GEOSPATIAL DATA COLLECTION AND ANALYSIS AS CRUCIAL PROCESSES IN AN INTEGRATED CENSUS

Olivia Blum and Rinat Calvo
Israel Central Bureau of Statistics

Abstract

In an integrated census, data collection is a combination of an indirect enumeration process of records in administrative registers, and a direct data collection using sample surveys. These heterogeneous sources of information imply selective coverage, different physical and analysis units, different formats and hence, not a straightforward connectivity or linkage of information and not a smooth warp and woof data blanket. Geospatial analysis is enacted in this realm to serve three special and unique functions.

First, it is an anchor for census data and therefore a data integrator. All records having any spatial anchor can be translated into a GIS layer through projection. Spatial analysis while drilling down through several layers enables the integration of data from different sources and of different attributes, while creating complex spatial entities.

The second function is supplying the processes of the integrated census with an infrastructure. The most prominent one is creating the census frame. Unlike the traditional census, where the coverage of the defined area and the coverage of the defined population are done simultaneously in the enumeration process, the integrated census implies that only the boundaries of the area and the definition of the population are defined but not the content. Drawing partial populations from administrative sources means that the samples for the supplement and correction surveys have to be drawn from an unknown census frame. Hence the use of spatial analysis is a source of synthetic items within the area boundaries.

Finally, in a multisource data collection, geospatial analysis is a source of independent, and at times unique, information, serving to reduce response burden created by direct interaction with the population. If parameters of distance, area, volume or any other attribute or function are needed, spatial analysis rather than the target population can serve as a source of the needed information.

Geospatial data collection and analysis anchor each data element in space and create complex entities in order to facilitate and enable systematic data integration and a defined coverage of relevant units. Proximity, partial overlapping, arial gravity centroids and other spatial relationships are functions used to define the input for the database that in turn, generates unusual census outputs.

Keywords: Integrated Census, Spatial Analysis, Spatial Sampling, Frame Development, Complex Entities

1 Introduction

1.1 The Integrated Census

Population and housing censuses involve data collection of all society members. These members are often defined as de-jure population of a country, meaning all people who usually reside within defined political geographic boundaries. In most countries, data is generated via a thorough scrutiny of the area and a direct interaction with the target population. This is the traditional way of data collection. However, more and more countries are trying to use existing files and databases for census purposes, in an indirect data collection process. The Nordic countries are more successful than others; Denmark and Finland perform a pure register based census, which relies solely on data collected by public agencies, usually governmental ones.

Norway and Sweden prepare an administrative census, by performing an integrated census where the source for persons data is administrative and the source for household data is a product of a traditional enumeration process. Israel is on the same path, yet the integration of data collection processes, planned for the next census, is bringing together a different ensemble, although the music is similar.

The idea is to use, rationally, existing sources of information for all census units and items, and to correct and supplement them by designated surveys (Blum,1999). This process introduces different meaning of a census, a statistical one, in which the population is represented although not all individuals are enumerated. The notion of a census as an actual count is lost in its fundamental attributes; the direct data collection and the inclusion of records of each individual in the database. Moreover, in a realm of high quality registers, the whole population is identified and recorded on the individual level and the need for samples is restricted to part of the attributes. It means that the character of the results is similar to those generated by the traditional process; almost all individual records are identified and have demographic and geographic location attributes, while all or part of the socio-economic traits are collected from a sample. However, when the available administrative records are of a mediocre quality and the population included is partial, sampling aims to individuals and attributes, demographic and socio-economic ones. Hence, the success of the integrated census depends on a very careful planning of the combination of administrative and surveys' data, bearing in mind the ability to reach estimates for small population groups and small geographic areas.

