Integration of Spatial Descriptive Statistical Data and Geographic Information

M. Sabrakos, I. Filis, T. Tsiligiridis

Informatics Laboratory, Agricultural University of Athens, 75 Iera Odos, 118 55 Athens, Greece

Abstract. This paper presents an integrated information tool able to display on a map, accurately, the combined spatial descriptive statistical data along with the geographical information of an area of interest. The user is able to relate data from the different sources in order to find the best matching and reach to conclusions about data correctness. All the statistical information is provided by means of the 'NewCronos' (NC) database, while the geographical 'Corine' Land Cover (CLC) is the basis for the spatial and geographical redistribution of the Farm Structure Survey (FSS) data. The proposed system uses techniques of Relational Database Management Systems (RDBMS), Geographic Information Systems (GIS) and Object Oriented Programming (OOP), enabling the development of some spatiotemporal entities that allow complicated analysis of potential scenarios to be carried out in a landscape study. To test the interface a small pilot area of Greece is chosen. After the reclassification (where needed) of the above data, common classes are created and presented on a map using an embedded GIS environment.

1 Introduction

Most of the statistical data used for policy purposes by Eurostat and other organizations are related to populations, activities, features and other events, which are by their nature, spatial in form. The management, the process and the display of statistical data associated with spatial locations that vary geographically is therefore, largely, a spatial process.

From the agricultural point of view, an important development nowadays is that agricultural activities are more and more combined with other activities such as environmental care, maintaining the landscape in good condition, forestry, preserving recreational and tourist areas as well as small scale of agricultural products, aquaculture, fisheries, etc. A sustainable reform should keep productivity high, so that, farmers remain competitive. Assessing the agricultural policies and their impact on the countryside is still a crucial factor. Thus, there is a strong need for statistical data on rural population, and particularly, on landscape and land use. In agricultural terms, the management of agricultural resources is increasingly complexed as conservation and environmental concerns play an expanding role in making conclusions. In this respect, GIS is considered necessary for the production of census maps, for dealing with census logistics, for monitoring census activities, and for data dissemination (U. Deichmann 1997).

Nevertheless, regional statistics of the EU are managed on the basis of the NUTS (Nomenclature des Unites Territoriales Statistiques) system. Generally, NUTS regions are the administrative regions of a country. For many statistical purposes, the NUTS framework is clearly robust and adequate. However, a general problem with the NUTS system arises from the fact that in many cases the administrative divisions do not coincide with the division by the nature, or by other phenomena of interest. As a result, it cannot be applied in its present form to units that are more relevant from a geographical point of view, such as drainage areas, landscape units, biotopes, etc.

Nowadays, with the advent of GIS, an extremely wide range of spatial analysis methods has been developed for carrying out data transformations between different spatial structures. These methods help to present the data in a more meaningful and consistent manner and enable different data sets, based on different geographical units, to be brought together and overlaid. Spatial transformations vary widely and may be described as processes of aggregation or disaggregation, or as surface modeling on the basis of point, line or polygon data. The classification and the terminology of these methods are not well established but available methods include: point in polygon process, areal weighting, modified areal weighting using control zones, modified areal weighting using regression relationships, optimization, simulated annealing etc. These different methods are based on different assumptions about the underlying spatial distribution of the data and are subject to different types of error and approximation. The above methods facilitate the spatial analysis of the statistical data required in the development and/or calculation of some more reliable indicators for the determination of the state and quality of the environment, able to measure the effect of the agricultural economy, across regions and countries. For this reason, a new software tool is required, able to query a database, aggregate / desegregate the data and plot the results on a map.

This paper presents a methodology, which helps in the development of an interface between the spatial descriptive statistical data from ‘New Cronos’ (NC) database and the geographical information provided by the ‘Corine’ Land Cover (CLC) of the area of interest. The software tool we have developed takes advantage of the Relational Database Management System (RDBMS) and Object Oriented technology in order to expand ESRI ‘MapObjects’ capabilities, providing new properties and methods. In particular, the User Interface (UI) is MS-Windows based and it has been developed using the Sybase ‘PowerBuilder 7.0’ software. The source code is independent of the actual RDBMS used, which in our case is Sybase. Finally, the combined data is presented on a map using GIS. Since the use of conventional GIS software is complexed and requires special skills, the ESRI ‘MapObjects’ are approved as a basis of the GIS software. Note, that the ESRI ‘MapObjects’ are a class of objects allowing further maps to be added and managed within an application.

