EUROPEAN FORUM FOR GEOSTATISTICS

SYSTEMS OF GEOSTATISTICS AND PROCESSES REQUIRED TO PRODUCE THEM

SUMMARY

This is the first essay on behalf of the Geostat project to produce a general description of the processes leading to an integrated system of information that may serve as a foundation for all efforts to achieve a sustainable process of development.

This paper represents an effort to present a general description of a hierarchy of spatial statistics (a GGS a Global system of Geo- Statistics, and EGS (European system of Geo- Statistics) or a NGS (National system of Geo- Statistics) as a part of a full- fledged statistical system GSS (Global statistical system), ESS or NSS, the processes required to produce them. The description proposed is of course not the only "true" description, but the important observation is that it "works". However, it is hoped that it may be improved in the iterative evolutionary process required from all so-called Darwin systems. According to this proposal, we have to discuss the quality of both the product and the production process as one side of the same coin. The quality assessment for a system of spatial statistics and the production process designed and implemented to produce it, will according to this proposal consist of a compound two step process starting a) by judging each of these steps separately, and b) judging all parts together. The proposal is based on the idea that in order to develop a GGS, EGS or NGS, we may use assume that in spite of the hierarchical structure, one eight step process is required.

1. Data model
2. Data capture
3. Object- based statistical databases
4. Non- spatial, Spatial and Temporal- analysis
5. Compilation of National, European or Global Geostat dataset(s) for dissemination
6. Dissemination
7. Feedback on real user needs
8. Overall quality assessment

For the production of a system of information of the kind that we assume respond to real user needs. These are of two quite different kinds; information (or rather structured information or knowledge) in response to for public authority needs and a great host of various information products in response to private sector needs. In this connection we are primarily concerned with the former. Both however should be regarded as products of the one and same process described here. Finally it is important to note that this description should be regarded as the foundation for a host of secondary processes and in particular 1) the all important quality assessment and/or 2) processes to assess the processes compliance with public law (as for instance confidentiality rules).

Stockholm Monday 10 December 2012

Lars H. Backer
lars.backer@mdmapping.se
Valhallavägen 145
115 31 Stockholm
Sweden

TABLE OF CONTENTS

INTRODUCTION

In this paper we will discuss the production the interdependence of the ideal of harmonised ESS and a harmonised system of processes used for its production. The Geostat project has as its objective to produce the prototype for a harmonised point- based system of spatial statistics for Europe. All previous experience from the production of complex systems, is that in order to succeed you will have to start with a good design for both the system and the processes to produce it, and then build an iterative process that will enable it to evolve over time. We have previously referred to this principle as the construction of a Darwin machine here for the production of a qualified system of spatial statistics for Europe.

1. The first principle to agree upon is that complex systems like the NSS (national statistical systems) , the ESS (European system of statistics) and the GSS (global system of statistics) and the processes used for their production are consciously designed and built to satisfy user needs. As our philosophers would say, "they are made, not found". In order to improve over time, they have to evolve in a (never- ending) process of design, build, use, evaluate, re- design, re- build, re- use, re evaluate, etc. ad infinitum.
2. The second principle is that in order to achieve real improvements over time the process must be organised as a classical project. That starts with evaluating (or re- evaluating)the qualities and problems, or the current state A of the system, then it goes on to evaluate (or re- evaluate) the ideal state B of the system that we want to achieve, and then evaluate (or re- evaluate) the actions C we have designed and built to improve the state A in the direction of B.
3. The third principle is that "if you cannot describe it, you cannot manage it". This means that the success of the process depends on qualified descriptions of a) the current "product" state A of the system, b) the ideal state B envisaged for the system, and last but not least the production processes C that will help us to re- build the current system A into B.
4. The final principle is that the quality of a system state A has to be measured, or assessed in comparison with the "ideal" state B.

