STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 36
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Methodological Issues Involving the Integration of Statistics and Geography
(Neuchâtel, Switzerland, 10-12 April 2000)

Topic (ii): Spatial database management and (geo-)data warehousing

FROM CENSUS MAPS TO THE CENTRAL SPATIAL DATABASE

Submitted by the Statistical Office of Estonia[1]

Contributed paper

ABSTRACT

The Statistical Office of Estonia (SOE) has completed the mapping programme for the 2000 Population and Housing Census. The programme was launched in 1995 and the database was completed in 1999. The delineation of enumerator areas and the printing of enumerator maps were based on this database. However, for the post-Census data processing and analysis, the database structure had to be adjusted. The software for scanning and processing the Census questionnaires and an anonymous Census database were developed in Oracle relational database management system (RDBMS). Thus the decision was made to convert the existing GIS database (in Mapinfo and ArcView formats) to Oracle Spatial. At the time of writing this paper, the database structure has been designed and data conversion will begin after the Census.

I. BACKGROUND OF THE EXISTING SPATIAL DATA AND CENSUS INFORMATION SYSTEM

I.1 Description of Census mapping

1. SOE launched the mapping programme for the 2000 Population and Housing Census in 1995. After completing the test areas the specification of the digital Census maps was compiled. According to the Specification, 1:50 000 maps in rural areas and 1:5 000 maps in urban areas were drawn. The specification was optimised to create a cartographic basis for the Census planning (Census area (CA) delineation) and the Census itself (maps for enumerators, maps for supervisors, etc.).

2. SOE delegated the Census mapping process to two companies – one in urban, another in rural areas. The production methodology was different. In rural areas, paper maps of the 1989 Census were used as a source material for buildings, digitised by the mapping company,

[1] Prepared by Teet Jagomägi and Inge Nael

overlaid to existing 1:50 000 base map and updated by local governments. In urban areas, the existing maps and orthophotos were used as a source and the maps were updated by the mapping company. For rural and urban areas the municipalities compiled household lists including the number of inhabitants in each building or apartment. The purpose of household lists was to provide information about the number of inhabitants for the delineation of enumerator areas (EA).

3. SOE stores digital maps in urban areas in Mapinfo, in rural areas in ArcView software and household lists in Foxpro software. The Census maps were ready in December 1999.

4. The next task was the planning of Census areas (CA). SOE assigned to each settlement the number of EAs based on the population count and other parameters. The aim of planning was to delineate the predetermined number of EAs for each region.

5. For towns (56 settlements) a dedicated software was created which uses Census map data in MapInfo format and outputs the new layers with all three levels of CAs (EA, Census districts and supervisor areas). The work of the software was semiautomatic - the task of the operator was to run automatic enumerator area delineating algorithm, and to verify the intermediate results of the automated process. The software was sent to most of the towns in the hope that local expertise would help achieve a more logical delineation. For the remaining towns the job was done in SOE.

6. The conclusion of the semiautomatic delineation was that the CA generation software was efficient - the whole town was done in less than one working day, adjusting input parameters enabled rapid testing of several versions, output data was used directly in enumerator map printing without further processing, etc. The campaign was faced with a number of problems – computer literacy in towns was not very high, it was necessary to train a lot of operators during a short period of time, the algorithm was sensitive to minor irregularities in the maps, household lists were not up-to-date, etc.

7. In rural areas (205 local governments) and smaller settlements of urban type (172 in total) the CA delineation was done by hand, based on the Census maps and the printed lists of households. The work was undertaken by regional bureaus of SOE in close cooperation with local governments. The paper maps with EAs drawn on them were sent to mapping companies for digitizing.

8. The delineation of CAs was completed on 4 February 2000. In total, there are 5,323 enumerator areas, 995 Census districts (collection of 4-6 EAs), and 165 supervisor areas (collection of Census districts).

9. The printing of enumerator maps began in January 2000 and was completed in March 2000, one week prior to the critical moment of the Census. Four types of maps were printed:

Map type	Scale in urban areas	Scale in rural areas
Enumerator maps	1:3 000 to 1:5 000	1:5 000 to 1:50 000
Census district maps	1:3 000 to 1:22 000	1:10 000 to 1:100 000
Supervisor area maps	1:5 000 to 1:22 000	1:20 000 to 1:120 000
Wall maps for local Census offices	various scales	

For this task two A3-size colour laser printers were used by subcontractors, the third kept in reserve in SOE as a backup for possible equipment breakdowns.