The structure of the paper is as follows: Section 2 describes the NC database, section 3 describes the CLC database, section 4 presents the application development, while section 5 presents the results from the comparison between the related databases. Finally, the conclusions of the system developed are presented in section 6.

2 Description of the ‘NewCronos’ database

NC is one of the main public distribution statistical information database of Eurostat. It is a dynamic database, constantly updated and re-organised. Datasets generally cover the Member States (MS) of the European Union (EU). Depending on the variables selected, data is available from 1960 to present day, can be annual, biannual, frequent every quarters or month and are expressed in several types of units (indices, absolute values, percentages, etc.) depending upon the indicator.

NC structure is hierarchical, and it has four main levels and uses special software to exploit the multidimensional tables [fig. 1]. The dimensions of these tables specify the country, the holding areas, the unit, the periodicity etc.. More than 160 million items of data in this macroeconomic and social database are available to all people who need high-quality statistical information for making decisions. Without loss of generality and for the purposes of this work only a small part of this data is used. In particular, from the nine (9) themes of level 1 available we have selected the 5th theme, named ‘Agriculture, Forestry and Fisheries’ and from that theme we have selected the ‘Eurofarm’ domain (level 2).

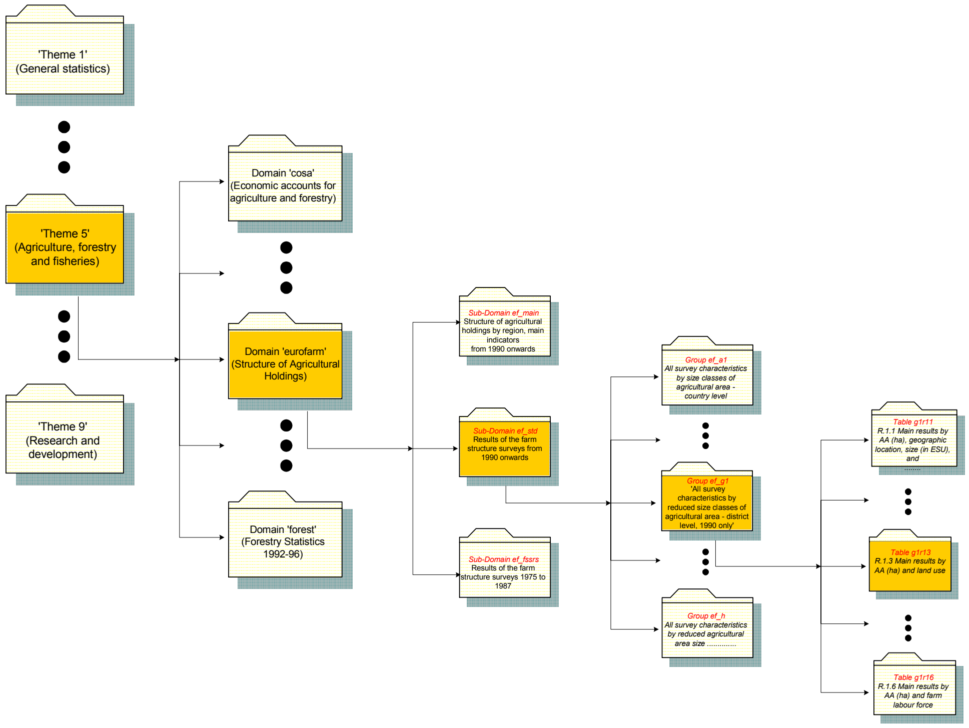

'Theme 1'
(General statistics)
'Theme 5'
(Agriculture, forestry and fisheries)
'Theme 9'
(Research and development)
Domain 'cosa'
(Economic accounts for agriculture and forestry)
Domain 'eurofarm'
(Structure of Agricultural Holdings)
Domain 'forest'
(Forestry Statistics 1992-96)
Sub-Domain ef_main
Structure of agricultural holdings by region, main indicators from 1990 onwards
Sub-Domain ef_std
Results of the farm structure surveys from 1990 onwards
Sub-Domain ef_fssrs
Results of the farm structure surveys 1975 to 1987
Group ef_a1
All survey characteristics by size classes of agricultural area - country level
Group ef_g1
'All survey characteristics by reduced size classes of agricultural area - district level, 1990 only'
Group ef_h
All survey characteristics by reduced agricultural area size ....
Table g1r11
R.1.1 Main results by AA (ha), geographic location, size (in ESU), and ....
Table g1r13
R.1.3 Main results by AA (ha) and land use
Table g1r16
R.1.6 Main results by AA (ha) and farm labour force