1.2 The GIS in the Integrated Census

In past censuses, Geographic Information System (GIS) has played an important supporting role in many different tasks: mapping localities, creating basic layers of buildings, addresses, street lineaments and statistical boundaries, creating maps in several cartographic forms and geo-analyzing population and household data. A more sophisticated use of the system had been introduced to the 1995 census in Israel including redistricting of enumeration areas, based on polling districts and the Central Population Register (Ben-Moshe,1997; Calvo,1998), management and control of the fieldwork operation, geocoding of returned questionnaires (and the data they carry) to the building level, imputation of missing data by proximity and more. The end result is a detailed spatial database, up to the building level, which enables a flexible spatial analysis.

The integrated 2006 census leads to further developments; heterogeneous sources of information imply selective coverage, different physical and analysis units, different formats and hence, not a straightforward connectivity or linkage of information and not a smooth warp and woof data blanket. Hence, multisource data collection requires some new and unique functions:

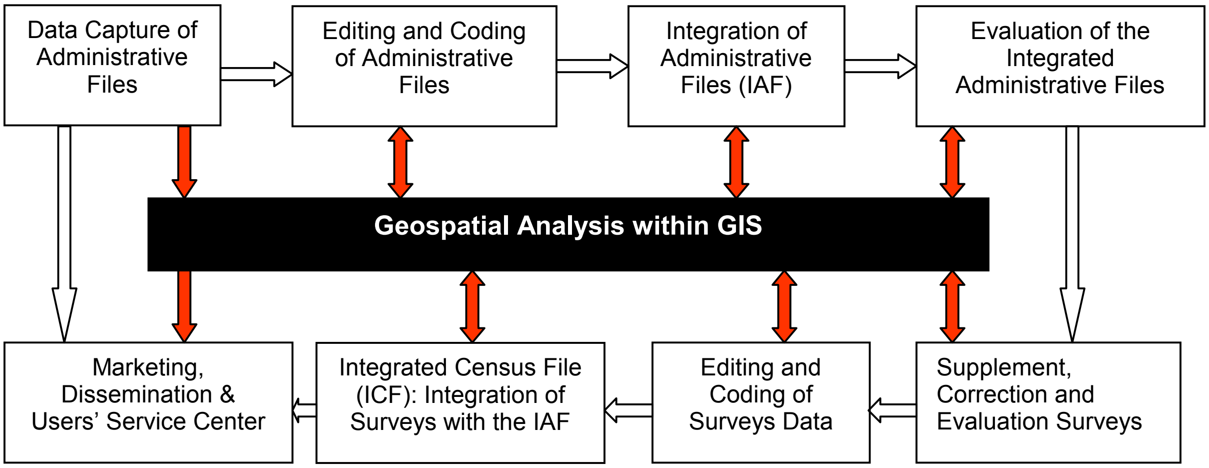
- an anchor for census data and therefore a data integrator;
- supplying the processes of the integrated census with an infrastructure;
- a source of an independent, and at times unique, information.

Data collection and geocoding anchor each data element in space and create complex entities in order to facilitate and enable a systematic data integration into one geo-referenced database and a defined and identified coverage of relevant units. Establishing a data infrastructure comes along with traditional database advantages: data integrity, quality assurance and tools for management and security. GIS provides data continuity and geographic reference, hence, a pretty good proxy for a census frame in which geographic relations are introduced as existing between entities such as buildings, city blocks, statistical areas, parcels, water meters, post office mailing addresses, zip codes and regional centers. The beauty of this idea is that it is not a prerequisite that entities would be nested in each other or would have any kind of ordered associations. They can be of a different shape and form and still have spatial relations via xy-coordinates within the same geospatial database. This feature enables the set up of a very complex set of entities within a GIS warehouse, after a preliminary stage of external integrity tests.

The ability to integrate information within geographic boundaries entails the use of tailored applications that enable the comparison between data input and the structure of the spatial database. Areas detected as carrying no data at all or carrying incomplete data, would require further treatment either by surveys or by imputation. The structure of the spatial database that refers to the xy-coordinates within geographic boundaries means that the GIS provides the census with two important infrastructure capabilities: drawing the census frame in geospatial terms and controlling the enumeration process by geospatial follow-up.