We are used to laude the principle of diversity. But if we agree that national systems of statistics belong to a hierarchy of statistics that need to be harmonised from the bottom and up, we will have to find ways to treat the NSS's the ESS and the GSS as one (however imperfect) integrated hierarchal system. This system may then be harmonised over time in a deliberate iterative process of constant evolution according to the principles mentioned above.

In the first paper in this series we have identified the generalized product / process as follows:

1. Data model
2. Data capture
3. Object- based statistical databases
4. Non- spatial, Spatial and Temporal- analysis
5. Compilation of National, European or Global Geostat dataset(s) for dissemination
6. Dissemination
7. Feedback on real user needs
8. Overall quality assessment

From this we may draw the conclusion that in order to develop a statistical system (or any complex system for that matter) we need three sets descriptions (designs):

A. A description, design (or re- description re- design) of the current Geostat system of point- based statistics. (The method for describing A is the same as that used for describing B below.)
    1. Current Data model

2. Current Data capture
3. Current Object- based statistical databases
4. Current Non- spatial, Spatial and Temporal- analysis
5. Current Compilation of National, European or Global Geostat dataset(s) for dissemination
6. Current Dissemination
7. Current Feedback on real user needs
8. Current Overall quality assessment

B. A description (or re- description) of the ideal state of the system (the vision). (The Geostat 1A project has presented a vision for an ideal system for Geostatistics to use as a point of departure here.)
   1. Ideal Data model
   2. Ideal Data capture
   3. Ideal Object- based statistical databases
   4. Ideal Non- spatial, Spatial and Temporal- analysis
   5. Ideal Compilation of National, European or Global Geostat dataset(s) for dissemination
   6. Ideal Dissemination
   7. Ideal Feedback on real user needs
   8. Ideal Overall quality assessment

C. A description (or re- description) of the process of production (improvement of A in the direction of B) The current paper is intended to serve as an effort to describe a general process dedicated to the production of geostatistics based on the methods generally used by European NSI's
   1. Actions to improve Data model
   2. Actions to improve Data capture
   3. Actions to improve Object- based statistical databases
   4. Actions to improve Non- spatial, Spatial and Temporal- analysis
   5. Actions to improve Compilation of National, European or Global Geostat dataset(s) for dissemination
   6. Actions to improve Dissemination
   7. Actions to improve Feedback on real user needs
   8. Actions to improve Overall quality assessment

CONCLUSION

The conclusion of this perspective is that the evaluation of a project of this sort should be focussing on two types of descriptions. The first is a description of the dataset that will have to fit into varying types of modelling efforts.

1. Descriptions of system states (EGS dataset)
   In this scheme the descriptions A and B are parts of the model to which the data contribute. In the context of the Geostat project, these data are assumed first of all to contribute to model human societies from two perspectives; the socio- cultural- and political system (the night- time population) on the one hand and the socio- economic- and financial- system (the day- time population) on the other.
2. Descriptions of transformation processes (EGS production process)
   In this scheme the description C has as it object the description of a series of actions that transform the state A of the system modelled into state B. These are process descriptions of more or less well coordinated (orchestrated) system of actions.

THE PRODUCTION PROCESS ISSUE FOR ISSUE

I think that this is a good beginning but should be further deceloped. I think it should be seen in the light of the general process used for the production of statistics where geostatistics is only one (albeit important) part. Please consider this:

1. DATA MODEL(POINT- BASED OBJECT FOUNDATION)

This point according to the general outline of the process as defined in paper 1 in this series.

1. Modelling (of a dataset in response to real user needs)
   The building of a statistical system must be regarded in the perspective of an integrated system of information needed for all levels of actions from local to global. This modelling effort will require the active cooperation between a" more practical science" and a "more scientific practice". The modelling effort will for obvious reasons have to be based on the object approach. Its scope must be holistic. This means that be useful for the description of the Earth as an integrated man environmental system. It must respond to the observation that we experience our world as a 4D universe, an interaction of objects in space and time.
   1.1. Modelling as product
      The result of cooperation between a more practical science and a more scientific practice
   1.2. Modelling as a process
      The modelling of man environmental systems as a process of continuing development
   1.3. Quality control of the modelling product and processes
      1.3.1.Output (product)
      1.3.2.Production Process

The data model used by NSI's should traditionally be a system that effectively may be used for modelling human societies. This is the foundation for all modelling of man environmental systems in general.. Unfortunately, there is not much agreement on a common approach to modelling e.g. human societies. However there is at least a general agreement that whatever its form, any modelling effort should be object-based. We hope that the modelling issue will become the focus of the GGIM process so that we may correct some of the crudest mistakes made by the INSPIRE project.