Fig. 1 NewCronos structure

The Eurofarm system is actually a network of databases, aiming to facilitate the evaluation of the community surveys on the structure of agricultural holdings in all the MS of the EU. Even though all MS must obligatorily carry out a basic survey (agricultural census) every 10 years, they can also conduct sample surveys at intermediate intervals. The data for basic surveys is available in a three-level geographical breakdown of the whole country, the regions and the districts, while for intermediate, surveys data are only available upon the two levels of country and regions.

3 Description of Corine Land Cover database (CLC)

CORINE (Co-ORdination on INformation of the Environment) Land Cover (CLC) is a geographic land cover/land use database encompassing most of the countries of the European Community, aiming to gather information associated with the environment on certain priority topics. It describes land cover (and partly land use) according to a nomenclature of 44 classes organized hierarchically in three levels. There are five major classes (Level 1), namely, *Artificial Surfaces, Agricultural Areas, Forest And Seminatural Areas, and Wet-*

lands and Water Bodies, which fall into 15 subclasses (Level 2). These subclasses are further divided into the above mentioned 44 subclasses (Level 3). A nomenclature is a list of categories summarizing information in a highly reduced form, while attempting to maintain maximum information content. It normally covers a particular field of interest.

CLC's elaboration was based upon the visual interpretation of satellite images (*Spot*, *Landsat TM* and *MSS*). The smallest surfaces mapped (mapping units) correspond to 25 hectares. Linear features, which are less than 100m in width are not considered. The scale of the output product was fixed at 1:100.000. Thus, the location precision of the CLC database is 100m.

The CLC database has recently become available for most of the territory of the EU and several PHARE countries (AL, PL, CZ, SV, RO, HU, BG, SI, EE, LV, LT). Although its exploitation has just started, it offers the potential for a wide array of uses. It can be used on its own for simple cartographic or statistical presentations and as a base for European-wide landscape analyses or more generally in combination with other data sets (spatial analysis, modelling, etc.).

4 Software Development

The main problem in the development of this application is that data is not distinct and as a result the linking of the two databases cannot be achieved easily. For example, it is not clear if the 3.1 field of the CLC, named, as "Forests" is identical to the 136_G1R13 field of NC, named as "Woodland". To facilitate the decision making process, the user may associate easily the NC's descriptive spatial data with the corresponding CLC's geographical data. In addition the associated data is plotted on a map and then it is compared among each other.

The application consists of the following parts:

- A relational database
- The class of objects for data manipulation
- The class of objects for GIS manipulation
- The main body of the application software containing the above items along with the functions required from the end user.

First of all, a step-by-step analysis of the software design is required. However, for the purpose of this research it is assumed that the pilot area is already known. Then, the appropriate design steps are as following:

1. On the CLC's geographic layer of the area of interest we add the remaining geographic characteristics (contour lines, roads, cities, lakes, rivers etc.). This will help to understand the exact location of the CLC data.
2. From the NC database we select only the "Theme 5" named as "Agriculture Data", and from this theme we select the sub-themes, which are associated with agricultural products. The data selected is at prefecture level, in thousands of hectares of agricultural products, as reported in the 1991 census. In addition sampling data may be combined with census data for validation and / or prediction purposes.
3. The data provided by the NC and the CLC databases has been studied in order to develop the entity relationship model and then develop the database system of the application.
4. CLC data has been stored in some database tables of the application, using especially developed programs, while NC's data stored manually. NC provides also the appropriate DLLs in order to develop programs for automated data transfer.