As for GIS as an independent source for census information in itself, it comes to serve one of the justifications to use existing data; the reduction of response burden created by direct interaction with the population. If parameters of distance, area, volume or any other spatial attribute or function, like spatial imputation, are needed, spatial analysis rather than the target population can serve as a source for needed information (see more details in section 2.3).

Schematic Description of the GIS location in the Integrated Census


Data Capture of Administrative Files
Editing and Coding of Administrative Files
Integration of Administrative Files (IAF)
Evaluation of the Integrated Administrative Files
Geospatial Analysis within GIS
Marketing, Dissemination & Users' Service Center
Integrated Census File (ICF): Integration of Surveys with the IAF
Editing and Coding of Surveys Data
Supplement, Correction and Evaluation Surveys

2 The Embedded Spatial Functions

The role of the geographic information system and the uses of multidimensional geospatial analysis are altered altogether in an integrated census. It is the combination of needs and embedded abilities that brings about an unexpected spatial census system. As shown in the above scheme, the GIS system is the nerve center of the integrated census and as such, performs diversified tasks: spatial record linkage used for data integration (between administrative files and between them and surveys' data), enumeration follow-up (in both data collection types, direct and secondary), evaluation, data editing and imputation, and spatial analysis for, inter alia, generating spatial census data (like the area of buildings' roofs), census frame, statistical record linkage of individuals or groups of individuals, and imputation. In return, the geospatial database is improved, updated, enriched and becomes a better tool for further production processes of statistics.

2.1 An Anchor for Data Integration

GIS was perceived as one of the most important technologies for integration of information in the new millennium (Dangermond, 1999). It is a powerful way to integrate data and furthermore, a powerful way to integrate sciences (Goodchild, 1999). Antenucci (1999) evaluates the GIS integration ability as a quantum change and Schaefer (1999) talks in terms of Jeffersonian Technology as a powerful tool for synthesizing and organizing information for the benefit of the lay man. Integration, synthesis and organization, using spatial references are abilities to be used *within* the next Israeli population census. What can be done between data sets and disciplines can also be done between census information sources.

Administrative registers and files, coming from different public and private organizations, may have different formats due to their unique orientation, may carry different identification indicators for people, such as CPR ID number, personal military number, employee's number, and may carry different identification indicators for other census entities, like localities, buildings and dwelling units. It forms several challenges when integrating data from different sources; there is a need to identify and isolate the potential spatial items and to determine how to best geocode them to the GIS while finding a common ground for all spatial indicators. Moreover, there is a need to find common personal indicators in all records of the same person. While geographic referenced system is built via spatial analysis, the personal indicators may enjoy the possibilities opened by this system, i.e. using shared spatial attributes to identify records as belonging to the same person.

This idea leads to the unique place and position of the GIS in the integrated census; each data item is introduced to its spatial attributes at each step along the census process. Once a data file or a database is captured, it is geocoded into the GIS and continues the census path while touching the spatial base at every step: evaluation of one file by another, integration of administrative data, follow-up of the fieldwork operation, integration of surveys data with administrative data, integration of census and non-census data, along time, for dissemination purposes. By localizing each individual in an identified point in a relevant space, the analysis of small area is possible, and hence the production of census information.

2.1.1 *Profiles of Spatial Record Linkage*

Record linkage is possible when the required link is between entities that share common attributes. In a world of integrated census, these shared attributes are heterogeneous. They can be of different definitions inspite of identical names or of different names inspite of identical definition, they can be census variables or linkage supporting ones, raw or calculated variables, with or without spatial pointers or identifiers, and collected for different purposes.