OBJECT APPROACH TO MODELLING

The Geostat position is stated in the Vision part of the Geostat 1A project. According to this approach the individual citizen should be the fundamental object for describing human societies could be used as a point of departure for our discussion here. In a point- based approach all object observations are provided with

1. Object-id (System, object class and object instant reference)
   1.1. Attributes (statistical variables)
   1.2. Spatial reference (point coordinates)
   1.3. Temporal reference (time- stamp)

We are here discussing the data model with a focus on the spatial or geographical reference. From the perspective proposed by the Geostat 1A project the modelling of a society (of societies) could look something like this:

A SPATIAL HIERARCHY OF SOCIAL SYSTEMS

One of the key tasks of a GGIM, the EGIM and perhaps National systems of spatial (and temporal) information is to try to build a shared object strategy for modelling the Earth as a hierarchical system of descriptions reaching from Global to Local.

1. Locality
2. Municipalities
3. National Regions or small states (e.g. Regions in France or Bundesländer in Germany)
4. Nations or large states (like UK or Sweden)
5. Global Regions (like the EU 27+)
6. The Earth

MODELLING WITH STATISTICS

We are, in the Geostat projects, primarily concerned with data to model human societies with the data available through the NSI's. These data systems are traditionally, but of course not exclusively, focussed on the description or modelling this in terms of Day- time (Socio- economic and financial systems) and night- time populations (socio- cultural and political systems).

DATA MODELLING PROCEDURES

We take it for granted that the "raison d'être" of official statistics is to contribute to the compilation of qualified models of single regions, countries etc. seen as integrated man environmental systems. The Geostat project is concerned with modelling the territory and social systems within the EU 27+ area.

1. Data modelling(point- based object foundation)
   - 1.1. Modelling human societies (Focus of the NSS)
     - 1.1.1. Model of socio- cultural and political systems
     - 1.1.2. Model of socio- economic and financial systems
   - 1.2. Modelling physical environment
     - 1.2.1. Models of the built environment (technosphere)
     - 1.2.2. Models of the natural environment

MODELLING HUMAN SOCIETIES

We have elsewhere elaborated the topic of modelling man environmental systems in general and human societies in particular. The reason for this is naturally due to the fact that these systems are mainly dependent on information of the type that is productd by the NSS (national statistical systems)

1. The Socio- cultural (and political) system
   Models of socio and cultural systems are based on a hierarchy of public authorities on all (generally 6) levels from Global to local
   - 1.1. Social systems
     - 1.1.1. Global system (Society of societies)
     - 1.1.2. Global Region(E-g Europe)
     - 1.1.3. National system
     - 1.1.4. Regions
     - 1.1.5. Municipalities
     - 1.1.6. Localities (Municipality sub- division)
   - 1.2. Individual / family
     - 1.3. Household
       (The ideal basic unit for geographical reference: Street address with coordinates.)
     - 1.4. The individual citizen
       - 1.1.1. Individual as political citizen (Night- time population)
       - 1.1.2. Individual as employee (Day- time population)
         (This unit provides the connection between the Socio- cultural system and the socio economic system)
2. The Socio- Economic system

2.1. Economic systems
2.2. Primary Business & Industrial category
2.3. Secondary B/I category
2.4. Tertiary B/I category
2.5. Quaternary B/I category
2.6. The business and industrial enterprise
2.7. The work/ production unit /site
(The ideal basic unit for geographical reference: Street address with coordinates.)
2.8. The individual employee
2.8.1. Individual citizen as employee (Night- time population)
2.8.2. Individual as political citizen (Day time population)
This unit provides the connection between the Socio- cultural system and the socio-economic system

3. Human societies
Modelled with data based on the citizen in two roles. First as a part of socio- cultural systems (Night-time population and secondly as part of the financial- economic- system as part of the day- time population.