5. We pointed out the appropriate functions and queries and we developed object classes to satisfy the requirements for uniformity in both, user and developer levels.
6. We developed an application in which the RDBMS, the GIS and the pre-mentioned object classes are used. The basic capabilities offered by this application are the following:
   - Ability to compose (aggregate) a new NC theme by selecting one or more CLC classes and vice versa.
   - Ability to decompose (disaggregate) an existing NC theme to one or more CLC classes and vice versa.
   - Ability to correspond (relate) NC themes to CLC classes.
   - Ability to classify (sort) the results either by date, or by county (region), or by CLC class.
   - Ability to observe the results plotted on the map and to classify these by means of geographical characteristics such as allocation of the selected growth by elevation.

4.1 The Relational Data Base

The GIS tool used by Eurostat for CLC database construction is the ESRI 'ArcInfo' software. This tool stores a set of tables in DBF format, containing both the spatial as well as the descriptive information about map's features, which have been logically organized into themes of information. Each theme consists of topologically linked polygons along with the associated descriptive data. Generally, X-Base formats, such as DBF, DBT, MOD, DIF, SDF, etc., cannot easily aggregate, desegregate, isolate and combine CLC data with other sources. Furthermore due to severe limitations associated with the temporal component of data in the GIS raster databases, a comparison between geographical data obtained in the past is very difficult in practice, (S. Dragicevic et al, 1998).

To support the exchange of heterogeneous data into an integrated database environment a conceptual model is required (C. Parent, 1998). The design of such a model has to take into consideration the loading and refreshing of the descriptive geographical data for each attribute of the GIS, at any time it may be required and then to link these data with the information derived from other sources such as NC data.

The conceptual model of the proposed database is described in [fig. 2]:

In the following points, a brief description of all the entities of the proposed conceptual model is given

- The entity "fss_area" contains the geographical data of the CLC database
- The entity "poly" contains the descriptive data of the CLC database. In case a geographical feature changes an automated procedure is raised, which updates the corresponding row of the "poly" entity. The attribute used to link the geographical information with the database is the "PolyKey". This attribute is used like a Foreign Key between CLC and the database.
- The entitiy "CorDescr" contains the basic agricultural classes of the CLC database.
- The entitiy "fssmaster" contains the basic agricultural categories of the NC database.
- The entity "fsscover" contains agricultural items each one of which is stored per year and per geographical area.
- The entity "fss_clc" contains temporary data designed to support the aggregation and desegregation functionality between the NC and CLC databases.

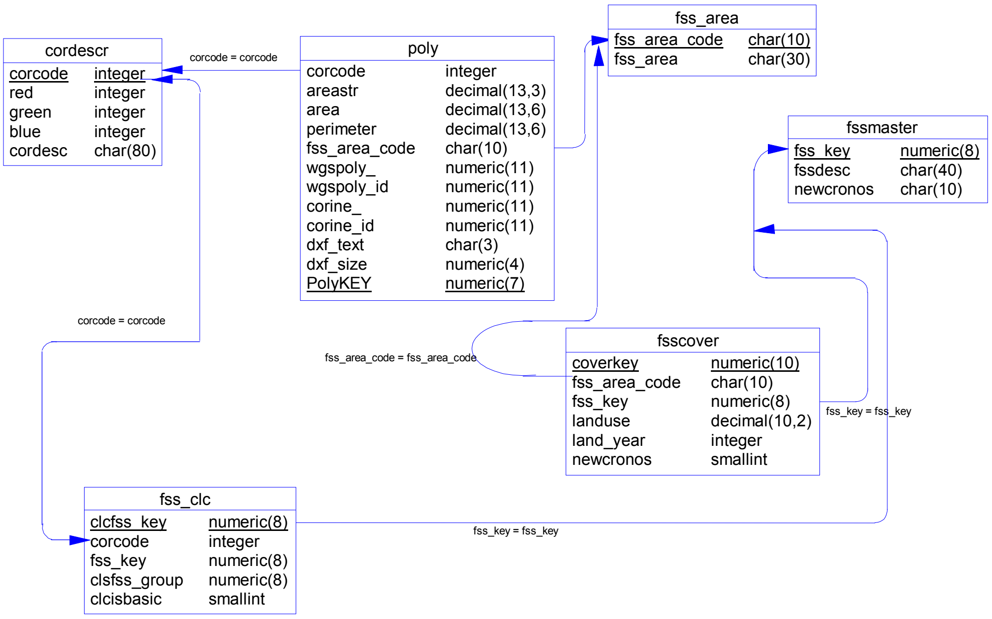


Fig. 2 Conceptual Model

4.2 The class of objects for data manipulation

The class of objects is based on a PowerBuilder object called DataWindow (R. Chandak et al, 1999). This class provides a simple way of retrieving, displaying and updating data from a specified data source. Although the data source is usually a database, it can also be a text file or other data structure. The class of PowerBuilder DataWindow Objects (PB-DWO) inherits the basic functionality and encapsulates the ability to dynamically, at run time, bind and combine data from different sources.