10. In parallel to the enumerator area map printing, the household lists were cross-matched with the Population Register and Building Register data. Selected columns from the resulting database were printed for each enumerator as auxiliary information to speed up the completing of Census questionnaires. The cross-matching was relatively labour consuming since, at present, there is no reliable identifier system to build up the relationship between different registers. Automatically, only 1 to 70% of database rows were matched using addresses, the percentage being higher in urban areas. The relationship between registers was created by local governments using their expertise and information about local inhabitants.

11. As a result of mapping effort, SOE has created a data set of about 400,000 buildings from approximately 300 urban settlements and about 200 rural municipalities, digital maps are associated with alphanumeric data – household lists, which in turn are associated with data from the Building Register and the Population Register. The data set is unique in Estonia in terms of accuracy, completeness, modernity and scope, which is worth maintaining in a better IT-environment than has been feasible so far.

12. The way in which the Census map database was processed until the completion of map production was not the most "high-tech" method, but was completely appropriate for purpose. However, the following disadvantages may hamper further development:

- different software environment for storing the Census data and spatial data would be difficult to handle;
- in case of paper maps the overlapping areas around urban areas were unavoidable, but in the spatial database it creates unnecessary duplication;
- in GIS data files is difficult to ensure data consistency and security;
- the data split between a number of files and file formats is difficult to analyse and use for generation of small-scale maps.

I.2 Description of Census data processing

13. The Census questionnaires will be scanned with two high-volume Fujitsu scanners and interpreted with Eyes&Hands software. During interpretation, operators will resolve the characters or words which cannot be recognized by the OCR/ICR software. The output of interpretation is text files.

14. Text files will be coded using the Oracle-based software. If fields in text files do not match with pre-defined data dictionaries, operators will be asked to resolve the situation. More complex situations will be forwarded to chief operators to handle. The process is self-learning, i.e. typical answers will be added to dictionaries automatically. In addition, context verification rules will be applied (i.e. child should not be older than parents, etc.). The whole process will be monitored from a security aspect – all data access events will be logged and suspicious transactions can be cancelled at any time.

15. The full database will be filed and stored in a highly secured archive. During the next stage, primary and secondary person identification information will be removed. As a result, an anonymous Census database will be created. The anonymous database will be processed by dedicated software for the generation of tabulations and to provide open access to the database. The anonymous database can be used with Oracle Reports, Oracle Discoverer and a number of other tools, including GIS software.

16. The link between GIS and the anonymous database processing software has been designed. For example, the list of buildings can be given as a selection criterion for the software. If the list contains buildings around power lines, the tabulations for population most likely affected by an electromagnetic field will be calculated. However, the anonymous database does not include geographic information at present. The map data must be stored and processed by GIS software (see Figure 1).

17. The Census data processing software is powered by Sun Ultra Enterprise 450 server and Oracle ver. 8.0.5.

18. To improve possibilities of analysing Census data and to overcome drawbacks in the current method of storage of spatial data, SOE launched a project to design a central GIS database.

