WORKING PAPERS AND STUDIES

ISBN 92-894-3430-9
ISSN 1725-0595
Cat. No. KS-AN-02-001-EN-N

2002 EDITION

TANDEM_GIS I

A (feasibility) study towards a common geographical base for statistics across the European Union

1

EUROPEAN COMMISSION

eurostat

THEME 1
General statistics

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (http://europa.eu.int).

Luxembourg: Office for Official Publications of the European Communities, 2002

ISBN 92-894-3430-9

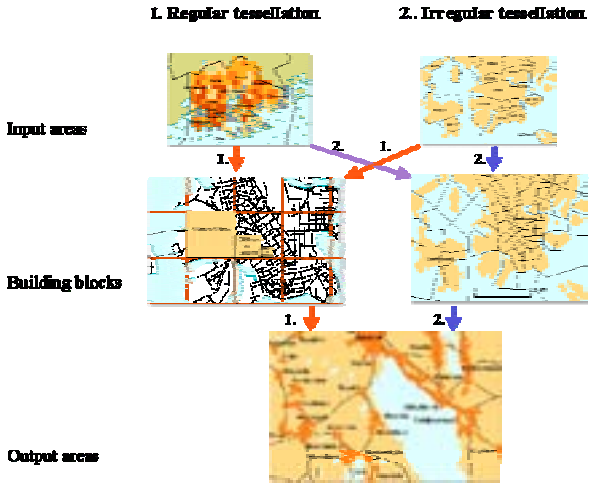
INDEX

The Tandem Consortium

TANDEM_GIS I

A (feasibility) study towards a common geographical base for statistics across the European Union.

Lars H. Backer[1], Statistics Sweden
Marja Tammilehto-Luode[2], Statistics Finland
Philippe Guiblin[3], The Office of National Statistics (UK)



Abstract

This is the final report of the Tandem project for studying the feasibility of improving comparable statistical territorial data to be used as a foundation for a system of small area statistics for the European Union.

---

[1] Lars H. Backer is a senior GIS- R+D- expert, since 1992 connected to the program for regional planning and natural resources at Statistics Sweden. He holds graduate and postgraduate degrees in architecture, urban- and regional- planning from ETH (Zurich) and Nordic universities (Oslo, Stockholm). Mr. Backer is specialised in the construction and use of static and dynamic digital models for urban and regional development. He has extensive International experience from Teaching, R+D and consultancy in his field.

[2] Marja Tammilehto-Luode is a Senior Adviser at Statistics Finland (SF). She is responsible for developing regional statistics and geographic information systems at SF. She has worked as head of the GIS team since 1992. Her work areas have spanned from GIS policy issues to new information technology application. Mrs. Tammilehto-Luode has extensive international experience in working groups on GIS, regional and urban statistics. She holds a M.Sc. in Geography.

[3] Philippe Guiblin is a Statistician Methodologist at the Office for National Statistics (UK), since 2000. He completed a Ph.D. in Geostatistics in 1997 (École des Mines de Paris, France) in which he explored different ways of applying the geostatistical methods to fish stock assessments within an EU funded project. He is now involved in several ONS methodological projects covering spatial modelling issues, spatial statistics and small area statistics.

The Tandem Consortium

The report is divided into 3 parts preceded by a "Management Summary".

- The first part (Part 1) is an effort to analyse the professional context relating to the current problem with a double objective. The objective of this section is first to make the conceptual description of a "vision" into a possible solution to the current problem. This concept is to be based on a discussion of some central aspects of the professional context. The second, central objective of this part is to describe a method for implementing it.
- The second part (Part 2) is an effort to build and test a "prototype" for a system of small statistical areas. The study suggests alternative methods for constructing comparable building blocks with the first practical responses of a combined grid-based and region-based approach. A case study focused on the delimitation of urban areas for the Cardiff region (UK) and the Helsinki region (Finland) has been carried out to not only illustrate and compare the results but to also meet the demand for statistics on such “structural” and “functional” regions.
- The third part (Part 3) is an effort to summarise the results of the first part and to formulate a set of recommendations for future work.

The figures mentioned in chapters 2.3.2 and 2.3.4 have for technical reasons been taken out. They can be obtained from the GISCO team in Eurostat (anette.bjoernsson@cec.eu.int). The references to the figures are kept in the report as footnotes in order to avoid distortions of the reading.

The Tandem Project

Knowledge in the form of an informational commodity indispensable to productive power is already, and will continue to be a major, perhaps the major stake in the world-wide competition for power. It is conceivable that the nation-states will one day fight for control over information, just as they battled in the past for control over territory, and afterwards for the of access to and exploitation of raw materials and cheap labour. A new field opened for industrial and commercial strategies on one hand, and political and military strategies on the other

Jean Francois Lyotard[4]

"Those who explore an unknown world are travellers without a map; The map is the result of the exploration. The position of their destination is not known to them, and the direct path that leads to it is not yet made."

Hideki Yukawa[5]

Of course, The entire effort is to put oneself
Outside the ordinary range
Of what are called statistics.

Stephen Spender[6]

An "after"-word

This is a report from an "emergent" project in progress. As Hideki Yukava so pointedly puts it we have been exploring an unknown world. The current report is the result of the exploration when the direct path to the result was not known. Looking back, we see that the road was not by any means straight according to linear logic. After the first phase in this quest for a system of small statistical areas, we have to admit that the results are not as clearly stated, and the arguments not as well put we would have preferred. However, due to iterative method we have chosen, we have reason to believe that everything will improve considerably with subsequent iterations.
At the end of the first "turn of the wheel", the Tandem effort to test the feasibility for creating a system of small statistical areas for EU we believe it is valuable to return to the initial question: Why bother?

Opportunities

There are many arguments in favour of this endeavour for this but we will focus on three points

1. First is the argument that statistical information on crude systems of administrative areas is not very useful as a foundation for action or designs for action as development planning and the like. The reason for this is mainly due to a lack of resolution, but also a result of the lack of comparability between such regions or "irregular tessellations" in terms of both size and the "population" of observations. A system of small areas designed as a system of small area statistics promises to meet the demand for better information for projects to explore opportunities and counter threats.
2. Secondly, a system of small statistical areas may greatly improve the comparability of statistics across the EU and open the doors for more and

[4] (Lyotard 1979) *The Post-modern Condition: A Report on Knowledge*
[5] (Hall 1993) *Mapping the next Millennium*
[6] As Quoted in (Gleick 1987) *Chaos; Making a New Science*

better aggregations, dis-aggregations, and re-aggregations of data in a network of systems of statistical areas.

3. Thirdly a system of small area statistics may prove very valuable to produce better high resolution data sets by improving the quality of high resolution data sets when used in combination with methods like sampling, small area estimation and the like.

Risks

Well then, this sounds fine, but ever student of dialectics will know that every coin has two sides so what about the risks?

1. The main risk for this type of projects is that it does not deal with an established problem where well known and well-tried solutions are ready at hand and applicable. We are here confronting an emerging problem whose feasibility may only be established when we have a solution. The main risk is therefor that which is always connected with R+D work, and described so well by the Touring principle that you do not know whether there is a solution to a mathematical problem, or how long it will take to solve it before you have tried.
2. However there are some projects that are more likely to succeed than others. The risk here seems largely to depend on the soundness of the premises for this project. How far are the vision, the prototype and the method selected to produce the desired results appropriate for the task?

We believe that we have reason to believe that we have done what is needed to reduce the risks for failure on both of these scores.

Conclusion

Although we argue that in a question like this, where we are dealing with an emergent problem, we cannot suggest a better method to test the feasibility of a system of small statistical areas and a system of small area statistics is to start building one.

We do not know for sure that we will succeed, or in what period of time, but we have reason to believe that this is by no means a "mission impossible" We believe that a solution is at hand, and we have made considerable progress on a path to find it.

The project co-ordinator

Lars H. Backer
For the Tandem consortium

The Tandem Consortium

Tables of contents

Figures and Tables

List of figures

List of Tables

A management summary

Lars H. Backer, Statistics Sweden
Marja Tammilehto-Luode, Statistics Finland
Philippe Guiblin, The Office of National Statistics (UK)

Abstract

This section contains a short summary of the report from the "Tandem" projects inquiry into the feasibility of developing a system of small statistical areas based on existing infrastructures planned or used by NSI's. The consortium believes that a project to this end is both feasible and desirable and recommends the commission to encourage further iterations in order to extend and improve the existing embryo into a full fledged prototype for a system of small area statistics for Europe.

Introduction

Over the recent years most NSI's have noticed an increasing demand for high quality statistics, with higher resolution, disseminated with increasingly higher frequencies and harmonised over ever larger areas.
Eurostat/GISCO has for some time argued for the need for "Basic statistical areas" that could be used to improve the resolution and comparability of "area-based" statistics within the EU. 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[7].
In response to papers submitted by United Kingdom[8], Sweden and Finland[9] 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 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 were invited to apply for a commission grant to study these questions further.
The demand for a new geographical base for statistics in Europe is not only emerged from a general need to improve the quality of the classical system(s) of official statistics. The main reason is an answer to recent increasing importance of information and knowledge systems in the new process oriented methods adopted in the production, distribution and consumption of goods and services in our societies. These methods are spreading at a quick rate to all institutions and organisations that act through hierarchies of "projects".
Projects contributing to the development of societies are no longer limited to the development of infrastructures within administrative borders, but has

[7] See (Rase 2000) *Technical Specifications; Improving comparability of statistical data across EU Member States*
[8] See (Wagget 1999, (20-21 October)) *Towards improved statistical comparability across Member states- A better geographical framework*
[9] See (Backer, Tammilehto-Luode and Rogstad 1999) *The Use of Grids to Improve the Comparability of Statistical Data.*

been forced to shift their focus to the development of networks whose output patterns are not satisfactorily captured by crude systems of "large area statistics".
This and other developments strengthen the demand for a new system of "small area statistics" for the EU. This system should not be regarded as substitute for the classical system of "large area statistics" but should rather be seen as a valuable extension to, or upgrading of, existing practices.

Problem statement

This project is not concerned with "large area statistics" as those aggregated on the NUTS hierarchy of administrative units but will focus on the feasibility of providing a system of "small area statistics" for the EU.
The critical part of such a system is to find areas that in their scale are close to that used for data capture, and therefore not far from that of man and his artefacts. It seems self-evident that a system of small area statistics for EU may not be built from scratch. We have therefore taken as our first basic assumption that the new system must be developed from an "ad hoc" solution compiled from existing systems of small statistical areas planned or already used by NSI's.
It is well known that there exists in Europe two traditions, or rather practices used for data capture and analysis; point and area based methods. The "Tandem" consortium derives its name from our second basic assumption stating that a system of small statistical areas must necessarily consist of two parallel and complimentary systems of input areas (Building blocks, (BB)), a system of regular and a system of irregular tesselations. For both of these systems we believe that we have formulated four fundamental requirements.

1. They should be as small as possible
2. They should be comparable in terms of population and/or area
3. They should eventually have a complete coverage across the EU
4. They should be linked to an adequate amount of statistics.

The main objection to the use of any system of even small areas for spatial analysis has been that they as building blocks are useless without pre-processing. In this process BB's are clustered to form systems output areas (OA's). Each system of OA's is unique in regard to the data set processed and the problem addressed. It is therefore our third basic assumption that a system for small area statistics must also consist of a series of methods dedicated to the production of output areas by clustering and/or aggregating statistical building blocks.

Method

Given an ad-hoc relatively heterogeneous system of small statistical areas it is imperative to design a process that will gradually transform this embryo into a well functioning system responding to user needs. This situation is not by any means new and we therefore suggest adopting a well-established iterative method for research and development.

1. Study user needs (Case studies)
2. Formulate (Re-formulate) a vision (The desirable)
3. Design (Re-Design) of a prototype
4. Test and evaluation (Searching the feasible)
5. Recommendations (for next iterations)

We have in the current project finished one full turn of this "wheel".

Results

It will not come as a surprise that the Tandem consortium has yet to produce a well functioning prototype to a system of small area statistics for Europe. However, we have produced a first embryo to this end by employing the method above.

1. User needs (Case studies)

In this case, where we are discussing a non mass-product that does not exist, it does not make sense to study user needs with marketing methods. User needs in this and similar cases are believed to be best explored by the use of "case studies" or "case study-" techniques. In the current iteration the consortium have built the first version of a prototype that may be used to delimit urban areas by clustering input areas with high population densities according to a crude case study based on a study of needs for information in the ESDP[10]. From this not very well-researched "case study" we have arrived at the conclusion that a well functioning system of small area statistics might offer users three major promises:

1. Better data
   - Better small area estimations.
   - Better sampling.
2. Better aggregations and dis-aggregations.
   - Statistics on functional and structural as well as administrative regions
3. Better infrastructure for doing spatial analysis.
   - Systems of high-resolution small statistical areas consisting of both regular and irregular tessellations bringing spatial analysis beyond the production of thematic maps on administrative areas.

It is frequently argued that the "project" is the basic unit for action within both public and private organisations. Projects vary considerably in terms of form as well as function from organisation to organisation. In principle however, they are either designed to exploit opportunities or counter threats. In our case this may be illustrated by typical situations that are acted upon with the support of geo-statistics;

1. Exploit (create) opportunities (Urban and rural (spatial) development)
   - Social integration
   - Urban-Rural Infrastructures
   - Water management
2. Counter threats (Catastrophe management)
   - Social conflicts
   - Infrastructure failures
   - Environmental catastrophes

In all these cases there will be many different actual and potential user groups within public institutions and business/industrial enterprises that might profit from the use of statistical information in their projects. Here there are two obvious users and uses.

1. Serve projects (Forward information for projects)
2. Evaluating results (Feedback information on projects)

For historical reasons NSI's have focused on "accounting" statistics to serve for the evaluation of past policies rather than providing information to initiate

[10] European Spatial Development Perspectives (ESDP)

and implement new "engineering" projects. Currently however, statistical systems are increasingly called upon to serve users with qualified information for both of these perspectives.

2. A Vision (A design representing the desirable)

There has been a lot of discussion around the need for a system of small area statistics in recent years, but very little consensus as to what is implied with this notion. We have therefore, based on an inquiry into the professional context connected with this topic made an initial tentative definition of the term "small area statistics" and "small statistical areas".
A system of small area statistics (SSAS) is:

1. A system of knowledge built from statistical micro data, designed as an "open" system of data (statistics), geographical features (regular and irregular tessellations) and methods (manual and automatic processes). A SSAS consists of, and is in turn itself a part of and designed to fit into, a constantly changing hierarchical network of processes dedicated to the production and analysis of qualified spatial information.
2. A system of small area statistics (SSAS) processes information both on- (aggregations, benchmarks) and for- (data- spatial- and temporal-analysis) to improve the results of overriding projects to counter threats and exploit opportunities in view of private and collective efforts to improve the human condition.

This crude definition is not yet self-explanatory but will have to serve until we have time to make it more short and precise. It consists of two parts, the first describing "formal" aspects and the other "functional".

3. A Prototype (An embryo to a solution)

The data used for the development of the "Tandem" prototype focuses on "urban-rural" questions is described in the case study using data from Finland (Helsinki) and UK (Cardiff). In the current project this is a very simple data set consisting simply of the sum of natural persons registered to the system of input areas.
We have then compiled a system of regular and irregular tessellations (Building blocks) linked to the statistics as required by the case study and methods needed for their manipulation.
We have a well-tried method for clustering and/or aggregating input areas (Building Blocks) by employing methods derived from different sources depending on the problems stated by the case studies.

4. A method for test and evaluation (Defining the feasible)

We will in further iterations develop a system of benchmarking processes that may be used to test and evaluate results. In the current tentative phase of the project we have limited our evaluation to non-formalised professional critical appraisals.

Recommendations

1. Regarding user needs (Case study)
   We recommend to expand and deepen our knowledge of "user needs" with further inquires into issues related to spatial development with a focus on urban areas.
2. Regarding the design (The Definition)
   We recommend that the design for a system of small area statistics is

based on the draft for the "Tandem" definition for a system of small area statistics, that is to be further developed in response to the results from future iterations.

3. Regarding the prototype (Features, data and methods).
   We recommend encouraging the further development of the "Tandem" embryo into a proper prototype to an open SSSA for Europe, hereby using existing components (statistics, systems of both regular and irregular tessellations and methods), structured according to object methods and technology.
4. Regarding further development of the prototype
   We recommend to develop and expand, the open Prototype for a SSSA according to an iterative R+D method that allows all suggested improvements to prototype be tested in the perspective of changing user needs and improvements to the design.

Lars H. Backer, Statistics Sweden

Abstract

The purpose for writing a Professional context for a professional paper like this is twofold. First of all it is to underline the importance of the "World view" the author(s) have adopted for the current study and secondly to provide a method in accordance with this perspective.

1. On knowledge and development
   In this chapter we are considering systems of small area statistics as a “Experiment” or a system of knowledge where all key information needed to understand and develop this tool is discussed and accumulated over time. The structure considered is not taken for granted, but developed as a consequence of an inquiry into modes of information production that are emerging as we develop the informational society.
2. On Processes
   As a consequence the inquiries discussed under the topic “Knowledge and development” this section is exploring consequences of the new modes of production that is emerging as a consequence of the recent change of paradigm.
3. On "Case studies" and user needs
   This section puts the user’s need in the focus for any discussions related to the “Form” and “Function” of a future system of small area statistics. It is argued here that the only effective method at hand is to study these (user needs) through a series of well researched “case studies” or case studies. In the current project we are studying the commotion’s need for qualified information to support projects aiming at spatial development.
4. On systems of small statistical areas
   In this section we discuss the fact that the single most important factor needed for the production and analysis of high-resolution spatial information is a system of small statistical areas.
5. On systems of small area statistics (a vision)
   As a first summary of the sections (1-4) above, we have formulated a preliminary description, or “definition” of the term small area statistics. Though long and not very elegantly formulated, we hope that this description may serve as a conceptual “Design” for a system of small areas statistics that may be used as a guide-line for the building and development of a “real” working prototype.
6. On a Method to this end
   As a second summary this section focuses on the second objective for this section; to suggest an effective method to develop a working prototype/product from the given point of departure. For this purpose we have suggested version of the iterative R+D method adopted in all kinds of development projects dealing with “emerging” problems, similar to those discussed here.

1.1 A Professional Context (WP 1.0)

1.1.1 Introduction

In order to help the reader we would like to make some comments about the current section (Part 1) that is dedicated to the professional context.
The goal or objective of this part of the work was to describe the task in its context(s). It will come a surprise to some, if not to the professional information engineers that the current section contains frequent reference to concepts like "information systems and networks", "object technology" or "processes" and "value chains". We will first try to explain why.
At the outset, the problem was quite simple. Just study the feasibility for building a system of small statistical areas for Europe that may serve to improve the comparability of statistics.
Gradually this simple scheme has become more complicated due to a series of challenges.

1. A future system of small area statistics
   First of all it was realised that in order to be useful a system of small statistical must be seen in relation to its immediate context: a system of small area statistics. In this context we are not only talking about a system of geographical features with statistics but also referring to a system of processes to generate a given set of structural and functional patterns of interest to our customers, in addition to data on administrative areas as customary today.
   In a future situation it will be necessary to depend on data from NSI's on a level of higher resolution than is produced and delivered to Eurostat at present. In order to function the system must therefore depend on a series of standard methods representing "best practices" that will be used by all NSI's and sent Eurostat as "pieces in a puzzle".
2. Demands for International standardisation work.
   Secondly, due to the fact that in order to be useful our statistical data must fit together with a host of other types of information in order to function in modern processes for information production (as for instance decision support systems related to development issues, or catastrophe management). Here we seem to be forced to co-operate with organisations like the "Open Gis Consortium" in order to define our data in harmony with those of other data producers. This means, we believe, that our data must be defined described according to "object methods" according to UML standards.
3. Exploit established methods
   Thirdly, it will seem evident to anyone involved in the production of a system for processing and refinement of data, that it makes sense to follow the state of the art as implemented by the majority of the information industry. Seen in this perspective it will be beyond doubt that a system of small area statistics should be designed and structured according a "process" view of the information society and based on "object" methods for data modelling and a process strategy. By choosing this path, we may not only produce a system who's output will be compatible with other information products (see above) but we may also

profit from a sea of expertise that has been developed in this field over recent years.

1.1.1.1 A point of departure

Within the context of the "Tandem Projects" it seems that 2 main goals or ends are of central importance for our efforts to develop a better geographical base for statistics in Europe.

1. A vision
   First of all we need a vision of a system of small statistical areas and a system of small area statistics. In a way this definition will have to serve as a hypothetical or conceptual prototype until we have a working "practical" solution.
2. A Method
   Secondly we are in need of a reliable method that will help us to systematically design, build and develop a working prototype from the given premises.
   Here we do not only mean a method, that may be used to test the feasibility of an effort to build a system for "small area statistics" for EU, but rather a method that may be used for the systematic development of the geo-statistical system in Europe in general.

The current section (Part 1) will focus on this end.

1.1.1.2 A better statistical system for Europe

The current project is intended to contribute to the development of an emerging statistical system for the EU. This statistical system will be directly dependent on the success of two sub processes.

- Contribute to the integration of the European statistical systems.
- Contribute to the development of an integrated European statistical system.

If we would choose to disregard the idea of a paradigm shift as described above it would only be possible to argue in favour of the current project in terms of an improvement of the classical "accountant" type statistical systems. According to the vocabulary used to describe societies in the industrialised area "Small area statistics" would only provide a better resolution for "benchmarking" type statistics and nothing more. This however will not suffice at present. "Small area statistics" are rather an answer to the need for better use of the information available in statistical databases for the projects that are designed to deal with threats and exploit opportunities.
In the old paradigm "accountant" type statistics geographical reference and spatial analysis are not of vital importance to the statistical system. With the new situation the statistical system(s) will have to answer to the need for this type of information or leave the task to others and thus fail to respond to a major opportunity.
We have with other words, in the current historical situation, to consider a more complex state of affairs than just adding a component to an existing stable structure, we have in the previous section explored the consequences of the emergence of the "Informational" society, for the statistical systems of Europe.
In the section dedicated to “knowledge and development” below we will argue that the production of statistics within the new paradigm is intimately connected to the development of processes rather than structures. In fact it

seems to make much more sense to regard statistical systems as processes rather than structures.
We have also argued that a statistical system, according to this new perspective may be seen as a feedback system of the type that is well known from systems theory (theories). As such it consists in its simplest form of two processes; a forward process and a feedback process.
In the production of development in a society, the forward processes are all macro processes that contribute to the production, distribution and consumption of goods and services, and the feedback process all benchmarking efforts constructed to measure its performance. Classical statistics are in this perspective closely related to the benchmarking idea.
The tandem project(s) are direct reflection of this new state of affairs it focuses on the task of improving our ability to provide qualified information needed to describe and explain complex processes that we (through our customers) have to understand in order to cope with problems and exploit opportunities.
This does not mean that the classical "benchmarking" type of statistics accumulated on administrative areas has become obsolete. On the contrary they are as important as ever.

1.1.1.3 A Geographical base for statistics

The paradigm change described here has affected all aspects of our production, distribution, consumption and recycling of goods and services in modern society, also including statistics.
In the old days of classical modernism, before the breakthrough of the new paradigm, statistics did not primarily serve the production processes themselves but limited its field of work to making inventories according to methods copied from "accountants" registering the output from a society conceived as a "machine".
With the emergence of the new paradigm all processes seem to accelerate and primary concern moved from the relatively stable systems to those that change fast. Where we before seemed satisfied with "input-output" models and planning systems, we now need dynamic models and interactive cybernetic approaches to achieve "development"
We suggest here that any “modern” statistical system will have to consist of two central components. A “horizontal” system of information, based on what is generally described as “small area statistics”, and a “vertical” “benchmarking” system of information used to measure the “state” of complex systems at given points in time.
In this perspective traditional statistics are primarily used and produced for benchmarking purposes. These structures and processes are therefore generally well developed. We have, on the other hand, less developed structures and processes that may be used to describe and explain many of the important large-scale processes that demand public attention. These may be very different ranging from natural catastrophes (Floods in Poland, FMD in United Kingdom), inadequacies and breakdown in the man-made environment (Urban development problems, provision of energy, water and the disposal of wastes) etc. or vital changes in socio-economic and cultural structures and processes. All these are intimately connected to geography.

1.1.1.4 Three opportunities

4. First is the argument that statistical information on crude systems of administrative areas is not very useful as a foundation for action or designs for action as development planning and the like. The reason for this is mainly due to a lack of resolution, but also a result of the lack of comparability between such regions or "irregular tessellations" in terms of both size and the "population" of observations. A system of small areas designed as a system of small area statistics promises to meet the demand for better information for projects to explore opportunities and counter threats.
5. Secondly, a system of small statistical areas may greatly improve the comparability of statistics across the EU and open the doors for more and better aggregations, dis-aggregations, and re-aggregations of data in a network of systems of statistical areas.
6. Thirdly a system of small area statistics may prove very valuable to produce better high resolution data sets by improving the quality of high resolution data sets when used in combination with methods like sampling, small area estimation and the like.

Stockholm August 2001

Lars H. Backer
Statistics Sweden

1.2 On Knowledge and Development

Abstract

This chapter makes an effort to describe the value of a dynamic reference system that constantly adapts to changing conditions in the systems environment (or context). It also points to the fact the we have radically changed our use of such information systems as a result of the digital revolution that have catapulted us into the informational society with its risks and opportunities.
For our purposes in the current project this means that we must learn to see the statistical system no longer as a structure of data aggregated on different levels but rather as a system of processes that produce such information.

1.2.1 Introduction

It seems essential, in the context of this as in any ambitious project, to reflect upon the general context, or the overriding purpose the contribution strives to serve.
The European project as a whole, carries at its core an ambition to contribute to collective and individual actions trying to prevent and relieve the suffering brought about by centuries of wars and conflicts.
The belief in the need for a shared European "project" that if not may replace traditional national (or nationalistic) projects, with a project that will focus on an effort to realise common objectives for the region as a whole. This idea rests firmly on the idea (that we will touch later), that the whole is (much) more than a sum of the parts.
Formulated in positive terms it seems proper to coin the ordeal ahead of us as an effort to develop our common cultural heritage into a genuine pan-European civilisation[11] that may serve as a model for other similar regions elsewhere.

1.2.2 Development

Development is generally taken to refer to changes in a society in the direction of that society's conscious or unconscious "projects". Traditionally societies relate to the world through two "overriding projects " in their efforts to develop their culture.

- A material "Survival" project (Having)
  The first would regard "suffering" as a "materialistic" problem of "survival" with no existential implication. This strategy would stress the importance of materialistic consumption, and regard society mainly as a means to achieve material welfare. This strategy is firmly dependent on knowledge produced and practised to a high degree through science and technology.
- A cultural "Identity" project (Being)
  The first strategy regards "suffering" as an existential problem to which no satisfactory "materialistic" solution may be found. This strategy would

[11] According to the Webster-Merian Dictionary we find: civ·i·li·za·tion: 1 a : a relatively high level of cultural and technological development; specifically : the stage of cultural development at which writing and the keeping of written records is attained b : the culture characteristic of a particular time or place, 2 : the process of becoming civilized, 3 a : refinement of thought, manners, or taste b : a situation of urban comfort
For a further discussion of the attainment of civilisation as a goal for human action please see the chapter "A changing vocabulary " in (Braudel 1993) *A History of Civilizations*

stress the importance of truth, justice, identity, social competence, refinement of thought, manners and taste and other values studied by the humanities.

When living bellow subsistence levels we would however expect that the problem of "having" would be of central importance, and when well above we should expect a greater emphasis on problems of "being". Most cultures however are not radical in their relation to either of these projects, but seem to develop both together.

Manuel Castells, the famous author of the influential work "The Information age: Economy, Society and Culture" noted in an interview that he found it alarming that the European Union put so much emphasis on the "Survival" or production project and had relatively little to show in terms of cultural integration.

Whether or not this materialistic orientation will remain dominant or that we will also stress other aspects of human existence as the European project evolves, will depend on how we succeed in our efforts to integrate a large number of European cultures into one civilisation. And this goal is not possible to reach unless we learn to share if not identical so at least compatible "views of the world".

1.2.3 The role of Knowledge

Independent of how the European project in the coming years will focus on "survival" or the "identity", we as statisticians will have to deliver a not insignificant part of the information needed to develop the region according to our overriding projects.

It is a well documented fact the we, as human beings all spend most of our lives in a constant and more or less conscious effort to build and maintain a model of our world in our minds that we use as a frame of reference for all actions. Without a working model of this kind, we would meet with all kinds of difficulties in our dealings with the world around us[12]. One of the most important carriers of such collective information is of course language. Similarly societies and cultures may be described as groups of human being that share mental models of this kind. "Outsiders" are thus judged according to lacking capability to respect this share worldview.

As the world "stirs and changes" this view of the world stirs and changes too" Occasionally some things are dropped and new things imported in processes that extend over generations.

For most of human history it seems that cultures worked on these models of reference more or less consciously, due to the fact that they are generally locked up in our minds and are only indirectly revealed through our speech and behaviour.

We have a long tradition, especially since the emergence of the art of writing, for trying to express the foundations of our cultures in words and letters. With the discovery of the scientific method, however, it gradually became clear that we could profit very much from consciously building a more qualified collectively shared model that mirrors the world around us.

[12] As a matter of fact, many mental disorders may be described as disorders of this type. See for instance (Bandler and Grinder 1975) *The Structure of Magic; a book about language and Therapy*

It was with great clarity shown that that there is no method to achieve development more effective than to design and execute actions by using reason and information produced according to the scientific method
Differently put, this means that ideally, it was discovered that ideally all action should be controlled by reason and based on qualified knowledge. We, as producers of qualified knowledge cannot influence our customer's rationality, but we may guarantee that they are provided with qualified knowledge. We may help our customers to build better reference systems to improve their actions.

1.2.4 A Paradigm shift

It remains without doubt that the European cultural heritage at least as it has manifested itself in the last century has been dominated by materialistic values. For most if not all of the current member states, it has been taken for granted that the main project for a "modern" is not to develop its cultural identity, but rather to change it into an effective machine for the efficient production, distribution and consumption of goods and services.
At the present stage in history we seem to be conscious that "development has not been linear, but has progressed over a series of more or less dramatic breakthroughs. Against this background it seems evident for many generally critical observers that we presently may look back upon a recent change in this tradition, a change that is so profound as to deserve the notion of a "paradigm shift[13]"

- The first Industrial Revolution (industrial society)
  During the first industrial revolution the focus was on the advantages gained by the development of hardware in the sense it has been described using the “machine” metaphor. This strategy, implemented in several stages over the past centuries, has resulted in a spectacular growth in scores of industrial nations. By the 1970’ies it is believed that this potential was spent, or was giving way to the idea that it was more productive to explore the improvement of processes controlled with information systems using computer technologies.
- The second industrial revolution (Informational society)
  The breakthrough of what we now call the second industrial revolution followed in the wake of the development of the digital computer and its integration into all aspects of modern life. The ideological foundation for this change had however emerged quite early in the century as a consequence primarily of the necessity to transcend classical physics[14].
  This paradigm shift involved the ability to add systems that process information to the machines that previously where limited to the processing of matter/energy. The industrial and other production systems produced using this technology are no longer mere extensions of our physical capabilities but our mental capabilities as well.

[13] For the discussion of this concept see (Kuhn 1962) *The Structure of Scientific Revolutions*
[14] For relections on this change please see my paper (Backer 1997) *Towards an integration of Space, Time and Statistics*

1.2.5 Systems theories[15]

One of the key differences between the applied and the theoretical sciences in the last half of the 20th century was a reluctance to drop the Cartesian idea of science being able to make "true" statements. Presently however it is generally accepted that all scientific concepts and theories can never be more than limited and approximate. This insight is crucial to all modern science, and makes it possible to turn systems thinking into science.
As a consequence of the paradigm shift above, we seem to be shifting our metaphor from a mechanistic "kit of parts" perspective compared to "machines", to a dynamic "integrated whole" perspective compared to the behaviour of "living systems[16]". This shift may also be described as a shift from a mechanistic thinking to a "systems" thinking.
In the new systems perspective there are 3 main characteristic of these "living systems:

- The first characteristic and the first criterion for describing and explaining the behaviour of "living systems" is a shift of focus from the parts to the whole. The living systems that are generally in the focus of our studies, are integrated wholes whose properties cannot be reduced to those of its parts.
- The second characteristic is the need to shift attention back and forth between system levels. The reason for this is the idea that all systems may be seen a parts of a hierarchy of systems. Everywhere we find systems embedded within other systems
- The third characteristic is a change in the relation between "objects" and "processes". Systems science has shown that due to the first characteristic, living systems cannot be understood by analysis of its parts. Instead a contextual approach is needed that focus on the idea that what we call a part is merely a pattern of relationships. "In the new systems thinking the metaphor of knowledge as a building is being replaced by that of a network" writes Capra in his "The Web of life."[17] As we perceive reality as a network of relationships, our descriptions too form an interconnected network of concepts and models in which there are no foundations.

1.2.6 From structures to processes

This man-environmental system that we want to "civilise" is, however, complex beyond human understanding and it is extremely difficult to approach this project in a rational manner without simplifications.
Any system may be described from two points of view; either as a system in terms of its parts (structure) or in terms of its processes. Nature on the one hand, may either be described in terms of the many structural forms with which she manifests herself, or be explained in terms of the wide range of underlying ecological and other processes. Culture on the other hand may be

[15] The thoughts reflecting a summary of later developments in systems thinking rests heavily on the excellent (Capra 1996) *The Web of life; A New Scientific Understanding of Living Systems*
[16] For a reference to the concept of "Living systems" please consult the classic (Miller 1978) *Living Systems*
[17] See (Capra 1996) *The Web of life; A New Scientific Understanding of Living Systems*.

described in terms of the numerous artificial structures, or explained in terms of the complex processes that produced them.
In the first industrial revolution the focus was quite naturally on the physical structures leaving all control over production processes to human beings.
Now the production processes sometimes integrated with physical structures came into focus as the most interesting field for improving productivity.
All processes of this type where described and often controlled by structural models and functional programs in computers. The production of information soon became many times over the most valuable part of any production process.
In this perspective, Processes controlled by knowledge and information takes a clear precedence over structures (hardware) thus adding new meaning to the functionalist device of classical modernism. “Form follows function”

1.2.7 An "Object" Approach

The idea that every system should be seen from two perspectives simultaneously is fulfilled in the "object" approach to information processing that now is the "state of the art" in all kinds of programming situations. Here every "object" may be either be called upon to reveal its properties or it may be called upon to respond with its "methods" or processes that serve as "actions" that change something in its environment.