QUALITY ASSESSMENT

2. DATA CAPTURE

This point according to the general outline of the process as defined in paper 1 in this series.

2. Data capture
2.1. Methods used for data capture
2.1.1. Registers
i Data capture tools and methods used in register systems
ii Critical production (quality?) bottlenecks
2.1.2. Census- based systems
i Data capture tools and methods used in census based systems
ii Critical production (quality?) bottlenecks
2.1.3. Sampling
i Data capture tools and methods used in sampling- based systems
ii Critical production (quality?) bottlenecks
2.1.4. Other
i Data capture tools and methods used in sampling- based systems
ii Critical production (quality?) bottlenecks
2.1.5. Mixed systems for data capture
i Mix 1
ii Mix 2.
iii Etc.
2.2. Cleaning (correction) of raw data
2.3. The production of statistical data tables
Building of Statistical database tables for integration national data into the NSS.(the object based statistical databases described in next section 2 below.). On the European level this process involves the building of the ESS from compiling information provided by the NSI's. On the European level this process involves the building of the GSS from compiling information provided by the NSI's as well.
2.4. Quality control of the data capture processes

2.4.1.Output (product)
2.4.2.Process

Data capture in response to the standard (?) data model recommended for Europe and the GGIM work. Censuses are just one of many methods used to capture the information needed for providing the attributes required by the model. All observation of attributes will be stored with the id- code, the spatial reference (coordinates) and the timestamp.
The critical issue in terms of quality here is the resolution of the smallest aggregation used. We have used the term location here. (Coordinates for Apartment, Street address, Building or real estate unit or if nothing else is available then the geometric centre of census areas)

DATA CAPTURE

Please consider this effort to provide an overview over the steps in the data capture process that

2. Data capture
   2.1. Methods for data capture
      2.1.1.Registers
         i Data capture tools and methods used in register systems
         ii Critical production (quality?) bottlenecks
            - Critical production cases 1
            - Critical production cases 2
            - Etc.
      2.1.2.Census- based systems
         i Data capture tools and methods used in census based systems
         ii Critical production (quality?) bottlenecks
            - Critical production cases 1
            - Critical production cases 2
            - Etc.
      2.1.3.Other method (s)
      2.1.4.Sampling
         i Data capture tools and methods used in sampling- based systems
         ii Critical production (quality?) bottlenecks
            - Critical production cases 1
            - Critical production cases 2
            - Etc.
      2.1.5.Mixed systems for data capture
         i Mix 1
            - Data capture tools and methods used in Mix 1 systems
            - Critical production (quality?) bottlenecks
         ii Mix 2.
            - Data capture tools and methods used in Mix 1 systems
            - Critical production (quality?) bottlenecks
         iii Etc.
   2.2. Cleaning (correction)
   2.3. Building of Statistical database tables for integration inn NSS, ESS or GSS .(the object based statistical databases described in section 2 below.)
   2.4. Quality control

3. OBJECT- BASED STATISTICAL DATABASES (MICRODATA DMBS)

This point according to the general outline of the process as defined in paper 1 in this series.

3. Object- based statistical databases (Microdata database system)
   - 3.1. Non- spatial attribute data (attributes to describe object according to model requirements)
   - 3.2. Spatial data (coordinates for observations)
     - 3.2.1. Smallest possible spatial reference
       - i Street Address
       - ii Building
       - iii Real Estate unit
       - iv Census are or similar Municipality sub- division
   - 3.3. Temporal data (timestamps for observations)
     (Not discussed here)
   - 3.4. Quality control of the Microdata system
     - 3.4.1. Output (product)
     - 3.4.2. Production Process

The next major step in the process is to build efficient statistical databases that may be related to the data model as implemented in harmony with the data model.(see step 1). It is essential that all object attributes are stored as point data. This ensures that most data processing can be accomplished outside GIS systems. These have generally relational database structure where all object tables, are connected with a geography table containing all geographical "part of-" , "belonging to-" and other relations. (e.g. object x belongs to; apartment a, address b, building c, real estate d, census area e, ..Municipality m, region, n...etc.). Both non-spatial and spatial selections should render the same result.