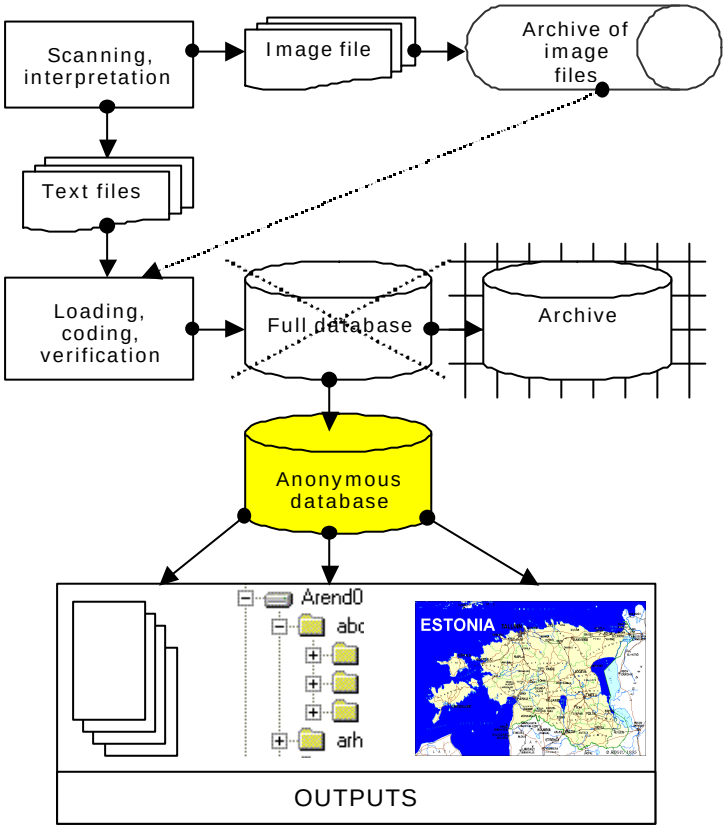

Figure 1. Simplified process diagram of Census data processing. The anonymous database is "GIS-supported", but does not include spatial data. The goal of the Central GIS database is to create tight association with anonymous database and collected map data.

II. BUILDING OF THE CENTRAL GIS DATABASE

19. The project for designing a central GIS database was launched in December 1999. The objective was to create a spatial database which would satisfy the following requirements:

- enables queries over the whole territory
- ensures user rights
- metadata is stored together with data
- data is not deleted, records are marked "not valid"

- data is suitable for a small-scale thematic mapping and large-scale analysis
- the database contains information of census maps and household lists
- the database avoids data duplication
- uses open data structure, permits use of as many GIS software applications as possible
- the database structure enables export of data for Web services in the future
- enables relation with an anonymous Census database
- enforces the technical quality and logical consistency of the data
- enables data exchange with municipalities.

20. Deliverables of the project were:
- reality model
- data model (ER-diagram)
- metadata specification
- resource estimation for conversion to the new Central GIS database structure.

21. The objective of the reality model is to guarantee the compatibility between different data sets on a logical level, i.e. to create a common definition of features in different data sets (e.g. to avoid cases where a shoreline in one data set is drawn at the average water level line, in another database the relief's 0-countourline, etc.). The reality model is independent of the hardware and software used for the implementation of the data model. Ideally, there should be one reality model for the whole country and SOE would use it. In Estonia, the generic reality model is not ready, therefore SOE had to take a lead in its scope of interest.

22. The reality model of the central GIS database consists of 21 feature classes. Some elements, which were shown on the Census maps, were not incorporated into the central GIS database (e.g. road closings or cliffs). It is quite likely that they are unnecessary for data analysis or are not feasible to keep up-to-date.

23. For each feature class the following information was described:

- definition
- the source of definition
- guidelines for recognizing and delineating the feature in the nature
- ID system of feature instances
- parameters of features (e.g. type and width of roads)
- relation (e.g. feature class "address" has a relation "uses" with feature class "street")
- mnemonic code
- version number
- remarks.

24. The database model was created based on the reality model. The database model is presented as an ER-diagram, a familiar tool in information systems development. However, for designing GIS databases this technique has been used relatively rarely. This difference will have to be decreased soon, together with eroding the difference of the subject of GIS into "mainstream" information systems. A "standard" ER-diagram was enhanced with additional columns for each entity and attribute, to accommodate GIS-specific information (like digitising rules, accuracy requirements, etc).

25. Oracle Spatial was chosen for storing spatial data. The main reasons were 1) the software for storing Census data is already Oracle and 2) general features of Oracle Spatial, described in Section IV, matched the requirements of Central GIS database.

26. Most of the GIS software packages do not support all standard features of relational database management systems. For example, only the most important subset of SQL is implemented. This imposes constraints on the database structure. In order to elucidate the widest range of GIS from which the client software could benefit from the spatial data stored in Oracle, the database structure was kept as simple as possible (for example, synonyms and composite keys were avoided). See Figure 2.