1.2.8 Action

"Nulla est causa philosophandi, nisi ut beatus sit" (there is no reason to reason without a view to happiness) say St. Augustin. If this reflects some truth then the main reason to engage in the current efforts is based on the belief that this information may lead to an improvement of the "human condition".
In a very general form any project or conscious collective action may be expressed as a recognition of a given situation as offering a series of threats or opportunities (seem in relation to our overriding "quest"). Based on this "insight" a new realistic "state" of the system is imagined and formulated as a goal for action. On this foundation a series of actions are planned that we have reason to believe will bring our "condition" closer to our preferred "state" of affairs. Then the action is implemented and the result observed and evaluated

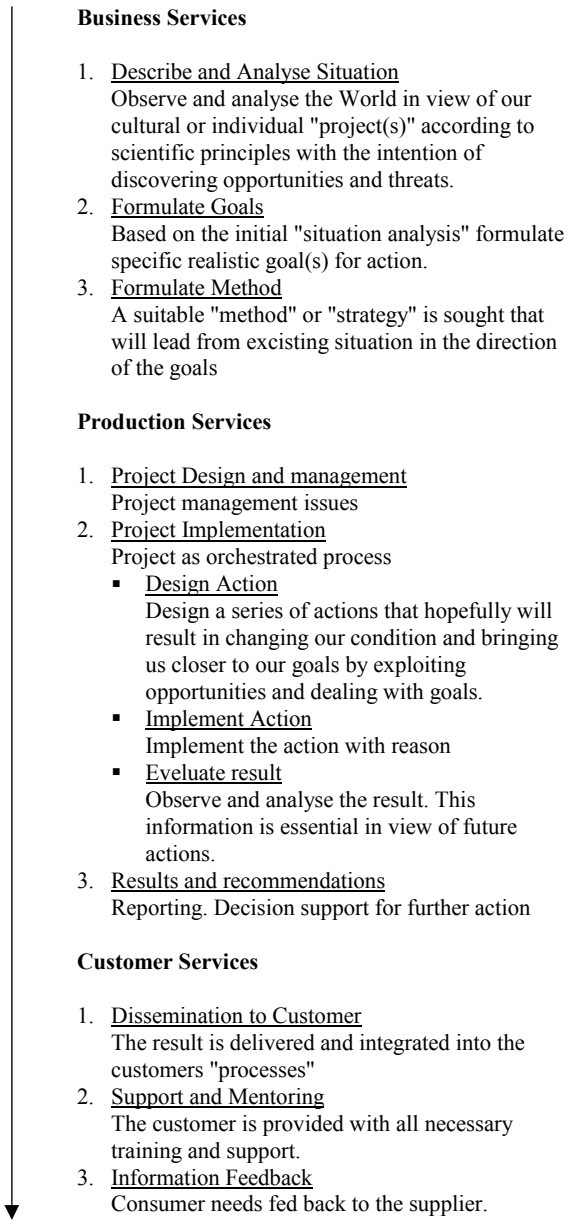


Figure 1: Business process

If the "Actor", that may be either a natural- or juridical-person, is a large institution or business Enterprise, its Activities may be seen as a system of large or small "vectors" with a flow of its processes going from the top down. In the case of systems that primarily process information, as is generally the case at the NSI's whose work is the issue here, the flows signify processes where information is transformed from "raw data" to information products. The goal here is naturally to deliver results that may significantly improve the performance of the customer's processes and the quality of the product he delivers to his customers "downstream". Hence these "vectors" may be rightly called "value chains". As the value of the thing processed increases as it flows on.

The theoretical sciences provide us with general information, and general models of the world that may be applied to most situations. The applied

sciences on the other hand strive to build all kinds of practical applications that apply these principles to specific situations.
It has frequently been argued that it is the task of the theoretical sciences to provide us with "true" models of the world that are independent of our practical "projects".
For the applied sciences however the focus is not on "truth" but "utility". It may even be argued that it is neither advantageous nor possible to create an "objective" true model of the world. All descriptions seem to be hopelessly bound to our "world view" project and our efforts to improve the human condition.

1.2.9 Conclusion(s)

1. That the idea of a system of knowledge and "Reference system" is valuable for the understanding of the "Process" approach to the use of information in both public and private institutions and organisations. This fact has become central for the production methods adopted by the paradigm shift that has resulted in the emergence of the informational society.
2. The condition for an effective use of information in computer systems is that the data used structured according to standards that are generally accepted and in professional use everywhere. This applies both to alphanumeric information and CAD/MAP models or features. In our current case we seem to be forced to adopt object methods and technology for this purpose and consequently model our data according to UML standards.
3. The increase in productivity offered by the informational society is based on the idea that products may be more efficiently designed and produced, constantly improved upon and changed in response to changing user needs.
4. The key to this control is not on the use of static models (Data and features) alone but to a high degree on the use of well described dynamic processes that constantly adapt to those of the systems customers.
5. In this perspective the local NSI's, Eurostat and its customers are linked in a system of "value chains" the constantly change and hopefully improve micro data over a series of interconnected value chains

1.3 On processes

Abstract

In the section called "Knowledge and development" it was argued that one of the keys to understand the transition from the "Industrial" society to the "Informational" society is to understand the shift in focus from "Structures" to "Processes". This radical change of perspective forces us to rethink some of our most cherished "taken for granted"'s. For instance, the fundamental idea that the concept "statistical system(s)" as a system of data (and eventually also geographical features) has to be reconsidered. In the new perspective the focus of this and similar systems should not the inputs or outputs but rather on the "value chain" that produces them. The idea is similar to the Buddhist law of dependent origination that states that "when the conditions (processes) are thus, this arises, when the conditions change this changes with it." We will here try to draft this idea in a way that may illustrate the use of this idea to organise systems for the production of statistics in general.

1.3.1 The process perspective

According to our draft for a principal (universal) project structure we have (above) described an action as a series of parts. According to current practices each and all of these parts are regarded as integrated sets of processes. Each process is generally the supra process of other sub processes and subsequently form great hierarchies of processes that interact in a network of exchange.

Each relatively independent actor will generally build one or several processes as part of its overriding "project(s)" One way to see this is to assume that a production process may have a theoretical and a practical side. There are two ways to discuss systems. Either to regard it from a structural perspective, and describe them as kits of parts, or to focus on their dynamic properties and regard them as interacting wholes. The former is the method applied by classical statistics regarding systems primarily as "machines" and the latter the perspective adopted by the "informational" culture and the currently dominating "state of the art".

The most important consequence of the paradigm shift for NSI's and other producers of information are the change of focus from structures only to a perspective contemplating both structures and processes. This fact is inherent in the fundamental difference between the complexity of these two perspectives. Whereas the structural perspective never can describe dynamic behaviour, the dynamic (process) perspectives reveal both structural and dynamic properties.

In the figure bellow the production process is rendered with double arrows to indicate the fact that most processes involve systems for processing both Matter/Energy and Information.

According to a process oriented approach to development (as opposed to a classical modernistic "planning" approach), there are two types of information needed. The first "forward" type of information describes distributive patterns of structures and processes in space and how they change over time. The other "vertical" or "feedback" type of information is generally aggregated onto the different administrative areas used in the management of different organisations and institutions[18].

[18] I use here the definition of these words according to Manuel Castells in (Castells 1996) *The Rise of the Network Society*. "by organisations I understand specific systems of means oriented to the performance of specific goals. By institutions I understand organisations

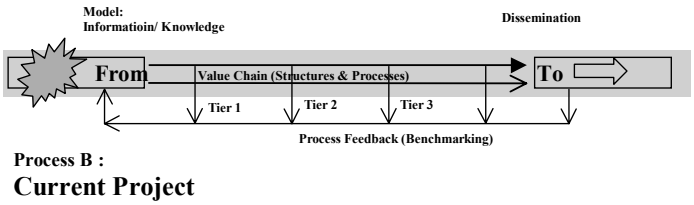


Figure 2: Process with feedback

In a production system that is built according to the "systems" paradigm, every project may be regarded as a feedback system in this sense. It may be described as a vector of flows from project the management and its project description to the dissemination of the result into the customer's processes. The process is described in terms of a value chain of sub-processes leading from (upstream) the production process downstream.
The output of a system like this could be either in a form that is produced by processing matter/energy or information or both.
In the new economy it is essential to be able to constantly change everything according to the changing customer needs, new discoveries etc. that is indicated by the results from the also changing benchmarking systems and its indicators.

1.3.1.1.1 Satisfying customer needs

Every production process is a vector aiming at producing a valuable input into another process. The value of the product may be measured in relation to the added value in the customers value chain (production process.
The supplier is thus a sub system to the customer in a downstream flow in the direction of added value.

invested with the necessary authority to perform specific tasks on behalf of the society as a whole."

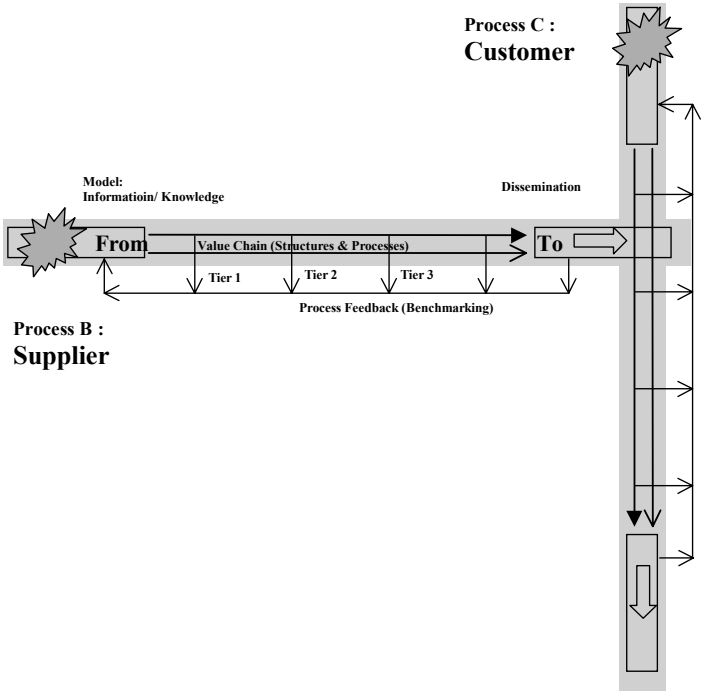


Figure 3: Production for a customer

In the current case Gisco is a supplier of qualified geo-statistical information to the Commission. The geo-statistical processes must therefore be very conscious of the customer's needs.

- This applies both to how the production process fit into the Commission's processes and
- How the product adds value to the commissions project.

1.3.1.1.2 Production with Suppliers and Customers

Every production process (Business project) is linked to the world having at least one "Supplier" and one "Customer". Each process, as well s each sub- and supra- ditto are thus interconnected in a web of processes from the smallest micro process to large industrial systems.

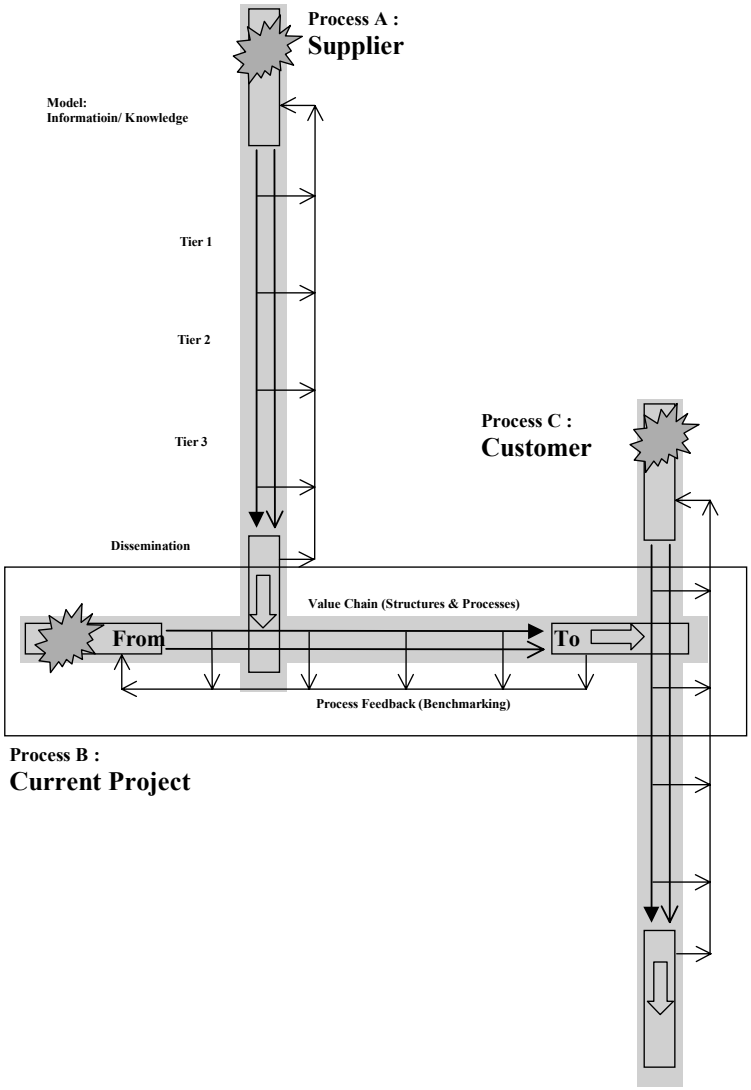


Figure 4: Production of knowledge (theory)

A project like the one discussed here is dedicated to the improvement the infrastructure for doing "Small area statistics" to improve the comparability of statistics across Europe. For this purpose a series of sub projects are constructed to become "best practices for the production of geo-statistics for EU.

1.3.1.1.3 Networks of processes

Every production process may in turn be regarded as a system of sub processes and in turn a part of a larger hierarchy of supra processes.

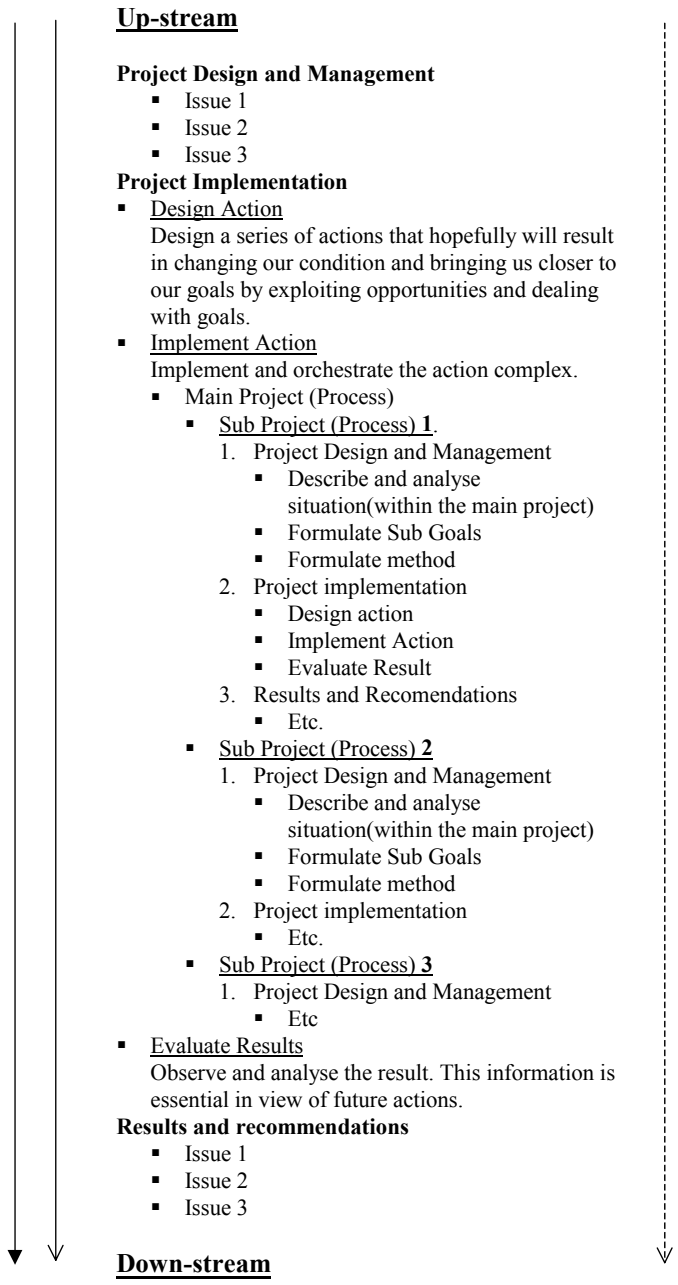


Figure 5: Hierarchies of Projects within projects.

In a feedback system there are two main production processes (value chains);

- A value chain regulating the production process, These processes are generally conducted according to "Engineering" practices.
- A benchmarking or evaluation (feedback) process according to "Accountant" methods.

Classical statistics are generally of the "Benchmarking" type. We have described this as information "on" development. We (the average NSI's) are quite good at producing this type of information. There are however two weak points:

- Although much work has bee invested in efforts to harmonise statistics here are few standardised production processes in geo-statistics. We need

to agree on “best practices” to look for candidates for methods that may later become standard for the whole union.

- Benchmarking results are generally accumulated to administrative areas. These are generally heterogeneous across the union, both in terms of size and the “population” of generally observed phenomena.

In classical statistics we are generally much less good at delivering information “for” development.

- In projects where statistical data are used “for” development, aggregations to administrative areas are not very useful. This is primarily due to the fact that problems seldom respect administrative borders. The obvious solution to this problem is to register all observations with geographical co-ordinates, or to construct a very small system of “irregular tessellations” that may be used for this purpose.
- This type of information is preferably kept lose to its micro state and aggregated to larger “masks” only when unavoidable.

1.3.1.1.4 Development

Neither production processes nor the end products are ever regarded as “perfect”. In this scheme all parts are in constant development through a constant iterative process. This is the key to the method used in this and all similar professional processes.
All processes may be constantly under observation through a benchmarking system and may be compared to similar processes elsewhere.
All possible improvements are constantly sought and new methods promising better results are tried. An iterative method for this is described under the section "Method" below.

1.3.2 Statistical system as a Process

If we agree that it is practical to adopt a systems approach to the description and development of statistics, then we suggest that we take the consequences of the current changes in other field of applied and theoretical science and start to shift our attention from seeing the statistical system as a "kit of parts" only, to a description of societies as networks of processes (the "integrated whole" perspective).
The statistical system may then be regarded as a system of structures (tables, diagrams etc.) or as a system of the production processes that produce them. From the process perspective, the statistical system as a structure has to be regarded as a product depending on the processes. When the processes change the structures change with them.
To harmonise statistical systems in this perspective means in order to effectively harmonise the structures (standardisation and semantic descriptions) means that we will have to harmonise the processes that produce them.

1.3.2.1.1 Developing the Statistical System

We are here very accomplished when dealing with “Benchmarking” statistical information produced and displayed according to “Accountant” traditions. Still there are many things here that could be improved. One obvious need is to be able to reliable statistics on other systems of regions than those used in public administration.

We are very much less good at providing reliable information for development. Here we must adapt to the needs of our customers and provide information that suits into the customers' processes here.

*1.3.2.1.2 Statistics **for** and **on** development*

Another perspective to the same situation is to focus on the use of statistics. The vertical and the horizontal may then be described according to their use for and on development efforts.

- Statistics **for** Development (Change)
  The use of statistics aggregated on to systems of small statistical areas to serve as point of departure for efforts to describe and explain structures and distribution patterns of phenomena who's development should be encouraged or discouraged.
- Statistics **on** Development (Change)
  The use of statistics aggregated onto hierarchies of administrative areas (as the NUTS system for public institutions within the context of the EU) to be used for "accountant" type decision support. "Statistics" in this sense are primarily useful for "benchmarking" purposes.

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[19].

We have elsewhere discussed the fact that this questions calls for a complimentary set of two systems of statistical areas; one system of regular- (equal areas) and another system of irregular tessellations (equal populations).

*1.3.2.1.3 Information **for** Development and Small area Statistics*

The problem that we are focussing on here, is not a question related to Statistics/Information on development (the benchmarking issue) only, but rather with the combination of both statistics for **and** on development.

The problem with this relation is that we have not explored the uses of statistics for development ("engineering" statistics) as much as we have studied the use of statistics on development ("accountant" statistics) in the benchmarking/accountant sense of the word.

In order to describe patterns of distributions there are two principal methods used; regular (grids) and irregular tessellations (blobs). Ideally a system of regular tessellations based on point observations are preferable here due to the fact that they may be aggregated to both regular and irregular tesselations.

When describing the use of information **for** development (in the engineering sense) information has to be expressed quite differently as in the case when statistics is used for benchmarking.

a.) In the benchmarking case we need to aggregate statistics onto areas representing, or corresponding to the system of administrative units used by the process owner.
b.) From the "engineering" perspective aggregations on administrative areas are of little or no use. Here we are interested in patterns of distributions.
    - Administrative units are generally not suitable for displaying information in order to study patterns of distributions. However, in

[19] The need for 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.

countries where no co-ordinates are available on micro level, small administrative areas, or data capture areas (like enumeration areas) are the only alternatives given.

- In some countries co-ordinates for micro data are available for all or parts of the data to be analysed. In these cases a standard system of grids or other regular tessellations are needed to map patterns that cross the borders between statistical systems.

1.3.3 Processes as Value chains

1.3.3.1.1 Value chains for producing geo-statistics

In the arguments above we have described two types of processes contributing to our customers processes. The first is used to provide information into processes **for** development (describe and explain actions to change existing systems using "engineering" methods), and the other to providing information **on** development (describe and explain states of existing systems using "accountant" methods). The first of these may be described as forward, or proper production processes and the other as Feedback processes.

It is interesting to note that in both of the cases discussed above the production processes may prove to be not very different. We shall see later that the main difference probably lies in the fact that in the two processes different infrastructure(s) and data sets may be used. The general structure of the process remains however very similar in both cases.

We believe that we in this case may suggest that a harmonisation project for geo-statistics could focus on the development of one "universal" value chain that could be very similar in all contexts where geo-statistics are produced and analysed.

1.3.3.1.2 A hypothetical Value Chain

A Value chain will obviously have to consist of both hardware and software. In this connection, differences in the hardware components used in a value chain are of little or no importance for the result. Much more critical for a high quality result are components like:

- The quality of the data used in the different processes
- The manual or automated transformations of the raw material that are used in the course of production or benchmarking.
- The information infrastructure used (for instance the systems of regular and irregular tessellations used for benchmarking or production)

Below is the draft for some of the main processes to be expected in a hypothetical Value chain for the production of Geo-statistics. The total process is here separated into a conventional string of sub processes. In this example we have in order to remain simple, refrained from describing the production system as it would be structured according to an "object" approach to process design, but used an old-fashioned linear process model to remain simple.

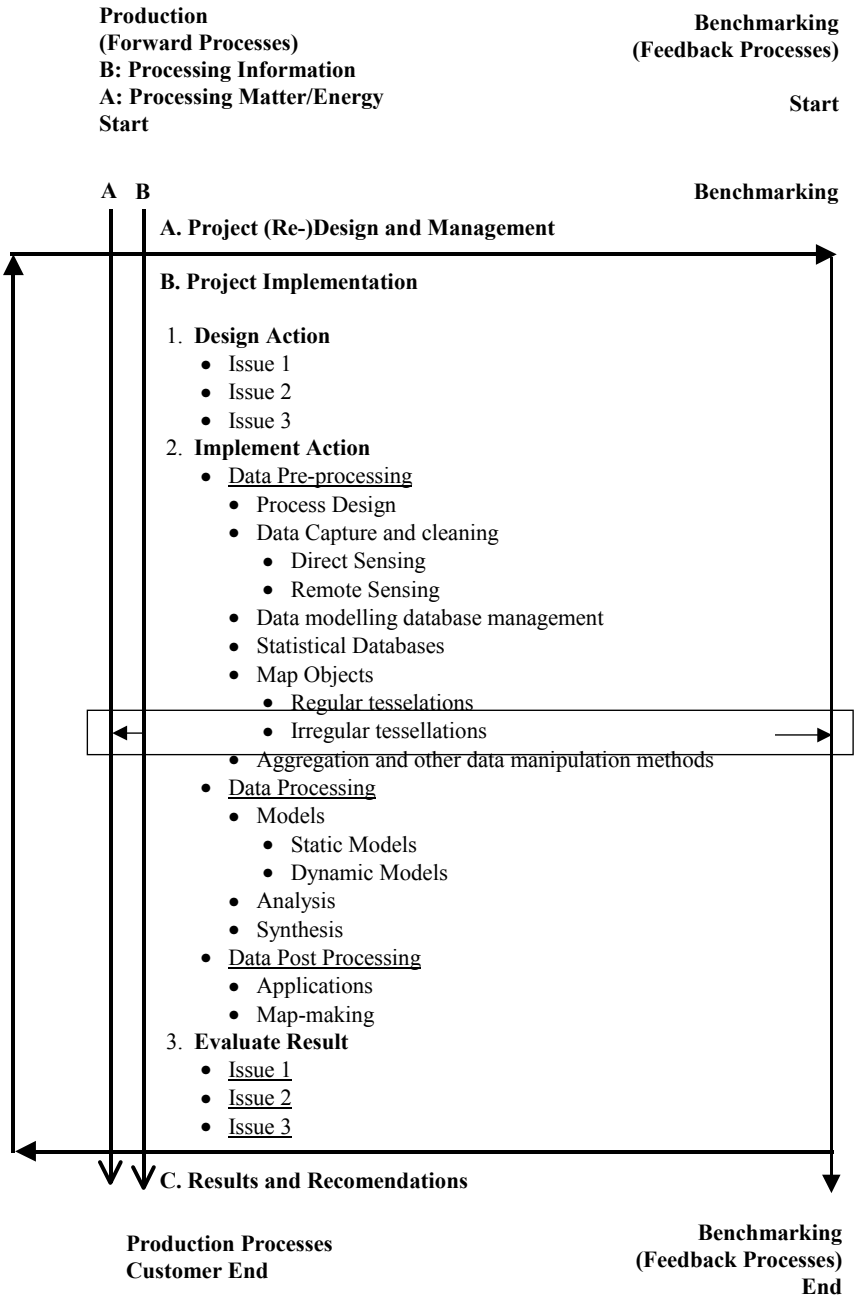


Figure 6: A rough draft for a "Universal Value Chain"

On the right side of the diagram are two parallel processes to indicate that most production processes require both hardware- and software- systems. (One could perhaps argue that a third could indicate manual processes)

1.3.3.1.3 Benchmarking Processes

Parallel to all Production processes there will be benchmarking processes that will calculate scores according to generally accepted rules. Parallel to important public projects such benchmarking is generally required over time to measure the effect of important actions.

Frequently Benchmarking results are onto administrative units, as in large projects where results are aggregated to administrative areas. Results are generally delivered to special evaluation process.

Each bullet may be seen as sub processes that may by further divided into new sub-processes ad infinitum. Ideally each sub process might produce results for both the production process and for the benchmarking process.
It is very important to understand that the value chain as a whole as well as each production and benchmarking process, must consist of both descriptions (detailed metadata) of the actual production processes and their control mechanisms (processes proper)!

1.3.3.1.4 To improve the quality of input data

I classical "accountant" type statistics the strategy for the improvement of a given statistical system or a group of statistical systems was to standardise the product (statistical data). In the current case this is not enough due to the fact that although two data sets are both produced according to the same standards, their quality will differ due to differences in the production processes used.

1.3.3.1.5 To improve the production processes

In order to improve the geo-statistical system of Europe, as well as those of individual member states, the issue is not so much a question of improving a data set, but rather a question of improving the “Value chains” that are used for their production.
The best strategy for improving Value chains is to benchmark similar systems. In our case we therefore could do much worse than try to work towards a more or less unified value chain for the production of central data sets needed at EU level. By using benchmarking techniques we could embark upon a development process where the individual NSI's could compete with another in search for "best practices". These could then be implemented throughout the system.
In this connection it is important to note that there are two obstacles that we have to cope with. Firstly we usually have no reliable descriptions of the processes that are used in the production and analysis of statistics as they are often integrated in the "of the shelf" software packages we use for most purposes. Without descriptions of this sort however we will have little possibilities to assure real improvements in the short as well as in the long run. Secondly we have generally no standard methods for evaluating the quality of our current processes (or products). The only rational way to do this seems to be that we will have to design a control system that measure the quality of our processes from raw data to the data sets that are used by our customers.

1.3.3.1.6 To improve information infrastructures

The most important of the information infrastructures needed for the production of high quality geo-statistics is the systems of regular and irregular tessellations used.
The emerging European statistical system is primarily produced through an integration of data and production practices from the whole of the EU. Even a casual glance on the NUTS hierarchy assembled for benchmarking purposes reveals that these systems are not very homogenous.
The lack of homogenity used for benchmarking purposes are generally very crude structures resulting from a long history of local power politics rather than from an ambition to provide comparable descriptions. The current

NUTS system is consequently a political issue and not open for development though iterative processes of the kind discussed above.
The lack of homogenity in the infrastructures used to describe the forward processes needed for development projects, however, is a different matter. Here we should meet no such obstacles.
If we choose to upgrade the current statistical system to provide data and infrastructure for producing small area statistics, we will have to compile a system of small areas that cover the whole are of the EU.

1.3.4 Conclusions

1. The key to any actor who wants to compete on the market in the informational society is to be "plugged in" on the Internet where all production (and consumption) is in the process of being organised in a network of networks where every unit has its suppliers and its customers.
2. The rule of this game is to tune your production processes to that of your customers and seek suppliers that are able to do the same for you.
3. In a world where every product may change from one instance to the next the key to increased productivity is not (merely) to improve the machines but (even more so) to build and manage effective information systems to control them.
4. An effort to produce a system of small statistical areas (a means of production), cannot (in this perspective) be separated from the idea of a system of small area statistics that represents the process. The device here is that "Form follows function" that means that if the system of small statistical areas do not suffice for the job, then the process have to seek new ways to satisfy customer needs. The focus is thus on the process.
5. The method used to study our own processes and adapt them to those of the customer's is to carry out "case studies" or "case studies". (see TQM and other methods)

1.4 On "case studies" and user needs

Abstract

This section is about the point of departure for any production of information; user needs. Without proper knowledge about our customers and their production processes we cannot be expecting to be able to serve them well. In this connection it is important to notice that there are two principal ways to approach customers depending on the type of product in question. We are here not discussing products for mass markets where marketing studies may provide the information we need. In this project we are discussing highly specialised tailored solutions targeted for a very limited collection of users. The method we suggest for studying their needs is not surprisingly the use of "case studies" or "case studies". The problem with this approach is that if we want to serve many customers with the same set of production processes we will have to test the system on a set of limited "case studies" that are the keys the functionality of all.

1.4.1 Introduction

Why you may ask, should we bother about "User needs"? Well, the reason is that we, in the current project are involved in the construction of a system where we have no precedence. For this reason we have proven solution to the task of designing and building a system of small statistical areas, and subsequently a system of small area statistics. This is an emerging problem where we hope that a viable solution to our task hopefully will present itself for us as consequence of a selected set of actions.
Experience has shown that the best method to adopt under these circumstances to start with one user needs as described in a "case study" and develop a solution for him. By studying more carefully selected "case studies" we may later adapt the result to a host of different situations.

1.4.2 The European project

It should not be difficult to see the European project as a system of processes designed and built to develop our realm by exploiting opportunities and countering threats all according to the overriding project formulated in the "Maastricht agreement" and elsewhere.
If our reflections on the paradigm shift and the need for a system of knowledge to build and develop stands to reason, then there are a host of potential "users " for small area statistics to be found within the divers projects funded to develop most aspects of the "man-environmental system(s) in Europe.

1.4.3 Information for European development

As suggested in the Introduction to the Professional context all actors, either we are talking about individual human beings, business or industrial enterprises, or individual or groups of national states, are dependent upon building of effective reference systems.
In these reference systems, especially those that are constructed according to the rules of the scientific method, are highly depended on qualified information to be delivered by NSI's (and their sub-contractors). We often talk about problems connected with the development of established statistical systems, but seldom about their relation to the other systems of knowledge with which they will be combined in our customers systems. Although essential to our customers, these issues are seldom discussed, but taken for

granted also within our professional context. Without this qualified knowledge and constant attention, we may easily become alienated from the needs of our customers and opportunities and threats in our relation to other information providers.

1.4.3.1.1 A Systems Approach

From a Systems point of view the task at hand is to strengthen Europe as a (living) system with its own identity and distinct properties, not only to be described in terms of the sum of its parts.
Systems like these are however complex beyond human understanding and we will have to break the whole down into a series of sub-systems that may be approached separately.
According to this strategy we might chose to describe Europe as a living system consisting of two types of fundamental processes acting within a context, one that process matter/energy and another that processes information.
A "systems" approach to an integrated reference system make us see beyond the limitation of traditional roles of this kind, and provides us with both a theoretical framework for our work and the practical tools needed to control and constantly develop the systems we depend upon for our welfare.

1.4.3.1.2 3 sub projects

Systems, seen as integrated wholes, are generally complex beyond human understanding and require that we divide them into hierarchies of interacting subsystems if we shall be able to describe and explain them.
As man environmental systems are complex beyond human it is essential that the whole be divided into parts according to a pre arranged system of modelling.
Following this strategy we will regard the parts and the whole territory of the EU as a Man-Environmental-System (MES) consisting of three macro "complexes" that are relatively independent in their relations. These should be well known to those involved in the development of urban and regional development strategies.

a.) The Socio-cultural and economic complex. (Virtual structures and processes). This group involves the interaction of groups of natural and juridical persons acting through systems of projects channelled through information networks. Economic and cultural development is here seen as a process involving the processing and communication of information.
b.) Infrastructure systems (Physical (Real) structures and processes). This group is dominated by artificial systems that process matter/energy, and are generally change and/or developed as a result of changes in the system that process information. In this sense artificial or "man-made" physical systems are to be seen as expressions of socio- cultural and economic processes as "form follows function"
c.) The Natural Environment (Structures and processes relatively peripheral to human utilitarian projects. Context in general). Natural systems are, or have at until quite recently been, relatively independent of direct human interaction. Changes in natural systems have however been indirectly dependent on human action

It is however of little value to remain on this conceptual level if we are to go beyond the idea of seeing our information merely as a system of data. We

therefore have to describe this overriding project as system of processes as below.

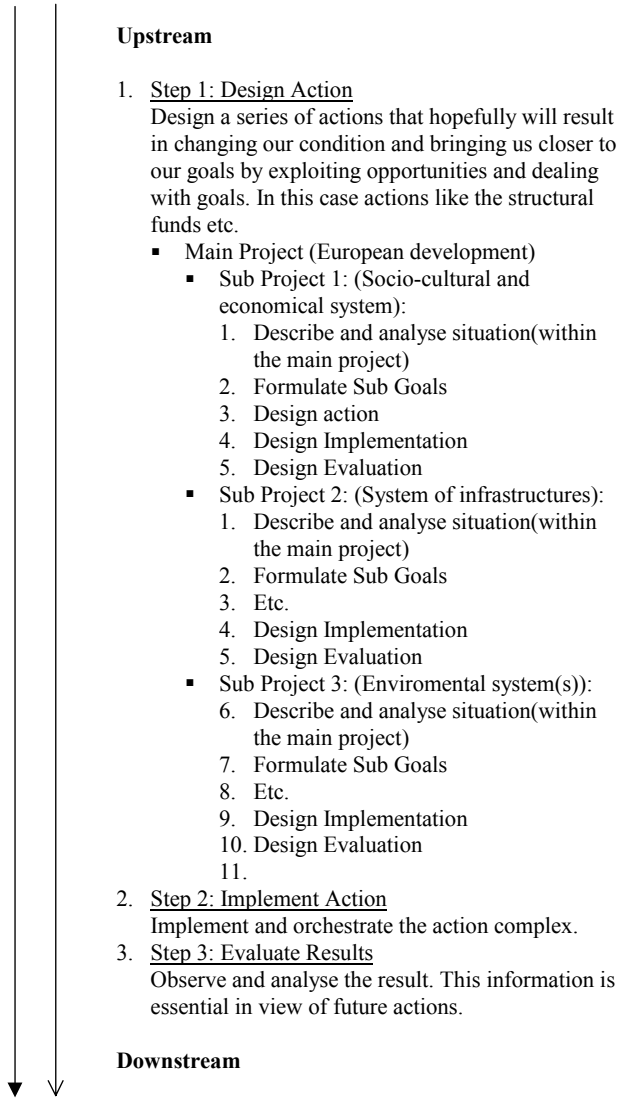


Figure 7: A Value chain for the production of information for spatial development

In this crude scheme the important thing is not the contents but the principle. That if we want to improve our contribution to this project (system of processes) then we could do much worse that to study how we might contribute to this "value chain".

1.4.3.2 Developing Infrastructure Systems

Of the three complexes mentioned here, we would like to focus on an example from infrastructure systems. These structures may be seen as the physical expression of socio- cultural and economic processes. Infrastructure systems may partly be seen as networks of "lines" (as transportation lines) and "nodes" (as nodes or stations/urban areas) used for the production, distribution and consumption of goods and services and dealing with the consequences.
It is one of the most central political objectives on behalf of the European Union to develop this infrastructure through projects partly funded through the structural funds responsible for the structural cohesion of the European

territory. The structural funds consisted originally of 3 parts: The Regional Development Fund (ERDF), the European Social Fund (ESF) and the European Guidance and Guarantee Fund (EAGGF)
Efforts to improve the spatial cohesion in Europe are, at least partly, orchestrated through co-ordination efforts formulated with the auspices of the ESDP (European Spatial Development Perspectives).
In this scheme EPSON (European Spatial Planning Observatory Network) will be responsible for data modelling and analysis for this purpose. In these and similar development projects, statistics could pay a better contribution than is usual at present.