A standard national statistical system may be seen a fundamental system of microdata organised in correspondence with the data model used. In Europe we are working in the direction of agreeing upon one shared data model for all EU 27+ countries.

Form this fundamental database we build processes to extract the data needed for different purposes. One such dataset is the information that will serve as the foundation for the European hierarchy of statistics on grids. We will call this the Geostat dataset. Thus we might consider;

INTEGRATION OVER THREE LEVELS

1. A National Geostat dataset (NGD)
   is achieved by integrating regional datasets:
   - 1.1. National region (or set of regions) A
   - 1.2. National region (or set of regions) B
   - 1.3. Etc.
2. An European Geostat dataset (EGD)
   is achieved by integrating national datasets:
   - 2.1. Geostat Czech Republic
   - 2.2. Geostat France
   - 2.3. Geostat Switzerland
   - 2.4. Geostat Norway
   - 2.5. Etc.
3. A Global Geostat dataset.(GGD)
   by integrating datasets provided by various global regions:
   - 3.1. Geostat Europe
   - 3.2. Geostat Australia and Oceania
   - 3.3. Etc.

We are at present concerned with the building of an EU grid dataset. This dataseet is produced through a integration of (ideally) all EU 27+ NSI datasets into one harmonised whole.

4. DATA- SPATIAL- AND TEMPORAL- ANALYSIS

This point according to the general outline of the process as defined in paper 1 in this series.

4. Spatial and temporal analysis
   - 4.1. Analysis of non- spatial object attributes (according to conventional statistics) (Not discussed here)
   - 4.2. Analysis of the spatial dimension of object attributes (focus on the spatial dimension)
     - 4.2.1. Introducing new grid data sets
       - i Bottom- up procedures
         - Table aggregation method
           - Method 1 (SAS method)
           - Method 2 (Mapinfo script method)
           - Etc.
         - GIS spatial method for aggregation
           - Method 1 (ArcGis method)
           - Method 2 (Mapinfo method)
       - ii Top- down procedures
         - Disaggregation method 1(Corine data)
           - Method 1(JRC method 1)
           - Method 2 (JRC method 2)
         - Disaggregation method 2 (soil sealing)
         - Etc.
       - iii Hybrid procedures
     - 4.2.2. Transformation of existing grid data sets
       - i Data transformations
         - Transformation method 1
         - Transformation method 2
         - Etc.
   - 4.3. Temporal analysis (Not discussed here)
   - 4.4. Quality control of the Microdata system
     - 4.4.1. Output (product)
     - 4.4.2. Production Process

Analysis procedures are focused on providing a full, high quality dataset through the integration of data captured by the NSI's themselves with data produced by other institutions e. g. Mapping agencies.

PRODUCTION CASES

1. Spatial and temporal analysis
   - 1.1. Attribute or non- spatial data analysis (according to conventional statistics) (Not discussed here)
   - 1.2. Spatial analysis (focus on the spatial dimension)
     - 1.2.1. Introducing new grid data sets
       - i Bottom- up procedures
         - Bottom- up production cases 1
         - Bottom- up production cases 2

- Etc.

ii Top- down procedures

- Top- down production cases 1
- Top- down production cases 2
- Etc.

iii Hybrid procedures

- Hybrid production cases 1
- Hybrid production cases 2
- Etc

1.2.2. Adoption of existing grid data sets

i Data transformations

- Transformation production cases 1
- Transformation production cases 2
- Etc

1.2.3. Temporal analysis

(Not discussed here)

DATABASE VS. GIS METHODS FOR AGGREGATION

It has been argued that the production methods used are highly dependent on the tools used for the analysis. It is generally assumed that there should be as much producer control over the processes executed by the hard- and software tools used in the process.