4.3 The class of objects for CLC manipulation

Since the CLC contains only agricultural classes able to be mapped on one or more regions, it will be interesting to relate these classes with other geographical features such as roads, lakes, contour lines etc. This requirement suggests the development of a class of objects that inherit the properties, methods and functions required to process geographical data. This class of objects encapsulates more functions and customized events to finally communicate with the database, and vice versa. In the sequel this class will be called “interoperable GEO-Object". For the development of this class the ESRI ‘MapObjects’ has been used.

4.4 The application software

As it has been pointed out, this application is computer-based software and it is able to display maps and descriptive data in a tabular form. This has been achieved by using geographical information from CLC database linked with tabular information of the multidimensional tables of NC. The user becomes part of the GIS without the necessity of specific skills and intimate knowledge of the data used.

To test the application, the pilot area of the Hellenic Island of Crete has been selected, because it combines different farming during the year with manifold terrestrial particularity. The island of Crete consists of four prefectures; Chania, Rethimno, Iraklio and Lasithi. Each prefecture (nomos) is a different Arc Shape file. The CLC database has been constructed using the Hellenic Geodetic Reference System 1987 (HGRS 87). Any additional geodata used such as roads, lakes, contour lines have been constructed using the World Geodetic System 1984 (WGS 84).

Because CLC is a geographic land cover / land use database encompassing most of the countries of the European Community, the first task is to transform the national coordinates into the global system WGS 84. To solve the problem of geodetic datum transformation without making changes in the application source code a map layer object is added. This object has a property to specify the path of the ASCII file, which contains the appropriate transformation parameters. Thus, in the pilot case, the transformation file from HGRS 87 to WGS 84 is described below:

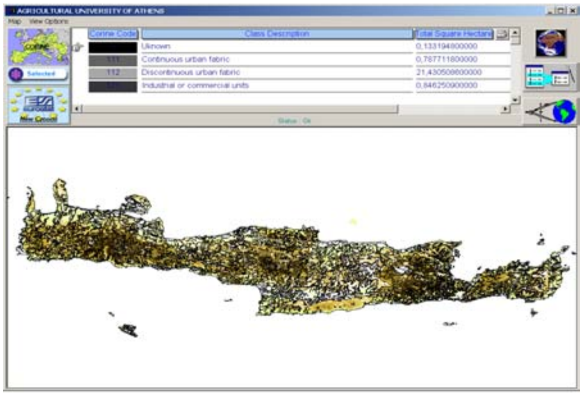
INPUT
PROJECTION TRANSVERSE
UNIT METERS
SPHEROID GRS80
PARAMETERS
0.9996 (Scale Factor)
24 0000 (Central meridian [λ_0])
00 0000 (Standard Parallel φ_o)
500000 (false easting)
00 (false northing)
END

OUTPUT
PROJECTION GEOGRAPHIC
UNITS DD (Decimal Degrees)
SPHEROID WGS84
PARAMETERS
1 (Scale Factor)
00 0000 (Central meridian [λ_0])
00 0000 (Standard Parallel φ_o)
000000 (false easting)
00 (false northing)
END

The basic geographical layer has been constructed using detailed geographical data, such as coastlines, contour lines, roads, airports etc. To verify an identical matching between this layer and the transformed CLC data added on the top the following checks have been made:

- A visual check showed a coincidence of polygon lines in a zoom out at about 400%.
- A numeric check showed that the area of each prefecture of both, the basic layer and the CLC layer are equal. Note that in order to obtain a prefecture area, one has to add the appropriate CLC polygons
- A geo-reference check showed that the maximum difference measured between the geometrical data of the geographical layers is less than 100 meters, which is the maximum precision allowed for products with scale 1:100.000.

The main application window includes the standard GUI controls (menu and buttons) as well as the PB-DWO and the interoperable geo-object. The PB-DWO contains the rows of the entity cordescr matching the selected area. The interoperable geo-object displays the corresponding polygons of the above entity [fig 3].