1.4.3.2.1 A method for mapping population densities

We have in the current project selected a case study of the methods used to provide better information on the human habitat. This is an important question since high-density areas contain up to 80% of the human population. The delimitation of urban/rural areas is therefore a central problem for most projects aiming at spatial development. We believe that we could make an important contribution here especially due to the rising interest for statistics on these and similar structural and functional regions.
I a delimitation of urban areas based on "night-time" populations we may map the parts of our total habitat that is used for "living", and later we may do similar analyses using information on "day-time" populations and high resolution travel to work studies etc.
Delimitation is of course a line dividing areas with different population densities. Depending the stratification used it should be easy to draw iso-lines that delimits areas with comparable population concentrations.
The local NSI's have for this purpose either no working system for delimiting their urban areas or have applied very different and individual approaches to this problem. The methods differ from manual delimitations to machine methods based on co-ordinates from real estate or building registers analysed with statistics.

1.4.3.2.2 Delimitation of urban areas

The simple solution to the problem of providing a comparable method for the delimitation of urban areas is a clustering method using Nuts5 areas. This approach follows the idea that urban concentrations may be seem as clusters of rather large administrative units that have average densities above a certain threshold. In is however evident that this approach is not satisfactory due to two main reasons:

- The administrative areas are harmonised in terms of neither population nor area.
- Administrative areas are not delimited according to the density of populations.

We believe that this is an example of a very general principle that administrative areas are seldom, if ever, useful for practical purposes of this kind. Administrative areas are designed for administrative purposes and useful for this purpose only.
If we want to describe variations in the settlement patterns in a country a different approach is needed. And the obvious solution would be to use a system of smaller statistical areas, or even enumeration areas or other regular or irregular tessellations used for data capture. In this example, however, it seems evident that most local NSI's will be reluctant to send micro-data, or

aggregated data with relatively high resolution beyond their national borders. In order to function a system of small area statistics covering the territories of more than one member state, we will have to provide a solution (to the delimitation of urban areas) where the national NSI's will be responsible for the delivery of processed information that may be combined by a third party.

1.4.3.3 Conclusions

1. We believe that a system of small area statistics should be designed and built according to user needs. These however, cannot be adequately well described using market analysis and similar methods. The best method, we believe, to study emergent problems is to study how the new system could compare to current methods based on relevant "Cases studies" (case studies)
2. We have from these and similar arguments decided to study, for the current project to study the feasibility of a system of small statistical areas for EU on its usefulness to the well-known problem of urban delimitation.
3. According to our idea that a system of small area statistics is not s set of data or features but rather a system of processes, it is essential that the "Cases studies" will focus on the production processes used by suppliers on the one hand and customers on the other. The idea here is improve the customer's processes by designing and building a process (value chain) that aims at improving the customers results.

1.5 A system of small statistical areas (features)

Abstract

The purpose of this section is to try to focus on the core of the problem that we seek to solve. Its central argument must be that there is a growing need for qualified information that may be used to respond to both emergencies and opportunities. It has been repeatedly shown that traditional statistics aggregated on administrative area are not suitable for this task. In order to respond effectively to such problems effectively, we need to describe and analyse phenomena if possible in a non-aggregated state, as clusters or swarms of individual occurrences (or if that is not possible, small-scale aggregations).

1.5.1 The Problem

During the industrial paradigm we were led to believe that it was rational to make a distinction between structures and process that change quickly and those that change at a slow pace. Long term processes that involved constant attention over long time periods were often related to the man-made physical environment and therefore orchestrated and co-ordinated through physical planning. On the other hand, processes that changed quickly as infrastructure failures or natural catastrophe's where treated quite differently and handled by different experts groups. In so far as statistics were effectively used in these instances, we may guess that it was not especially suitable for either. Statistics was used primarily in order to describe the "state" of different administrative units over time.

During the last decade however, we have seen a considerable change in the way we set about to maintain and develop man-environmental systems. One such change is a need to integrate the tasks approached by previous planning efforts with those that handle emergencies. The reason for this is that we are forced to take "holistic[20]" and process oriented perspective due to an accelerating pace in all important processes that makes it much more difficult to regard anything as constant or unchanging of long periods of time.

The systems of administrative areas are still very important especially for accounting purposes but they are becoming less useful in "engineering" processes aiming at the development of systems for production, distribution and consumption of goods and services as these have grown as webs without respect for such increasingly "virtual" limitations.

In our efforts to find and develop effective methods to provide such systems with qualified information we will have to integrate all types of information available. This means that all data regardless of the purpose to which it was originally produced all data must comply to the same standards

1.5.2 A system of Small statistical Areas needed

In order to map a pattern of systems belonging to the same class of objects the best thing would be to have geographical co-ordinates for every known instance. This is in many countries impossible for many important variables. The next best thing therefore is to have, for each phenomena, a system of

[20] Ho lis tic (according to the Merriam-Webster's Collegiate Dictionary). 1 : of or relating to holism. 2 : relating to or concerned with wholes or with complete systems rather than with the analysis of, treatment of, or dissection into parts <holistic medicine attempts to treat both the mind and the body> <holistic ecology views man and the environment as a single system>

statistical regions where every one unit ideally either contain the same "population" of the current object or, that they occur in the same frequency. It has been argued elsewhere that every collection of systems, tend to be distributed in space according to a unique pattern. Due to differences in scale they will also emerge on different levels of resolution. From this demand follows the idea that no one system of regions can ever be adequate for the analysis of all types of statistics. Every phenomenon will need its own unique system of regions.
From these and similar facts follow that we in order to map different patterns of distribution, we either need co-ordinates for each occurrence of the object we study or a system of very small areas of similar size, that in turn may be clustered according to demands of each collection of objects studied.
We have no such harmonised system of small statistical areas for the EU

1.5.3 A small area infrastructure for EU

From the above, it will be apparent that the Process perspective to geo-statistics puts new demands on the statistical system. We must now provide qualified information not only for benchmarking to describe the development of the societies we observe, but also provide better information to serve as a point of departure for direct actions that be implemented to improve the human condition.
We have elsewhere[21] described what this difference in orientation implies. In short it calls for an integrated and effective system for doing small areas statistics for the whole of the European region.
According to the problem statement formulate at the start of the current project the Tandem consortium argued for the need for a system of statistical areas that fulfilled 4 central criteria:

- Roughly equal in terms of Population or Area
- As small as possible
- Homogenous across the EU
- Statistics

Below we shall try to discuss each of these criteria in turn

1.5.4 Roughly equal in terms of Population or Area

The main problem with defining statistics that are useful for describing and analysing the spatial distribution of phenomena in space is that we generally will want to base our analysis on a system of regions that are either equal in terms of area or population.
We are consequently looking for two systems of small statistical regions that are to be used in combination. Each of these systems has their own special advantages and problems.

a.) Regular tessellations (Grids) (Roughly equal in terms of size)
As for the regular tessellations we have little problem with the demand of equal size. The problem here lies with the statistical variables available for doing geo-statistical analysis on grids.

b.) Irregular tessellations (Blobs) (Roughly equal in terms of population)
In the case of the irregular tessellations the demand for a system of regions with roughly equal “population” poses some major problems as it (the population) that will vary according to the variables studied. Thus a

[21] Please see the paper (Backer 2001) *Accountants and Engineers; why the difference?*"

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.

It seems natural that a complete system of small statistical areas must consist of a solution for both of these cases. We need an integrated system of both regular and irregular tessellations.

1.5.5 As small as possible

Regarding the relative size of the areas in an integrated system of regular and irregular tessellations it is important that they are as small as possible in relation to size of the territory that must be analysed. This seems, in terms of irregular tessellations, to indicate that they should ideally be much smaller than the smallest NUTS level.

The reason for this lies in the fact that the system of small irregular tessellations will seldom be used as such but only after they have been clustered to achieve the criteria of equal population, or equal area for each subset.

Candidates for a system of small statistical areas must be sought among the very smallest statistical units available in each national statistical system.

1.5.6 Homogenous across the EU

The integration of geo-statistics across the EU will naturally demand that analysis should be comparable across regional and national borders.

Due to the fact that traditional statistics are generally presented on systems of administrative areas that differ widely in size (and population) over the territory of the union it is very important that the statistical "seed" areas are assembled from units that are devoid of political prestige.

The homogenity criteria means simply that the quality of any cross border analysis will be no better than the weakest subset of "seed" regions. The more homogenous the finer will the resolution of the analysis prove.

It is to be hoped that the smallest statistical area available in most cases is based on similar data capture techniques (using relatively comparable blocks or enumeration areas). To fulfil this criteria it seems evident that either an increasing amount of all statistics are captured with co-ordinates on micro level, or that there is a intentional policy on behalf of the members of the union to further the establishment of a common system harmonised small areas generated with machine (Thiessen polygons etc.) or manual (digitalisation of enumeration areas) techniques.

1.5.7 Statistics

The need for statistics is obvious. Without statistics no analysis.

In the typical situation we have statistics available on relatively large administrative areas only. This may or may not work well with analysis's involving the whole of the human population, but with virtually all other data sets, the distribution of populations will differ radically.

There are many methods available to either aggregate data from registers with coordinates on micro level, or providing by disaggregating statistics from larger to smaller units (small area estimation techniques etc.).

1.5.8 Hypothesis

1.5.8.1 The Point of departure

The main problem for displaying and analysing the spatial distribution of statistics describing properties of systems on different scales and densities is the system of regions used for analysis.
Obviously we cannot construct a system of regions for each purpose. Instead we might start from a system of small areas and the design a method for clustering these statistical seed areas according to current needs.
When a suitable clustering of the areas is available, the pattern of their distribution may be mapped and the result analysed.

1.5.8.2 The hypothesis

The hypothesis suggested to create a system of statistical seed areas is that the smallest system of statistical regions in each country may very well serve as a crude start for an iterative harmonisation process.

1.5.8.3 The start (an intermediate solution)

It seems evident that we will have to start with an existing system of small statistical areas used in the different member states and the start a project in order. This system will have to consist of a collection of the smallest system of statistical regions available in each country.
The current project will, as a feasibility study test the method described here on a limited system of regions. (Data from Finland and UK). In the belief that these two countries may represent the two main statistical cultures on Europe, it should be possible to make at least some general remarks on the feasibility for a larger project for the whole of the EU
In the next stage a process of harmonisation should follow in order to improve the all-over quality of the data set. Large regional differences are to be expected. If so it might be advisable to employ a strategy where a system of "seed" areas are developed first for each relatively homogenous region separately. Later these regional systems will have to be integrated in a further stage.

1.5.9 Conclusions

1. There are two "traditions" for small area statistics that may be referred to as "point" and "area" based statistics.
2. Those countries that have point-based statistics have co-ordinates on micro-level (or close to it), will tend to use regular tessellations for small area statistics, or construct irregular tessellations from them using Thiessen- polygons or other methods.
3. The critical groups of NSI's from the perspective of small area statistics are the countries that have most of their data on relatively large administrative areas using census and sampling methods. For these countries it seems reasonable that the systems of enumeration and other small statistical areas may be used for a system of small area statistics as discussed here.
4. Countries using area-based statistics have little choice but to stick to their irregular tessellations (or make them as small as possible) and dis-aggregate these data to regular tesselations when needed.

5. However, most statistical systems have a varying amount of both kinds of statistics. It seems also safe to assume that there will always be a need for both systems in a future system of small area statistics.

1.6 A system of small area statistics

Abstract

In this section we are trying do create a vision of a system of small area statistics along the lines sketched in the previous pages. This effort is important in order to define a "virtual" conceptual "model" to serve as a guiding star in our efforts to assemble a working prototype to a system of small statistical areas. We have for this purpose chosen to use a simplified method to make a semantic description of a working system based on the empirical and theoretical information at hand. It is clear that a definition like this must be used with great care, but if properly argued may serve as the key to the noble art of laying an egg without an egg.

1.6.1 Introduction

What is, or what do we mean by, the terms "A system of small area statistics" and a "A system of small statistical areas" in this perspective? In the following section we will make a draft towards preliminary sketch for a system of small area statistics to serve as a hypothetical conceptual prototype to a solution. We have chosen to try this in the form of an outline for a definition for this concept.

1.6.1.1 Focus on the connotation

- The Notation (sign)
  "A system of small area statistics" and a "A system of small statistical areas"?
- The Denotation (meaning 1)
  We will not dwell on this issue due to the fact that these two concepts are well on this issue because it has no bearing on the discussion we seek here.
- The Connotation (meaning 2)
  The connotation of these terms are of course much more complex than directly indicated by the denotation of the term and highly dependent on the reference system held by the reader. In order to reach an understanding on the current and potential uses of this term we will have to embark upon a semantic analysis of some kind. As always you need an "egg to get a hen". Therefore we will suggest an "embryo" to a definition and hope it will trigger off a fruitful discourse into the subject matter.

1.6.1.2 The Notation

"A system of small area statistics" and a "A system of small statistical areas" both of these terms are intimately and inseparably related when we are talking about geo-statistics or spatial statistics. The one focuses on the data the other on the features.

1.6.2 The denotation

The denotation of the Notation to be discussed must be concentrated around the three central words; System, Small and Statistics.

1.6.2.1 System

In order to discuss and develop the concept I believe it is essential to describe its con-notations and de-notations with reference to a relevant field of knowledge. From my point of view it is natural to describe what we are

discussing as a "System" of small area statistics. This assumption leads to two obvious fields of knowledge:

1. System theory
   We find that it is natural that when discussing a small area statistics we are constantly forced to relate our discussion to an empirical situation. In our case most of our efforts are related to "problems" in connection with our efforts to develop societies and their habitat. We suggest that the currently best ways to discuss these issues be related to the use of systems theory. There might be problems here, because these theories have in the past been criticised for its generalisations.
   After it has been widely accepted that no knowledge may be assumed to be universally "true", and that all theories are only valid within a given set of border conditions, we are forced to accept also "systems theory" as a sound cornerstone for the scientific system.
2. Object methods.
   In this case we need a proper professional language and a system of knowledge to be the theoretical "home" for this subject. The methods and technologies related to the description and manipulation of systems may be discussed using a series of approaches more or less directly related to the current subject (the definition of a SSSA). We will, as it seems close at hand in relation to this subject to choose the quickly maturing and widely used concepts and methods related to the so-called "Object" approach. In order to make quick advance in this as in other fields we will have to build on what exists.

1.6.2.2 Small Areas

Here the difference seems to be quite obvious that these terms relate to another complimentary set of statistics and statistical areas; "Systems of large area statistics (SLAS)" and "Systems of large statistical areas (SLSA)". The most obvious example of a system of large statistical areas are the NUTS system in Europe, or any other system used to present official statistics in most countries. We are not concerned with these systems here although both systems are mutually interdependent.

1.6.2.3 Statistics

I will not venture into a discussion of what we mean by statistics[22] as defined in my Webster-Merrian dictionary:

1. a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data
2. a collection of quantitative data

1.6.2.4 Small area statistics, vs. small statistical areas

The first term is focussing on the facts, quantitative and qualitative data, the other on the fact that mast aggregation are related to a system of areas of irregular (blobs) or regular tessellations (grids).

1.6.3 Draft for a semantic definition

It will be self evident that we cannot provide more than a first tentative set of notes related to this subject. This field of work is still to little explored and

[22] The Etymology: German Statistik study of political facts and figures, from New Latin statisticus of politics, from Latin status state

need to mature through a professional discourse. Still we would like to make some remarks as a contribution to this discussion.

A system of small area statistics (SSAS) is:

1. *A system of knowledge built from statistical micro data,*
2. *Designed as an "open" system of data (statistics), geographical features (regular and irregular tessellations) and methods (manual and automatic processes).*
3. *A SSAS consists of, and is in turn itself a part of and designed to fit into, a constantly changing hierarchical network of processes*
4. *dedicated to the production and analysis of qualified spatial information,*
5. *both on- (aggregations, benchmarks) and for- (data- spatial- and temporal- analysis).*
6. *A SSAS is dedicated to the improvement of the results of overriding projects to counter threats and exploit opportunities in view of private and collective efforts to improve the human condition.*

1.6.4 An illustration

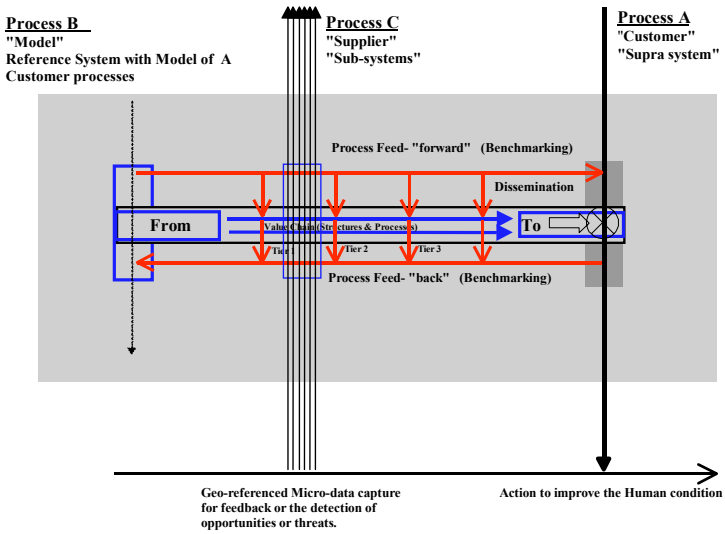


Figure 8: Model of a standard system for the production of small area statistics

In the illustration above is depicted three vertical processes. Process A represents the customer (Here the commission) who needs the product to decide on projects to further development in Europe. The process C represents various sources of information captured by different mean (remote sensing etc.) and delivered as input to Process B that represents the system of small area statistics discussed here. It is a system of knowledge designed to support action (process A) this information is needed on two resolutions: Large area statistics on administrative areas (NUTS etc.) and small statistical areas to simulate functional and structural regions. The system is open in the sense that it is provided with feedback processes that evaluate the quality of the processes, and forward processes that make improvements continuously (red arrows). The system consists mainly of data and features (provided by process C) and processes organised as a value chain (B).

1.6.5 Notes towards a connotation

Let us have a closer look at each of these six theses in turn.

1.6.5.1 A system of knowledge built from statistical micro data

1. *A system of knowledge built from statistical micro data,*

1.6.5.1.1 A System of knowledge

The idea that a SSSA should be described as a "system of knowledge" is perhaps not so obvious. Still, as will become apparent from the discussion below, that if we agree about the need to regard "small area statistics" for a region as a system that can adapt to changing conditions we have already accepted the idea that it is a system of knowledge. The reasons for this is simply that adaptation calls for "systems of reference" that are used to model the world we want to change.
This knowledge is accumulated as a side product from the "systems" never ending efforts to describe and adopt to the networks it serves, and partly is being incorporated in the manual and automatic methods (processes) used. Thus designed, a System of small area statistics consists not only of the 3 components Data, Features and methods. But of an extensive system of knowledge that is used to describe and discuss different aspects of its construction and use.

1.6.5.2 An "open" system consisting of data, features and methods.

2. *designed as an "open" system of data (statistics), geographical features (regular and irregular tessellations) and methods (manual and automatic processes).*

1.6.5.2.1 Established and emerging problems

Is it acceptable to talk about systems of small area statistics? Well, we believe that it depends on your point of view. If you think in terms of the statistical system as a "closed" and *established* structure of information you are likely to miss the point related to the need to call it a system of small area statistics. The reason for this is related to the ideas described here and elsewhere as the result of the paradigm shift in the aftermath of the second industrial revolution.
If you accept this development as a fact it is important to accept that everything has to be regarded as *emerging* and described as "flowing", and constantly changing and hopefully developing.
If that is so then all things has to be designed as to adopt, or perish. Systems theory is therefore needed to deal with the problem of change of dynamics. This however, does not mean that we have a great need for "closed" systems. We need both but also a method to see the difference.

1.6.5.2.2 We have suggested regarding systems of small area statistics as "open".

Her we must first note that we are not using this word here in the sense usual for systems theory. There the terms "open" and "closed" are used contrary to the intuitive meaning and therefore should be corrected[23].
"Open" systems in this sense may develop in the sense of adapting to changing conditions, as "closed" systems don't. We would therefore point to the important observation that such systems have three important properties:

- Systems as processes
  They may be described primarily in terms of a system of processes.
- A hierarchical network of interconnected processes.
  Processes (that in turn may be described as systems of processes), may be linked in a hierarchical network of interconnected processes.
- Processes change (adapt)

A system of small area statistics must in order to be complete, display and properly function in terms of al these three properties. In order to function well, this single perspective does not suffice. We must be able to regard any system from at least three perspectives: (1) the system itself (2) the system as a system of sub systems and (3) as a part of a supra system.
Similarly, a system of SSSA for the EU, should be regarded as an integrated system of national subsystems (of small statistical areas) and as a part of a future global, or at least supra-regional system of small statistical areas.
A SSSA, as seen in isolation (see the European SSSA), must according to the object method suggested show all the three main properties demanded of an "open" system, if we follow the recommendation(s) indicated in this comment. This means that it must show two important properties;

1. Fit into existing networks of production and other processes
   It must fit as an effective link in a network of processes where it will be used as a part of larger processes. This ability is what we have elsewhere have described under discussions around "networks" of processes (systems). We have in the papers related to this note suggested that any SSSA should in this sense be suited to "user needs" in the sense that they should fit into different places in such networks. The best way to do this, we suggest, is apply the technique known as "Use Cases[24]". We have suggested talking about use cases as expressions of the "desirability" of a process or system.
2. Must be able to adapt to changing conditions
   The other central aspect related to "open" systems is that we will need methods to be able to respond to changes in the networks the system is connected to. The best way to do this seems to be to adopt what is generally know as an iterative R+D method. This approach is of course generally accepted as "state of the art" for keeping all products under constant development. We have in this suggested that the test of the "feasibility" of a project or a product is its ability to dynamically change and adapt to the constraints presented by both sub- and supra- systems.

---

[23] We have problems with several other related concepts as well. See for instance problems in connection with the word "information" that is used very differently in discussions around cybernetic systems (Norbert Wiener etc.) and the "information" concept used in the telecom industry (Shannon information) for this discussion see for instance (Nörretranders 1991) *Märk världen; En bok om vetenskap ock intuition*

[24] For discussions on the theory and practice of "use cases" see for instance (Jackobson 1994) *The object advantage, Business Process reengineering with object technology*

1.6.5.2.3 A system of small area statistics described with OO[25] methods

If we agree to discuss small area statistics as a "system" according to the methods suggested by "object theory" as suggested in the notes on the denotation we will have to think about a system consisting at least the three following components.

1. Data
   A data set aggregated to a system of regular or irregular tessellations as mentioned under pnt. 2.
2. Features
   A system of (relatively small) features (Tessellations) to which the statistics are related.
3. Methods
   A system of processes needed to manipulate data sets and features according to specifications demanded of the system.

In this list the "methods" is the only less obvious component for those who think about statistical systems as "closed" and not developing. For them the methods, or processes are of less importance than the output (products that generally consists of data and features only)

1.6.5.3 A hierarchy of networks

3. A SSAS consists of, and is in turn itself a part of and designed to fit into, a constantly changing hierarchical network of processes

This section is a consequence of defining a system of small area statistics as a hierarchical network of processes is evident to anyone involved in TQM and other project aiming at the "re-engineering" of their production processes.
The task of shifting the focus from structures and products to our own and the customers processes is fundamental to any "learning" organisation.

1.6.5.4 Production and analysis

4. dedicated to the production and analysis of qualified spatial information,

The idea that a system of small area statistics should be used for both the production of and the analysis of spatial information should not be objectionable to anyone working with spatial statistics.
Still is seems that most activities are limited to the production of spatial information end less to their analysis. The reason for this is not very clear, but it seems that this has something to do with the scale the data refers to. It is simply impossible to do spatial analysis on systems of large statistical areas.
It is the objective of SSAS (small area statistics) to focus on the NSI's ability to expand and improve their input into our customers processes in this respect.

[25] OO stands for Object Orientation, and relates to the language or system of concepts and methods accumulating around efforts to describe, build and develop systems both within manufacturing and information industries.

1.6.5.5 For and on Development

5. *both on- (aggregations, benchmarks) and for- (data- spatial- and temporal- analysis).*

In this section we try to refer to the beneficiaries of our efforts. Here it seems that we could extend our activities services to supply both information on and for overriding projects.

1.6.5.5.1 Two uses, two users

We have elsewhere[26] discussed the idea that there are two main uses and users of statistics. The first of these is the classical use of statistics for accounting purposes. This use and these users have a very long tradition in the average statistical office, as statistics like public incomes and expenditures as well as inventories and other records were naturally aggregated in this manner. For contemporary use, however, this perspective is becoming less important as other and new- uses of our information are emerging.

The most important of these new uses is the need for information that may serve as a basis for describing and analysing information about phenomena that cross all local, national and International borders. Examples here are

1. Socio-cultural and economic catastrophes like wars creating large local problems and refugees flooding large regions.
2. Large Structural breakdowns as the Chernobyl meltdown spreading fallout over large areas.
3. Large natural catastrophes as the results of flooding in Poland and elsewhere, or the dangerous epidemics like the FMD in the UK recently.

In all these cases, it is clear that conventional reporting with statistics aggregated on administrative areas will not help those involved in counter actions. For this purpose new users must be allowed to demand better products from the NSI's responsible for these emerging uses.

1.6.5.6 To improve the human condition

6. *A SSAS is dedicated to the improvement of the results of overriding projects to counter threats and exploit opportunities in view of private and collective efforts to improve the human condition*

This section refers to the overriding projects these systems will (hopefully) serve. We believe that there are two important overriding projects. The first are the projects concerned with the development of new and existing modes of production (the way we produce, distribute and consume goods and services). The other overriding project is concerned with a cultural "identity" project not primarily concerned with survival, but rather with finding answers to questions about reference systems related to values and questions of identity.

This might be taken for granted, but we are also here reminded of the destructive uses to which such systems may be put.

[26] See for instance the paper (Backer 2001) *Accountants and Engineers; why the difference?*

1.6.6 Conclusions

1. The method suggested here is to provide a conceptual model of the system needed is based on using the idea of a semantic definition. This definition (or rather description), is based on a theoretical and an empirical description of the "unknown". This "model" will eventually, and over time, develop into a conceptual specification that will serve as a guideline for the experimenters in their effort to build and develop prototypes.

1.7 A Method

Abstract

In order to test the hypothesis we stress that this is an exercise belonging to the sphere of the applied sciences. This means that: “the proof of the apple is in the eating”. It is not the issue to explore new theories in this field but rather to design (compile), test and develop a system of statistical according to benchmarking, and similar methods generally used in industrial R+D projects.

1.7.1 Method

It is not very difficult to suggest a method for developing a system of small area statistics according to the given premises and the conceptual model described above. The obvious choice is to design an iterative "evolutionary" model where a prototype is first designed and built according to an theoretically argued and empirically founded model or hypothesis, that is tested and redesigned over a series of iterations where the hypothesis and the prototype emerges guides by customer needs.
The most difficult part of this process is the start. As we always will need an egg to get a hen, to lay a new egg…., we will start with a creative phase as described in the current report.
It is very likely that both theoretical arguments and practical solutions will be crude and not well formulated, but as one iteration follows the next all aspects of the system will emerge from the rough contours barely discernible in the first effort.

1.7.1.1.1 Preparatory

1. Study the need for a SSSA[27]
2. Describe professional context and a method
3. Describe user needs with case studies

1.7.1.1.2 The first Iteration (current project)

4. Design and build a prototype
5. Test and evaluate
6. Formulate recommendations for further work (iteration 2)

1.7.1.1.3 The second iteration

7. Redesign a new version of the prototype
8. Test and evaluate.
9. Formulate recommendations for further work (iteration 2)

In its present phase, the Tandem project is aiming to prepare for a second iteration that will try to define a "full fledged" prototype for a system of small area statistics although still on a limited scale.

1.7.2 Step 1: Design and build a Prototype

1.7.2.1 A Realistic problem (Case)

The Experiment is based on needs defined in accordance with the “Case study” and will challenge the current best practice for the delimitation of urban areas in Europe using nuts5 areas.

[27] A System of Small Statistical Areas.

The Test will focus on using the method developed here to make better delimitation using the smallest statistical areas available in the three countries represented in the consortium; Finland, United Kingdom and Sweden.
The experiment will focus on data from 3 urban areas in the 3 countries involved. The Urban areas are Helsinki, Cardiff and Stockholm.

1.7.2.2 A suitable data set

1.7.2.3 Build Testbed

1.7.2.4 Processes as Value chains

Although this is not a project in the tradition of the theoretical sciences but rather a “professional” project according to methods current in the applied sciences we are working according to the principles of the scientific method. This means that processes and products are regarded as “hypothesises” that are tested by comparing to “best practice” by using standard benchmarking indicators.
The context for doing geo-statistics in this sense is a series of dedicated processes that may be generalised into a "universal" value chain for the production of geo-statistics. There is not yet any discussion about the need for a system of best practices that may be challenged in a general effort to develop the geo-statistical system as a process.
We might however try to design a value chain according to generally accepted ideas about what is considered to be the current "best practices" in different areas by reference to the professional peers in this field. As always this structure is open to critical appraisal and may be used to accumulate knowledge here.
The method should be constructed according to the rules of "Scientific method" where it is tested

a.) In relation to a theoretical framework accumulated for this context (the theoretical assessment).
b.) In relation to empirical data collected for a problem that may serve as a proper example for the current context (the practical assessment)

This method applies to the development of a general supra processes for the production of geo-statistics, as well as for the development of much smaller sub processes.
One of the most important demands on current production processes is that they should be flexible and easy to change according to changing customer needs.

1.7.2.5 A value chain

In order to stage an “experiment” to test the usability of a hypothetical system of tessellation’s we will design a “universal” value chain that lies as close as possible to the generally accepted methods for doing geo-statistics in this and affiliated areas.
This value chain is focusing the further development of a section of the larger value chain (above). In the current version of this process we are here focusing on nothing but modelling and analysing spatial date using Irregular and regular tessellations.
We will use a standard 3-tier approach to this problem.
The first tier or data management. (Here we start with a given data set consisting of a comparable combination of tessellations and statistics for the

3 test regions.) This is the critical phase in this experiment. Here the regions in the "seed" are clustered according to parameters suitable for the delimitation problem. Here we will found our experiments on theoretical and practical work by Martin[28] and Openshaw[29].
The second tier or data processing (In this tier we have a system for clustering the given system of regions to serve the purpose expressed in our experiment)
The third tier or presentation and evaluation. (In this tier we will try to display and evaluate the quality of the result according to a simple "Benchmarking method)

1.7.2.6 Design and build benchmarked process

1.7.3 Step 2: Test and Evaluate

1.7.3.1.1 Run tests with parameters to suit aims.

1.7.3.1.2 Displaying and format Results

1.7.3.2 The Practical Assessment

Practical assessment of the production processes.

a.) The practical assessment will focus on the practical foundation for the construction of a value chain. As an integrated process. This means practical considerations regarding longer value chains as they are generally designed in practice.
b.) The practical assessment will then go to individual parts and discuss the theoretical references central to each sub-system. Here the focus is on the parts or smaller sections of the production process as they are designed in practice.

Practical assessment of current benchmarking practices.

d.) Here a practical assessment of the benchmarking system seen as a system of components.
e.) Here a practical assessment of each component seen in isolation.

1.7.3.3 The Theoretical Assessment

Theoretical assessment of the production processes. Here we are discussing theoretical work discussing different strategies used for the production of a product. There are two approaches to this:

a.) The theoretical assessment will focus on the theoretical foundation for the construction of a value chain as an integrated process. Here the focus is on the system approach along the crude lines discussed above.
b.) The theoretical assessment will then go to individual parts and discuss the theoretical references central to each sub-system. Here the focus should be on relatively isolated sub-processes. As they are treated by the academia.

Theoretical assessment of current benchmarking practices. As in the case of the production processes we should agree on benchmarking processes that are suitable to judge the quality of a production process. There are also here two approaches to this:

[28] See for instance (Martin 1991) *Understanding socio-economic geography from the analysis of surface form.*
[29] See for instance (Openshaw 1991) *Developing appropriate spatial analysis methods for GIS*, (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

a.) The theoretical assessment of benchmarking practices as applied to the current situation.
b.) The theoretical assessment of special aspects, or parts of the current process.

1.7.4 Step 3: Formulate results and recommendations

At the end of each iteration all-over results are presented and recommendation(s) formulated for the next iteration.

1.7.5 Conclusions

1. The conclusion of this section is that there exists a simple and well-tried method for designing and developing solutions to emerging problems. The iterative method is universally used almost everywhere.
2. An essential part aspect for using these methods is the call for well designed benchmarking practices that may measure the quality of critical sub-processes (as well as the result). These system indicators are invaluable for drawing conclusions in order to recommend actions for further iterations.
3. In this perspective it seems evident that classical statistics on administrative areas are primary useful as benchmarks indicating the state of a society or a group of societies.

Part 2. In search of a Prototype

Marja Tammilehto-Luode, Statistics Finland
Philippe Guiblin, The Office of National Statistics (UK)
Lars H. Backer, Statistics Sweden

Abstract

In this section is described how the conceptual "design" t (the definition) and the "method" is transformed into o a real experiment.

1. In the first chapter, or introduction the method is directly related to the practical task of setting up a “experiment” to serve as the experimental environment where ideas to build or improve the design will be tested.
2. In the chapters on theoretical and practical assessments it is demonstrated that we have a crude and simple but working embryo that shows great potential for development.

The key to the method adopted for this purpose is inspired by idea that the design for a solution should be founded on both theoretical argument(s) and practical experiment(s). Consequently the embryo to a prototype is assembled according to an ad-hoc approach and gradually improved over series of iterations, where each “solution” is evaluated according to theoretical- and practical- assessments. The main objective for this discourse is to formulate recommendations for new iterations.

2.1 To build an experiment

Lars H. Backer, Statistics Sweden

2.1.1 Introduction

The main reason for building a Prototype is not merely to produce a "System of small statistical areas" but rather to build an experiment in the form of a complete "system of small area statistics" that will be developed along with the system of statistical areas (regular and irregular tessellations) into a well functioning integrated system. In this sense the real objective of this project is to develop a system of small area statistics, first to serve as an experiment for experiments with small statistical areas, and later to serve as a system of small area statistics.
The method designed to this end is a conventional iterative R+D method, where a prototype is designed based on empirical and theoretical information, and developed over a series on iterations that alternate between theoretical and practical assessments. For each iteration our practical and theoretical knowledge increases and increases the chances for making effective improvements.

Stockholm August 2001

Lars H. Backer
Statistics Sweden

2.1.2 Method

The method selected and used in our effort to define s prototype to a system of small area statistics for Europe is neither new nor spectacular. The idea is simply that staring with a study of "User needs" in the light of the current "Professional context" we will formulate a conceptual "Vision" of the solution we are seeking. In a creative "leap" we will then make an attempt to build an "ad-hoc" practical solution (a Prototype) that as close as possible imitate the performance described in our definition. Using the "Vision" as a guiding star, the prototype is developed over a series of iterations that alternate between theoretical and practical assessments were the prototype is changed according to a set of recommendations produced as a result of tests (benchmarking) and evaluations

2.1.2.1.1 The case study

Everything starts from user needs. The key to understanding user needs is a well researched "Case study" that serves to describe a real situation where a solution like the one we are contemplating will make a "real" difference.
In the current project we have focussed on the need to provide qualified information on and for "Spatial development" in Europe to support political action funded though the structural funds.
Close to 80% of the population in Europe is at present living and acting in urban areas, or at least areas with relatively high population densities, that represent a very small part of the total territory in most member states. Community development therefore is closely related to qualitative and quantitative aspects of such areas. It is therefore a key problem for geo-statistics to contribute to our knowledge about these functional and structural areas in way that are comparable across the EU.
We have consequently, as the first "Case" to be solved with the prototype for a system of small statistical areas tried to use existing data sets, features and methods to make comparable delimitations of urban areas for our test regions.

2.1.2.1.2 The professional context

In order to find a solution it is essential that we use the rich source of knowledge accumulated in what we call the professional context. This heterogeneous system of theoretical and practical knowledge represents a wealth of knowledge that may help us to make a design that is in the forefront of current practice, and help us to avoid costly mistakes.