1. Data Aggregation tools (used for both bottom- up and top- down processes)
   1.1. Data analysis (Database systems)
      1.1.1. +Excel
      1.1.2. + SAS
      1.1.3. +SQL
      1.1.4. ArcGis relational (non- spatial) database system
      1.1.5. Mapinfo relational (non- spatial) database system
      1.1.6. Etc
   1.2. Spatial analysis (GIS)
      1.2.1. ArcGis spatial database system for spatial analysis
      1.2.2. Mapinfo spatial database system for spatial analysis
      1.2.3. Etc.
   1.3. Temporal analysis (Simulation software)
2. Data Disaggregation (Top- down procedures)
   2.1. Spatial disaggregation of

This part of the process is processed with the help of the use of GIS tools. Usually this may-, when processing the information stored in dedicated statistical databases-, be accomplished without any dependence on other (e.g. NMA-) datasets. Such data are used for orientation only. However, NMA datasets may prove very valuable for processes to compensate for the lack of a high resolution object base-. Analysis here is generally concerned with the production of analysis types that are to be considered as parts of a data infrastructure (e.g. the delineations of urban areas). One of the most fundamental system analysis required to demonstrate the hierarchal structure of aggregations needed. This relates to hierarchies of both regular and irregular tessellations and their relationship.

5. DATASET PREPARED FOR DISSEMINATION

This point according to the general outline of the process as defined in paper 1 in this series.

5. Compilation of National, European or Global Geostat dataset(s) for dissemination
   - 5.1. Contents in response to real user needs
     - 5.1.1. Data for public authority use
       - i Indirect action (information required for the design, implementation and evaluation of national laws)
       - ii Direct action (information required for the design, implementation and evaluation of direct action (e.g. the building of physical infrastructures)
     - 5.1.2. Data for private sector use
       - i Data needed e.g. for production process
       - ii Data needed e.g. in the products themselves
   - 5.2. Quality control of the Microdata system
     - 5.2.1. Output (product)
       - i Control of the quality of the dataset (compared with promised specifications)
       - ii Confidentiality control (or compliance with other institution regulations, rules or national laws)
       - iii etc.
     - 5.2.2. Production Process

DATA FOR PUBLIC AUTHORITY USE

Check for use for the planning, design, implementation and evaluation of both direct and indirect public authority action

1. Indirect action (legislation)
2. Direct action (direct action to change MES)

DATA FOR PRIVATE SECTOR USE USE

1. Product
2. Production process

CONFIDENTIALITY CHECK

So much have been said about confidentiality checks. It seems that this is the most difficult part of the process. However, it seems that the best way to approach this is to make one proposal for a chnfidentiality check that we regard as aequate for most situations. This system may then be used for benchmarking and as a foundation for negociating the issue.

COMPLIANCE WITH OTHER RULES AND REQUIREMENTS

This might involve checks concerning compliance with e.g. INSPIRE standards.

6. DISSEMINATION

This point according to the general outline of the process as defined in paper 1 in this series.

6. Dissemination for national and international use
   - 6.1. Spatial data infrastructure (for spatial statistics) for dissemination of statistics over the internet
   - 6.2. Business model
   - 6.3. Quality control of the Dissemination system
     - 6.3.1. Output (product)
     - 6.3.2. Production Process

Dissemination of data is of two types. a) Standard data collections that follows a given specification or standard. b) dataset that are produced for a customer according to varying specifications. Dissemination

also depends on it use in two very different user groups; public authority (e.g. data needed for a GGI or GGIM system) use on the one hand and private use as dominated by specifications

1. Dissemination
   1.1. Analogue dissemination
   1.2. Digital dissemination
      1.2.1. Dissemination over the Internet
         - Proper object structure
         - Spatial data infrastructure (for spatial statistics)
         - Business model
         - Etc.