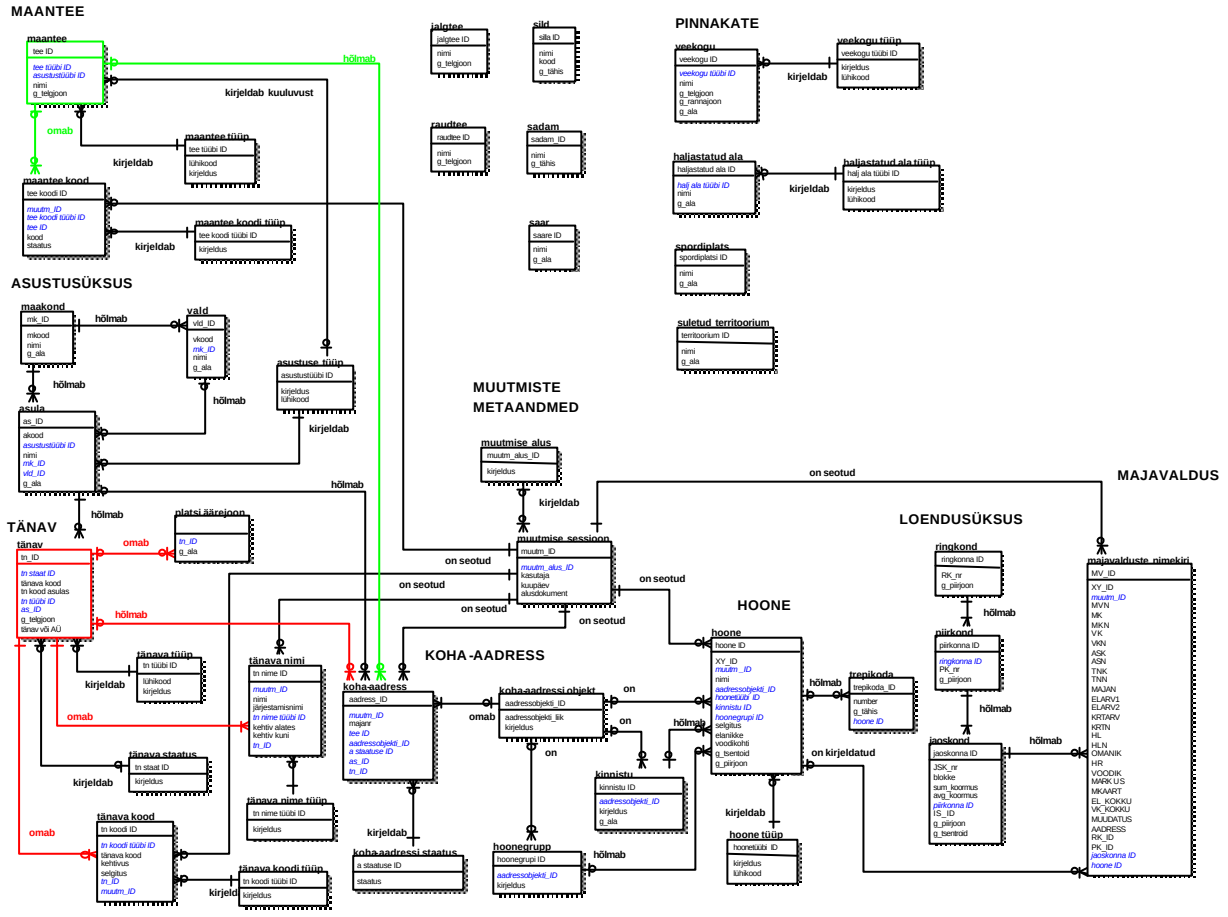

Figure 2. Feature classes were modelled with 1 to 6 entities in Oracle. Entities with attribute starting with "g_" (e.g. "g_ala") can be mapped. Oracle Spatial technology stores location and shape of the geographic object just as another attribute.

27. Time and cost estimation for data conversion to the new structure was made. The most labour consuming part is the quality control of the data to be converted.

28. For implementing the Central GIS database, it is planned to use 10 licenses of Oracle 8i Enterprise Edition and Oracle Spatial. GeoMedia, GeoMedia Pro and Mapinfo Professional will be used as client GIS software for handling spatial data.

III. FUTURE WORK

29. Conversion of data to the central GIS database will start as soon as possible after the peak workload of the Census (Census date being 31 March 2000) is over; conversion is expected to be finished in 6 months. After this deadline the Statistical Office of Estonia will be able to carry out spatial analysis in the same IT-environment, as it is common for alphanumeric data.

30. Enumerators are instructed to mark corrections on the EA maps. These corrections will be entered into the Central GIS database. The result will be the most detailed and up-to-date database of Estonian buildings and roads.