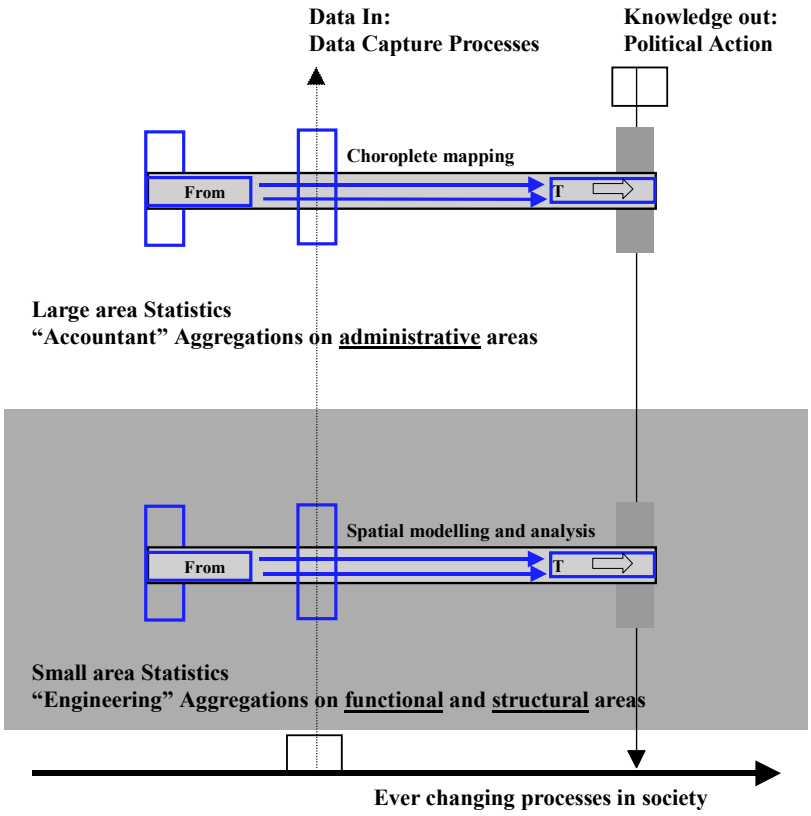


Figure 9: One statistical system, two dimensions

2.1.2.1.3 The vision

The vision is the conceptual "design" for a system of small area statistics as a solution that promises to meet user needs not only in the short term but also in a larger perspective. As long as we have no Prototype to a system of small area statistics we will have to build one based on rational reflection.

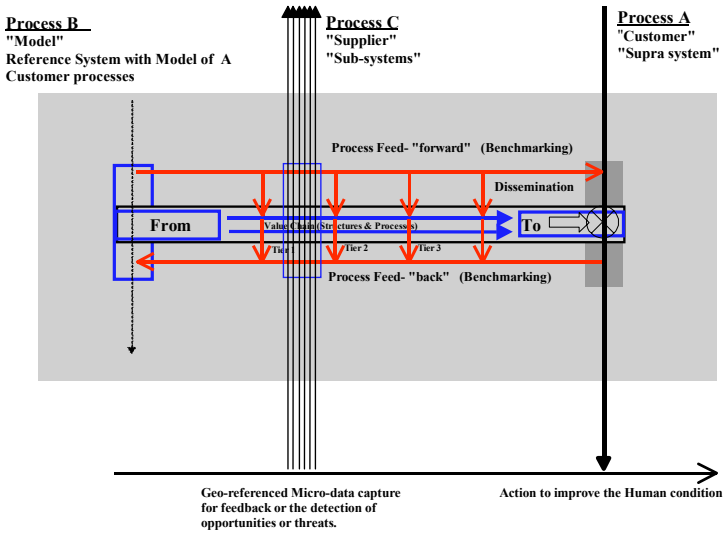


Figure 10: The experiment

In the first part of this report we have made an effort build a conceptual "definition" of a system for small area statistics based on the "state of the art" in the production of information.
The main characteristic of this definition is that it extends the ideas of a statistical system from a simple combination of data and geographical features as in classical statistics. In order to cope with new and constantly changing user needs, this "static" notion of statistics has to be expanded to take full notice of the central importance of production processes, both on the supplier and the user side.
This makes the "definition" or "vision" suggested a little complicated (especially at this early stage), but we believe that this along with other aspects of this quest will like in the Pygmalion myth emerge from this crude "subject matter" as the project advances. (for the formulation please see Part 1 of this report)

2.1.2.1.4 Basic Assumptions

1. Start with what exists.
   A system of small area statistics cannot be built from scratch, We have to do with what is given, This implies the use of existing systems of small statistical areas as enumeration districts etc.
2. Improved successively
   It is assumed that due to the heterogeneity of the data-sets initially presented by the member countries a system of small are statistics must emerge from a diligent harmonisation development process spanning over years.
3. Locally produced, centrally assembled (and developed)
   It is assumed that the member countries NSI's will be reluctant to send micro-data beyond their borders. This implies that a system of small area statistics must be pre-processed separately in each country and assembled centrally like pieces in a puzzle.
4. Both point and area based statistics
   It is assumed that Compiled by a system of both regular and irregular tessellations. This is not only due to the fact that we will probably always have to deal with parallel geo-statistical systems based on both point and area statistics

2.1.2.1.5 The "design"

We have chosen to adapt the method a design method that rests on the principle that all products produced according to the principles rendered by the scientific method should stand on two feet:

1. Theoretically well founded, which means that it should be based on a system of knowledge (ideas) that is "sound" in relation to a selection of common rules
   - No internal contra-diction's
   - Public (open to critical appraisal)
   - Fit into a wider system of knowledge (ideas)
   - Etc.
2. Empirically founded, which means that it should have a bearing on situations in the "real" world. In our case this means that a design should fulfil a demand for "Utility". We have in the first section argued that it is essential that the solution sought be soundly founded on user needs described through real "Case studies" or case studies.

- That the experiment should be repeatable that it should function on different data sets or situations.
- Etc.

2.1.2.1.6 A method for development

The method for development adopted by this project is well known and frequently used wherever products have to adapt to a constantly changing world. In such situations no solution is final and the method for development is an integral part of the production process.

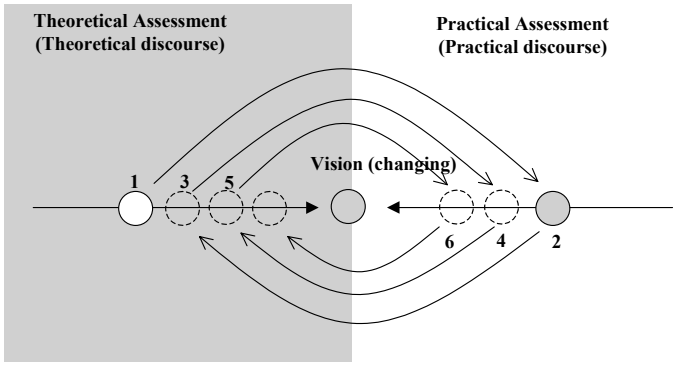


Figure 11: The development of a conceptual or "real" prototype through a constant dialogue/discourse between a theoretical and a practical assessment.

Each full iteration between a theoretical and a practical assessment will hopefully bring the prototype and our theoretical understanding of the questions involved, ever closer to the concept we have in "mind" (our vision). For practical reasons we have to start with a relatively limited empirical situation. As the project evolves we will have to expand this "view" aiming at a solution that eventually will cover the whole of the EU. We are in this report referring to this as the result of the fist major iteration. This does not mean that we have not used the iterative method to develop our "embryo" to a prototype through a series of internal iterations, but only that the project has in this period not changed its focus beyond the problem of delimiting urban areas. The work has therefor been focused on making an "ad-hoc" solution work with especial focus on "methods".

2.1.3 The Result

The result of this "first" iteration is that we believe that we have an "embryo" to a working solution. It has been tested on just a single "case study" using data from very limited "situation". Still, it seems to be ready for the next challenge that may scale up the "case study" and expand the system of "methods" to more applications.

2.2 A Theoretical Assessment (WP 2.0)

Marja Tammilehto-Luode, Statistics Finland

Abstract

This work package suggests alternative methods of constructing comparable building blocks, geographical units for a European-wide geo-statistical system. Alternative methods are studied from two different approaches. The regular tessellation approach aims at a system of statistics by geo-referenced grid squares and the irregular tessellation approach aims at the harmonised small area statistics with irregular geographic units. This work package discusses theoretical background and practical implications of these two approaches.

2.2.1.1 Introduction

Regional statistics of the European Union are managed on the basis of the NUTS (Nomenclature des Unites Territoriales Statistiques) system. The system of the NUTS areas, and administrative areas in general, is far from ideal for flexible and comparable regional statistics. The system faces the problem of what has been termed the "modifiable area unit problem" (MAUP, e.g.)[30]. However, little effort has been made internationally to achieve better spatial comparability in terms of regional statistics. The harmonisation of statistics has, more or less, concentrated on the concepts, definitions and classifications of data.

Each country has its own building blocks by which regional statistics are compiled. Administrative areas are not always the only ones. There is a strong need for a system of territorial units with relatively small populations (statistical blocks) that could be used in combination with the NUTS hierarchy. The problem is how to harmonise these very diversified systems of small area statistics.

This study tries to find alternatives for a statistical system with comparable geographical units. The aim is to draft a theoretical system for relevant building blocks for displaying, analysing regional statistics. The drafted system has to be usable in the contexts and analyses of Geographic Information Systems. The drafted building blocks need to be powerful not only for the description of structural differences of regions, but also for the description of processes and their changes between regions.[31].

The final objective of this study is to select potential methods to be tested with test data and define a relevant case study to be performed with test data. This is part of the " Feasibility Study Towards a Common Geographical Base for Statistics Across the EU" made by a consortium of three partners; Statistics Finland, Statistics Sweden and the Office of National Statistics in the United Kingdom. Statistics Finland is responsible for this part of the study.

Helsinki August 2001

Marja Tammilehto-Luode
Statistics Finland

[30] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*
[31] (Backer 2001) *Accountants and Engineers; why the difference?*

2.2.1.2 Problem description

There are two major geometrical data models. One is based on a regular, usually square, tessellation subdividing the space into cells of a regular size and shape. Another is based on an irregular tessellation with an irregular subdivision of space.[32]. Along with the usage of GIS (Geographic Information Systems) -tools, the opportunity to be able to use both data models has become increasingly important.

There are two main traditions of collecting and geo-referencing statistics within the Member States. The statistical systems may be classed as either register-based, area-based or some mixture of the two. Finland and Sweden both belong to those countries in which statistics production is essentially based on comprehensive registers. The approach to geo-code statistical units is point-based. Statistics Finland has geo-coded buildings, and Statistics Sweden real estate, for linking with other statistical entities. This could be the main reason why statistics by relatively small areas, such as grid squares of 1 km x 1 km, or even smaller, have been produced and used quite widely in both countries. The United Kingdom is an example of those countries, where data collection is mainly based on areas (such as enumeration areas in Censuses) and geo-references of data units are made by defining the boundaries of collection areas or the reference points in them. This approach allows for a different accuracy of spatial configuration depending on the size of the statistical area unit. Usually statistics are produced by different areas, by irregular tessellation.

The current system of regional statistics of the European Union is managed on the basis of the NUTS (Nomenclature des Unites Territoriales Statistiques) system. These units were established by Eurostat to provide a single uniform breakdown of territorial units for the production of regional statistics for the European Union. It has the benefit of being well established, reasonably stable, hierarchical and, in general, well matched to the national statistical regions. The NUTS framework is, however, far from ideal for many applications. The system is proving insufficient due to differences in the sizes of the units at the same level – such as, for example, the NUTS 3 regions in Finland and Belgium. The system faces the “modifiable area unit problem” (MAUP etc.)[33].

There has been a constant need for more comparable regional statistics for the whole of Europe, or even for the individual countries for different purposes[34]:

- for more effective visualisation of the data
- to combine or compare data on different spatial units
- to combine or compare data on different spatial scales
- for statistical analysis of the data (to test spatial patterns and trends)

The aim of the current work is to summarise the studies of theoretical assessments of systems of statistics by regular and irregular tessellation. The purpose is to assess the theoretical synthesis of different approaches for a geo-statistical system that could lead to an improvement of the comparability of statistics across the EU.

---

[32] (Frank and Mark 1993) *Language Issues for GIS*
[33] (Martin 2000) *Towards the Geographies of the 2001 UK Census of Population*
[34] (Eurostat 1999) *GIS Application Development. Final Report. SUP_COM 1997 -LOT 3*

However, this is not a project to extend or change the NUTS hierarchy of administrative areas. This is rather a project aiming at improving the comparability of statistics by studying better means for aggregating or dis-aggregating statistics on other hierarchies of regular and irregular tessellation. This is a project to study alternative building blocks for describing patterns of distribution of phenomena better than the traditional systems of administrative areas, such as NUTS.

2.2.1.3 Methods

2.2.1.3.1 The framework for methods

This study is carried out parallel to two approaches; the regular tessellation approach and the irregular tessellation approach. Statistics Finland is responsible for the regular tessellation approach and the Office of National Statistics is responsible for the irregular tessellation approach.

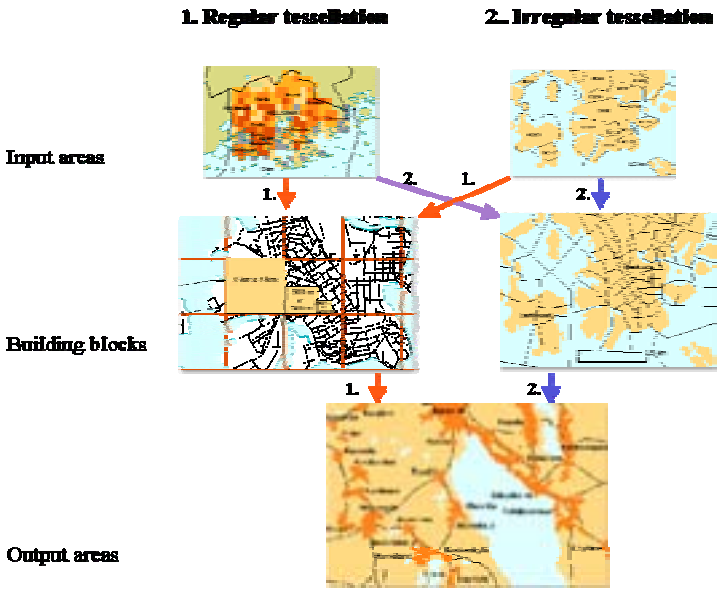


Figure 12: Regular tessellation approach and irregular tessellation approach from diversified input areas to harmonised building blocks and comparative output areas.

The key question is how should the geo-statistical systems be constructed if they include optimised building blocks of either regular tessellation or irregular tessellation. The fact is that input data for both kinds of building blocks are alike – in the European Union input data are very diversified including point-based and area-based geo-references. A final goal is that both approaches, the regular tessellation approach and the irregular tessellation approach, would create building blocks, which are valid for comparative studies of spatial phenomena all over Europe. The results of spatial analysis by using these building blocks should be more relevant and of a better quality than those by using the traditional system of NUTS areas (see Figure 12 above).

These two approaches are studied separately in order to design two theoretical benchmarking systems of comparable building blocks for further spatial analyses. The practical and realistic methods of constructing the prototypes are chosen according to the related literature and empirical experiences of the consortium members. A case study to test the usability of the two prototypes was chosen after several discussions and inspired by ESDP[35] and Eurostat (1998). The prototypes are used for the delineation of urban areas of test areas. This is to compare the two systems of building blocks and their use in spatial analysis as well as the results with the same kind of delineation by data using the NUTS system.

2.2.1.3.2 Regular tessellation approach

The critical points for constructing a grid-based statistics concern
1) The input data itself
2) The size of the grid cells and
3) A projection system

There are several methods for compiling point-based or polygon-based data to grids. Different methods give different results. The results are highly dependent on the characteristics of the input data and the amount of information available[36]. There are specified methods for both point-based and polygon-based data. There are different methods for data with categorical and numeric variables or extensive and intensive variables[37].

The Nordic system of grid-based statistics was chosen as a candidate for the best practice concerning grids constructed from point-based input data. Three candidate methods for converting polygon-based data to grid data were chosen for further testing. The Finnish data give an excellent opportunity for testing different methods. It is possible to compare "real grids" to estimated grids. The real grid-based statistics are constructed by using micro data linked to accurate reference points of the centroids of buildings. Estimated grids, on the other hand, are constructed from polygon-based information, in this case statistics by postal code area using estimation methods.

All candidate methods are applicable to ArcInfo and ArcView software.

An optimum size of grid cell in a grid-based statistical system is mostly dependent on the quality of the input data and the scale of analysis in which the data are used. It is often mentioned that the cell resolution should be the same as, or coarser than, the input data or that to ensure compatibility between countries the ideal cell size should be equivalent to the larger regions[38].

Often the optimum size of a grid cell is also the minimum size of the grid cell, which is defined not only by quality reasons, but also by confidentiality reasons. In sparsely populated countries like Finland, grid-based statistics face confidentiality problems especially in rural areas. Increasing the size of

[35] (European Commission 1999) *ESPD. European Spatial Development Perspective 1999. Towards Balanced and Sustainable Development of the Territory of the European Union.*
[36] (Briggs 2000) *Spatial Transformation Methods for the Analysis of Geographic Data*
[37] (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.*
[38] (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.* and (ESRI 2000) *ArcInfo 8.2: Cell-based Modelling with GRID, ARC/INFO User's Guide*

the cell does not properly solve the problem. There is a need for spatially oriented disclosure control methods.
The size of the cell should also be considered by studying the processing speed and disk space of the potential gridded data. If the disk space is unlimited and the processing speed irrelevant, the analysis should determine the cell size. For instance, it is not practical to have 100-m resolution for data that are examine the whole European population structure.
In this project, the hypothesis of an optimum size of grid is tested by samples of data from two countries. The conversions from polygons to grids are made by using two different cell sizes. The results of the conversions are compared by using Finnish real grids and estimated grids. Comparisons are made by studying the statistical variables of resulting data sets, the confidential problems of data sets and the usability of data sets for spatial analysis.
The production of grid-based data is sensitive to map co-ordinate references, their location and projection system. If the origo is different from the same size grid-nets, the configuration of the data in grid cells is different[39].
Construction, visualisation and analysis of grid-based data require a rectangular co-ordinate system, which does not distort regular polygons, or distorts them as little as possible The projection system has to be a map projection with distance units in metres or kilometres. The idea of grid – based statistics, as an effective spatial database is highly dependent on the option to use distance units in analyses. For a European-wide system, the projection has to be universal and mathematical transformation from national systems needs to be available.
A hypothetical projection system for the whole of Europe is the Universal Transverse Mercator (UTM) –system, which is widely used in grid-based systems of nature and land use information [40]. A common projection for the Nordic grid[41] is also UTM. The UTM system includes sixty zones which keep the distortions reasonably small even on medium and large scale maps. For the test data, the UTM –projection system was chosen, including different zoning for the UK and Finland; It was agreed to use UTM 30 for the UK and UTM 35 for Finland.

2.2.1.3.3 Irregular tessellation approach

Several methods have been developed to produce better statistics based on input data by irregular tessellation. The methods propose different ways of re-aggregating data from a low level of statistical areas to a higher level of areas. The results include changed output areas (a "better output geography"), which are constructed using optimisation algorithms based on consistent criteria.
The current zoning design methods developed by academics rely on the use of one fundamental algorithm, **the automated zone design program (AZP**). This algorithm was originally developed by Openshaw [42] who explored modifiable area unit effects. Subsequently, it was used to experiment with

[39] (Grasland 2000) *Spatial Homogenity and Territorial Discontinuities.*
[40] (EEA. European Environmental Agency 2001) *CORINE Land Cover*
[41] (Tammilehto-Luode and Backer 1999) *GIS and Grid Squares in the use of Register-based Socio-economic Data*
[42] (Openshaw 1977) *A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning, and Spatial Modelling*

zone design. The AZP algorithm is used to improve the consistency in displaying data at one geographical level.
Two British groups have created software packages in which a number of region-building procedures have been implemented, called ZDES (Zone Design System) developed by the Leeds University Department of Geography in the UK and SAGE (Spatial Analysis GIS Environment) developed by the Sheffield University Department of Geography in the UK. These two systems are implemented on a Unix linked with the GIS ArcInfo.
The original AZP technology has been extended to form the ZDES system based on the work of Openshaw and Rao[43], but subsequently extended and improved by Openshaw and Alvanides [44]. ZDES supports three variants of a heuristic search process for finding "optimum" solutions for the target functions. It is a zone-design system for Arc/Info within which the linkages between the GIS and AZP functions are completely transparent to the user. ZDES uses a set of macro programming facilities (AML) for linking functions within and outside Arc/Info and also allows the users to build the menu-driven applications.
SAGE (Spatial Analysis in a GIS Environment) is another system based around Arc/Info, which provides a wide range of facilities for the analysis of area-based health data. The system consists of graphical and statistical software, which calls for ArcInfo running as a server.
To implement the method, this study chose to use Fortran and
Visual basic routines developed on the same principles as ZDES but independently. It seemed more convenient at this stage to use 'non-dependent GIS' software.
The aim is to find optimum groupings of the output regions for any of three exclusive aggregation criteria[45]:

- aggregated regions having near equal values in terms of a selected variable;
- aggregated regions having a similar degree of heterogeneity with respect to the values recorded for their constituent zones
- a location-allocation approach to the problem, in which the arrangement of the input zones, within each aggregated region, reflects an attempt to maximise the weighted accessibility calculated on the specified input variable.

However, the algorithm can also be used to help answer another question of fundamental interest to the Tandem Project; namely, how to compare the statistical properties of systems of irregular tessellation with those of regular tessellation.
A given set of areas (for a given geography) design functions can be calculated using.

- Equal population design function
- Shape function
- Homogenity function

[43] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*
[44] (Openshaw, Alvanides and Whalley 1998) *Some further experiments with designing output areas for the 2001 UK Census*
[45] (Openshaw, Alvanides and Whalley 1998) *Some further experiments with designing output areas for the 2001 UK Census*

For each of these functions, the minimum characterises a state of stability. Regarding the principles of AZP, it produces the best geography when the weighted sum of those three functions leads to a minimum. At the same time, those functions give a statistical characterisation of the state of "stability" of a given system of areas. The interest here is how to use the performance of the AZP algorithm run only once on each of the two systems to be compared and define a set of comparison criteria.
The principal criticisms of the method can be summarised as follows: [46]
There is no assurance that any or all of the design functions will meet whatever minimum or maximum limits are placed upon them;
The quality of the solutions depends on the respective weighting given to the design functions and how each of them is scaled or standardised;
Using multiple objective functions in zone design makes it more difficult to find a good optimum. Alternative objective functions have to be used to find a stable solution. This process might be complex and involve interactive trade-offs between different design functions.
The zoning algorithm can theoretically be used for any level of geography

- Address level data
- Small areas
- Small grid squares providing that these basic requirements are met

Contiguity information and Population/household or other counts for the building bricks – the counts may be a block with address level and grid squares.
However, the numbers of input areas/building blocks per processing unit, such as a ward, have to be considered. The more areas, the greater the number of possible combinations and, thus, the more time needed for processing.
Openshaw and Rao give elements of the performance of the different algorithms involved in a study using 1991 British Census data for 2926 EDs. They raised the following points:
The basic AZP algorithm is good for a small number of OA and starts to deteriorate if more than 100 OA are involved;
The tabu version is better, but less reliable, for more than 150 OA
The simulated annealing algorithm version is the best.

2.2.1.4 Conclusions

The data collected and maintained by National Statistical Institutes are enormously diversified according to spatial configurations. Small area statistics differ from country to country and sometimes even inside a country depending on the type of the data concerned. Some kind of harmonisation of the input data is greatly needed before they can be used for further, comparable analyses. This is especially true in studies concerning the spatial characteristics of different variables and countries.
This work suggests alternative methods of constructing comparable building blocks, geographical units for a European-wide geo-statistical system. Alternative methods are studied from two different approaches. The regular tessellation approach aims at a system of statistics by geo-referenced grid squares and the irregular tessellation approach aims at the harmonised small area statistics with irregular geographic units.

[46] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

The Tandem Consortium

There are several methods for converting point-based or polygon-based data to grids. On one hand, there is a need to find the best practices for producing grid-based statistics. This is especially true with register-based, accurately geo-referenced data. On the other hand, there are also methodological problems. Different methods of converting polygon-based information to grid-based information give different results. The results are highly dependent on the characteristics of the input data and on the amount of information available. The Nordic system of grid-based statistics is chosen as a candidate for a best practice concerning grid-based statistics constructed from point-based input data. Three candidate methods for converting polygon-based data to grid data are chosen for further testing.

The optimum size of the grid cells in a grid-based statistical system is mostly 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 a minimum size, which is determined not only by quality reasons but also by confidentiality reasons. In this project, the hypothesis of an optimum size of grid tested by samples of data from two countries. The conversions from polygons to grids will be made by using at least two different cell sizes.

A grid-based statistical system is a geographical information system. A common reference system for geographic information is needed to ensure that the data are compatible across Europe. Grid-based data require a rectangular co-ordinate system that does not distort regular polygons, or at least distorts them as little as possible. A hypothetical projection system for the whole of Europe is the Universal Transverse Mercator (UTM) –system, which is widely used in grid-based systems of nature and land use information (EEA 2001). The UTM system includes sixty zones which keeps the distortions reasonably small even in medium and large-scale maps.

The UTM –projection system is recommended for further test data. There is a need to use different zoning for data from the UK and Finland. It was agreed to use UTM 30 for the UK and UTM 35 for Finland.

Several methods have also been developed to produce better statistics based on input data by irregular tessellation. The methods propose different ways of re-aggregating data from a low level of statistical areas to a higher level of areas. The aim is to find the optimum grouping of the output regions for certain aggregation criteria.

There are two British research groups, which have created software packages in which a number of region building procedures have been implemented, called ZDES (Zone Design System) developed by the Leeds University Department of Geography in the UK and SAGE (Spatial Analysis GIS Environment) developed by the Sheffield University Department of Geography in the UK.

For the implementation of the method, this study chose to use Fortran and Visual basic routines developed on the same principles as ZDES but independently from GIS software. Professor David Martin from the University of Southampton provided the software and much useful advice during the completion of the project.

The zoning algorithm can theoretically be used for any level of geography providing that some basic requirements are met. However, the numbers of input areas/building blocks per processing unit have to be considered. This is why the test data at this stage of the project should not include too many

polygons. Anyhow, it is suggested that the method be tested on two different categories of data: polygon-based data and grid-based data.

2.2.2 Systems of irregular tesselations (WP 2.1)

2.2.2.1 Introduction

This work-package proposes the evaluation of zoning methods able to provide different output areas for geographical statistics. The System described in this work-package tries to give the best possible presentation of Zoning Design Methodology. Theoretical background and practical implications are discussed.
The aims of this work-package are to propose different alternatives for building output areas for European statistics and to evaluate their statistical characteristics. Several methods have been developed in order to produce better statistics based on the production of statistics on irregular tessellation. Methods and algorithms are presented below. They propose different ways of re-aggregating data from a low level of statistical areas to a higher level of areas. The result is a changing of output areas (a "better output geography") using optimisation algorithms based on consistent criteria.

2.2.2.2 Problem description

2.2.2.2.1 Professional context

The statistical system used within the European Union is the NUTS system (Nomenclature des unités territoriales statistiques. It provides a single uniform breakdown of territorial units for the production of regional statistics for the European Union.
For many statistical purposes, the NUTS framework is far from ideal. There is therefore a need for spatial transformation of statistical data. Conversion of the data from one unit to another (for mapping of the data, to aggregate or dis-aggregate data to different spatial scales or different geographies, for the purposes of statistical analysis, to provide estimates at unsampled or unmeasured locations…).
For many purposes, statistics need to be generated for spatial units that do not conform exactly to the NUTS regions.

- The differences in the size of the NUTS regions between countries may create misleading maps.
- Differences in characteristics of the areas (e.g. population or age structure), which might be used as denominators, generate variations in rate or ratio data and may produce unstable patterns in the geographical distributions[47].

All these factors create the need for spatial transformation of statistical data: that is the conversion of the data from a spatial unit to another. This is necessary, for example:

- For more effective visualisation (mapping) of the data;
- To combine or compare data based on different spatial units;
- For the purpose of statistical analysis of the data (e.g. to test for spatial pattern or trends);

Several methods have been developed within the last ten years. Many research studies have been developed in the academic framework (e.g. for census purposes). The objectives of this work-package are to propose an illustration of what it is feasible to do in the context of European statistics.

[47] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

Objectives

The objectives of this study are to construct a prototype of region-based system with the test data of two different countries.
A construction of region-based statistics from point-based and from polygon-based source data is described.
This Work-package proposes to present 'Automated Zoning procedures' developed by Openshaw[48] and Martin [49]. This method is tested on real data sets in the work-package WP3.1.

2.2.2.3 Method

2.2.2.3.1 The key points in automating design

The basic algorithm

Basically the current zoning design methods developed by academics lie on the use of one fundamental algorithm, the automated zone design program (AZP). This algorithm is due to Openshaw [50]who originally developed it to explore modifiable area unit effects. Subsequently it was used to experiment with zone design. The AZP algorithm is used in order to improve the consistency in displaying data at geographical level.

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 (M<N). 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 to 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". If F(Z) decreases (increases) sharply to the absolute minimum (maximum) the algorithm is called a "steep descent (ascent) algorithm".

Optimisation under constraints

For the purpose of the study, the user must assign weighted importance to the available design constraints. In a census context for example, the geographer wants to respect a statistical consistency between the size of the output areas, or the nature of the tenure description. Those conditions guided by the study and mathematically defined by the user are called "constraints". There are implicit constraints on Z such that each of the original N zones have 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 may 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.

---

[48] (Openshaw 1977) *A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning, and Spatial Modelling*
[49] (Martin 1998) *2001 Census Output Areas: from Concept to Prototype,* (Martin 2000) *Towards the Geographies of the 2001 UK Census of Population*
[50] (Openshaw 1977) *A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning, and Spatial Modelling*

Openshaw writes: “this optimisation task might be categorised as a constrained non-linear integer optimisation problem. It can only be solved via heuristic methods that may not find the global optimum result; indeed, there is no way of knowing whether there is a single global optimum result to find! The view here is that finding a global optimum may be less relevant than finding extremely 'good' results, however 'goodness' is to be measured”.
The following section describes the contents of the AZP in more details.

Remark: terminology
We deliberately in this report have used the terminology defined below in order to distinguish the different steps of the aggregation process.

- Input areas - Seed areas, are small (usually the smallest in a country) areas which are geometrically and geographically defined and to which data are or can be aggregated
- Building blocks are a set of areas, which are aggregated from input areas with a common criterion of comparable areas.
- Output areas, are results of re-aggregating or clustering building blocks by any spatial analysis

The theoretical literature uses the term Output Area (OA) for the areas as a result of “elementary” aggregation process. We prefer the term Building Blocks (BB) and keep the term Output Areas for larger domains as a result of aggregation of BB according to the purpose of the study (e.g. delimitation of urban areas, an urban area is an Output Area).

2.2.2.3.2 Algorithms, objective functions, constraints and software

The AZP algorithm

The principles

The optimisation is driven by three principles:

- Given a study of N communes 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;
- this resultant set of pseudo-regions (regionalisation) provides new sets of output areas (OA) at a higher level of geographical definition.

The AZP (automated Design Procedure) algorithm: a local boundary optimiser

Here is presented the AZP algorithm [51].

- *Step 1*: Start by generating a random zoning system of N small zones into M regions, M < N
- *Step 2* : Make a list of the M regions
- *Step 3* : Select and remove any region K at random from this list
- *Step 4* : Identify a set of zones bordering on members of regions K that could be moved into region K without destroying the internal contiguity of the donor region(s)
- *Step 5* : Randomly select zones from this list until either there is a local improvement in the current value of the objective function or a move that is equivalently as good as the current best. Then make the move, update the list of candidate zones and return to step 4 or else repeat step 5 until the list is exhausted
- *Step 6* : When the list for region K is exhausted return to step 3, select another region, and repeat steps 4-6

[51] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

- *Step 7* : repeat steps 2-6 until no further improving moves are made.

Comments made by the authors of the AZP.

Positive points

The AZP algorithm can work with any type of objective function that is sensitive to the aggregation of data for N zones into M regions.

Negative points

1. A mildly steepest descent algorithm. The AZP algorithm is not the steepest descent direction because the search is the geographically focused on a selected region. (The steepest descent version would scan the global list of all possible single moves that could be made for all M regions and then select the best which is less effective than the basic AZP,[52];
2. The optimum found might be a local sub-optimum and not the overall optimum. Some portions of the graph of the objective function F(Z) might present erratic variations and then temporary optima which might mislead the algorithm to a premature conclusion;
3. The performance of the simple AZP algorithm seems to diminish as the N-into-M aggregation problem becomes harder both because of increasing N and also because of irregularities in zone size and shape (e.g. the number of local sub-optima rapidly increases). The complexity and hardness of the optimisation task appear to interact with the nature of the function being optimised [53].

Variants of the AZP algorithm

In order to improve the performance of the AZP three modifications of the AZP have been considered[54].

Using a "Simulated annealing algorithm"

Under conditions of simulated annealing, a certain level of sub-optimal moves may be accepted according to the overall "temperature" of the solution. The temperature is set to an initial high value and gradually decreased through successive iterations through the algorithm. It can be used in order to reduce the possibility of being trapped by a local sub-optimum. The idea is to extend the search process in considering two, three or more moves at a time in the search for a local improvement.

Using "a tabu search heuristic"

This method allows the search process to escape from local optima whilst avoiding cyclical behaviour specifying that some moves are not allowed by the algorithm. If a zone *I* is moved from region *K* to region *L*, then the reverse move is prohibited (tabu) for *R* subsequent iterations. The key algorithm parameter is the length of a tabu period, *R*.
This value seems to be problem specific.

Using "Parallel algorithm"

This method allows large numbers of zones to be handled or more complex zonal optimisation problems. Implementation of AZP as a parallel algorithm is problematic due to its local sequential search for improvements in the objective function.

2.2.2.3.3 Objective function

[52] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*
[53] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*
[54] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

What are the properties required by an ideal output geography?
Cliff et al (1975) suggest that in general a good zoning system should be

- as simple as possible,
- homogenous and
- compact.

There is a more extensive list of 'tests' for census geography in
Coombes, M. [55] addressed the issues in a census framework. Considering the main points he raised in order to produce a better output geography here are the objectives this study will be focused on:
The building blocks should be as small as possible given confidentiality constraints;
The building block system should provide a consistent basis across the across Europe;
Do the output areas represent 'real world' entities, such as settlements, that can be recognised? (see WP3).
Do these areas allow comparison with previously published data (see WP3)?
Is the whole land surface covered without gaps (see WP3)?
Are the boundaries available in digital form (see WP3)?
Can they be readily and accurately linked to the geography codes used in other data sets? (see WP3)
Wise et al (1997) use similar criteria to (2) and (3) and suggest that the zones should also be of equal size. Martin[56] uses population size, shape, and homogenity,
and defined the objective function as the weighted sum of up to three different design functions**.**

Design functions

The three main design functions can be summarised as:

Equality population zoning:

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

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

or:

$$F_{pop}(Z) = \sum_{j}^{m} abs\left(\sum_{i}^{n}\delta_{ij}P_i - T_j\right),$$

where

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

P_i is the number of people in zone i,
T_j is the target size for region j.
The two design functions presented 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**;**

[55] (Coombes 1995) *Dealing with census geography: principles, practices and possibilities*
[56] (Martin 1997) *Implementing an automated census output geography design procedure*

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

Shape design function

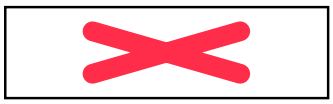
The function to minimise is:

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

BB_{ascj} is the shape score for area (BB) j,
H_{scj} is the hexagon score j.
The shape score is the total squared boundary length of the building block (or hexagon surface) divided by area.
Remark
The hexagon is used as the most compact shape capable of completely filling a 2-D area. The score for a square with unit sides would be boundary = 4 and area = 1, so score = 4*4/1 = 16. A hexagon has a score of $8\sqrt{3}$ approximately = 13.9.
The GIS calculates the Output area shape score.