7. USE (IN RESPONSE TO "REAL" AND "ASSUMED" USER REQUIREMENTS)

This point according to the general outline of the process as defined in paper 1 in this series.

7. Feedback on Real user needs
   7.1. Public authority use
      7.1.1. Direct action
      7.1.2. Indirect action
   7.2. Private sector use
      7.2.1. Data needed for production process
      7.2.2. Data needed in the products themselves
   7.3. Quality control of the Feedback system
      7.3.1. Output (product)
      7.3.2. Production Process

This type of spatial analysis is done in response to specific, not general data requirements. We often refer to these as "Use Cases".that may serve as a general illustration of the use of e.g. spatial statistics.

8. QUALITY- AND OTHER CONTROLS CONTROL

This point according to the general outline of the process as defined in paper 1 in this series.

8. Overall Quality Control
   8.1. Quality control issue by issue
      8.1.1. Quality control of the modelling product and processes
         i Output (product)
         ii Production Process
      8.1.2. Quality control of the data capture processes
         i Output (product)
            - The statistical data tables
         ii Production Process
      8.1.3. Quality control of the Microdata system
         i Output (product)
         ii Production Process
      8.1.4. Quality control of the Microdata system
         i Output (product)
         ii Production Process
      8.1.5. Quality control of the dataset compiled for dissemination
         i Output (product)

- Control of the quality of the dataset (compared with promised specifications)
- Confidentiality control (or compliance with other institution regulations, rules or national laws)
- etc.

ii Production Process

8.1.6. Quality control of the Dissemination system

i Output (product)

ii Production Process

8.1.7. Quality control of the Feedback system

i Output (product)

ii Production Process

8.2. Overall Quality control

8.2.1. Output (product)

8.2.2. Production Process

Every full iteration should end with a quality assessment not only relating to the finished product, but also the whole process

1. Quality controls
   1.1. Quality of Data model
   1.2. Quality of Data capture procedures
   We might consider the quality of the datasets resulting from the use of these methods separately or in a mix (as is generally the case. This will be difficult but not impossible. Anyway it is way beyond the
   1.3. Quality of core population datasets
   The quality of the European Geostat dataset will depend on the degree of harmonisation achieved among the EU27+ meber NSI's. This will mean that in the long perspective we need to agree upon one common datamodel and data structure that will produce a good result. I will call this the "Ideal" National Geostat Dataset. This may be used for benchmarking purposes.
   1.4. Quality of analysis procedures
   1.5. Quality of the dataset prepared for dissemination (The output)
   1.6. Quality of dissemination (The Dissemination dataset etc.)
   1.7. Quality according to Users
2. Quality of evaluation procedure

APPENDIX 1: TOOLS

TOOLS 1: MICRODATA (DATABASE MANAGEMENT SYSTEMS)

This includes all attribute, spatial reference data and timestamps generally in alfa- numeric form. These databases are generally conventional database systems where the time and spatial references are processed with conventional management systems or other tools like:

1. +Microsoft SQL server
2. + Oracle
3. +Sybase
4. Etc.

TOOLS 2: CONVENTIONAL NON- SPATIAL OR TEMPORAL DATA ANALYSIS

Conventional attribute analysis without direct consideration for the spatial or temporal dimension(s) in statistical institutes are done in different computer environments. Some of these are;

1. SAS
2. Statistica
3. Excel
4. Etc.

It is important to note that many GeoStatisticians use some or all of these tools extensively prior to their spatial analysis work. It seems that SAS is often preferred for more heavy work for instance to extract work databases to serve as a foundation for a project. Others will prefer to do this, and all spatial analysis work in the database management environment provided by the GIS software.

The use of scripts to document procedures for the iterative development of production processes are widely used in all of these environments.

TOOLS 3: SPATIAL ANALYSIS

Work with non- spatial and spatial analysis is genrally implemented in dedicated GIS environments. The most popular on the “high end” is ARCH GIS, and on the “middle end” Mapinfo”. In addition a wide array of differnet tools are used. These include open source, public domain and proprietary solutions.

