. Helsinki 1965.

Grid Database. http://stat.fi/tup/ruututietokanta/index_en.html

GEOSTAT 2010-2011. Representing census data in a European population grid. ESSnet project. Grant agreement for an action with multiple beneficiaries. Agreement number 50502.2009.004 -2009.860. Eurostat.

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

INSPIRE Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). http://inspire.jrc.ec.europa.eu/

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, Kari (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

Rusanen Jarmo (2009). List of publications. http://www.oulu.fi/geography/staff/professors/publications/rusanen.pdf

Tammilehto-Luode, M. (2001). Disclosure control for demographic statistics. Statistical Journal of the United Nations Economic Commission for Europe. Volume 18:4. ISSN 0167-8000.

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.

Vaattovaara, Mari (1998). Residential differentiation within the Metropolitan area of Helsinki. Pääkaupunkiseudun sosiaalinen erilaistuminen. City of Helsinki Urban Facts Research Series 1998:7.