Homogenity design function

This function is also called by some authors, the tenure homogneity constraint (if tenure is the census variable used to compute it). Conceptually it could be based on any social classification.
It measures the squared difference from a completely uniform social composition (measured in terms of the proportion of households falling into various tenure classes, for example):



The function to minimise is:
BB_{dhj} is the dominant tenure homogenity proportion of area j,
If $BB_{dhpj} = 1$, only one tenure class is present in the BB j, the BB j is at its highest level of homogenity. The further to 1 BB_{dhpj} the less homogeneous the BB j.
Example:
Here is one example of how to determine the dominant tenure homogenity proportion. In a census context, it is determined by four tenure types (owner, rented, council, and non-permanent). For each BB dominant tenure group is calculated, so if a BB has say 60 households with:

- 45 owner
- 10 rented
- 3 council
- 2 non-perm

then "owner" is the dominant tenure of consisting of 45 owner households, and the dominant proportion is 45/60 = 0.75
The three functions can be weighted according to user criteria and added to ensure overall optimisation.

Other possible design functions

- Correlation function**:** of interest mainly to demonstrate aggregation or modifiable areal unit effects.

- Distance function: fit a distance decay function to the data with respect to a user-selected central point.
- Spatial autocorrelation function**:** of interest because it represents a measure of global map pattern.

Constraints

As presented above constraints are mathematical elements of the objective function defined by the user in a second place. Shape and Homogenity design functions may, for example, be categorised as shape and homogenity (design) constraints if the user prefers this terminology.

Other constraints may be defined in order to respect:

- a population size limit for each output area;
- characteristics related to the contiguity of areas to be merged.

Remark: constraint of compactness

David Martin (1997) uses a constraint of compactness, which he defines like this:

"A compact zoning system will tend to have a minimum sum of within zone travel distances around the point of minimum aggregate travel. This can be simply expressed as a set of M separate P-median problems, one for the members of each of the M output zones. Optimising the sum of these M P-median problems will tend to produce naturally compact zones that automatically adapts to the local distribution of address points".

As an example, in a census context the main constraints used are:

- No BB to be below threshold (say, 50 population and 25 household): in a census context, obtaining identically sized zones is often less important than ensuring that the smallest one exceeds a specified minimum size;
- All BBs must consist of contiguous postcodes (unless offshore islands are involved);
- All parts of a split postcodes must be placed in same BB;
- Urban and rural postcodes (defined in terms of address point density) ideally must not be placed in same BB. This constraint might be ignored if it would mean violation of one of the other constraints.

Penalty functions

Another method for handling constraints in such an optimisation problem is to add a penalty function to the objective function that reflects constraint violations. The weighting given to these constraint violations is then gradually increased so that a sequence of unconstrained optimisations will gradually move towards a solution of the original constrained problem.

Powell-Fletcher method

The method due to Powell[57] is regarded by geographers[58] as a "far superior alternative to handling constraints". He developed a modified penalty function approach that has two sets of controlling parameters. This was later generalised to handle inequality constraints; see Fletcher [59].

Available software

Two British groups have created software packages in which a number of region building procedures have been implemented, called ZDES (Zone Design System) developed by the Leeds department of Geography and SAGE (Spatial Analysis GIS Environment) developed by the Sheffield department

[57] (Powell 1969) *A method for nonlinear constraints in minimisation problems*
[58] (Openshaw 1996) *Developing GIS Relevant Zone Based Spatial Analysis Methods.*
[59] (Fletcher 1987) *Practical Methods of Optimisation*

of Geography. These two systems are implemented on a Unix workstation linked with the GIS ArcInfo.

ZDES: an ArcInfo Interface

The original AZP technology has been extended to form the ZDES system based on the work of Openshaw and Rao[60], but subsequently extended and improved; see Openshaw and Alvanides [61].

The authors write:

The aim is to find optimum groupings of the output regions for any of three exclusive aggregation criteria:

- aggregated regions having near equal values in terms of selected variable;
- aggregated regions having a similar degree of heterogeneity with respect to the values recorded for their constituent zones
- a location-allocation approach to the problem, in which the arrangement of the input zones within each aggregated region reflects an attempt to maximise the weighted accessibility calculated on the specified input variable.

ZDES supports three variants of a heuristic search process for finding "optimum" solutions for these target functions.

A zone-design system for Arc/Info within which the linkages between the GIS and AZP functions is completely transparent to the user.

It involves the use of a set of macro programming facilities (AML) for linking functions within and outside Arc/Info and also allows the users to build the menu driven applications.

Choice of the design function and implications within ZDES

The algorithms proposed by ZDES can solve the problem of the optimisation of one objective function (one "equality zoning " design function + "size" and "homogenity" design functions) subject to constraints.

In ZDES some constraints are implicit in the algorithm and can never be violated (e.g. contiguity).

Other constraints are defined by the user and can be applied to each of the individual M output zones.

The principal criticisms can be summarised as follows[62].

There is no assurance that either any or all of the design functions will meet whatever minimum or maximum limits are placed upon them;

The quality of the solutions depends on the respective weighting given to the design functions and how each of them are scaled or standardised;

Using multiple objective functions in zone design increases the difficulty to find a good optimum. Alternative objective functions have to be used in order to fin a stable solution. This process might be complex and involve interactive trade-offs between different design functions.

A simpler strategy is to select one of the design functions as the objective function and treat the others as equality or inequality constraints, setting realistic, explicit, and attainable target values for them. [63]

---

[60] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

[61] (Openshaw, Alvanides and Whalley 1998) *Some further experiments with designing output areas for the 2001 UK Census*

[62] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

[63] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

The Tandem Consortium

Remark: Operational Information

Openshaw and Rao give elements on the performance of the different algorithms involved on a study using 1991 British Census data for 2926 ED's. They raised the following points:

The basic AZP algorithm is good for small number of OA and starts to deteriorate if more than 100 OA are involved;

The tabu version is better but less reliable for more than 150 OA

The simulated annealing algorithm version is the best;

Compute times in seconds (for a Sun-Supersparc):

Number of OA	Basic AZP	AZP-Tabu version	AZP-Simulated Annealing
10	730	5169	19575
75	2010	5684	10493
100	2053	5169	12579
200	2229	7875	131766

SAGE (Wise and al.)

SAGE is another system based around Arc/Info, which provides a wide range of facilities for the analysis of area-based health data. The system, called SAGE (Spatial Analysis in a GIS Environment), consists of graphical and statistical software, which calls ArcInfo running as a server. The SAGE system provide regionalisation tools able to attempt to find the optimum solution to any or all of three aggregation criteria:

- to minimise the within-region variance of one of more than one variables;
- to minimise the differences between the value for a variable within any region and the overall mean regional value – this is comparable with ZDES;
- a simple attempt to maximise regional compactness.

Remark

Sage only works on a Sun Unix workstation.

Using AZP to compare different system of output areas

The AZP algorithm is used in order to create a new system of output areas from a given system. However the AZP algorithm can also be used to help to answer another question of fundamental interest to the Tandem Project, namely how to compare the statistical properties of systems of irregular tesselations with those:

- of grid squared;
- of grid based systems.

Regarding the paragraph 2.1 for a given set of areas (for a given geography) design functions can be calculated.

- Equal population design function
- Shape function
- Homogenity function

For each of these functions the minimum characterises a state of stability. Regarding the paragraph 2.1 AZP produces the best geography when the weighted sum of those three functions leads to a minimum. At the same time those functions give a statistical characterisation of the state of "stability" of a given system of areas. The interest here is how to use the performance of the AZP algorithm run only once on each of the two systems to be compared and define a set of comparison criteria.

To what kind of lower level units can AZP be applied?

The zoning algorithm can theoretically be used for any level of geography

- Address level data
- Small areas
- Small grid squares

providing that these basic requirements are met:

1. contiguity information and ;
2. population/household or other counts for the building bricks - the counts may be a block with address level and grid squares.
3. The numbers of building blocks (BBs) per processing unit, such as ward, have to be considered, the more such BBs, the greater the number of possible combinations, and thus the greater time needed to process.

This is something need to consider especially if you wish to use address level BBs.

2.2.3 Systems of regular tessellation (WP 2.2)

Marja Tammilehto-Luode, Statistics Finland

2.2.3.1 Introduction

This work will focus on the problem of defining a system of regular tessellation that may be used to increase the comparability of statistics across the European Union. The aim is to design a draft of the best current alternatives for a framework of statistics by regular tessellation in Europe. The purpose is also to discuss the problems and advantages of such a grid-based system.
The aim is not to suggest a final solution but rather a seed from which a satisfactory solution may emerge after a series of iterations. The objective of this work is to study methods that may be used for constructing a prototype of a grid-based statistical system for Europe. The purpose was also to design tests and a case study for a prototype of a grid-based statistical system.
This work is part of the study of " A Feasibility Study towards a Common Geographical Base for Statistics across the European Union" carried out by the Tandem Consortium. The Tandem Consortium comprises representatives from Statistics Sweden, the Office of National Statistics in United Kingdom and Statistics Finland. Statistics Finland is responsible for this study.

2.2.3.2 Problem description

2.2.3.2.1 Professional context

Although most National Statistical Institutes use grid –based data in one form or another, there exists no generally accepted system of regular tessellation, which is adopted when, for example, comparing statistics in cross-border situations. For larger areas, consisting of several countries, systems for grids are often for environmental issues (ETC/NC, CORINE Land Cover, and GRID-Arendal). However, grid squares have also been used to model population globally. The Gridded Population of the World (GPW, www.sedac.ciesin.org/plue/gpw) data includes population estimates for 1990 and 1995 on a 2.5' grid.

A short history of grid-based statistics

In Nordic countries, grid squares have been used for displaying population distribution for more than 100 years. The first known grid map was published in the mid-19th century, when the public railway was being planned in Sweden[64]. The purpose of the map was to calculate the number of people living within various distances from alternative railway routes.
Grid maps were increasingly used in the 1960s as a basis for calculating the population in areas, which were defined functionally rather than administratively (Claeson[65], General Census of Population in Finland[66].
Finland was the first country in Norden to put accurate geo-references to statistical units in the 1970 Population Census. These geo-coded points (centroids of the buildings identified by map co-ordinates) and their links to statistical units make the production of small area statistics quite flexible.

[64] (Öberg and Springfeldt 1991) *Befolkningen*
[65] (Claeson 1964) *A Chorological Public Analysis*
[66] (Statistics Finland 1960) *Non-Administrative Urban Settlements and their Boundaries, etc Helsinki, 1965, 1960.*

The Tandem Consortium

Data improvements and an implementation of GIS tools in the early 1990s made the use of grid-based statistics increasingly popular. International Grid-based environmental and population atlases were provided with databases for GIS-users (ETC/NC, GWP). At the same time in Norden, socio-economic statistics were increasingly made by grid-geography for researchers and, later, for a private sector for such as market planning (Muilu [67], Räisänen [68], Rusanen *et al.* [69], Vaattovaara [70], Öberg, and Springfeldt. [71]).

Advantages and disadvantages of statistics by grid squares

There are many reasons why grid-based databases are constructed and used considerably, especially internationally.

The traditional advantage of grid-based spatial databases is that they are easily stored and manipulated using a computer. A 'map algebra', which is the kernel of many GIS-packages, is originally based on a raster-based data model[72].

One of the basic advantages of GIS is to combine data by their location. Grid squares are often used for harmonising different types of data sets. Many environmental data sets are, for example, stored as gridded data, including, for example, climate and elevation indicators or pollution data. Analysis of population and environmental data together may be done only by storing both types of data by grid squares.

Since the grid is often a relatively small statistical unit compared to the 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 by choosing small, evenly sized area units like grids[73].

Grid squares have the great advantage of being stable over time. Traditional statistical areas, administrative areas, tend to change in every country, in some more than others. The grid squares are spatially stable and do not move from one year to the next. Grid-based statistics are independent of any regional changes. Spatial changes in the data can be traced easily between different time periods. This makes temporal analyses easier by grid data than by traditional polygon-based data.

An advantage of grid-based statistics is their ability to present a "unpopulated" land or other non-existent variables in space. It would be difficult, if not impossible, for example, to describe depopulation in the Nordic countries if statistical figures were available by administrative units only[74]. Besides studies on depopulation, grid square data have been used in

---

[67] (Muilu, Rusanen, Naukkarinen and Colpaert 1999) *Local Poverty in Finland 1995*

[68] (Räisänen, Rusanen and Naukkarinen 1996) *Socio-economic grid data and GIS for analysing changes in the Finnish countryside.*

[69] (Rusanen, Naukkarinen, Muilu and Colpaert 1996) *Asutus Ruotsissa Suomea keskittyneempana*

[70] (Vaattovaara 1998) *Pääkaupungin sosiaalinen erilaistuminen. (Residential differentiation within the metropolitan area of Helsinki, in Finnish with Abstract in English).*

[71] (Öberg and Springfeldt 1991) *Befolkningen*

[72] (Tomlin 1983) *A Map Algebra*

[73] (Martin 1991) *Understanding socio-economic geography from the analysis of surface form.*

[74] (Haarala and Tammilehto-Luode 1999) *GIS and Registerbased Population Census.*

Finland, for example, in studies concerning regional incidence of diseases[75], unemployment [76], changes in rural industrial structure[77], regional segregation [78] and migration[79]. Regular subdivision of space allows a hierarchical structuring of areas of varying size. Data by small grids can easily be summed up to form larger areas. New statistical areas and different hierarchies of output areas can be compiled.
However the grid cell is an abstract, artificial, somewhat inhuman, spatial unit, which the inhabitants themselves cannot recognise. It is hard to understand the character of grid cells or describe it to those unfamiliar with the system. Often it is necessary to visualise the results by identifying the location of the grids, at least on a small scale. For statistical tables, grid squares need to be aggregated to an administrative or other well-known regional level.
Grid squares seem to be more sensitive to changes in projections. A set of administrative areas (NUTS) and a system of grids will both be distorted when displayed in another projection, but we will react more strongly to the distorted grids, as they will lose their regularity.
Finally grid-based statistics are highly dependent on source data. If source data is already aggregated to polygons, there is a need for estimation methods.

2.2.3.2.2 Problem statement

A construction of grid-based statistics (data are aggregated to grids, which are geo-referenced) faces some major critical points. One needs to recognise at least:

1) Characteristics of input data such as
   - accuracy of geo-references in input data (sample points, address points of each houses…)
   - types of geo-references (point-based or polygon-based already aggregated data)
   - types of variables (categorical, numerical, extensive, intensive)are in input data
2) What is an optimum size of a grid cell for?
   - quality reasons
   - confidentiality reasons
   - scale of the analysis
   - efficiency of processing and analysing
3) What is an optimal projection system for?
   - transformation of input data
   - comparing data of different countries and different scales
   - displaying results internationally

The aim of this work is to study the critical points mentioned above.

---

[75] (Karvonen, Rusanen, Sundberg, Colpaert, Naukkarinen, Tuomiolehto and Group 1997) *Regional differences in the incidence of insulin-dependent diabetes mellitus (IDDM) in Finland during 1987-1991*
[76] (Muilu, Rusanen, Naukkarinen and Colpaert 1999) *Local Poverty in Finland 1995*
[77] (Räisänen, Rusanen and Naukkarinen 1996) *Socio-economic grid data and GIS for analysing changes in the Finnish countryside.*
[78] (Vaattovaara 1998) *Pääkaupungin sosiaalinen erilaistuminen. (Residential differentiation within the metropolitan area of Helsinki, in Finnish with Abstract in English).*
[79] (Kauppinen, Rissanen, Rusanen, Naukkarinen, Muilu and Colpaert 1997) *Migration as a function of population*

The problem is to find the best available solutions for constructing grid-based statistics for Europe.

2.2.3.3 Methods

2.2.3.3.1 General research interests

The purpose of this work is to draft a prototype for a grid-based statistical system (for the whole of Europe) and to plan a method of testing it. The prototype will be constructed using Nordic experience and data together with existing knowledge and applications with different data structures and methods in other countries.
The primary research interest is to find the optimal territorial division in the form of grid squares – geographical configuration of population data – to analyse similarities and differences in areas between different countries. One purpose is to test candidate methods of creating grid-based statistics. A prototype of a grid-based data set will also be tested in a real user case: to define comparable functional areas of urban and rural characteristics.
The focus will be on comparability and usability of spatial data structure in GIS analysis.

2.2.3.3.2 Methods of creating grid cells

There are several approaches for creating grid cells from unit data (e.g. Ohtomo 1997[80]). The most precise results will be achieved if individual statistical units 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 is simply aggregated microdata by point in the polygon method or even simpler calculated frequencies within grid cells by using the map co-ordinates of each unit point[81].
If the input areas are polygons, such as enumeration areas or postcode areas, which do not fall exactly into grid squares or the areas overlap several grid cells, an estimation process is needed to change an irregular polygon structure into a regular grid structure.
One option is to allocate input areas to a grid square, if more than half the area falls into that cell. On the other hand, if an input area is larger than a grid square, data can be assigned in total to a grid cell, which contains the centroid of the input area. Alternatively, the data can be distributed evenly across all grid cells that fall into the input area.
The centroids or representative points can also be used directly to assign input data to grid cells. A method using population-weighted representative points will generate better results than one using GIS-calculated geometric centroids [82].
There are also more complicated areal interpolation methods, which can be used to estimate polygon-based data for grid cells. Many of these methods are

[80] (Ohtomo 1997) *Small Area Statistical Databases*
[81] (Tammilehto-Luode and Backer 1999) *GIS and Grid Squares in the use of Register-based Socio-economic Data*
[82] (United Nations: Department of Economic and Social affairs 2000) *Handbook on geographic information systems and digital mapping*

described by Eurostat [83], Flowerdew and others[84], Goodchild and others [85] and Fisher and Lagford [86]. Some of these methods can be implemented using standard GIS overlay functions. Due to the drawbacks of standard methods, Eurostat has also developed special programs to transform data for zones to a coverage of grid squares of a desired size[87].
The conversion of polygons to grid square data can also be guided using additional information such as land use, road infrastructure, settlement patterns (Bracken and Martin[88], Martin[89], Langford and Unwin[90] and Deichmann [91]). This was, for example, the case when GRID- Arendal estimated the population density for the Baltic Sea drainage basin region by 5 km x 5-km grids. Urban –rural data sets were generated from the Land Cover data set to guide the various probabilities of population density for urban and rural grids. Population statistics were used in conjunction with the administrative units, as well as rural and urban data sets (www. grida.no.baltic/arcview/popd_gis.txt)

2.2.3.3.3 Candidate methods for the construction of a prototype

Points to grid cells

An aggregation of point-based information to grid cells is a similar procedure to the aggregation of point-based information to any polygon. However, as regular tessellation, the grids include information that simplifies the process considerably. In theory, one first makes a polygon-on-point overlay between the point theme and the grid cell theme. The attribute data of the grid cell theme will be assigned to the point theme. As a result, each observation gets a unique grid cell number, which can then be used as a common denominator of aggregate information within each grid. In practice, map co-ordinates themselves can be used for aggregating geo-referenced points, as has been done in Finland and Sweden. The presumption then is that the co-ordinate system used for the geo-referencing is originally rectangular and, thus, provides an undistorted grid structure.
In Finland and Sweden, the geo-reference for each cell in a grid-net is agreed to be the bottom left-hand corner of a cell as defined by the co-ordinate pair. In our national co-ordinate system, two 7-digit numbers are needed to denote a 1m x 1m grid. Combining the first digit of the x co-ordinate with the first digit of the y co-ordinate gives us a (2-digit) code for a 1,000km x 1,000km

---

[83] (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.*
[84] (Flowerdew, Green and Kehris 1991) *Using areal interpolation methods in Geographic Information Systems*
[85] (Goodchild, Anselin and Deichmann 1993) *A general framework for the areal interpolation of socio-economic Data*
[86] (Fisher and Langford 1995) *Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation*
87 (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.*
[88] (Bracken and D 1989) *The generation of spatial population distributions from census centroid data*
[89] (Martin 1991) *Understanding socio-economic geography from the analysis of surface form.*
[90] (Langford and Unwin 1992) *Generating and mapping population density surfaces within GIS*
[91] (Deichman 1996) *Smart interpolation. A review of spatial Database Design and modelling. Use of GIS in Agricultural Research*

grid. By adding the second x co-ordinate and the second y co-ordinate we get a 100km x 100km grid inside the latter. Therefore, 10 digits are needed to build a code for a 1km x 1km grid. To divide this further, we could cut the co-ordinates to 500m, 250m, 100m, and so on. A similar system can be constructed for any rectangular co-ordinate system (Tammilehto-Luode et al. 2000).

The practical way to calculate statistics by grids with geo-referenced micro data is to

Sort out the data by co-ordinates into descending order

Round the map co-ordinates by the size of the grid cell

Aggregate the data by similar map co-ordinates

The default system of map co-ordinates is a geographic system and each zone must be processed separately.

In the Nordic system, a map co-ordinate is the primary identifier of a grid cell usually acknowledged as the bottom left-hand corner of a grid, as mentioned above. An indexing system of grid cells can also be introduced. One way of indexing is to number the grid cells by the extent of the data when the minimum and maximum of the co-ordinates (geo-referenced points) are known.

The formula for the numbering is as follows:

Index = (x max - x min)/ a *((yi –ymin)/a + (xj-xmin)/a)) +1

where x and y are the map co-ordinates of the cells and a is the size of the grid cell.

A hierarchical indexing system of grid cells is sometimes valuable, for example, for making hierarchical queries more efficient (Tammilehto-Luode et al. 2000). The hierarchical system of grids in Finland is based on the minimum grid size of 250m x 250m. Larger grids are re-aggregated from these grids, meaning that the sides of the larger grids need to be divisible by 250m.

In the system described above, gridded data are made by calculating frequencies within the grid cells with the information of the map co-ordinates of each unit point. Producing numeric data by grid cell only requires an additional 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, for example, a major value, an area-weighted value, and so on. A rather simple approach is to create a cross-table containing the grid cell identity code as the primary key, with additional columns representing the unique values of the case items.

Polygons to grid cells

In this case it is important to differentiate between intensive and extensive variables (Goodchild, Lam 1980)[92]. 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.

[92] (Goodchild and N. 1980) *Areal interpolation: a variant of the traditional spatial problem.*

Intensive variables are easier to handle, because the original polygons are derived from the spatial variation itself [93]. The classic method for converting extensive data is based on combining 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.
In this study, three methods were chosen for further testing. All of them can be made with the ArcInfo software. Two of them are ArcInfo's standard algorithms; PointGrid and PolyGrid[94]. The third method is developed by Eurostat and written in the ArcInfo macrolanguage, AML [95].
The PointGrid algorithm converts data associated with point features to the 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 different 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 falling within their boundaries. The data are interpolated to a new grid square using the following equations:

Extensive $Xj = n\Sigma i=1\ (Xi\ (Aji/Ai))$

Intensive $Xj = n\Sigma i=1\ (Xi(Aji/Aj))$

where

Xj = data for grid square j
Xi = data for region i
Aji = area of intersection of grid square j and region i
Ai = area for region i (extensive)
Aj = area for grid square j (intensive)
n = number of regions

2.2.3.4 Optimal size of a grid cell

When accurate geo-referenced point data are available (as it is in the Nordic countries), the optimal size of a grid is dependent on the scale of the study as well as the amount of confidential grid cells.

---

[93] (Hansen 2001) *PSSD-Planning System for Sustainable development. The Methodological Report*
[94] (ESRI 2000) *ArcInfo 8.2: Cell-based Modelling with GRID, ARC/INFO User's Guide*
[95] (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.*

In practice most Finnish customers who use grid data in their own GIS would prefer grids as small as possible, because of their flexibility. It is easy to aggregate smaller grids into larger grid cells or polygons if a study needs it. On the other hand, there are also data in more detail level available for studies on a larger scale.
The size of the grid in socio-economic statistics in Finland varies from 250 m x 250 m to 5 km x 5 km. The smaller grid size is only used for urban area data. In rural areas, where population density is low, data by smaller grids often face disclosure risk.
In other countries, the grid cell sizes range from 100 metres used in Sweden and in the United Kingdom, to 1 kilometre grids for Japan and the Republic of Korea, to 5 kilometres used for some international databases (United Nations 2000)[96].
In countries where gridded data are estimated using polygon-based information, a size of the grid is theoretically optimal when the transformation from polygon-based to grid-based data structure is most accurate. The quality of the output data is usually a prior problem. In practice, the size of input areas (in terms of surface area and population) varies considerably from country to country. To ensure compatibility between countries, the ideal cell size should be equivalent to the larger regions of input data [97].
It must be recognised, however, that spatial patterns and geographical analysis will give different answers on different scales, and even for different zonal systems defined by the same scale[98]. The modifiable areal unit problem is insoluble since the results reflect the reality of geographical processes. An aggregation of data to neutral units as grid squares is only one way to smooth the effects of different output area sizes The size of the grid cell should also be considered from the processing speed and disk space points of view. 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 Europe.

2.2.3.5 Optimal projection system

A grid-based statistical system of regular tessellation is, by definition, constructed using equal areas. The construction, visualisation and analysis of grid-based data requires a rectangular co-ordinate system that does not distort regular polygons, or at least as little as possible.
The production of grid-based data is also sensitive to the location of the origo. If the origo is different from the size of the grid-nets, the configuration of the data in grid cells is different[99].
A grid-based statistical system is a geographical information system. A common reference system for geographic information is needed to ensure that the data are compatible across Europe. For the time being, each country has its own map projection and their data are locally defined. For a common system of grids, there should be an international projection for the whole of

[96] (United Nations: Department of Economic and Social affairs 2000) *Handbook on geographic information systems and digital mapping*
[97] (Eurostat 1998) *Urban database.*
[98] (Grasland 2000) *Spatial Homogenity and Territorial Discontinuities.*
[99] (Grasland 2000) *Spatial Homogenity and Territorial Discontinuities.*

Europe for small-scale mapping. However, the strength of the grid systems lies in medium and large scale studies which need several projections with either sectors or zones to keep the distortions of grid squares as small as possible. Anyhow, a common reference system has to be universal, and mathematical transformations from national systems need to be available
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 [100]. The UTM co-ordinates 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[101].

2.2.3.6 Tests with empirical data

Concerning the critical parts of the study, four different tests are chosen to be done with empirical data:

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

2.2.3.6.1 Tests of the candidate algorithms for converting polygon-based data to grids

In this study, three methods were chosen for further testing. Two of the algorithms represent standard applications in GIS-software (PointGrid and PolyGrid in ArcInfo[102]). The third method is developed by Eurostat and written in the ArcInfo macrolanguage, AML [103]. In order to make the tests sufficiently relevant, there is a need for different geo-referenced data with different variables.
The Finnish data give an excellent opportunity to test different methods. It is possible to compare “real grids” to estimated grids. The real grid-based statistics are constructed using micro data linked to accurate reference points of centroids of buildings. Estimated grids, on the other hand, are constructed from polygon-based information, for example, statistics by postal code areas, using estimation methods.
In this study, the Finnish polygon-based data (postal code area data) are converted to grids using different candidate algorithms and the results are compared to each other and the real gridded data. At least one extensive and one intensive variable are needed in the test data. Comparisons are made using statistical values (mean, maximum, minimum.) and visually, mapping the results using a converted variable.

2.2.3.6.2 Tests of optimum size of grid cell and optimum projection system

In this study, the hypothesis of an optimum size of grid will be tested using samples of data from two countries; one of which has polygon-based input

---

[100] (EEA. European Environmental Agency 2001) *CORINE Land Cover*
[101] (Tammilehto-Luode, Backer and Rogstad 2000) *Grid data and area delimitation by definition towards a better European territorial statistical system*
[102] (ESRI 2000) *ArcInfo 8.2: Cell-based Modelling with GRID, ARC/INFO User's Guide*
[103] (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.*

areas and the other, which has point-based, accurately geo-referenced input data.
The conversions from polygons to grids are made using two different cell sizes. The results of the conversions are compared using the Finnish real grids and estimated grids. The comparisons are made by studying the statistical variables, confidentiality problems, and usability of the resulting data sets for spatial analyses.
The UTM projection system is chosen for the test data, but with different zonings for different countries. It was agreed that UTM30 would be used for the British data and UTM35 for the Finnish data. The grid-based maps will show a potential distortion of grid squares.

2.2.3.6.3 Tests of usability of the prototype for visualising population structure

A hypothesis that grid-based statistics describe a population structure better than NUTS-based statistics is tested simply by visualising the results of prototypes of two different countries' gridded data against their correspondent NUTS5-level data.
Visualisations of population densities by NUTS 5, 10km x 10km and 1km x 1km grid square are made. The visualisations of gridded data are made as typical choropleth maps without any smoothing methods.

2.2.3.6.4 Tests of usability of the prototype for delineating urban areas.

Finally, a case study is made with prototypes of test area gridded data. The prototypes of gridded data of both countries are used for delineating urban areas of these test areas. The delineation is made using an urban area definition originating from the Labour Force Survey of Eurostat (1998) and a method compiled of ArcInfo standard routines. The results of the delineation, using two sizes of grid cells, are compared with the results using polygon-based, NUTS 5 data. The comparison is made visually and statistically.
The definition of an urban area in this application is:

- **a densely-populated area,** a contiguous set of local areas, each of which has a density of over 500 inhabitants per square kilometre, where the total population for the set is at least 50,000 inhabitants[104].

The method of delineation by grid-based statistics is an application of Statistics Finland for delineating localities. It is accomplished using the ArcInfo 8.2 software[105].
The basic idea is to create zones from selected grid-polygons with population densities equal to, or above, 500/km^2 and then join the population data (summarised by zone) to these zones. The zones whose populations equal or exceed 50,000 are selected.

The method, step by step:

The data to be used include two polygon grids, one for the population density (= **density_grid**) and the other for the total population (= **total_grid**).

- Step 1: Create point coverage **point1** from **total_grid** where each point represents one 1km x 1km or 10km x 10km grid polygon. This step eliminates the risk of redundancy error in the data in for example coastal areas.
- Step 2: Select from **density_grid** the polygons where the density of population is equal to, or above, 500/km^2. Then reselect overlapping

[104] (Eurostat 1998) *Urban database.*
[105] (ESRI 2000) *ArcInfo 8.2: Cell-based Modelling with GRID, ARC/INFO User's Guide*

polygons from **total_grid**. Convert the selected polygons to a new coverage.

- Step 3: Create buffers around the features in the new coverage created in "Step 2". First create the buffers outside the polygons and then, using the same buffer distance, create another set of buffers inside the already buffered features in order to preserve the original area. Output is **buffer_zone**.
- Step 4: With IDENTITY, overlay **point1** and **buffer_zone** to create a new point coverage containing the <buffer_zone>-id. Output is **point2**.
- Step 5: With FREQUENCY, summarise the total population in **point2** by <buffer_zone>-id to create INFO table **FRQ**. Join **FRQ** to **buffer_zone**.
- Step 6: Finally, from **buffer_zone** select the zones with total populations equal to, or above, 50,000.

The delineation of urban areas by using NUTS5 -level data is used applying the above method as much as possible.

2.2.3.7 Conclusions

Although most National Statistical Institutes use grid-based statistics in one form or another, there is no generally accepted system of regular tessellation that is adopted when comparing statistics in a crossborder situation. This study has focused on the problems of defining a system of regular tessellation that may be used to increase the comparability of statistics across the European Union.

Nordic countries have a long experience of displaying population distribution by grid squares. The statistics are compiled with point-based source data, where established links from statistical units to geo-referenced points are available. These point-based source data give the most precise results for grid-based statistics. However, there are also methods of converting polygon-based data to grids, which have been applied in cases when the input data are already aggregated (Ohtomo [106] and Eurostat [107]). Some of these methods can be implemented using standard GIS overlay functions. Different methods are based on different assumptions about the source data and are subject to different types of error and approximation[108].

If the system of grids should cover the whole of Europe, it is obvious that the source data for grids differ from country to country. The methods used to convert source data to grids depend primarily on characteristics of the source data and on spatial characteristics of phenomena being analysed [109]. In this study, three of the "benchmarking " methods were chosen for further testing with empirical data.

Another critical point, when grid-based statistics are constructed, is the definition of an optimal size of the grid cell. An optimal size of the grid cell is also highly dependent on source data and its spatial structure. If accurately geo-referenced source data are available, the optimal size of the grid cell is dependent on the scale of the study, where grid-based data will be used, as well as the amount of confidential grid cells in the output data. In the countries where gridded data are estimated by polygon-based information, the

[106] (Ohtomo 1997) *Small Area Statistical Databases*

[107] (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.*

[108] (Briggs 2000) *Spatial Transformation Methods for the Analysis of Geographic Data*

[109] (Eurostat 1999) *GIS Application Development. Final Report. SUP_COM 1997 -LOT 3*

quality of results after conversion is also critical. The size of a grid cell is theoretically optimal only when the conversion from polygon-based to grid-based data structure is most accurate. To ensure compatibility between countries, the ideal cell size should be equivalent to the larger regions of input data[110]. The size of the grid cell should also be considered from the processing speed and disk space point of view. In this study, two different cell sizes were chosen for further testing with empirical data.

The third major critical point when constructing grid-based statistics is an optimal projection system. On one hand, a common geographic reference system is needed to ensure that the data are compatible across Europe. On the other hand, grid-based data require a rectangular co-ordinate system that does not distort regular polygons, or at least as little as possible. There should be an international projection for small-scale mapping for a European system of grids. However, the grids are most efficient in medium and large-scale studies, which place the demand on several projections. Anyhow, a common reference system has to be universal and mathematical transformations from national systems need to be available. The projection system must also be a map projection with distance units in metres or kilometres. The idea of grid squares as an effective analytical tool relies on distance units being used within the data.
The location of the origo in the grid-net is also very crucial. If the origo is different than the size of the grid-nets, the configuration of the data in grid cells is different [111].
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[112]. In this study, the UTM projection system was chosen for further tests by empirical data.
One of the major advantages of grid-based statistics is that they describe a spatial distribution of phenomena far better than traditional statistics using different sizes of polygons (e.g. administrative areas). The problems known as a modifiable areal unit problem (MAUP) and ecological fallacy difficulties [113] can be partly avoided by choosing small, evenly sized area units like grids [114]. This hypothesis was chosen for further testing by comparing prototypes of grid-based statistics to be made by empirical data with NUTS 5-based data to describe the population structure for test areas.
Grid-based statistics should also be especially good for describing continuous phenomena in space. This is why a case study for delineating urban areas with grid-based statistics was chosen for further testing. The NUTS 5-based delineation of urban areas by Eurostat's definition[115] is used as a reference.
It must be recognised that spatial patterns and geographical analysis will give different answers on different scales and for different zonal systems defined

[110] (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.*
[111] (Briggs 2000) *Spatial Transformation Methods for the Analysis of Geographic Data*
[112] (EEA. European Environmental Agency 2001) *CORINE Land Cover*
[113] (Robinson, Morrison, Muehrcke, Kimerling and Gubtill 1995) *Elements of cartography*
[114] (Martin 1991) *Understanding socio-economic geography from the analysis of surface form.*
[115] (Eurostat 1998) *Urban database.*

using the same scale[116]. An aggregation of data to neutral units of grid squares is only one solution for smoothing the effects of different sized output areas. It has many advantages. But a user of such data has to be aware of the quality and structure of input data and the requirements and expertise of the end-user. However, if grid-based statistics are to be comparable across Europe, there must be standardised, easy-to-use methods of compiling statistics with a common reference system.

[116] (Grasland 2000) *Spatial Homogenity and Territorial Discontinuities.*

2.3 A Practical Assessment (WP 3.0)

Philippe Guiblin, the Office of National Statistics (UK)
Marja Tammilehto- Luode, Statistics Finland

Abstract

The work-package 3.0 provides the first practical responses to the main objectives the project:
to develop a combined Grid-based and a Region-based approach able to tackle the limitations inherent in the NUTS system;
to implement two parallel approaches: Two types of methods working on two types of system of data
This work-package is divided in two sub work-packages. Analyses within each work-package were carried out separately in parallel. The WP3.1 proposes to apply the automatic zoning procedure to both of grid system and system of areas in order to propose a system irregular areas. The WP3.2 proposes to apply methods focused on he production of a system of regular areas. The work-package WP3.0 summarises the work carried out within each work-package.
In order to illustrate and compare results from both approaches a case study focused on the delimitation of urban and rural areas for Cardiff (UK) and Helsinki (Finland) has been carried out.
In terms of results, the practical part of the project has shown on one side that:
It is feasible to produce a better system of small areas;
It is possible to implement the two methodologies working in parallel on the two systems (regular and irregular);
A convergence of results is demonstrated on the case study considered.
And on another side that:
There is a need for more tests in order to compare and combine more efficiently the capacities of each of both approaches. .