1. Arch GIS
   1.1. Non- spatial database management
   1.2. Spatial database management
2. Mapinfo
   2.1. Non- spatial database management
   2.2. Spatial database management /simulation
3. gvSIG
4. etc.

TOOLS 3: TEMPORAL ANALYSIS (SIMULATIONS ETC.)

(Not discussed here)

APPENDIX 2: COMMENT APPROACHING THE DATA MODEL

SEE THE DOCUMENT "GEOSTAT_2011_DATASET.DOC"

The proposal is to collect and disseminate data per country. The data will also be disseminated in a European dataset. In order to keep both datasets aligned we propose to make the combination GRD_ID and CNTR_CODE the primary key of the data.

The starting point is a European-wide grid net of 1 km^2 cells corresponding to the INSPIRE specifications[1] and covering EU27 + EFTA countries. All grid cells intersecting the landmass of the countries concerned, including all inland waters The grid net can be divided into country nets allowing for easier data handling. The grid net files will be named Grid_ETRS89_LAEA_1K_CC whereby CC is the ISO country code, so Grid_ETRS89_LAEA_1K_SE for Sweden. The European grid net file will be named Grid_ETRS89_LAEA_1K_EU. Those grid nets represent the framework for the integration of national grid data. The actual grid dataset consists of .csv text files with the unique INSPIRE grid cell code as reference to the grid net. The conventions for naming the national grid dataset file and the variables are as follows:

Name of the file	GEOSTAT_grid_POP_1K_CC_YYYY (CC: country code, YYYY: ref. year, e.g. GEOSTAT_grid_POP_1K_SE_2011
Column Names	GRD_ID, METHD_CL, YEAR, POP_TOT

Where:

GRD_ID	Identification code of the grid cell (lower left-hand corner) according to INSPIRE
METHD_CL	Method used for the grid cell; A (aggregated), D (disaggregated) and M (mixed)
YEAR	Reference year of the data
POP_TOT	Population count of the grid cell rounded to integers (in the case of border cells the cell contains the share of the population for the country in CNTR_CODE).
CNTR_CODE	ISO code of the country in which the grid cell is located (in the case of border cells the grid cell is reported from all neighbouring countries with the same GRD_ID but different CNTR_CODE attribute.
DATA_SRC	For national datasets the country code; for the European disaggregated datasets the data source. For border cells a similar approach as for CNTR_CODE is adopted.

[1] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_SU_v3.0rc2..pdf.

Cells with zero total population are preserved (POP_TOT = 0), for example:

GRD_ID;METH_CL;YEAR;POP_TOT;CNTR_CODE;DATA_SRC
...
1kmN4101E4453;A;2006;**0**;NO;NO
...

Regarding the METHD_CL attribute for each grid cell, there can be various combinations of actual methods combining aggregation, disaggregation and even other estimation methods. For the sake of simplicity the goal is to differentiate cells which have been merely disaggregated using a dasymetric approach ('D'), from those grid cells which are simply 'point in polygon counts' based on detailed georeferenced source data ('A') and from those which apply various methods and data sources to estimate and model the population figure ('M'). For further details metadata will be used.

As an illustration, a shared border cell between Sweden will be reported by SE as:

GRD_ID	1kmN4101E4453
METHD_CL	A
YEAR	2011
POP_TOT	2
CNTR_CODE	SE
DATA_SRC	SE

and by NO as:

GRD_ID	1kmN4101E4453
METHD_CL	A
YEAR	2011
POP_TOT	4
CNTR_CODE	NO
DATA_SRC	NO

In the integrated European dataset organised in line records per grid cell the data will look as follows:

GRD_ID;METH_CL;YEAR;POP_TOT;CNTR_CODE;DATA_SRC
...
1kmN4101E4453;A;2006;4;NO;NO
1kmN4101E4453;A;2006;2;SE;SE