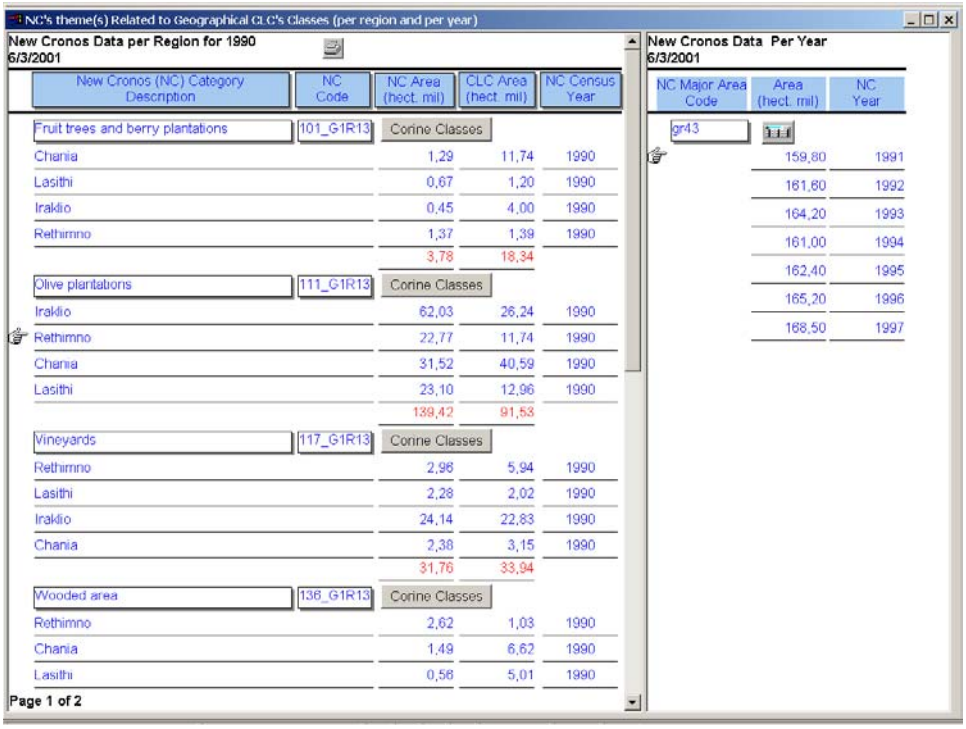
Corine Code	Class Description	Total Square Hectare
	Uknown	0,133194800000
111	Continuous urban fabric	0,787711800000
112	Discontinuous urban fabric	21,430508800000
[illegible]	Industrial or commercial units	0,846250900000

Fig. 3 Main Window

4.4.1 Functions

All the functions supported by the commercial GIS such as pan, zoom, spatial queries, distance and bearing calculator etc., are included in the interoperable geo-object as methods performed by menu selections.

The basic objective of the system concerns Temporal GIS and Spatiotemporal Modeling capabilities using the interoperable geo-object as well as the dynamic creation of thematic maps sowing the themes of the 'New Cronos' per region and per year [fig.4].



NC's theme(s) Related to Geographical CLC's Classes (per region and per year)

New Cronos Data per Region for 1990
6/3/2001

New Cronos (NC) Category Description	NC Code	NC Area (hect. mil)	CLC Area (hect. mil)	NC Census Year
Fruit trees and berry plantations	101_G1R13	Corine Classes		
Chania		1,29	11,74	1990
Lasithi		0,67	1,20	1990
Iraklio		0,45	4,00	1990
Rethimno		1,37	1,39	1990
		3,78	18,34	
Olive plantations	111_G1R13	Corine Classes		
Iraklio		62,03	26,24	1990
Rethimno		22,77	11,74	1990
Chania		31,52	40,59	1990
Lasithi		23,10	12,96	1990
		139,42	91,53	
Vineyards	117_G1R13	Corine Classes		
Rethimno		2,96	5,94	1990
Lasithi		2,28	2,02	1990
Iraklio		24,14	22,83	1990
Chania		2,38	3,15	1990
		31,76	33,94	
Wooded area	136_G1R13	Corine Classes		
Rethimno		2,62	1,03	1990
Chania		1,49	6,62	1990
Lasithi		0,56	5,01	1990

Page 1 of 2

New Cronos Data Per Year
6/3/2001

NC Major Area Code	Area (hect. mil)	NC Year
gr43		
	159,80	1991
	161,60	1992
	164,20	1993
	161,00	1994
	162,40	1995
	165,20	1996
	168,50	1997

Fig. 4 NC Themes - CLC Classes (per region and per year)

4.4.2 Aggregation – Disaggregation – Isolation

The basic requirement of this application is the aggregation and disaggregation of data, namely the ability to stack up and break up the CLC classes, the NC themes and finally to associate them.