2.3.1.1 Introduction

This work-package provides the first practical responses to the main objectives the project:

- To develop a combined Grid-based and a Region-based approach able to tackle the limitations inherent in the NUTS system;
- To develop the methodological framework for a consistent European Geo-statistics

It has to be reminded that spatially referenced statistics refer

- on one hand to masses of individual spatially referenced observations,
- and on the other to collections (aggregations) of data or data sets corresponding to different systems of tesselations[117].

There are generally two types of tesselations (regular and irregular) that are used for binning or aggregating statistics.

Irregular tesselations are generally used to compare regions with equal or comparable populations, whereas **regular tesselations** are used to compare regions with equal or comparable areas.

London August 2001

Philippe Guiblin
Office of National Statistics (United Kingdom)

[117] (Merrian 1971) *Webster's Third New International Dictionary*

2.3.1.2 Problem description

2.3.1.2.1 Professional context

This work-package presents a practical pilot study. The key objectives of the project in the short perspective are:

Improving the System of Irregular Tesselations

There are differing views as to whether there is a need to improve of the NUTS hierarchy as such. This project will focus on the possibility of defining a small geographical "Statistical block" (or system of building blocks) to be used in combination with the NUTS hierarchy. Programs developed in UK will be applied on small data region- or gridded data sets.

The statistical system used within the European Union is the NUTS system (Nomenclature des Unités Territoriales Statistiques). It provides a single uniform breakdown of territorial units for the production of regional statistics for the European Union.

For many statistical purposes, the NUTS framework is far from ideal. Therefore is a need for spatial transformation of statistical data. Conversion of the data from one unit to another (for mapping of the data, to aggregate or dis-aggregate data to different spatial scales or **different geographies**, for the purposes of statistical analysis, to provide estimates at unsampled or unmeasured locations…).

Theoretical advantages of irregular systems are presented in the work-package 2.1. The work-package 3.1 proposes a case study illustrating the interest of using methods focused on the optimisation of region-based boundary systems.

2.3.1.2.2 The advantage for European statistics

With the emergence of the EU the incomparability of these statistical system poses a great challenge to both Eurostat and countries involved.

- The first effort to improve the comparability of geo-statistics across EU was the implementation of the NUTS system;
- Over time the NUTS hierarchy has proven to be valuable but not ideal as a system for comparing statistics. There are two main objections
- It is not a homogenous system of regions with comparable populations.
- There exists a strong need for a system of territorial units with a relatively small population (a statistical block) that may be used in combination with the Nuts hierarchy.

In this project we will not focus on the problems connected with the NUTS system as in point 1 but try to improve it by looking for a new small building block that may be better suited for the comparison of regions.

Defining a System of Regular Tesselations

There is a need for a hierarchy of grids that may be used to compare the distribution of observations across the whole area of the EU. Programs developed in Sweden and Finland will be applied on small data region- or gridded data sets.

The grid-based system of statistics may integrate a wide range of different kind of spatial data. However a production of harmonised grid-based statistics for a whole of Europe needs harmonised methods to aggregate and dis-aggregate different types of input data. There is lots of methods available even by using standard GIS tools, but each tool may give different results (Briggs 2000). Most of the tools are designed for a certain type of data and

they are applicable only under certain conditions. It is important for example to differentiate between intensive and extensive variables and numerical and categorical variables, which may need different kind of treatment when conversions from one data set to another are needed.
Both approaches face problems linked to confidentiality issues. These issues are discussed theoretically (WP2.2) and illustrated in practice (WP3.2).

The advantage for European statistics

There exists no established system of grids to map and display data on micro-level or to bin them on to regular tesselations
There exists no established standard for the use of projections datum or other technical solutions that are needed for the use of equal area spatial units.
In this project we will focus on some of the fundamental questions in connection with the establishment of an equal-area system suitable for the aggregation of statistics

2.3.1.2.3 Objectives

The Objectives of the practical assessment of this project is to use the methods described in the theoretical assessment in order to produce:

- a prototype of region-based system with the test data of two different countries.
- a prototype of grid system with the test data of two different countries, two different geo-reference systems and two different categories of variables.

2.3.1.3 Methods

2.3.1.3.1 The grid-based method

The Nordic database[118].

Finland, Norway and Sweden belong to a group of countries in which statistics production is essentially based on comprehensive registers. The approach of the geo-coding of statistical units is point-based instead of area-based in these three countries.
Sources of geo-coding:

- centroids of buildings (Finland);
- centroids of real estate properties;
- digital road and street networks with addresses are available for linking statistics by their address location to the geography.

The geo-coded-points and their links to statistical units make the production of small area statistics flexible in the Nordic countries. Statistics by relatively small areas, such as grid squares of 1 km x 1 km or even smaller, have been produced.

General Research interests

Such a flexible data resolution provide adequate material for[119]:

- Analysing the similarities and differences in the spatial structure of the population between different countries;

[118] Register-based systems and geo-coded points as basic statistical units (see (Tammilehto-Luode, Backer and Rogstad 2000) *Grid data and area delimitation by definition towards a better European territorial statistical system*
[119] (Tammilehto-Luode, Backer and Rogstad 2000) *Grid data and area delimitation by definition towards a better European territorial statistical system*)

- Analysing the functional distribution of the population in large areas containing both urban and rural districts;
- Analyses focused on the "night-time" population, daytime population and workplaces;
- A better practice of the use of GIS software.

Combination of grids and regions are widely used in Sweden and Finland. The two systems are regarded as complementary. For the purpose of creating a harmonised geo-statistical system for the Nordic countries the NUTS regions can be used in most cases.
Practically, 1,000-m grids can be used for large areas and 100m grids for small areas to effectively describe the spatial distribution of phenomena.

One application to delimitation of urban areas[120]

Finland and Sweden have developed automatic methods for the delimitation of urban area. The theoretical aspects of this method are presented in the WP2.2.
Boundaries are originally designed on analogous maps by the local authorities and digitised later on by mapping agencies.
An automated delimitation process is more objective and less costly.
Three factors support the methods:

- Map co-ordinates are assigned to buildings, real estates or addresses in official databases where other statistical information can be linked to them;
- Digital, vectored basic maps to proper scale are available;
- Availability of GIS software.

The centres of buildings shall be registered as an urban settlement if it is inhabited by at least 200 persons (60-70 dwellings). The distance between the buildings shall normally not exceed 50 metres.
Automatic delimitation of urban areas is done in two separate steps.

- The number of residents is geographically distributed to the co-ordinates of address buildings;
- A GIS is used to aggregate polygons of urban settlements according to an agreed set of criteria.

Unit-specific data that are definable by building co-ordinates have been summed up to grid-squares. Grid square data are regarded as the most flexible statistical geographical data. Data by small squares can easily be summed up to for larger areas.

2.3.1.3.2 The region-based method: production of blobs

The UK system for area statistics

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. There were 155 180 of these small areas in the UK with an average 1991 Census population of about 366. However, these units do not provide a spatial system that satisfies all users of Census data, and thus there is demand for alternative geographies. The British contribution to the project is to propose different alternatives for building output areas for European statistics. Several methods have been

120 (Tammilehto-Luode, Backer and Rogstad 2000) *Grid data and area delimitation by definition towards a better European territorial statistical system*

developed in order to produce better statistics based on the production of statistics on irregular tessellation (region-based method).
Methods and algorithms are presented in the WP2.1 and summarised below. They propose different ways of re-aggregating data from a low level of statistical areas (Input Areas) to a higher level of areas. The result is a changing of output areas (**the building blocks**) using **optimisation algorithms** based on consistent criteria. Resulting building blocks form the basis for further aggregation in order to display statistics or geographical patterns.

The key points in automating design

The practical assessment of the region-based approach consists of the application of 'automated zoning procedures' (AZP) due to Openshaw [121] who originally developed the basic algorithm. This AZP algorithm is used to improve the consistency of displaying data at different geographical levels.
The basic algorithm was developed by Openshaw [122] in order to explore modifiable area unit effects. This AZP algorithm is used in order to improve the consistency of displaying data at geographical level.
The AZP (automated Design Procedure) algorithm: a local boundary optimiser
The aim is to optimise (under constraints) a function of the data generated by a zoning system defining an aggregation of N original zones into M regions or output zones (M<N). Details are presented in WP2.
Martin,[123] who developed the algorithms in Fortran and Visual Basic, suggests the use of three constraints:

- Equality population zoning: a basic and frequent function used in many census applications by geographers. Regions are devised that are equal or near equal value in terms of a selected variable (e.g. population size or numbers of economically active people).
- Shape design function: minimisation of perimeter squared divided by area, which maximises the compactness of the output tracts.
- Homogenity design function: measures the similarity of values of variables within any area of interest. This constraint allows for incorporating social constraint within the algorithm. 2.2.3 Application example: Reengineering census EDs (for Merseyside)

Although the use of those three constraints is now available within the software used in the framework of the Tandem GIS project, only population and shape constraints have effectively used.

2.3.1.4 Tests with empirical data

2.3.1.4.1 Work plan for analyses

The purposes of the case studies are:

- to propose a methodology to create homogeneous building blocks;
- to provide an example of how to use it for the delimitate urban/rural areas
- densely populated areas;
- intermediate areas;

[121] (Openshaw 1977) *A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning, and Spatial Modelling*
[122] (Openshaw 1977) *A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning, and Spatial Modelling*
[123](Martin 1997) *Implementing an automated census output geography design procedure*

- thinly populated areas
  in using population and/or shape constraints within the Automated zoning approach and the grid-based approach.
- to mention other possible purposes (definition of census, small areas for sampling…);
- to present a methodological work;

It is not the object of the project to advise on policy involvement.

2.3.1.4.2 Test data

The actual system of European statistics

Within the Member States, statistical systems are classified as being either register-based, area-based or some mix of the two.

Register-based systems use registers created for administration purposes. In the Nordic countries, for example, they contain accurate and up to date information as well as geo-references at the individual address level. These geo-references provide ultimate flexibility of referencing, subject to data confidentiality.

In area-based systems, the geo-reference relates to some small census area, such as collection area, output area or census tract. Other Member States may use data collected for various purposes, and therefore have a variety of geo-references such as commune or low level administrative areas or grid-based systems.

Although originally the statistical data might have been gathered from either register-based or area-based inputs, the final structure of the small area statistics will be area based.

The Test data for the Tandem GIS project

Analyses will be focused on the urban/rural dilemma. Two metropolitan areas will be considered: Cardiff (UK) and Helsinki (Finland).

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 urban areas:

The Metropolitan Helsinki Region (Finland)

The Helsinki data sets cover the city of Helsinki and contain the 1998 census population counts. Both grid-based data and postal code data (polygon-based data) were calculated by using data on geo-referenced buildings. Population density was calculated for postal code areas by using information of postal code polygons. The data were delivered on Arc/Info coverage –formats. The original amount of 1 km x 1km grids is 13003 and amount of postal code areas is 392.

Cardiff region (United Kingdom)

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

The data were delivered in MapInfo-formatted files. The original data includes 817 polygons covering the county.

(**Remark:** a typical British ED would contain about 200 households. ED resolution is smaller than NUTS 5 level).

Optimal projection system

A system of grid –based statistics – system of regular tessellation is by definition constructed by using equal area. Construction, visualisation and

analysis of grid-based data require a rectangular co-ordinate system, which does not distort regular polygons, or distort them as little as possible. For a European wide system it has to be universal and mathematical transformation from national systems need to be available.
A hypothetical projection system for the whole of the Europe is Universal Transverse Mercator (UTM) –system, which is widely used in grid-based systems of nature and land use information (EEA 2001). UTM co-ordinates define two dimensional, horizontal positions. UTM system includes sixty zones which keeps the distortions reasonable small also in medium and large scale maps.
For the test data UTM –projection system was chosen including different zoning for UK and Finland; It was agreed to use UTM 30 for UK and UTM35 for Finland.

2.3.1.5 Requirements for the analyses

2.3.1.5.1 Region-based approach : Production of an optimised boundary system (work-package WP3.1)

Requirements for analyses

- Use of the AZP algorithm (Openshaw, Rao, 1994)[124];
- Use of Fortran and Visual Basic routines written by Martin, 2000[125]
  - Inputs: vector based data + contiguity files created by Arc/Info for each data set
- Use of a Fortran Compiler / Visual Basic
- Use of MapInfo and/or ArcView to display the results
- The initial boundary data are in **Arc/Info format.**

Remark: 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 homogenity function will be used in the analyses.

Output

- Production of a set of optimised boundary system depending on the target variable;
- Production of output statistics and patterns at a higher resolution level.

Input for the method

- Choice of a target variable (ex: population counts);
- Definition of objective functions and choice of constraints;

Remarks:

- For practical reasons it may be required delimiting separated sub-areas in which the rezoning will operate separately in order to respect geographical or administrative constraints. This may improve the speed of the calculations but requires the preparation of a constraining polygon file.
- In the framework of this project, neither geographical nor administrative constraints have been taken into account. The statistical building blocks are independent from the NUTS system (no need to fit within predefined boundaries such as NUTS 4, for example).

Grid based approach (see work-package 3.2)

Requirement for analyses

- use of vector and raster data;

[124] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*
[125] (Martin 2000) *Towards the Geographies of the 2001 UK Census of Population*

- use of point-based and polygon-based data;
- use of categorical, intensive and extensive variables;
- use of micro data and aggregated data;
- use of different map projections;
- use of different size of grids – different scale;
- use of standard and applied algorithms to produce grid net and gridded data.

Output

- Production of grids from different set of input data;
- Comparison of different methods of production of grids by real and estimated grids;
- Comparison of displaying test data by NUTS 5 and grids;
- Further use of grids: Delineation of urban areas; comparison of results by using NUTS 5-data and gridded data.

Input of the method

- Choice of the projection system
- Choice of the size of the grid cell
- Choice of methods to produce grid-net (by ArcInfo or ArcView)
- Choice of methods to convert of input data (by ArcInfo or ArcView)
- Choice of methods to display and analyse the data.

Purpose of analyses

The Purpose of empirical tests is to compare the algorithms to produce grids from different kind of input data and to construct the prototype of a European gridded data set for further use.
The delineation of urban areas (with the Eurostat definition) by the data set will be performed and compared with the delineation performed by NUTS 5 data set.

2.3.1.6 Analyses, results and discussion

The objective of this section is to present a case study analysing the functional distribution of the population in large areas containing both urban and rural districts within the region of Helsinki and the County of South Glamorgan.

Running the AZP algorithm: production of blobs

Focus: 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.
The results are presented in the work-package 3.1.
The algorithm used on British ED population data was used to produce a set of optimised boundary set of 527 building blocks (to compare with 817 initial input areas). 463 building blocks can be aggregated in order to define an urban area.
The algorithm used on Helsinki post-code population data was used to produce a set of optimised boundary set of 266 building blocks (to compare with 436 input areas). 119 building blocks can be aggregated in order to define an urban area.

The AZP algorithm used on Helsinki gridded population data need to be amended in order to deal with the enormous quantity of data. Hopefully the computer limitations will be sorted before the release of the second version of this report.
Only equal population zoning and one single shape constraint (perimeter squared/area) were used.
The interesting results are:

- the three delimitation zones look fairly more distinguishable on the 'optimised' map;
- the transition from densely populated zones to thinly populated zones is better designed;
- within each of the three groups of zones the areas are of nearly equal size.

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 practical investigation.

The Grid-based approach
Focus: Production of grids from different set of input data
Comparison of different methods of production of grids by real and estimated grids. Comparison of displaying test data by NUTS 5 and grids
Further use of grids: Delineation of urban areas; comparison of results by using NUTS 5-data and gridded data.
The Purpose of empirical tests is to compare the algorithms to produce grids from different kind of input data and to construct the prototype of a European gridded data set for further use.
The delineation of urban areas (with the Eurostat definition) by the data set will be performed and compared with the delineation performed by NUTS 5 data set.
A summary of the results is presented in this chapter, more detailed exercises are presented in the annexed documents.
The results presented in WP3.2
Different algorithms to convert the polygon based input data to grid data were tested against the hypothetical real grids (the Finnish gridded data made by using data on reference points of buildings). The Finnish postal code data were converted to grids and results were compared together and against the real gridded data.
Visual comparisons of population density by NUTS 5 and by 1 0 km x 10 km and 1 km x 1 km grid.squares are visualised are then performed.
Visualisation of gridded data is made as typical choropleth maps without any smoothing methods.

The main results are:

- Grid-based applications should be more realistic, because they do not support ecological fallacy, which may be a problem with large NUTS 5 areas (Martin 2000).
- Urban area defined of Finnish test data by 1 km x 1km grids is 45% smaller than by similar definition of NUTS 5 areas. In British test data the differences of areas are even bigger; by 1 km x 1km building blocks area is considerably smaller than by NUTS 5 areas;

- There are big differences in population figures as well. In Finland population in urban areas defined by 1 km x 1km building blocks is 10 % smaller than ones defined by NUTS 5 building blocks;

2.3.1.7 Conclusion of the analyses

Conclusions and recommendations are expressed in each of the WP3.1 and WP3.2 work-packages. 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.

1. 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**.**

2. A system of grid-based statistics has a lot of advantages, which has been discussed in Work packages 2.2. 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 (see WP2.1 and WP3.1).
- 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 (see WP3.1, WP3.2). 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.

2.3.2 Test Runs 1 (WP 3.1)

Philippe Guiblin, The Office of National Statistics (UK)

2.3.2.1 Introduction

This work package is a part of the study of " A feasibility study towards a common geographical base for statistics across the European Union" carried out by the Tandem Consortium. In the Tandem Consortium there are representatives from Statistics Sweden, the Office of National Statistics in United Kingdom and Statistics Finland. The Office for National statistics Finland carried out this work package.
The aim of this work package is to describe and test the feasibility of solutions for constructing regional-based statistical system. The purpose is to construct a prototype of a region system with the test data of two different countries. The prototype will be evaluated by using technical and statistical and cartographic criteria. A case study for delineating urban areas of test areas is performed.
The test data represent two types of potential input data. The data set from United Kingdom is by Enumeration Districts, which are in practice the smallest dissemination areas of British census data. It represents typical polygon-based input data. The data from Finland are compiled from geo-referenced centroids of buildings and thus represent point-based input data.
The tests were made in cooperation with methodologists and GIS experts from the Office for National Statistics. Special thanks to Kerry Ellis, Bruce Mitchell and Patrick Heady. Also a special thanks to Professor David Martin, University of Southampton, who provided software and much useful advice during the completion of the project.

2.3.2.2 Problem description

2.3.2.2.1 Professional context

The statistical system used within the European Union is the NUTS system (Nomenclature des unités territoriales statistiques. It provides a single uniform breakdown of territorial units for the production of regional statistics for the European Union.
For many statistical purposes, the NUTS framework is far from ideal. Therefore is a need for spatial transformation of statistical data. Conversion of the data from one unit to another (for mapping of the data, to aggregate or dis-aggregate data to different spatial scales or **different geographies**, for the purposes of statistical analysis, to provide estimates at unsampled or unmeasured locations…).
For many purposes, statistics need to be generated for spatial units that do not conform exactly to the NUTS regions.

- The differences in the size of the NUTS regions between countries may create misleading maps.
- Differences in characteristics of the areas (e.g. population or age structure), which might be used as denominators, generate variations in

rate or ratio data and may produce unstable patterns in the geographical distributions.[126]

All these factors create the need for spatial transformation of statistical data: that is the conversion of the data from a spatial unit to another. This is necessary, for example:

- For more effective visualisation (mapping) of the data;
- To combine or compare data based on different spatial units;
- For the purpose of statistical analysis of the data (e.g. to test for spatial pattern or trends);

Several methods have been developed within the last ten years. Many research studies have been developed within academic frameworks (e.g. for census purposes[127]. The objectives of this work-package are to propose an illustration of what it is feasible to do in the context of European statistics.

2.3.2.2.2 Objectives

The objectives of this study are to construct a prototype of region-based system with the test data of two different countries.

A construction of region-based statistics from point-based and from polygon-based source data is described.

Tests are made by using the 'Automated Zoning procedures' developed by Openshaw[128] and Martin [129]. The method is tested on two different categories of data:

- Polygon based data
- Finnish postal code data;
- British Enumeration District data.
- Grid-based data
- The Finnish gridded data.

This methodology requires the same algorithms but starts with data at very high spatial resolution and is able to provide very small synthetic areas for setting up a building block system that can be applied to any country able to provide data at a very small geographical level.

There will also be a case study to delineate urban areas based building blocks.

2.3.2.3 Methods

2.3.2.3.1 The key points in automating design

The theoretical assessment of the region-based approach consists of the application of 'automated zoning procedures' (AZP) due to Openshaw [130] who originally developed the basic algorithm. The description of the algorithm is given in the Work-package 2.1.

---

[126] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

[127] (Martin 1998) *Optimising Census Geography: the Separation of Collection and Output Geographies*

[128] (Openshaw 1977) *A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning, and Spatial Modelling*

[129] (Martin 1998) *2001 Census Output Areas: from Concept to Prototype,* (Martin 2000) *Towards the Geographies of the 2001 UK Census of Population*

[130] (Openshaw 1977) *A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning, and Spatial Modelling*

The basic algorithm was developed by Openshaw[131] to explore modifiable areal unit effects. This AZP algorithm is used in order to improve the consistency of displaying data at geographical level.

The AZP (automated Design Procedure) algorithm: a local boundary optimiser

The aim is to optimise (under constraints) a function of the data generated by a zoning system defining an aggregation of N original zones into M regions or output zones (M<N). Details are presented in WP2.1.

Martin [132] who developed the algorithms in Fortran and Visual Basic suggests the use of three constraints:

- Equality population zoning: a basic and frequent function used in many census applications by geographers. Regions are devised that are equal or near equal value in terms of a selected variable (e.g. population size or numbers of economically active people).
- Shape design function: minimisation of perimeter squared divided by area, which maximises the compactness of the output tracts.
- Homogenity design function: measures the similarity of values of variables within any area of interest. This constraint allows for incorporating social constraint within the algorithm

Although the use of those three constraints is now available within the software used in the framework of the Tandem GIS project, only population and shape constraints have been used within the case studies.

2.3.2.3.2 Testing the prototypes

In this study two types of tests were chosen for usability tests of prototypes of region-based statistics.

Choice of constraints and zone design function (optimality constraints)

According to WP2.1, Martin [133]suggests the use of three constraints: population size, shape, and homogenity constraints.

In effect in all these zone design applications the objective function is the weighted sum of up to three different design functions**.**

According to the remarks expressed by Openshaw and Rao[134] and presented in the WP2, a few technical points must be considered carefully before defining the objective function. Two main points were retained in the completion of this case study.

1 'To ensure that all the output zones exceed a minimum comparable degree of compactness and homogenity':
   - Here, the median and the mean of the studied variable were used to fix the threshold and the target population values;
   - Only the equal population zoning- and shape constraints were used. Therefore no social constraints have been added to the objective function.

2 'The shape constraint. It is difficult to know what limits to use and it may not matter much, or at all, if occasionally a strangely shaped zone is produced'.

[131] (Openshaw 1977) *A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning, and Spatial Modelling*
[132] (Martin 1997) *Implementing an automated census output geography design procedure*
[133] (Martin 1997) *Implementing an automated census output geography design procedure*
[134] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

- a single weighted shape constraint was used. The quality of the result was assessed by the consistency of the outputs produced by the programs (values of the objective functions, visual checking of the results, maps).

Further investigation need to be done in order to check the accuracy of the checking criteria. Nevertheless the objective of the project was to use the AZP method and to compare its outputs with a method based on the production and analyses of grid.

Visualisation of region-based statistics

Harmonised region-based statistics should make statistics by different territorial divisions comparable especially from the visual point of view. In this study prototypes of region data of both test areas are visualised and compared with Input statistics of the same areas. Visualisation is made by using MapInfo 6.0

Delineation of urban areas

A problem of delineating urban areas was chosen for a user case. With reference to Eurostat studies a hypothesis of this study is that the new building blocks, 'blobs' created by more detailed source data are better for this task than NUTS 5 areas. The definition for urban area in this 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[135].

The method for delineation by region -based statistics is an application from another method of Statistics Finland for delineating localities (see Work-Package WP3.2). It is completed by using MapInfo 6.

The basic idea is to create zones from selected building blocks with population density equal to or above 500/km^2 and then join the population-data (summarised by zones) to these zones. Then finally those zones whose population are equal to or exceed 50000 are selected.

2.3.2.4 Optimal projection system

A system of grid –based statistics – system of regular tessellation is by definition constructed by using equal area. Construction, visualisation and analysis of grid-based data require a rectangular coordinate system, which does not distort regular polygons, or distort them as little as possible. For a European wide system it has to be universal and mathematical transformation from national systems need to be available.

A hypothetical projection system for the whole of the Europe is the Universal Transverse Mercator (UTM) –system, which is widely used in grid-based systems of nature and land use information (EEA 2001). UTM co-ordinates define two dimensional, horizontal positions. UTM system includes sixty zones which keeps the distortions reasonable small also in medium and large scale maps.

For the test data UTM –projection system was chosen including different zoning for UK and Finland; It was agreed to use UTM 30 for UK and UTM35 for Finland.

[135] (Eurostat 1998) *Urban database.*

2.3.2.5 Test with empirical data

2.3.2.5.1 Test data

Different levels of resolution are considered in this work-package according to different European geographies and purposes.

2.3.2.5.2 The Helsinki metropolitan Region (Finland)

The Helsinki data sets cover the Metropolitan Helsinki area and contain the 1998 census population counts of three age groups. Two level of resolution are considered (one for each data set):

- Population at post-code level;
- Population on 1km x 1km grids.

Both grid-based data and postal code data (polygon-based data) were calculated by using data on geo-referenced buildings. Population density was calculated for postal code areas by using information of postal code polygons. The data were delivered on ArcInfo coverage –formats. The original amount of 1 km x 1km grids is 13003 and amount of postal code areas is 392.

2.3.2.5.3 Cardiff region (United Kingdom)

The British data set covers the county of South Glamorgan (NUTS3 containing the City of Cardiff and the Vale of South Glamorgan) and contain the 91 census population counts at Enumeration District (ED) level.
The data were delivered in MapInfo-formatted files. The original data includes 817 polygons covering the county.

Remark:
a typical British ED would contain about 200 households. ED resolution is smaller than NUTS 5 level.
Three exercises are considered and developed below:

Construction of a set of building blocks from British Enumeration Districts (Eds)

- As such it was traditionally the smallest geographical entity for which census data could be provided in UK.
- Case study: South Glamorgan ED census population data.

Construction of a set of building blocks from Finnish Post-codes

- A very low level of data resolution. A very flexible geography to provide aggregated data at a higher level;
- Case study: Helsinki population data.
- in order to produce output statistics data have to be aggregated at ‘ward’ level (currently the smallest "official" geography in Europe)

Construction of a set of building blocks from 1km gridded Finnish data

2.3.2.5.4 Results

Testing the algorithm

The algorithm used is the basic Automatic Zoning Procedure described by Openshaw and Rao (1995). The versions developed by Pr. David Martin in Fortran and Visual Basic programming language were used to produce the outputs. An important part of the work accomplished during the project was the implementation of these methods within the ONS.
The final outputs were performed using the GIS software MapInfo.

Analyses, results and discussion

Focus: 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 on 1km x 1km grid and at postcode level for the area of Helsinki. *AZP on ED data: South Glamorgan area*

2.3.2.5.5 Creation of a new boundary set of building blocks

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).[136] Initially the 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 aspect 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 note 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, 1995).

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[137]. The results, basics statistics and comparison with the input data set are presented in the table 1.

Input Areas						
Number of EDs	Minimum Population / non zero	Maximum Population	Median/ Average population	Minimum Area m^2)	Maximum Area (m^2)	Ppp Average Are Area (m^2)
818	0 / 64	1030	492 / 477.384	10,285.31	13,575,61	510,904.34
Set of optimised Building Blocks:						
Population Threshold: 500; target population: 500; shape constraint: on; homogenity 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

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 those examples, we have entered minimum values (threshold) close to the median to ensure a substantial ratio of aggregated areas and a target population close to the mean The table below gives the same statistics at the NUTS 5 (Ward) level.

NUTS 5						
Number of wards	Minimum Population	Maximum Population	Median/ Average population	Minimum Area (m2)	Maximum Area (m2)	Ppp Average Are Area (m2)
47	1,892	16,965	8,030/ 8,356	792,191.18	44,153,93.92	8,867,277

Table 2: Population and geographical statistics at NUTS 5 level.

[136] The set of ED boundaries is presented on the Figure 13 and covers the 2 districts of the city of Cardiff and the Vale of South Glamorgan.

[137] The result is shown on Figure 14

A set of 527 optimised building blocks such as this presented above allows for a more flexible system.

2.3.2.5.6 Delimitation of rural/urban areas

The use of the AZP algorithm seemed 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[138]. The input areas (Eds for Cardiff and theVale of South Glamorgan) have been divided into the following 3 classes of populated areas[139]:

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

Only equal population zoning and one single shape constraint (perimeter squared/area) were used.

The interesting points are:

- the three delimitation zones look fairly more distinguishable on the 'optimised' map[140], as if the data were 'smoothed';
- the transition from densely populated zones to thinly populated zones seems better designed;
- 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, 1995)[141].

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

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. *AZP on the Helsinki postcode population data*

2.3.2.5.7 Creation of a new boundary set

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). The Helsinki area displays certain features, which contrast with the South Glamorgan area:

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

[138] Maps of population densities are presented on the Figure 13 and 14. They were produced using the GIS software MapInfo.

[139] Figure13 where the 3 colours are representing the 3 population classes.

[140] A final map presenting the delineation between urban zones (area containing population of more than 50,000 people) is presented on Figure 14

[141] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

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 3 below)

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

Table 3: Distribution of "empty " zones

Input Areas						
Number of postcodes	Minimum Population / non zero	Maximum Population	Median / Average population	Minimum Area (m2)	Maximum Area (m2)	Ppp Average Area Area (m2)
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; homogenity 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 4: Comparison of basic statistics for population and area sizes for the Input and AZP-output set of areas

The case presented here takes the total population as a target variable, results are presented below. Results are summarised on the table 3.[142]

2.3.2.5.8 Delimitation of urban/rural areas

Here again only equal population zoning and a shape constraints (perimeter squared/area) were considered, the threshold was chosen close to the median and the target population close to the mean of input population data.
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.[143]

AZP on gridded data: Helsinki Region, 1km x 1km grids
Creation of a new boundary set
Applying the AZP algorithm on gridded data involved adaptations in the Visual Basic program and to reconsider the data set on which the program

[142] The boundary sets are shown on Figure 15 and 16.
[143] A final map presenting the delineation between urban zones (area containing population of more than 50,000 people) is presented on Figure 17.

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 was 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.
Since the grid size is 1 km x 1 km the density of each grid cell is equal to the population. A map of the density of the population of the study area is presented [144]where three colours have been used to represent three class of density of population and one colour used particularly to represent the empty cells (195). 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[145]:

Input Areas						
Number of grids	Minimum Population / non zero	Maximum Population	Median / Average population	Minimum Area (m^2)	Maximum Area (m^2)	Ppp Average Area 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; homogenity 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 5: Comparison of basic statistics for population and area sizes for the Input and AZP-output set of areas

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. 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.[146]

2.3.2.5.9 Delimitation of urban/rural areas

The results of the delineation between urban zones (area containing population of more than 50,000 people)[147] and a more general overview[148] (the 'chunk' within the Helsinki region). The results look fairly similar to those obtained elsewhere. Nevertheless the use of AZP lead to the creation of zones with intermediate populated density in the middle of densely populated zones.[149] 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.

[144] Figure 18
[145] Figure 19 shows the result of the aggregation process
[146] Figure 20.
[147] Figure 21
[148] Figure 12
[149] Figure 22

2.3.2.6 Conclusions

2.3.2.6.1 Discussion

This case study gives an illustration of the advantages of the Zoning design approach.
The technical feasibility of such an approach applied at a large scale like the European Union is demonstrated. Zoning design methods offer a flexible way of aggregating data in order define the most meaningful set of areal units which best suits a particular analysis objective. This allows for the dissemination of statistics for areal units small enough and homogenous enough to be really useful spatial building blocks (see Openshaw, Rao, 1995[150]). This also allows for purposes like identifying geographical patterns such as 'urban/non urban areas'.
It is obvious that the choice of the level of initial input areas (Eds or post-code) is crucial. The data (e.g. population counts) are aggregated (not estimated!) to a higher level. The question of the relevance for each European country has to be considered carefully.
It was a part of the project to organise the technical implementation of the system. The software presented in the Work-package 2.1 is powerful but GIS dependent. For technical reasons the option chosen in the practical assessment of the project was to work on GIS-free software. The programs (Fortran and Visual Basic) offered by Dr. David Martin provide such flexibility. The GIS (MapInfo here) used aftermath provide the visualisation tools. Nevertheless it has to be noted that Arc/Info is still required for the production of contiguity files as input to the analyses.
Remark:
The Java application of ZDES, known as ZD2K should also be available soon and will offer software-independent tools for this kind of work, but wasn't available to you within the timescale of the Tandem GIS project.

2.3.2.6.2 Recommendations

The zone design approach is a way to produce 'harmonised data' in the sense of 'producing more homogenous data'. Or preferably 'aggregating data in a more meaningful boundary set'.
The relevance of the results depends on three points:

1. The provision of a 'good' initial set of areal units.;
2. The confidentiality restrictions applied to the dissemination of statistics;
3. The performance of the AZP algorithms.

The points 1 and 2 are country dependent.
As for the point 3, Openshaw and Rao mentioned (1994) that the user has to be aware of "possible risks of zoning anarchy as a result of more flexible zonations": 'Zone design can be used to destroy or discredit the results of spatial analysis as well as to provide a powerful new tool able to enrich geographical study' (Openshaw and Rao, 1994)[151]. "The nature of the units being studied is of a great concern. The user has also to be aware of the uncertainty due to the MAUP (Modifiable Areal Unit Problem) 'under his or her control'; the multiple possibility of final results to the different objective functions the AZP program seeks to optimise".

[150] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*
[151] (Openshaw and Rao 1995) *Algorithms for re-engineering 1991 Census geography*

2.3.3 Test runs 2 (WP 3.2)

Marja Tammilehto-Luode, Statistics Finland

2.3.3.1 Introduction

This work package is part of "A feasibility study towards a common geographical base for statistics across the European Union", carried out by the Tandem Consortium. The Tandem Consortium has representatives from Statistics Sweden, the Office of National Statistics in the United Kingdom and Statistics Finland. Statistics Finland carries out this work package.
The aim of this work package is to describe and test the feasibility of solutions for the constructing of a grid-based statistical system. The intention is to construct a prototype of a grid system with the test data of two different countries, two different geo-reference systems and two different categories of variables. The prototype will be evaluated using technical, statistical and cartographic criteria. A case study of delineating urban areas in test areas will be performed.
The test data represent two types of potential input data. The data set from the United Kingdom is by Enumeration Districts, which are in practice the smallest dissemination areas of British census data. It represents typical polygon-based input data. The data from Finland are compiled from geo-referenced centroids of buildings and thus represent point-based input data. The Finnish test data are compiled to grids and, for comparison, to postal code areas.
The software that are used for the constructing and testing of grid-based systems are ArcInfo [152] or ArcView [153]. These are the most widely used software for Geographic Information Systems among National Statistical Institutes [154].
The tests were conducted in co-operation with the GIS team of Statistics Finland. Special thanks to Liisa Kanerva and Jaakko Suikkanen.