31. As a result, SOE will be able to perform a detailed GIS analysis of Census data as well as generate various thematic maps based on the tabulation data. A powerful and flexible database system gives an opportunity to provide services for studies, initiated by scientists outside SOE, as well as various on-line services in cooperation with other organizations.

IV. STORING OF SPATIAL DATA IN ORACLE RDBMS

32. Oracle Spatial is the foundation technology for the development and implementation of spatial data warehouses. It allows storage and manipulation of both spatial and alphanumeric data in a single database. Any mix of standard Oracle8i tables and spatial data tables can be used with a standard method to retrieve data — SQL. Oracle Spatial is not a GIS software. For displaying graphical data, creating thematic maps, printing maps, etc., the GIS client software is necessary. All major GIS software vendors support Oracle Spatial to a greater or lesser extent.

33. Oracle Spatial supports three basic geometric forms that represent spatial data: [1]

- Points and point clusters – The points may represent locations of buildings, fire hydrants, utility poles, etc.
- Lines and line strings – The lines may represent roads, railroad lines or utility lines.
- Polygons/complex polygons with holes – The polygons may represent outlines of cities, districts, and vegetation. A polygon with a hole may geographically represent a forest surrounding a clearing.

In addition, rectangle, circle and compound elements with and without an area are supported.

34. The interaction of various geometric features may be determined through the use of comparison operators, including touch, overlap, inside, and disjoint. This is common to GIS software, the difference is that data storage tier of information system can perform analysis prior to transporting huge amounts of data to the application (client) software. Oracle can perform also simpler spatial operations - length, area, buffers, polygon manipulations (union, intersection, difference, XOR.)

35. Performance is optimised through the use of a two-stage (in Oracle documentation two-tiered) query model. This model reduces load and query processing overhead and provides scalability as the spatial data volume grows. The first stage, or a primary filter, permits a fast selection of a small number of candidate records to pass along the secondary filter. The primary filter uses approximation to reduce computational complexity. The secondary filter applies exact computational geometry to the result set of the primary filter. These exact computations yield the final answer to a query. The secondary filter operations are computationally more

intense, but they are applied only to the relatively small result set from the primary filter. Queries can be spatially constrained, as defined by an “area of interest” chosen by the user. Eliminating data outside the area of interest from consideration during queries will ensure optimum performance levels.

36. Applications can be created using the client/server or three-tier architecture thus distributing the processing associated with the application across any number of software layers. This flexibility is provided to the user because Oracle Spatial maintains the understanding of spatial data within the database server itself.

37. Oracle Spatial takes advantage of all of the security features offered by the Oracle RDBMS server. Features such as user authentication, role authorization, role based security, integrated security auditing, secure client/server connections, data stream security, and standards compliance are all available to the database administrator when storing spatial data in Oracle. This is the major difference in comparison to standard GIS software packages.

38. Oracle introduced object/relational format of storing spatial data in version 8i, up to this version relational data format was used. The new format is faster, but creates compatibility problems with GIS software, they have started to support the new format gradually.

39. The main advantages of storing spatial data in Oracle RDBMS are:

- all data elements are in the same environment, increasing data security, ensuring data integrity, etc.;
- spatial data is stored in the format which enables its use by a number of GIS clients;
- application software developers can choose between several software tiers for optimising security, speed, ease of use and development flexibility;
- spatial data is documented, structured and organized similarly to the alphanumeric data.

40. The main disadvantages of storing spatial data in Oracle RDBMS are:

- the technology is new and technical specifications have been changed for a number of times (Oracle 7 Spatial Data Option, Oracle 8 Spatial Cartridge, Oracle 8i Spatial);
- the experience with storing spatial data in RDBMS is rare in the GIS community and the learning curve is steep;
- the client software versions are sometimes unstable;
- Oracle does not support multiple coordinate systems and raster data.

41. In conclusion, Oracle Spatial does not replace traditional GIS file formats, but provides valuable option for circumstances, which demand added value provided by Oracle and do not suffer from relatively young technology.

V. REFERENCES

[1] Oracle Spatial. Data Sheet, March 1999.

[2] Inno. V., Data Processing for Census 2000. Newsletter of Statistical Office of Estonia, 1999. (In Estonian)