As it appears from [fig. 4], the results from the association of the CLC aggregated data with the NC disaggregated data are displayed in the left part of the window, while the CLC and the NC data for the specified year are displayed in the right window. The census area is displayed in the "NC Area" column, while the area shaped after the aggregation and disaggregation is displayed in the "CLC Area" column.

The results of the composition-decomposition procedure are plotted on a map. Although some of the CLC and NC data relations are obvious, some other are not. Therefore this procedure is useful in making new relations.

Internally, the above procedure has been executed with the use of recursive queries, the results of which participate equivalently in the database relations in order to build new queries. The resulted queries feed some spatial queries, the results of which are presented by the interoperable Geo-object.

The following query presents how to obtain the results of [fig.4].

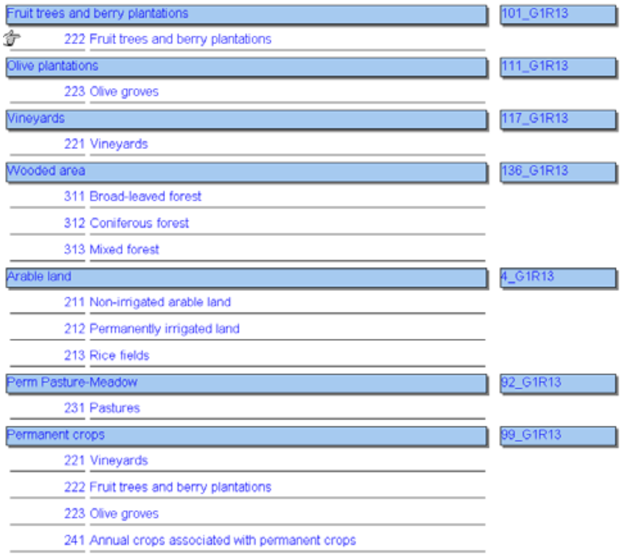
```
SELECT "fssmaster"."fssdesc", "fssmaster"."newcronos", "fssmaster"."fss_key",
       "fsscover"."landuse", "fsscover"."land_year",  "fss_area"."fss_area",
       "fss_from_clc"."total_area", "fss_area"."fss_area_code"
FROM {oj {oj "fssmaster" LEFT OUTER JOIN "fsscover" ON "fssmaster"."fss_key"
        = "fsscover"."fss_key"} LEFT OUTER JOIN "fss_from_clc" ON
        "fsscover"."fss_key" = "fss_from_clc"."fss_key"}, "fss_area"
WHERE ("fss_area"."fss_area" = "fss_from_clc"."fss_area") AND
         ("fss_area"."fss_area_code" = "fsscover"."fss_area_code") AND
         (("fsscover"."newcronos" = 0))    ;
```

Note that the entity 'fss_from_clc’ arises from the following query:

```
SELECT fss_area.fss_area, fss_clc.fss_key, Sum(poly.areastr)/10000000
FROM dba.cordescr, dba.fss_area, dba.fss_clc, dba.poly
WHERE (fss_clc.corcode=cordescr.corcode) AND
      (poly.fss_area_code=fss_area.fss_area_code)AND
      (poly.corcode=cordescr.corcode)
```

5 Results

The combinations of the NC themes and the CLC classes used for this application are presented in [fig.5]. From the combinations we may obtain that the “Vineyards” and “Olive groves” categories are discernible and directly comparable. The remaining combinations have been made using a NC theme associated with a ‘Level 2’ CLC class. For example, the NC theme “Arable Land” is associated with the CLC class “2.1 Arable land” (Level 2 class) which consists of three ‘Level 3’ classes (“2.1.1 Non-irrigated arable land”, “2.1.2 Permanently irrigated land” and “2.1.3 Rice fields”). Additionally, in this theme, we could easily associate one more class (e.g. “3.3.4 Burnt areas”), which belongs to “3.3 Open spaces with little or no vegetation” (Level 2) class, or even “2.4.4 Agro-forestry areas” which belongs to “2.4 Heterogeneous agricultural areas” (Level 2) class. The results for each category can be printed from the software. For example, using the above combina-