2.3.3.2 Problem description

2.3.3.2.1 Professional context

Grid-based systems of statistics may integrate a wide range of different kinds of spatial data. However, production of harmonised, grid-based statistics for the whole of Europe needs harmonised methods for aggregating and dis-aggregating different types of input data. There is lots of methods available even by using standard GIS tools, but each tool may give different results[155]. Most of the tools are designed for a certain type of data and they are applicable only under certain conditions. It is important, for example, to differentiate between intensive and extensive variables and numerical and categorical variables, which may need different kinds of treatment when conversions from one data set to another are needed.
Traditionally, data for point or polygon-based units are collected with different systems and accuracy of geo-references. Aggregating point-based information to grid cells is a more simple operation than dis-aggregating

[152] (ESRI 2000) *ArcInfo 8.2: Cell-based Modelling with GRID, ARC/INFO User's Guide*
[153] (ESRI 2000) *ArcView*
[154] (UN-ECE 2000) *Questionnaire on the Implementation of GIS in Statistics*
[155] (Briggs 2000) *Spatial Transformation Methods for the Analysis of Geographic Data*

polygon-based data to grid cells [156]. However, both approaches have to be available if data from different sources and from different countries are to be harmonised to a common system of grid-based statistics.
There is also a need to optimise the size of the grid cells because of quality reasons and confidentiality reason of the data set. The proper size of the grid cell is highly dependent on the source data and their quality and also on the purpose and scale of the analyses that are to be done with grid-based statistics. Confidentiality reasons may also affect the minimum size of the grids available.
Finally, a grid-based system of statistics is a spatial system with geo-references. Both location and attribute information on a particular object need to be read and visualised on a map. The distance units need to be usable for analyses of the data. However, each country and different data sources use different geo-reference systems. The spatial reference system has to be common for a European-wide grid-based system. The map projection of a grid-based system has to preserve as much as possible the shape of the grids. It also has to be so common that mathematical conversions are available from different national geo-reference systems.

2.3.3.2.2 The objective

The objective of this study is to construct a prototype for a grid system with the test data of two different countries, two different geo-reference systems and two different categories of variables.
The construction of grid-based statistics from point-based and polygon-based source data is described.
Tests are made with different methods for dis-aggregating polygon-based data to grids by using Finnish gridded data and Finnish postal code data. The Finnish gridded data (the construction of which will be also described) are considered as representing real gridded data because of their accurate, point-based geo-references. The Finnish postal code data will then be converted to grids by different methods to represent “estimated grids” and the results are compared with the real grids. The best methods for converting point-based and polygon-based source data to grids are chosen according to Nordic experiences with grids, on the one hand, and according to tests made with the Finnish test data, on the other.
Prototypes for gridded systems are first made out of point-referenced input data (Finland) and then from polygon-based input data (UK).
The prototypes are then used for comparing their use with similar data sets by NUTS5 levels. The purpose is to visualise the population structures in the test areas by using the new prototypes of gridded statistics and by using traditional statistics by NUTS5 levels. There will also be a case study of delineating urban areas by using gridded and NUTS–based building blocks.

2.3.3.3 The methods

2.3.3.3.1 Points to grid cells

Aggregating point-based information to grid cells is a similar procedure to the aggregating of points to any kinds of polygons. However, as regular tessellations the grids include information that simplifies the process

[156] (Hansen 2001) *PSSD-Planning System for Sustainable development. The Methodological Report*

considerably. In theory, one first makes a polygon-on-point overlay between the point theme and the grid cell theme. The attribute data of the grid cell theme will be assigned to the point theme. As a result, each observation gets a unique grid cell number, which can then be used as a common denominator of aggregate information within each grid. In practice, map co-ordinates themselves can be used for aggregating geo-referenced points, as has been done in Finland and Sweden. The presumption then is that the co-ordinate system used for the geo-referencing is originally rectangular and thus provides an undistorted grid structure.

In Finland and Sweden, the geo-reference for each cell in a grid-net is agreed as being the bottom left-hand corner cell as defined by the co-ordinate pair. In our national co-ordinate system, two 7-digit numbers are needed to denote a 1m x 1m grid while using the bottom left-hand corner cell co-ordinates for defining the grid. Combining the first digit of the x co-ordinate with the first digit of the y co-ordinate gives us a (2-digit) code for a 1,000km x 1,000km grid. By adding the second x co-ordinate and the second y co-ordinate gives us a 100km x 100km grid inside the latter. Therefore, 10 digits are needed to build a code for a 1km x 1km grid. To divide this further, we could cut the co-ordinates to 500m, 250m, 100m, and so on. A similar system can be constructed for any rectangular co-ordinate system.

The practical way to calculate statistics by grids with geo-referenced micro data is to

1) Sort out the data by co-ordinates to a descending order
2) Round the map co-ordinates by the size of the grid cell
3) Aggregate the data by similar map co-ordinates

The default system of map co-ordinates is a geographic system and each zone must be processed separately.

A map co-ordinate is the primary identifier of a grid cell, usually acknowledged as the bottom left-hand corner of a grid, as mentioned above. An indexing system of grid cells can also be introduced. One way of indexing is to number the grid cells by the extent of the data when the minimums and maximums of the co-ordinates (geo-referenced points) are known.

The formula for the numbering is as follows:

$$\text{Index} = (x_{max} - x_{min})/a * ((y_i - y_{min})/a + (x_j - x_{min})/a)) + 1$$

where x and y are the map co-ordinates of the cells and a is the size of the grid cell.

A hierarchical indexing system of grid cells is sometimes valuable, for example, for making hierarchical queries more efficient (Tammilehto-Luode et al. 2000[157]).

The hierarchical system of grids in Finland is based on the minimum grid size of 250m x 250m. Larger grids are re-aggregated from these grids, meaning that the sides of the larger grids need to be divisible by 250m.

In the system described above gridded data are made by calculating frequencies within the grid cells with the information of the map co-ordinates of each unit point[158]. Producing numeric data by grid cell only requires an addition procedure. To produce categorical data, more complicated

[157] (Tammilehto-Luode, Backer and Rogstad 2000) *Grid data and area delimitation by definition towards a better European territorial statistical system*

[158] (Tammilehto-Luode and Backer 1999) *GIS and Grid Squares in the use of Register-based Socio-economic Data*

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. A rather simple approach is to create a cross-table containing the grid cell identity code as the primary key, with additional columns representing the unique values of the case items.

2.3.3.3.2 Polygons to grid cells

If the collection units of input data are polygons, such as enumeration areas or post code areas, which do not fall exactly into grid squares or the areas overlap several grid cells, there is a need for an estimation process to change an irregular polygon structure to a regular grid structure.

There are several methods for creating grid-based statistics from polygon-based source data[159]. Some of the methods can be implemented using standard GIS overlay functions. Because of the drawbacks of the standard methods there are also special programs that have been developed by, for example, Eurostat.

In this case it is important to differentiate between intensive and extensive variables[160]. 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.

Intensive variables are easier to handle, because the original polygons are derived from the spatial variation itself (NERI Technical Report). 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.

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 [161]. The third method is developed by Eurostat and written in the ArcInfo macrolanguage, AML[162].

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

159 (Briggs 2000) *Spatial Transformation Methods for the Analysis of Geographic Data*, (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.* And (Eurostat 1999) *GIS Application Development. Final Report. SUP_COM 1997 -LOT 3*

160 (Goodchild and N. 1980) *Areal interpolation: a variant of the traditional spatial problem.*

161 (ESRI 2000) *ArcInfo 8.2: Cell-based Modelling with GRID, ARC/INFO User's Guide*

162 (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.*

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:

Extensive $X_j = {}^{n}\Sigma_{i=1} (X_i (A_{ji}/A_i))$

Intensive $X_j = {}^{n}\Sigma_{i=1} (X_i(A_{ji}/A_j))$

where

X_j = data for grid square j

X_i = data for region i

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

A_i = area for region i (extensive)

A_j = area for grid square j (intensive)

n = number of regions

2.3.3.3.3 Testing the prototypes

In this study two types of tests were chosen to assess the usability of the prototypes for grid-based statistics.

Visualisation of grid-based statistics

Harmonised grid-based statistics should make statistics by different territorial divisions comparable especially from the visual point of view. Different sizes of territories do not affect the image of spatial distribution. In this study the prototypes of gridded data for both test areas are visualised and compared with NUTS5–based statistics for the same areas. Two different sizes of grid cells are used: 1km x 1km and 10km x 10km. The visualisation is accomplished using ArcView 3.2 software.

Delineation of urban areas

Grid-based statistics are generally considered to be good for spatial analyses. The problem of delineating urban areas was chosen as the user case. With reference of Eurostat studies, the hypothesis of this study is 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 this 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)[163].

The method of delineation by grid-based statistics is an application of another method of Statistics Finland for delineating localities. It is accomplished with the ArcInfo 8.2 software.[164]

163 (Eurostat 1998) *Urban database.*

164 (ESRI 2000) *ArcInfo 8.2: Cell-based Modelling with GRID, ARC/INFO User's Guide*

The basic idea is to create zones from selected grid-polygons with population densities equal to, or above, 500/km^2 and then join the population data (summarised by zone) to these zones. The zones whose populations equal or exceed 50,000 are selected.
The method step by step:
The data to be used include two polygon grids, one for the density of population (= **density_grid**) and the other for the total population (= **total_grid**).
Step 1: Create point coverage **point1** from **total_grid** where each point represents one 1km x 1km or 10km x 10km grid polygon. This step eliminates the risk of redundancy error in the data in e.g. in coastal areas.
Step 2: Select from **density_grid** the polygons where the density of population is equal to, or above, 500/km^2. Then reselect overlapping polygons from **total_grid**. Convert the selected polygons to a new coverage.
Step 3: Create buffers around the features in the new coverage created in "Step 2". First create the buffers outside the polygons and then, using the same buffer distance, create another set of buffers inside the already buffered features in order to preserve the original area. Output is **buffer_zone**.
Step 4: With IDENTITY, overlay **point1** and **buffer_zone** to create new point coverage containing the <buffer_zone>-id. Output is **point2**.
Step 5: With FREQUENCY, summarise the total population in **point2** by <buffer_zone>-id to create INFO table **FRQ**. Join **FRQ** to **buffer_zone**.
Step 6: Finally, from **buffer_zone** select the zones with total populations equalling, or above, 50,000.

2.3.3.4 Optimal size of grid cells

It is often argued that the cell resolution should be the same as, or coarser than, that of the input data, or that to ensure compatibility between countries the ideal cell size should be equivalent to the larger regions [165].
In this project the hypothesis of the optimum size of the grid will be tested using samples of data from two countries. The conversions from polygons to grids are made using two different cell sizes. The results of the conversions are compared using the Finnish real grids and estimated grids. The comparisons are made by studying the statistical variables of the resulting data sets, the confidentiality problems of the data sets and the usability of the data sets for spatial analyses.

2.3.3.5 Optimal projection system

A grid-based statistical system of regular tessellations is by definition constructed using equal areas. The construction, visualisation and analysis of grid-based data require a rectangular co-ordinate system that does not distort regular polygons, or distorts them as little as possible. For a European-wide system it has to be universal, and mathematical transformations from national systems need to be available.
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[166]. UTM co-ordinates define

165 (ESRI 2000) *ArcInfo 8.2: Cell-based Modelling with GRID, ARC/INFO User's Guide* and (Eurostat 1997) *Geographical Information Systems in Statistics. Final Report of SUP.COM 95. LOT 15.*
166 (EEA. European Environmental Agency 2001) *CORINE Land Cover*

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 system was chosen for the test data, but with different zonings for the UK and Finland. It was agreed that UTM30 would be used for the UK and UTM35 for Finland.

2.3.3.6 Tests with empirical data

2.3.3.6.1 The test data

The test data contain population data from two countries:
The UK data comprise 1991 census count data by enumeration district (ED, irregular polygons) of the region of Wales. The population density was calculated using information on the ED polygons. The data were delivered in MapInfo format files. The original data included 6,376 enumeration districts (polygons). A subset of 817 polygons covering a part of the South Glamorgan County was also used.
The Finnish data contain the 1998 population counts by 1km x 1km grid and postal code area in the Helsinki region. Both the grid-based data and those by postal code area (polygon-based data) were calculated using data on geo-referenced buildings. The population densities for the postal code areas were calculated using information on the postal code area polygons. The data were delivered in ArcInfo coverage format. The original number of 1km x 1km grids was 13,003 and that of postal code areas 392.
The testing can be divided into three parts. One part consists of the testing of the algorithms for converting polygon-based data to grids. The quality of the results is studied by comparisons of the Finnish data; weighing real grids (compiled from point-based, accurately geo-referenced data) against the estimated grids (converted from polygon-based postal code area data) produced using three different algorithms. The tests will determine the selection of the method for converting polygon-based data to grid-based data.
In the second part of tests after prototypes of grid-based systems are made from test data of both countries the population structures are visualised by grid-based statistics. The quality of the results is studied by comparing them to same kind of visualisation of NUTS5 areas.
Finally, the prototypes are used for delineating the urban areas of the test areas. Again, comparing is performed by using NUTS5-level building blocks to do the same kind of delineation.

2.3.3.6.2 Results

Testing the algorithms

Different algorithms for converting polygon-based input data to grid data were tested against the hypothetical real grids (the Finnish gridded data made by using the data on the reference points of buildings). The Finnish postal code area data were converted to grids and the results were compared against each other and against the real gridded data.
To render the results of the tests comparable, the Finnish data needed to fulfil certain assumptions:
-The source data (seed data) are exactly identical for each conversion of polygon-based data and point-based data
- Each grid set is in the same projection system
- The spatial coverages of the data are exactly identical

- The origos of the comparative grid nets are the same
- Each grid cell has only one value for each variable

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" (Table 6, Table 7).[167]
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 quite well 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 6).
Using the other variable, i.e. population density, the variation in the results is smaller (Table 7). 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.
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.
Visualisations of the results of the different conversion procedures are contained in Figures 25-30. All visualisations are made by the results with the population count variable. From maps one can see that RegionGrid algorithm imitates the reality quite well at least in the Finnish case. The PolyGrid algorithm on the other hand transforms the polygon-based data so that it looks like the original data, if the size of a grid cell is small enough. In larger grids PolyGrid - and PointGrid –algorithms seemed to give more random results.

Visualisation of gridded data[168]

The visualisations of gridded data are made as typical choropleth maps without any smoothing methods.
With the Finnish test data it is obvious that small grids give the best result in terms of quality or reality. The 1 km x 1 km gridded data shows that there are variety in population density inside the municipalities and postal code areas, which is smoothed away with averages of polygon-based data. On the other

[167] Results were compared visually as well (Figures 25-30).
[168] Visualisations of population densities by NUTS5, 10km x 10km and 1km x 1km grid square are contained in Figures 31-33.

hand 10 km x 10 km grids seem to be too big to compete with postal code areas but seems to be better than NUTS 5 areas to describe population density variation inside the test area.[169] In the region Wales 1km x 1km gridded data follows the trends of data by enumeration areas. There is only one enumeration area in the test data, which is as large as 10 km x 10 km grid. This is why also in the sub-area of South Glamorgan 10 km x 10 km is far too big to illustrate population structure.[170]

Delineation of urban areas[171]

By comparing the test results to these reference maps one can detect a big difference[172]. 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[173]).

The basic statistics in the table attached (Table 8) 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 of areas of NUTS5 regions in the two different countries. The South Glamorgan test area is quite small to be used a grid size of 1 km x 1km to describe continuous phenomena – The whole region of Wales seem to be better to apply this size of grid cells. Also for the Helsinki region this size seems to fit quite well.

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 bigger when defined by 1km x 1km building blocks than when defined by NUTS 5 building blocks.

2.3.3.7 Conclusions

2.3.3.7.1 Discussion

The data collected and maintained by National Statistical Institutes are enormously diversified according to spatial configurations. Small area statistics differ from country to country and sometimes even inside a country depending on the type of the data concerned. Some kind of harmonisation of the input data is greatly needed before they can be used for further, comparable analyses. This is especially true in studies concerning the spatial characteristics of different variables and different countries.

One way to harmonise data is to produce statistics by grid squares. A system of grid-based statistics has a lot of advantages, which have been discussed in Work package 2.2. However, the production of grid-based statistics can be complicated depending on the input data. A European-wide system of grid-based statistics needs harmonised production methods.

---

[169] Figure 31.

[170] Figures 32-33.

[171] Visualisations of the results are contained in Figures 35-37. Figure 34 includes the "points of departure" of both the test areas and of the delineation of urban areas using NUTS5-based statistics.

[172] Figures 35-37

173 (Martin 2000) *Towards the Geographies of the 2001 UK Census of Population*

There are several methods for converting point-based or polygon-based data to grids. It may be a question of finding the best practices for producing grid-based statistics. This is especially true with register-based, accurately geo-referenced data. Nevertheless, there are also methodological problems. Different methods of converting polygon-based information to grid-based information give different results. The results are highly dependent on the characteristics of the input data and on the amount of information available. There are also several specific problems, which can or cannot be taken into account in standard conversion methods. These relate, for example, to the way the data near the outer boundaries of the study area, or the data on islands or other special physical structures, should be taken account.
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 reasons but also by confidentiality reasons. In sparsely populated countries like Finland grid-based statistics face confidentiality problems, which need to be solved with spatially oriented disclosure control methods. Even in the test data from the Helsinki region 40% of the 1km x 1km grids have fewer than 10 inhabitants. The size of the cell should also be considered from the processing speed and disk space points of view. 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 Europe.
A grid-based statistical system is a geographical information system. A common reference system for geographic information is needed to ensure that the data are compatible across Europe. For the time being, each country has its own map projection and their data are locally defined. For a common system of grids there should be an International projection for the whole of Europe for small-scale mapping. However, the power of the grid systems does lie in medium and large scale studies which need several projections with either sectors or zones for keeping the distortions of grid squares as small as possible.
According to Finnish experiences, there are many kinds of issues that have to be taken into account if data by grid square are to be delivered. Some of the most important of these are confidentiality problems, which are not studied in detail in this context. There is also a problem with diversified systems from the customers' viewpoint, if grid-based statistics are to be delivered. This, too, has not been discussed in this study. In this sense the objectives in the constructing of a grid-based statistical system are the same as in the constructing of any kind of new information system. The system should, for example, be equipped with universal and versatile technical solutions and delivery formats.
Grid cells can be encoded in raster or vector format. In Finland, vector-based cells are used more often, but it is rather easy to convert data from vector to raster, and vice versa. The main advantage in using vector-based grid cells compared to using raster-based cells is the possibility to assign multiple attributes to a vector-based grid cell. The most important disadvantage of vector-based grid cells is the lack of any compression techniques.

2.3.3.7.2 Recommendations

The Tandem Consortium

A harmonised, grid-based statistical system is a relevant alternative to polygon-based diversified national systems of territorial statistics. In this study, the intended purpose of a grid-based statistical system is to provide comparable building blocks for further analyses and for compilations of territorial divisions for statistics.
A European-wide system of grid-based statistics needs harmonised methods for converting data from different kinds of input data to grids. The Nordic system of grid-based statistics may be a candidate for the best practice concerning point-based geo-referenced data. A "benchmark" for the methods for converting polygon-based information to grids was tested in this study together with some standard GIS methods. The RegionGrid algorithm developed by a research group (Eurostat 1997) proved to be the best in this study, too. In general, when there are no additional data available the use of the RegionGrid algorithm is highly recommended for converting polygon data to grid data.
The optimum grid-cell size may vary depending on the quality of the data, their accuracy and on the scope of the analyses to be done. The smaller the grid cells the greater the resolution and accuracy, but the coding, database storage and processing speed for analyses are more costly. If a cell size is finer than the input data resolution no conversion will produce more accurate data than the input data. It is generally accepted that the resultant grid size should be the same or coarser than the one for the input data. The test with the Finnish data supports this notation.
The map projection system, UTM, proved to be good in this study. Different zones for the UK and Finnish data prevented distortion of the grid cells in maps. However, further studies are needed to test the effect of data transformation from one co-ordinate system to another. It is obvious that the input data should be transformed before it is gridded. The transformation can also only be done to gridded data. There are no estimates available of how much the order of the procedures affects the final results. There is also a need for case studies with a small-scale gridded data system for the whole Europe. If there is only one international map projection for small-scale gridded maps for the whole of Europe, how much will the grid cells be distorted in the Border States?
Visualisations of gridded data can be done with advanced smoothing techniques. These techniques are the subject of further studies. In this study, traditional visualisation techniques were used to prove the characteristic of gridded data of minimising the modifiable area problem (MAUP)[174]. The spatial units are not only comparable within a country but also between countries.[175]
One type of phenomenon that the grid-cell data structure is best suited to represent is continuous spatial data. To test this a case study of delineating urban areas was made with both the Finnish and the UK data.[176] When the building blocks are comparable the results of the delineation are comparable as well - provided the delineations in the different countries are made using the same method. The method that was chosen here is relatively simple and can be performed with standard GIS tools. Compared to the urban areas

[174] (Martin 2000) *Towards the Geographies of the 2001 UK Census of Population*
[175] Figures 31-33.
[176] Figures 35-37.

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 tell about misleading 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.

This study hopefully represents one step towards a common geographical base for comparing statistics across Europe. Further studies are needed before final recommendations for the construction of a European-wide, grid-based statistical system can be given. There is scope for methodological development in the areas of data security, conversion programs (polygons to grids) and database management (grid hierarchies), for example. More case studies are also needed with different kinds of input data, scopes of analyses and types of spatial analyses.

This study has ascertained that a grid-based statistical system may be an alternative for a common, European geo-spatial statistical system. Further studies are warmly recommended.

2.3.4 Figures for WP_3.2

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

	sum	count	mean	max	min	variance
Population 1998 (10 km x 10 km)						
Real grid squares	1 541 507	168	9176	290445	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 6: (Statistics Finland Table 1): The Finnish test data Comparison of real grid squares with estimated grid squares

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

	count	mean	max	min	variance
Population density 1998 (10 km x 10 km)					
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 7: (Statistics Finland Table 2): The Finnish test data Comparison of real grid squares with estimated grid squares

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

	Total population	Pop density	Area (km2)
Finland			
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

Table 8: (Statistics Finland Table 3): Statistics of urban areas defined by different kind of building blocks

Part 3. Results and Recommendations

Lars H. Backer, Statistics Sweden

Abstract

This Paper describes efforts on behalf of the "Tandem" project to inquire into the feasibility of building a system of small statistical areas (SSSA) for Europe. This problem has not proved to be straightforward in the sense that it may be solved but using well-known and established procedures. Instead of a non-existent ready solution, we suggest a method, from which a SSSA will emerge from the dialectics between "Desirable" and the "Feasible".

Our basic assumptions:

1. This project is, and should, be driven by user needs
2. That it is possible to build a system like this from existing (although .poorly harmonised) components
3. That it is not feasible to build a new system "from scratch".

What we have done:

1. We have formulated and discussed the idea that there are three fundamental promises that may be realised with a SSSA: Better data, Better spatial analysis and Better statistical aggregations between and onto new tessellations.
2. We have formulated and tested the idea that an "embryo" to a SSSA that may be compiled from existing data (Statistics), features (Regular and irregular tessellations) and methods (for aggregation and clustering)
3. We have formulated and discussed the idea that the best way to test the feasibility of this idea is to test the feasibility of this idea.

What we recommend to do:

1. Regarding user needs (Case study)
   We recommend to expand and deepen our knowledge of "user needs" with further inquires into issues related to spatial development with a focus on urban areas.
2. Regarding the design (The Definition)
   We recommend that the design for a system of small area statistics is based on the draft for the "Tandem" definition for a system of small area statistics, that is to be further developed in response to the results from future iterations.
3. Regarding the prototype (Features, data and methods).
   We recommend encouraging the further development of the "Tandem" embryo into a proper prototype to an open SSSA for Europe, hereby using existing components (statistics, systems of both regular and irregular tessellations and methods), structured according to object methods and technology.
4. Regarding further development of the prototype
   We recommend to develop and expand, the open Prototype for a SSSA according to an iterative R+D method that allows all suggested improvements to prototype be tested in the perspective of changing user needs and improvements to the design.

3.1 Results and Recommendations (WP_4)

3.1.1 Introduction

If we can trust the judgement of the influential historian Ferdinand Braudel[177], then we, for good and bad, are no longer living in closed worlds. With the emergence of the informational society, we are all participating in a wast network of open systems that seems to melt us all into one large socio-cultural and economical system of open interdependencies. This process is not very unlike the current European project, or the integration we had to endure within the emerging national states not so long ago.
If this prophecy becomes true, we will need qualified systems of knowledge for a hierarchy of local to global man-environmental systems to support decisions, to confront threats and exploit opportunities- not only to regions defined by borders, but to networks spanning large portions of the earth.
Within this context the Tandem project have tried to meet the challenge to contribute to the development of a system of small statistical areas for Europe.
The Tandem project is trying to improve the geographical base for statistics within the EU by exploring the feasibility of building a system of small statistical areas, it is not concerned with any efforts to change the existing NUTS system. A system of SSSA's, may serve, as we shall see, different purposes within a system of geo- statistics.

Stockholm August 2001

Lars H. Backer
Statistics Sweden

[177] See (Braudel 1993) *A History of Civilizations*

3.1.2 About this section

This part of the report is one of several work packages. Its purpose was intended as a more summary in order to describe the Tandem project as a whole. The result is not a straight set of recommendations, but rather the draft for an iterative project that will eventually lead to a working system of small statistical areas. The reason for this lie in the fact that we are not confronted with an established problem that may be solved using well-known methods. Instead, we are confronted with an emerging problem to which no ready solutions apply and to which the only method to test the feasibility is- to test the feasibility of it.
The paper therefore has to describe a result that in effect is a project description and not the explanation of a given set of result.
It is therefore a pity that the current paper does not give enough credit to preliminary results especially in connection with the theoretical and practical assessments described in the work packages 2 and 3 (produced by Marja Tammilehto-Loude and Philippe Guiblin). Their contribution(s), although fundamental to further work as it provides what we have called an embryo to a prototype that could be developed into a SSSA, is only mentioned indirectly in the current text.
The form of this and similar reports reflect, naturally enough, the sub-culture from which the main author has emerged. In this case the result is rather structural and technical, and perhaps not very readable at times. This will satisfy some readers but turn off those who are not interested in questions related to the design and construction of SSSA's.
We will, to this end, write a cross between a reading manual and an executive summary to improve the "communicative power" of our project.

3.1.3 The Task

We have in the preliminary document leading to the formulation of the current project argued for need to build or develop a system of small statistical areas where the components are:

1. Roughly equal in terms of population or area
2. As small as possible
3. Homogenous across the EU
4. May be described with statistics

3.1.4 Three perspectives

The Tandem project is rather an inquiry into 3 central issues related to the task to design, build and implement a system of small statistical areas (SSSA) for Europe in response to the belief that the possible is found somewhere between the desirable and the feasible.

1. The Desirable.
   The desirability of a system as related to the needs of major uses and user's of these tools and this information.
2. The Possible
   The formal and functional properties that should be used in the design, building and development of a prototype for a SSSA for EU
3. The Feasible
   The Feasibility of a project to develop a prototype to a SSSA into an efficient tool for meeting our customers needs

We decided that the best way, given the resources available, to explore these issues was to embark upon an effort to compile a small but representative embryo to a system of small statistical areas.

3.1.5 Two concepts

There are two concepts that are central to the understanding of this paper; the ideas reflected in the words open and closed systems. There are a series of different more or less elaborated versions of definitions to describe these concepts.
Both words are related to the idea of a system. The word system is derived from the Greek word Synhistanai ("to place together") To understand things systematically literally means to put things into a context. "Systems thinking" is based on the idea that the property of the whole arises from relationships between parts, and thus trying to understand a phenomenon within the context of a larger whole[178].
Systems are frequently classified into two groups; "open" systems and "closed " (or "feedback") systems. These terms are difficult because they tend to be counter-intuitive. In fact the term "open" and "closed" systems are frequently used to mean the opposite of what was originally meant with the concepts.
Forrester[179] has remarked that an open system is characterised by outputs that respond to inputs, but where the outputs are isolated from and have no influence on the inputs. A closed system or feedback system is in contrast influenced by its past behaviour.
In popular non-technical use, especially related to "living systems" the idea of openness relates to the idea that a system is able to develop to grow and evolve, in contrast to the idea of closed-ness of a machine, that reflects the idea of being closed, confined to a given set of conditions, and thus incapable of development[180].
This is not the place to discuss this topic further but we will resort to the popular use of the term leaning on the idea that the "proof of the apple is in the eating". We will in these documents retort to the idea of an "open" system as referring to a system that is able to react on positive and negative feedback systems that may grow and develop by learning from previous experience. We will conversely use the term "closed" systems to relate to systems that are designed and built to fit a given set of conditions. When these conditions change (as for instance when a machine runs out of fuel) the closed system is by itself no longer able to adapt and it closes down unless incorporated into an open system that refills the tank.

---

[178] This section is a rephrasing of definitions mentioned by (Capra 1996) *The Web of life; A New Scientific Understanding of Living Systems*
[179] See (Forrester 1968) *Principles of Systems; Text and Workbook*
[180] Examples of the use "open-ness" in this manner is frequently to be found in technical literature (Popper 1971) *The open Society and its enemies, Plato; Part 1,* (Popper 1971) *The open Society and its enemies, Hegel and Marx; Part 2* as well as in the public "discourse". We talk about "open-ness" in relation to the ability to adopt to changing conditions. We believe in keeping an "open mind" to be receptive for new ideas, and to learn from previous behaviour (feedback).

3.2 A. A Vision (formulating the desirable)

3.2.1.1 Start from what exists

The "desirability issue" is an inquiry into questions related to the need for a system of small statistical areas. This effort is not starting from a "Tabula rasa", it is rather a development that naturally must grow from what already exists.

3.2.1.2 Two users and two uses of geo-statistics

We believe that there are two important uses, and consequently two important user groups of interest to anyone who tries to improve geo-statistical systems in Europe.

1. The first user group will have benefits from a system for doing spatial analysis. As we have argued in this project it is not possible to do this on data sets aggregated on administrative areas not only due to their size, but mainly because we need a system of input areas that will have to be clustered prior to proper analysis.
2. The other user group will benefit a better tool for comparing aggregated statistics to complement the use of NUTS areas and equivalent.

It is our hypothesis that a new system of small statistical areas will have to meet these fundamental needs.

3.2.2 Hypothesises for the current iteration

In the current first iteration, the desirability issue was explored on a topic chosen from the work to serve as "decision support" in connection with urban-rural problems as they appeared in connection with the ESDP's (European Spatial Development Perspectives). Here it is essential that we have access to comparable structural (and possibly later) delimitations or delineation's of areas with comparable structural properties.
The experiments, although ambitious on this point, was promising but too limited to be of any direct practical use yet. But if further pursued we may hope for the fulfilment of three promises:

1. Better data availability, resolution and quality
   A system of small statistical areas promises us an infrastructure that may be used to improve the availability, quality and resolution of current statistics.
2. Better infrastructure for Spatial Analysis
   A system of small statistical areas, promises us an infrastructure for doing better spatial analysis based on existing statistical information.
3. Better tools to improve the quality of aggregations and dis- aggregations
   A system of small statistical areas promises us an infrastructure to improve our abilities to aggregate and dis-aggregate statistics within larger hierarchies of statistical areas (both regular and irregular tessellations).

3.2.3 Discussion

What we call user needs are studied differently in different professional traditions. In order to inquire about the desirability for a system that does not exist we frequently find ourselves caught in a "catch 22" situation; How can I inquire about the desirability of a product that does not exist? This is especially apparent in situation(s) where we are dealing with so-called open

systems that constantly change. The best way to solve this seems to be to study potential "users" and "case studies" that are constantly updated as customers change and with them the structures and processes we contribute to.

3.2.3.1 Case studies

So, whereas the demands for information and statistics in the first case may be described by market surveys and similar methods, we must in the latter case resort to other means.
The professional method to develop a complex system, is generally to study customer needs by referring to "case studies" to model the customers processes in order to provide services that will provide added value. In our case we are looking for case studies that not only give added value to a limited group of potential users, but also are suited for the development of the prototype in all of three central respects:

1. Projects to improve existing data
   We will need to explore possibilities for compiling small area data sets for the EU. Here it is often argued that the individual NSI's hardly will submit sensitive data on individual observations, or even aggregations on small areas. Still it is possible that harmonised data sets are produced and analysed simultaneously in all member states and aggregated at Eurostat.
   - The "Aggregation problem"
     Provide better statistics to more systems of areas use direct aggregation and dis- aggregation methods. In this case all aggregation(s) and dis- aggregation are dependent on a clean hierarchy of areas where statistics on the smallest areas may always be aggregated to any of the larger. The current projects is here involved in the quest for a system of small(er) statistical regions (areas) that may allow
     - aggregations onto more systems of regions, and
     - may be used for spatial analysis
   - The "Cutting problem"
     Provide better statistics to more systems of areas that demands the shifting from one hierarchy of areas to another. This is generally accomplished by using so called cutting methods. This problem involves the need for better statistics on areas that may not be described as aggregations of existing systems of statistical areas[181].(The current project is not involved in issues pertaining to this set of problem)
2. Projects to improve Features (geographical objects)
   We assume that it would in most instances be very costly to build a system of SSSA's from scratch. The result will depend on the availability of existing SSSA's that are currently used for this or other purposes.
   - Systems of regular tessellations
     To develop and improve systems of Regular tessellations
   - Systems of irregular tessellations
     To develop and improve systems of Irregular tessellations

[181] See the problem where statistics are given on a system of administrative areas and has to be transferred to a system of water catchment areas.)

3. Projects to improves methods
   We have in the course of the present project worked on the assumption that the methods needed to produce and test SSSA's are critical to the production, maintenance and development of a shared system of geo statistics.
   - Methods for creating improving data sets for SSSA's:
     - Improve the quality of small area estimations
     - Improve the quality of sampling
   - Methods for creating and improving tessellations for use with SSSA's:
     - Regular tessellations
     - Irregular tessellations
   - Methods for clustering SSSA's.

3.2.4 Recommendations for future work

3.2.4.1 Case studies

Method to verify or refute the assumptions that are used as a foundation for an inquiry into the "desirability issue" is not to produce market- or similar analyses, but to test a set of assumptions on empirical studies. We believe that the best methods for such inquiries are case studies as generally used in connection with object modelling and analysis.

3.3 B. A Prototype (building the possible)

3.3.1.1 The parts exist but do not fit together

The "form and function" issue is an inquiry into questions related to the form and function expected of a prototype to a qualified system of small statistical areas. It follows from the desirability issue that this Prototype should be built from existing components. There are two important observations here:
The first is that we assume that all the parts necessary for building a system of small statistical areas already exist in one form or another in the majority of all of the national statistical systems.
The other assumption is that for some of the central topics there are several solutions offered to solve similar problems. It is a central to the current project that a "dialectical" approach is used in these cases. This implies that when pairs of solutions appears, both solutions are encouraged and implemented. The obvious example for this is to include both systems based on regular tessellations and those using irregular tessellations, as these are to be regarded as complimentary.[182]

3.3.2 Hypothesis for the current iteration

We have in our reports discussed the structure of an open system of small area statistics, built according to the rules of the scientific method. We have here noted that any hypothesis (here prototype) should be well supported by theory as well as practice.

1. A prototype in the sense of a real world system for a practical assessment
2. A theoretical system of knowledge or model to describe the prototype.

[182] See here the discussion in the preliminary papers and the Project description (Backer 2001) *Tandem Wp_0: The Project Description*

In the current project we have made an effort to describe and discuss both of these parts in a theoretical and a practical assessment.

3.3.2.1 An ad-hoc approach for a practical assessment

In the situation where the embryo to a prototype has to be compiled from existing components it is reasonable to assume that an "ad hoc" approach to the development of a prototype is the most promising. This method simply implies that a system is assembled and crudely fit together in order to function, even with poor performance. This approach is generally used only in combination with an iterative development strategy.

3.3.2.2 The "object" approach for a theoretical assessment

In order to test the feasibility of an open system of the type we are seeking here we need a method to describe it that allows us to handle the difference between building closed and open systems. We believe that this demand calls for the use of an object approach.

3.3.3 Discussion

In the WP_1 (the "Professional Context"), we have argued for the need to reconsider many "views" that we have taken for granted regarding statistical systems in general, and geo-statistical systems in particular. We are here especially thinking about our tendency to regard "statistical systems" as closed structures that will improve but not change in any fundamental way over time.

It is our view the methods we use to produce, distribute and consume goods and services (including statistics) in so-called modern societies has changed so profoundly in later decades that it makes sense to talk about a paradigm shift. This is not an intellectual vision, but a pragmatic fact that we are forced to relate to. Due to the fact that the greater part of the world economy (and therefore also a majority of our users outside the public sector) has already changed in order to exploit the opportunities this shift offers, we have no alternative but to adopt. This change is not intellectual exercises to be disregarded at will, but rather changes to which we are forced to adapt.

3.3.3.1 An open system

The desirability of a solution is generally related to the utility that it offers its user(s) From the users perspective there are two types of problems that require different solutions, or that are better solved using different methods; "established" and "emerging" problems. We are not primarily concerned with established problems because they are generally reasonably well solved with existing methods. Instead we are concerned with an ever-increasing number of emergent problems with no traditional solution.

Whereas established problems are well known as well as those who try to solve them, we have to laboriously seek effective solutions to emergent problems.

The consequence of this apparent shift of paradigm, does not however imply that we have to re-start building our statistical systems from "scratch. It merely demands that we broaden our classical view of the geo-statistical system, from static to a dynamic perspective. An open geo-statistical system is not to be regarded as a machine in this sense, but rather as a constantly changing network of processes that perpetually adapts itself to changing conditions.

When regarding the world around us, it seems evident that some systems change very quickly indeed whereas others do not change perceptively at all. We must consequently be able to handle both situations with confidence, as both solutions are equally necessary.

3.3.3.2 Building a Prototype on what exists.

Our object here is not primarily to comment on problems related to different system(s) of small statistical areas, but to design and build one. The method we are suggesting here is to develop the new from components that already exists. One version of these methods is to adopt an ad-hoc approach where the new system is first built by compiling a system from existing components. The new system does not even have to function well at start but is regarded as a seed from which the new system may grow through a series of iterations.

1. An ad hoc conceptual sketch ("embryo" or "seed" solution) compiled from existing components.
2. A working Prototype. Developed through a series of iterations.
   - Start:
   - Re-design
   - Test and evaluate
   - Recommend improvements
   - Iterate:
3. A finished product to meet user needs

Generally speaking a product never leaves the prototype stage. It is constantly on the "workbench" in order to be improved and adapted to new and ever-changing user demands.

3.3.3.3 An "Open" prototype

The prototype for a system of small statistical areas may, be described and handled using three easily recognisable parts. The difference between an open system and a closed system lies in the dependence of the open system on a method for dealing with the need for constant adaptation to changing constraints (Point 3 below).

1. A working prototype (practice)
   First a workbench is needed. The workbench consists of the latest version of the product under development as well as all tools and personnel needed for its development. In the current project the Tandem consortium is responsible for providing the workshop and the workbench for the development of the SSSA.
   The whole purpose of the current exercise is to build and develop a system of small area statistics. It is not a design only, merely described in a set of reports. It is rather a system of data (statistics) geographical features, and methods (programs) that may eventually develop into a complete system
   - To describe and analyse the spatial distribution of phenomena across the territory of Europe. (analysis)
   - To aggregate statistics on range of administrative and non-administrative areas. (aggregations and dis- aggregations)
   - To serve as a tool to analyse and improve existing and new data sets for these regions. (improving data, features and methods))

2. A system of knowledge (theory)
   Secondly, we need to register all knowledge needed to describe and explain the project and its use. This system of knowledge is generally accumulated in a series of documents describing the system as a whole as well as a kit of parts.
   For the prototype to become useful for the user, a series of descriptions are needed. These documents are later used for the management of the prototype .
   - Recommended methods (manual for its use)
     Documents to describe the different methods used for different data manipulations needed in the three main processes
   - Specifications (Blueprints)
     Documents to describe the data, features and methods used.
   3. Methods for developing the prototype
      Thirdly, we need to define and develop a method for the development of the system of small statistical areas for EU. The only feasible way to develop, to "nurse" a prototype from a mere embryo to a well functioning product to apply an iterative R+D method.

3.3.4 Recommendations for future work

3.3.4.1 A prototype for an "open" SSSA

We cannot model the development of a complex open system, due to the fact that open systems are not designed in the sense that they are built, once and for all, according to a given set of plans. Each iteration improves the system, but always in relation to changes in what is desirable or feasible.
So, we need an embryo to a prototype for an open system of small statistical areas to test the feasibility of a SSSA that fulfils user needs concerning data, features for doing better spatial analysis and better aggregations.

3.4 C. Evolution (developing the feasible)

The Feasibility issue is an inquiry into questions related to the feasibility of a project dedicated to the task of developing a system of small statistical areas. We find that there is a fundamental difference between the task of testing the feasibility of a closed as compared to that of testing the feasibility of an open system.
In the first case we are dealing with established situations that may be described because they are known. I the case of open systems we do not initially know what tasks they have to perform. Open systems are on the other hand "self-developing" and may change fundamentally as time goes by.

3.4.1 Hypothesis for the current iteration

The hypothesis for the current iteration is based on the preliminary papers and other contributions written prior to and leading to the Tandem project.

1. Here the feasibility was studied with observations **on** systems of small statistical areas.
2. In the current stage of the project the feasibility is to be tested using information **for** the development of a SSSA for Europe. The idea now is to break the dilemma that you cannot really know whether it is feasible to build a SSSA- before you build a SSSA.

3.4.2 Discussion

The feasibility of an open system is dependent on how it stands up to a series of iterations, where the Prototype is evaluated as an efficient solution to problems defined through Case studies as described above.
In this procedure a project is feasible as long as it compete with "best practices" in a contest to meet customer demands. In the present case, where we have no existing best practices to benchmark against, each iteration competes with those that go before. The end of the exercise comes when we run out of funds or we are not able to provide a better solution to the case study problem.
The design of all efforts will in the initial stages be founded on the professional judgement of the developers. It is believed however, that in due time, when the product has found its way to the final user, the solutions offered will have to become subject to the users acceptance or refinance.
According to our judgement, arguments in favour of a system like this may be founded on three main points or arguments. We will accordingly look for case studies that give us an opportunity to develop the SSSA on all these scores:

1. The need to improve the availability, quality- and resolution- of statistical data sets related to small statistical areas.
   - Data (statistics)
     - Improve the quality of small area estimations
     - Improve the quality of sampling
   - Features
     - Regular tessellations
     - Irregular tessellations
2. The need for data, and infrastructure for doing comparable spatial analysis on small statistical areas for the EU
   - Map and analyse spatial patterns related to socio-cultural and economic processes.
   - Map and analyse spatial (man-made) infrastructures indispensable for development
   - Map and analyse the extent and consequences of Floods, storms and other natural catastrophes
3. The need to improve the statistical systems ability to aggregate and dis-aggregate statistics between systems of regions beyond limited to the administrative areas used primarily for public administrative purposes.
   - Improve the quality of aggregations (and dis- aggregations) on irregular tessellations that do not comply with NUTS or other systems of administrative areas.
   - Improve the quality of aggregations (and dis- aggregations) of data sets that require the calculations from one system (of regions) to another

3.4.2.1 Iterative R+D

The iterative method for R+D is based on a simple series of iterations of a type that is widely used in the information industry. It is based on the principle of accumulated growth where a prototype is developed over a series of iterations where-ever new ideas to improve the prototype are tested and evaluated. In this process all aspects of the project is constantly improved.

Not only the prototype itself, but also all other components as drawings and descriptions; the methods to test and analysis the results of each run, and the hardware- and software- tools used in the process.

3.4.2.1.1 Preparatory

10. Study the need for a SSSA[183]
11. Describe professional context and a method
12. Describe user needs with case studies

3.4.2.1.2 The first Iteration (current project)

13. Design and build a prototype
14. Test and evaluate
15. Formulate recommendations for further work (iteration 2)

3.4.2.1.3 The second iteration

16. Redesign a new version of the prototype
17. Test and evaluate.
18. Formulate recommendations for further work (iteration 2)

3.4.2.2 Preparatory work (preliminary iteration)

Most information here is based on studies described in the work package "Professional Context" (WP_1)

3.4.2.2.1 1. Study the need for a SSSA

We have not been able to locate any "in depth" studies describing the need for a system of small area statistics in Europe. We have however, based on our professional experience, in this and other reports argued that that there is definitely a clear need for an European hierarchy of small areas for statistics, based on similar existent or planned systems at national and local levels, and integrated with similar efforts on a global level.

3.4.2.2.2 2. Described the professional context (the theoretical foundation)

We have formulated a description of a professional context that may be used a theoretical foundation for the development of a SSSA. This foundation has served to provide a conceptual specification for a SSSA that may be constantly developed over time.

3.4.2.2.3 3. Described user needs from case studies (the empirical foundation)

We have formulated a description of the practical context for a SSSA's based on the study of a case study founded on real needs[184]. For the first steps towards the definition of a prototype we have focussed on the need for better delineation (delimitation) of urban areas.

3.4.2.3 The first Iteration (iteration 1)

As the work on the "Professional Context" was accumulative, and developed in parallel to the work on the work packages WP_2 to WP_3, the design did not fully respect the demand for strict adherence to the need for a consequent OO approach.

3.4.2.3.1 4. Designed and built an embryo for a prototype for a system of small statistical areas.

[183] A System of Small Statistical Areas.

[184] Here refer to Daniel's 5 areas where improved data is needed, and where this method may prove valuable, if not indispensable. (Mountain areas, Islands, Urban areas, Environmental statistics, Urban studies,)

We have designed and built an ad hoc prototype that may be developed into a working system of small area statistics.

- Urban statistics
- A system of small geographical areas

3.4.2.3.2 5. Test and evaluate the prototype

We have, in the course of the 2001 iteration (Mark_1) of the Tandem project, run a series of (sub-) iterations to get the embryo working and to test and evaluate the results. The results of these tests are described in the WP_2 and WP_3 of the Tandem report(s).

3.4.2.3.3 6. Present the results to Eurostat to decide upon further action

The current project is primarily planning for the next iteration Mark_2 (Iteration 2 / for the year 2002) but with a view to Mark_3 (Year 2003) and Mark_4(Iteration 4 / 2004).

3.4.2.4 Preparations for further iterations

We have presented our work in a series of reports to Eurostat, our beneficiary, along with recommendations for further iterations for two more years, or otherwise according to.

- Mark_2 (Iteration 2 / 2002)
- Mark_3 (Iteration 3 / 2003)
- Mark_4 (Iteration 4 / 2004)

3.4.3 Recommendations for future work

What we have accomplished so far produced one full iteration of the development process. This means that the embryo is not yet quite mature enough to be called a prototype. In order to get that far, we have to bring the R+D process through more iteration(s).
This means that we have, in addition to the iterative process itself, provided all initial information needed to start the Iteration process.

1. (Step 1) Design and Build a Prototype (Iteration 1)
   - (Step 2) Test and evaluate the Prototype (1)
   - (Step 3) Formulate recommendations for further work (Iteration 2)
2. (Step 1) (Re-) Design and (Re-)Build the Prototype (Iteration 2)
   - (Step 2) Test and evaluate the Prototype (2)
   - (Step 3) Formulate recommendations for further work (Iteration 3)
3. (Step 1) (Re-) Design and (Re-)Build the Prototype (Iteration 3)
   - (Step 2) Test and evaluate the Prototype (3)
   - Etc.

3.5 Results

Our present stage of the Tandem inquiry into the Desirability, and Feasibility of a project to develop a prototype for a SSSA may regard the three components discussed below as interconnected links in a never-ending process, that like an inward circling spiral that gradually improves the result with each iteration.

3.5.1 The Vision

3.5.1.1 The Basic assumption

The need for a better geographical base for statistics

The basic assumption(s) in connection with testing the desirability of a system of small statistical areas is that it should improve the comparability of statistics. This need is clearly reflected in the initiative leading to the current project.

3.5.1.2 The hypothesis for the current iteration

3 Promises

The desirability for a system of small statistical areas (SSSA) seems, according to our current knowledge, to be founded on three Promises;

1. Better data availability, resolution and quality
   A system of small statistical areas promises us an infrastructure that may be used to improve the availability, quality and resolution of current statistics.
2. Better infrastructure for spatial analysis
   A system of small statistical areas promises us an infrastructure for doing better spatial analysis based on existing statistical information.
3. Better tools to improve the quality of aggregations and dis- aggregations
   A system of small statistical areas, promises us an infrastructure to improve our abilities to aggregate and dis- aggregate statistics within larger hierarchies of statistical areas (both regular and irregular tessellations).

Our inquiries into this issue have not revealed any objections to this hypothesis.

3.5.1.3 Hypothesis for future work:

Depend on case studies

We believe that the best way to increase our knowledge about user needs in this type of situations is to study case studies as frequently described in professional literature[185].

In order to look for suitable cases, it could be advisable to choose those that contribute to development on most if not all of the 3 points mentioned above.

3.5.2 The Prototype

3.5.2.1 The basic assumptions

The components already exist, and should be used

The basic assumption with regard to what is possible in relation to a system of small statistical areas is that the components needed for such systems already exist. We believe that it is possible to assemble a working SSSA from these components, but believe that a lot of work ahs to be done in order to develop it from a prototype into a well functioning whole.

It is a central to the current project that a dialectical approach is used. This implies that when pairs of solutions appears, both solutions are encouraged and implemented. The obvious example for this is to include both systems

[185] Here see for instance (Jackobson 1994) *The object advantage, Business Process reengineering with object technology*

based on "Regular tessellations" and those using "Irregular tessellations", as these are to be regarded as complimentary.

3.5.2.2 The hypothesis for the current iteration

An Ad-Hoc method described and built with object methods and technology

In order to build a prototype for a SSSA, from components already in existence, that in time may develop into well functioning system we have to retort to a method that is suitable for the task. This calls for the use of an ad hoc[186] method in combination with an iterative R+D method for developing a compiled prototype into an interacting whole.

We have however frequently argued that when dealing with open systems of the kind we are looking for here, we have to build a parallel system of knowledge that describes it. For this purpose we have suggested an object approach widely used in the informational industries for this purpose.

3.5.2.3 Hypothesis for future work:

We suggest to build an open self-developing SSSA

We believe that, in order to develop a prototype to an open SSSA, systematically from existing components, we are well advised to apply object technology. This (crudely) means that the prototype, as well as a finished product could (in addition to supportive documentation), consist of 3 main components:

1. Properties
   A system of data (Statistics) derived from the case study(s) applied.
2. Features
   A system of geographical features (here a system of small statistical areas, either regular or irregular tessellations)
3. Methods
   A collection of methods needed to cluster regions and aggregate data along lines similar to those described in the WP_2-WP_3 the Tandem Reports.

3.5.3 The Evolution

3.5.3.1 The Basic assumptions

It is not feasible to build a closed SSSA

It is not feasible to suggest the development of solutions that requires the construction of a completely new system. It is a basic assumption that we have to develop the new system from existing components, already up and running although not harmonised or useful with equivalent systems in other countries.

This assumption, beside economical considerations, leads us to abandon the idea to build a closed system to be built as a finished product to be implemented in all of the member states at the same time. Instead we have to think in terms of a method for building an ad hoc open system, that is gradually integrated and harmonised over time.

[186] See for instance (Boguslaw 1965) *The new Utopians*

3.5.3.2 The hypothesis for the current iteration

We have to build a system to test its feasibility

There are principally two different methods used to write feasibility studies. The first, would be formulated as a study on the problem of building a system of small statistical areas. (We might suggest that the first, preparatory papers[187] are efforts to formulate a study of this type.) The second method is to refrain from theoretical speculations and argue that the best way to test the feasibility of a project is to implement it (on a small scale, if necessary) as far as it goes.

Another important point to mention here is the fact that as long as we are dealing with closed systems, then both of these two alternatives may prove useful. When discussing open systems, however, only the second alternative will be practical due to the fact that user needs and other constraints here change constantly. Although there are risks involved here, we believe that the best method to test the feasibility is to test the feasibility.

3.5.3.3 Hypothesis for future work:

The iterative R+D method

The Feasibility for a project to succeed that takes as its object to develop a SSSA according to the Tandem model would be considerably improved if organised according to state of the art methods. It seems obvious however that these components (data sets, features and methods) are assembled ad hoc and need to be developed into a working prototype over a series of further iterations. We have however a well functioning iterative R+D method that will serve this purpose.

1. (Step 1) Design and Build a Prototype (Iteration 1)
   - (Step 2) Test and evaluate the Prototype (1)
   - (Step 3) Formulate recommendations for further work (Iteration 2)
2. (Step 1) (Re-) Design and (Re-)Build the Prototype (Iteration 2)
   - (Step 2) Test and evaluate the Prototype (2)
   - (Step 3) Formulate recommendations for further work (Iteration 3)
3. (Step 1) (Re-) Design and (Re-)Build the Prototype (Iteration 3)
   - (Step 2) Test and evaluate the Prototype (3)
   - Etc.

3.6 Recommendations

The task of building a SSSA is not an established problem in the sense that the constraints needed to write a specification might be formulated in advance. Instead it is to be regarded as an emerging problem, whose solution is not known, and the direct path that leads to it is not yet made[188]

To this end we recommend an "evolutionary" approach to the development of a SSSA based on work done so far in the Tandem projects. We have in the

[187] See here in the Reference for a list of some of these contributions (Wagget 1999, (20-21 October)) *Towards improved statistical comparability across Member states- A better geographical framework*, (Backer, Tammilehto-Luode and Rogstad 1999) *The Use of Grids to Improve the Comparability of Statistical Data.* (Rase 2000) *Technical Specifications; Improving comparability of statistical data across EU Member States,* (Backer 2000) *Tandem GIS: A feasibility study towards a common base for statistics across EU. A management summary* etc

[188] Please see the correct citation by Hideki Yukawa as quoted in (Gleick 1987) *Chaos; Making a New Science* to be found at the top of this paper.

project described what we believe to be a feasible and well tried set of methods designed to seek a possible, well functioning solution(s) to satisfy the desirable within the limits of the feasible.

3.6.1.1 Explore the "desirable"

In order to explore our notions of the desirable we recommend that we make a continuos effort to explore user needs, not through market research or similar information **on** SSSA's, but rather through direct information **for** the satisfaction of these needs.
For this purpose we recommend the use of a few but carefully chosen case studies, designed and implemented in order to be used in a process to develop an emerging SSSA.

3.6.1.2 Build the "possible" prototype for a SSSA

In order to formulate a prototype making as far as possible use of existing components, we have built and tested a crude "embryo" according to this principle. It is our hope that this might serve as a first constructive step towards the development of a well functioning SSSA.
To achieve this objective we recommend applying an open system approach to the development of a genuine prototype to a SSSA by gradually extending and deepening the results from these first experiments.

3.6.1.3 Test against the "feasible"

For a design to become an acceptable, if only to serve as a preliminary prototype for an emerging SSSA, we need to study its feasibility on a well-functioning method for testing and evaluation. The fundamental principle here should be that any solution that scores best in a benchmarking based on user demands, is to be considered as "best practice" and used as the point of departure to judge further suggestions for improvements.
For this purpose we recommend the use of an iterative system of R+D to guaranty a systematic "never-ending" development process that constantly adopts to changing needs.

References & Bibliography

Lars H. Backer, Statistics Sweden

Abstract

These lists of references are intended to evolve into a system of source material related to issues related to small statistical areas and small area statistics. All cited documents are collected in an "Endnote" database available through the "Tandem" consortium.

References

Reports

The Results and recommendations resulting from the activities sorting under, or related to the Tandem project are reported in a series of reports that also have been further explored in a series of complimentary papers
For details please see the attached bibliography.

Background papers leading up to the current project

References (For details see the Bibliography)

- (Wagget 1999, (20-21 October)) *Towards improved statistical comparability across Member states- A better geographical framework*
- [(Backer, Tammilehto-Luode, and Rogstad 1999) *The Use of Grids to Improve the Comparability of Statistical Data.*
- (Rase 2000) *Technical Specifications; Improving comparability of statistical data across EU Member States*
- (Backer 2000) *Tandem GIS: A feasibility study towards a common base for statistics across EU. A management summary*

Primary Reports:

The work from the Tandem group has been prepared for publication in 5 reports (including the Project description), one for each of the main work packages (WP 1-5):
The Reports from the work produced by the Tandem consortium in this field has been summarised in three blocks:

1. A Professional context
   A professional context that has as its object to define a method to design and develop the theoretical context and method for compiling an embryo to a system of small statistical areas,
2. Experiments with a prototype for a prototype
   An experiment to design test the embryo according to the method designed by the professional context.
3. Results and recommendations
   A presentation of results with recommendation for that the embryo should be systematically expanded and developed through a series of iterations over the coming years.

At the end of a project all weaknesses and inconsistencies become apparent at least to the authors. In the current case however, we hope that the efforts described in the Tandem project will not come to an end here.
If the Commission, so decides, this project could be turned into an expanding project, where the embryo may develop through a series of effective iterations, into a well functioning system of small statistical areas for the EU.

References (For details see the Bibliography)

*
- (Backer 2001) *Tandem Wp_0: The Project Description*

*
- (Backer 2001) *Tandem Wp_1: The Professional Context*

*
- (Tammilehto-Luode and Guiblin 2001) *Tandem Wp_2.0: A Theoretical Assessment (Summary)*
- (Guiblin and Tammilehto-Luode 2001) *Tandem Wp_2.1: A Theoretical Assessment (Systems of Irregular Tessellations)*
- (Tammilehto-Luode and Guiblin 2001) *Tandem Wp_2.2: A Theoretical Assessment (Systems of Regular Tessellations)*

*
- (Guiblin and Tammilehto-Luode 2001) *Tandem Wp_3.0: A Technical Assessment (Summary)*
- (Tammilehto-Luode and Guiblin 2001) *Tandem Wp_3.2: A Technical Assessment (Test runs 2.)*
- (Guiblin and Tammilehto-Luode 2001) *Tandem Wp_3.1: A Technical Assessment (Test runs 1.)*

*
- (Backer, Tammilehto-Luode, and Guiblin 2001) *Tandem Wp_4: Final Results and Recommendations*

Supplementary Reports:

In order to discuss the questions studied in a wider forum supplementary papers have been prepared by three of the consortium members. The first for the Statistical Commission and Economic Commission for Europe; Conference of European Statisticians in Geneva in June 2001
The second for the world conference of the ISI (International Statisticians institute) in Seoul in August 2001. The third for the conference of the Statistical Commission and Economic Commission for Europe; Conference of European Statisticians; Work session on geographical Information systems in Tallinn, Estonia September 2001

References

- (Backer 2001) *Accountants and Engineers; why the difference?*
- (Tammilehto-Luode, Backer, and Guiblin 2001) *A Feasibility Study Towards a Common geographical Base for Statistics across Europe*
- (Guiblin 2001) *In Search for a common Geographical Base to compare Statistics across the EU: The Tandem GIS project*

Other related papers and "spin-offs":

In addition to the reports and papers mentioned above, a series of additional working papers and notes have been prepared to explore related questions in depth.

References (For details see the Bibliography)

- (Backer 2001) *Point- and Area- based Statistics; The Program for a Workshop in search of a system of an infrastructure for small areas for statistics.*

The Tandem Consortium

- (Backer 2001) *In search of a system of small areas for statistics; The background for two questions for discussion at the ECE Conference in Tallinn in September 2001*
- (Backer 2001) *Networking for Co-operation and Partnership: A Benchmarking Approach to the Development of Geo-statistics.*

Bibliography

Backer, Lars H. 1997. Towards an integration of Space, Time and Statistics. Paper read at Statistical Commission and Economic Commission for Europe; Conference of European Statisticians; Work session on geographical Information systems, 22-25 September 1997, at Brighton.

———. 2000. Tandem GIS: A feasibility study towards a common base for statistics across EU. A management summary. Stockholm: Statistics Sweden.

———. 2001. Accountants and Engineers; why the difference? Paper read at Conference of European Statisticians, June 8 2001, at Geneva.

———. 2001. In search of a system of small areas for statistics; The background for two questions for discussion at the ECE Conference in Tallinn in September 2001. In *The Tandem Project: Working Papers*. Stockholm.

———. 2001. Networking for Co-operation and Partnership: A Benchmarking Approach to the Development of Geo-statistics. In *The Tandem Project: Background Notes*. Stockholm.

———. 2001. Point- and Area- based Statistics; The Program for a Workshop in search of a system of an infrastructure for small areas for statistics. In *The Tandem Project: Working Papers*. Stockholm.

———. 2001. Tandem Wp_0: The Project Description. In *The Tandem Project*. Stockholm: Eurostat, Gisco.

———. 2001. Tandem Wp_1: The Professional Context. In *The Tandem Project*. Stockholm: Eurostat, Gisco.

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 of the National Statistical Offices and National Mapping Agencies, 20-21 October, at Luxembourg.

Backer, Lars H., Marja Tammilehto-Luode, and Philippe Guiblin. 2001. Tandem Wp_4: Final Results and Recommendations. In *The Tandem Project*. Stockholm: Eurostat, Gisco.

Bandler, R., and J. Grinder. 1975. *The Structure of Magic; a book about language and Therapy*. Palo Alto, California: Science and Behaviour Books.

Boguslaw, Robert. 1965. *The new Utopians*. New York: Prentice Hall.

Bracken, I., and Martin D. 1989. The generation of spatial population distributions from census centroid data. *Environment and Planning A* 21:537-543.

Braudel, Ferdinand. 1993. *A History of Civilizations*. Translated by R. Mayne. Harmondsworth, England: Penguin books.

Briggs, David. 2000. Spatial Transformation Methods for the Analysis of Geographic Data. Paper read at UN/ECE Statistical Commission and Economic Commission for Europe: Work Session on Methodological Issues Involving the Integration of Statistics and Geography, 10-12 April, at Neuchatel, Switzerland.

Capra, Fritjof. 1996. *The Web of life; A New Scientific Understanding of Living Systems*. October 1997 ed., *Anchor books*. New York etc.: Doubleday.

Castells, Manuel. 1996. *The Rise of the Network Society*. III vols. Vol. I, *The Information Age: Economy, Society and Culture*. Cambridge MA. Oxford UK: Blackwell Publishers.

Claeson, C-F. 1964. A Chorological Public Analysis. Doctoral Thesis, Sweden.

Coombes, M. G. 1995. Dealing with census geography: principles, practices and possibilities. In *Census Users' Handbook GeoInformation International*, edited by S. Openshaw. Cambridge.

Deichman, U. 1996. Smart interpolation. A review of spatial Database Design and modelling. Use of GIS in Agricultural Research: UNEP GRID Report.

EEA. European Environmental Agency. 2001. CORINE Land Cover: Topic Centre on Land Cover.

ESRI. 2000. *ArcInfo 8.2: Cell-based Modelling with GRID, ARC/INFO User's Guide*. Redlands, California: Environmental Systems Research Institute, INC. USA.

ArcView 3.2 (Mid-end software for processing geographical information). Environmental Systems Research Institute, INC. USA, Redlands, California.

European Commission. 1999. ESPD. European Spatial Development Perspective 1999. Towards Balanced and Sustainable Development of the Territory of the European Union. Potsdam: In-formal Council of Spatial Planning Ministers in Potsdam.

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

———. 1998. Urban database. In *Working document for the Meeting of the Working Party "European Infra-regional Information System and Urban Statistics"*: Eurostat.

———. 1999. GIS Application Development. Final Report. SUP_COM 1997 -LOT 3. Luxembourg: HTS Consultants in Association with Nene centre for Research.

Fisher, P.F., and M. Langford. 1995. Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A* 27:211-224.

Fletcher, R. 1987. *Practical Methods of Optimisation*: Wiley.

Flowerdew, R. F., M. Green, and K. Kehris. 1991. Using areal interpolation methods in Geographic Information Systems. *The Journal of the RSAI* 70 (3):303-315.

Forrester, Jay W. 1968. *Principles of Systems; Text and Workbook*. Second Preliminary Edition (1972) ed. Cambridge, Massachusetts: Wright-Allan Press.

Frank, A., and D. Mark. 1993. Language Issues for GIS. In *Geographical Information Systems, Volume 1: Principles.*, edited by D. Maguire, M. Goodchild and D. Rhind. Harlow: Longman Scientific Technical.

Gleick, James. 1987. *Chaos; Making a New Science*. New York: Penguin.

Goodchild, M. F., and Lam N. 1980. Areal interpolation: a variant of the traditional spatial problem. *Geoprocessing.* 1:297-312.

Goodchild, M.F., L. Anselin, and U. Deichmann. 1993. A general framework for the areal interpolation of socio-economic Data. *Environment and Planning* 25:383-397.

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

Guiblin, Philippe. 2001. In Search for a common Geographical Base to compare Statistics across the EU: The Tandem GIS project. Paper read at Statistical Commission and Economic Commission for Europe; Conference of European Statisticians; Work session on geographical Information systems, September 2001, at Tallinn, Estonia.

Guiblin, Philippe, and Marja Tammilehto-Luode. 2001. Tandem Wp_2.1: A Theoretical Assessment (Systems of Irregular Tessellations). In *The Tandem Project*. Stockholm: Eurostat, Gisco.

———. 2001. Tandem Wp_3.0: A Technical Assessment (Summary). In *The Tandem Project*. Stockholm: Eurostat, Gisco.

———. 2001. Tandem Wp_3.1: A Technical Assessment (Test runs 1.). In *The Tandem Project*. Stockholm: Eurostat, Gisco.

Haarala, R., and M. Tammilehto-Luode. 1999. GIS and Registerbased Population Census. In *Statistics, Registers and Science*, edited by J. Aho. Helsinki: Statistics Finland.

Hall, Stephen S. 1993. *Mapping the next Millennium*. First Vintage books ed. 1 vols. Vol. 1. New York: Vintage books.

Hansen, H. S., ed. 2001. *PSSD-Planning System for Sustainable development. The Methodological Report*. Vol. 351, *Neri Technical Report*: Ministry of Environment and Energy. Denmark.

Jackobson, Ivar. 1994. *The object advantage, Business Process reengineering with object technology*. Wokingham, England etc.: Addison-Wesley.

Karvonen, M., J Rusanen, M Sundberg, A. Colpaert, A. Naukkarinen, J Tuomiolehto, and for the DiMe Study Group. 1997. Regional differences in the incidence of insulin-dependent diabetes mellitus (IDDM) in Finland during 1987-1991. *Annales Medicin* 29:297-304.

Kauppinen, J., E. Rissanen, J. Rusanen, A. Naukkarinen, T. Muilu, and A. Colpaert. 1997. Migration as a function of population. *Nordia Geographical Publications* 26:17-27.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Langford, M., and D.J. Unwin. 1992. Generating and mapping population density surfaces within GIS. *The Cartographic Journal* 31:21-26.

Lyotard, J. F. 1979. *The Post-modern Condition: A Report on Knowledge*. Manchester.: Manchester University Press.

Martin, D. 1997. Implementing an automated census output geography design procedure, Department of Geography, University of Southampton, Southampton.

———. 1998. 2001 Census Output Areas: from Concept to Prototype. In *Population Trends: 94*.

———. 1998. Optimising Census Geography: the Separation of Collection and Output Geographies. *International Journal of Geographical Information Science* 12.

———. 2000. Towards the Geographies of the 2001 UK Census of Population. *Transactions of the Institute of British Geographers* 25:321-332.

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

Merrian, ed. 1971. *Webster's Third New International Dictionary*. Edited by G. a. C. M. Co. 3 vols. Vol. III, *Encyclopaedia Britannica*. Chicago, London etc.: Benton, Wilhelm.

Miller, J.G. 1978. *Living Systems*. New York: McGraw-Hill Book Company.

Muilu, T., J. Rusanen, A. Naukkarinen, and A. Colpaert. 1999. Local Poverty in Finland 1995. Paper read at The IGU Commission, at Albuquerque, New Mexico, USA.

Nörretranders, Tor. 1991. *Märk världen; En bok om vetenskap ock intuition*. Translated by j. Wahlén. 1993 bokförlaget bonnieer Alba, Stockholm ed: Bonnier Alba.

Ohtomo, O. 1997. Small Area Statistical Databases. Paper read at Second International Workshop on Population Databases and Related Topics, 14-19 January, at Jakarta.

Openshaw, S. 1977. A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning, and Spatial Modelling. In *Transactions of the Institute of British Geographers 2*.

———. 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.

———. 1996. Developing GIS Relevant Zone Based Spatial Analysis Methods. In *Spatial analysis: modelling in a GIS environment*, edited by P. Longley and M. Batty. Cambridge: GeoInformation International.

Openshaw, S., S. Alvanides, and S. Whalley. 1998. Some further experiments with designing output areas for the 2001 UK Census. In *The 2001 Census: What do we really, really want?*, edited by E. Rees. Leeds: University of Leeds.

Openshaw, S., and L. Rao. 1995. Algorithms for re-engineering 1991 Census geography. *Environment and Planning A* A 27 (27):425-446.

Popper, Karl R. 1971. *The open Society and its enemies, Hegel and Marx; Part 2*. Princeton, New Jerey: Princeton Univ. Press.

———. 1971. *The open Society and its enemies, Plato; Part 1*. Princeton, New Jerey: Princeton Univ. Press.

Powell, M. J. D. 1969. A method for non-linear constraints in minimisation problems. In *Optimisation*, edited by R. Fletcher. London: Academic Press.

Rase, Daniel. 2000. Technical Specifications; Improving comparability of statistical data across EU Member States, April 2000.

Robinson, Arthur, Joel Morrison, Philip Muehrcke, A. Jon Kimerling, and Stephen Gubtill. 1995. *Elements of cartography*. Sixth Edition ed.: J. Wiley & Sons.

Rusanen, Jarmo, Arvo Naukkarinen, Toivo Muilu, and Alfred Colpaert. 1996. Asutus Ruotsissa Suomea keskittyneempana. *Tietoaika* 1996 (2):15-18.

Räisänen, S., J. Rusanen, and A. Naukkarinen. 1996. Socio-economic grid data and GIS for analysing changes in the Finnish countryside. Paper read at Second joint European Conference & Exhibition on Geographic Information: Geographic Information from Research to Application through co-operation, 27-29 march 1996.

Statistics Finland. 1960. Non-Administrative Urban Settlements and their Boundaries, etc Helsinki, 1965, 1960. In *General Census of Population*. Helsinki.

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

Tammilehto-Luode, Marja, Lars Backer, and Philippe Guiblin. 2001. A Feasibility Study Towards a Common geographical Base for Statistics across Europe. Paper read at Conference of the International Statistical Institute, September 2001, at Seoul, South Korea.

Tammilehto-Luode, Marja, and Philippe Guiblin. 2001. Tandem Wp_2.0: A Theoretical Assessment (Summary). In *The Tandem Project*. Stockholm: Eurostat, Gisco.

———. 2001. Tandem Wp_2.2: A Theoretical Assessment (Systems of Regular Tessellations). In *The Tandem Project*. Stockholm: Eurostat, Gisco.

———. 2001. Tandem Wp_3.2: A Technical Assessment (Test runs 2.). In *The Tandem Project*. Stockholm: Eurostat, Gisco.

Tammilehto-Luode, Marja, and Lars H. Backer. 1999. GIS and Grid Squares in the use of Register-based Socio-economic Data. In *Bulletin of the International Statistical Institute. (ISI'99. 52nd Session)*. Helsinki: International Statistical Institute.

Tomlin, C. D. 1983. A Map Algebra. Paper read at Harvard Computer Graphics Conference, at Cambridge, Massachusetts. US.

UN-ECE. 2000. Questionnaire on the Implementation of GIS in Statistics. Paper read at Work Session on the Integration of Statistics and Geography, 10-12 April, at Neuchatel, Switzerland.

United Nations: Department of Economic and Social affairs, Statistics Division. 2000. *Handbook on geographic information systems and digital mapping*. Vol. 79, *Studies in Methods*. New York: United Nations.

Vaattovaara, M. 1998. Pääkaupungin sosiaalinen erilaistuminen. (Residential differentiation within the metropolitan area of Helsinki, in Finnish with Abstract in English). In *Tutkimuksia (Studies) 19998:7*. Helsinki: Helsingin kaupungin tietokeskus (City of Helsinki Urban Facts).

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.

Öberg, Sture, and Peter Springfeldt, eds. 1991. *Befolkningen*. Edited by L. Wastenson. 17 vols, *Sveriges Nationalatlas*. Stockholm: Sveriges Nationalatlas Förlag (SNA).