Fruit trees and berry plantations 101_G1R13
222 Fruit trees and berry plantations
Olive plantations 111_G1R13
223 Olive groves
Vineyards 117_G1R13
221 Vineyards
Wooded area 136_G1R13
311 Broad-leaved forest
312 Coniferous forest
313 Mixed forest
Arable land 4_G1R13
211 Non-irrigated arable land
212 Permanently irrigated land
213 Rice fields
Perm Pasture-Meadow 92_G1R13
231 Pastures
Permanent crops 99_G1R13
221 Vineyards
222 Fruit trees and berry plantations
223 Olive groves
241 Annual crops associated with permanent crops

Fig. 5 NC themes and CLC classes

tions one may verify that the difference between the NC and the CLC data is 6% for "Vineyards" and 35% for "Olive groves".

Note that a query is developed for the "Olive groves", in order to find the maximum elevation of the particular cultivation. As it appears, the difference is even greater because at about 10% of the cultivation seems to grow between 800 – 1200 meters, something which is not true.

6 Conclusions

The proposed application has been developed by using the RDBMS benefits and the OOP logic, having the advantage that many of the objects can be used in similar GIS applications with a little effort of maintenance. It is an easy-to-use tool, ideal for comparison of descriptive census results and interpreted geo-data, in order to conclude about the correctness of these data. If the expert combines the ability of simultaneous comparison and appearance of results of different years, the conclusions will be more reasonable.

Future research is to continue improving the idea of interoperable geo-object by adding methods and properties for uncertainty manipulation and to investigate requirements of GIS in a fuzzy object data model. Our final objective is to embody in the Geo-Object the ability to generate and visualize transitions from one state to another, using the rules of an expert spatiotemporal system.

References

1. Christine Parent, Stefano Spaccapietra, Esteban Zimanyi, Conseptual Modeling for Federated GIS over the Web. Ecole Polytechnique Federale de Lausanne – Laboratoire de Bases de Donees, 1
2. Dueker, K.J. (1979) Land Resource Information Systems: A Review of Fifteen Years Experience. *Geo-Processing*, 1, 105-128.
3. ESRI MapObjects (1999), Programmer's Reference.
4. UML Document E.S.R.C Economic and Social Research Council at http://reads.dur.ac.uk/newcronos/documentation/en/theme5/eurofarm/notmeth.htm
5. UML Document Eurostat at http://reads.dur.ac.uk/newcronos/nc.html
6. Lorentzos N, Sideridis A , Yialouris C, V. Kolias (1999), An integrated spatiotemporal system. Elsevier, Computers and Electronics in Agriculture 22 (1999) 233-242.
7. Ozemoy, V.M., Smith, D.R., and Sicherman, A. (1981) Evaluating Computerized Geographic Information Systems Using Decision Analysis. *Interfaces*, 11, 92-98.
8. R. Candak, P. Chandak (1999), Advanced in PowerBuilder 7 Techniques. Wiley. pp. 202 – 253.
9. Richard E. Plant, Marc P. Vayssieres (2000), Compining expert system and GIS technology to implement a state-transition model of oak woodlands. Elsevier, Computers and Electronics in Agriculture. 27 (2000), 71-93.
10. Uwe Deichmann (1997) Geographical information systems in the census process – Technology options, costs and benefits. Paper prepared for the Workshop on Strategies for the 2000 Round of Population and Housing Censuses in the ESCWA Region Cairo, 6-10 December 1997.
11. Valerie Cross, Aykut Firat (1998), Fuzzy objects for geographical information systems. Elsevier, Fuzzy sets and Systems 113 (2000) 19-36.
12. Vuzana Dragicevic, Danielle J. Marceau (1998), An application of fuzzy logic reasoning for GIS temporal modeling of dynamic processes. Elsevier , Fuzzy sets and Systems 113 (2000) 69-80.