On the one hand, this complex reality presents difficulties when trying to link records. The needed harmonization of multi-source variables are partial, if at all, the files produced for different purposes may share only core and very basic variables, the file units may be of different analysis levels which do not necessarily refer to each other in an hierarchical order. On the other hand, heterogeneity of sources and the ability to use non-census data included in the administrative file, pave an indirect path for linkage and integration. This indirect path can be mediated by non-census variables or by a third file as a broker, in a chain linkage. For example, the individual record from the Population Register is linked to the Dwelling Units register through the two sided links of the mediator, the Property Register, as done in the Nordic countries. Within this frame of reference, when the policy declared is to use the GIS as the nerve cord of the census, all data units are to carry spatial attributes as a result of georeference to the same system of spatial identifiers, like xy coordinates. It means that all data units are linked to one another in a direct or indirect process of matching characters.

The actual linkage can be a result of a deterministic process, in which individual units are linked by individual ID and a set of verifying variables (date of birth, gender, etc.) and both are geocoded if at least one of them carries a spatial anchor. A deterministic linkage can be also achieved by linking two individual records through the spatial attributes they carry, even if the personal attributes are not unique. Examples of such spatial profiles are:

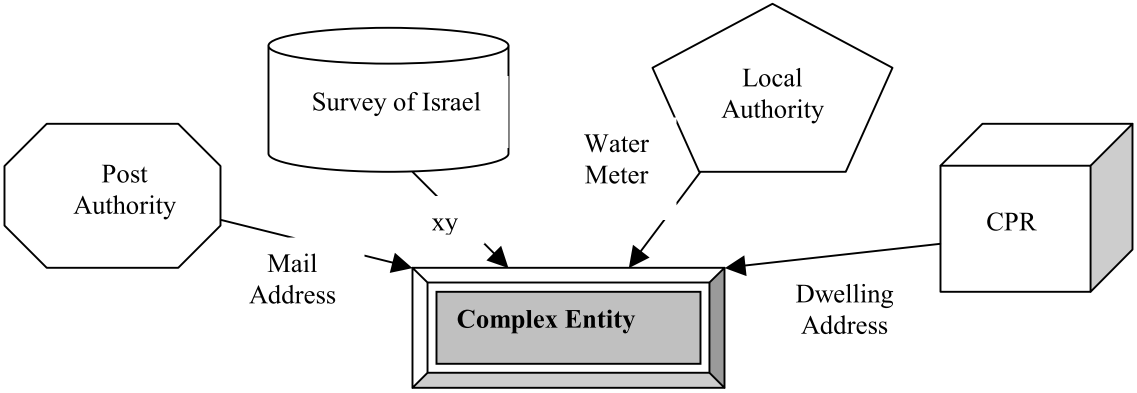
- locality code, street code, house number, first name, last name;
- water meter number, apartment number, first name and last name.

In these examples the GIS serve as a repository of links, however, it can also serve as a tool for linkage in a spatial-statistical process, when none of the attributes is unique, but some of the profiles are quite close to it. For example, locality code, street code, house number and first name to be linked with a record carrying water meter number, apartment number and first name or with a record carrying electricity poll number, house number and first name.

This sort of linkage is facilitated with an antecedent process of creating complex spatial entities through geospatial analysis. Once the electricity poll is correlated to the water meter and to the specific apartment within the building, the statistical apparatus is more reliable. Furthermore, the GIS functions provide modeling capabilities with high quality spatial referenced information for small geographic units (United Nations, 2000). Using this capabilities to link records can be done while creating an *ad hoc complex entity* rather than a permanent one; meaning that in different record linkage processes different profiles of spatial anchors are matched but not different individuals. When the goal is the population in its geographic location and not vice versa, it is worthwhile to reduce the demands for accuracy of the spatial location rather than the demands for the integrity of the data belonging to the identified population. In technical terms it means that the profile of the spatial entity is less refined, and therefore, it contains several smaller, not definable entities.

2.1.2 Reliability Levels

The process in which complex entities are created is a geospatial-statistical one. As explained above, linking records from different sources brings about new entities with multifaceted profiles; a house linked to a water meter to a mail address to a voltage wire, where each identifier has a tail of information with regard to people. These profiles are heterogeneous; their structure is diversified, they are stable or fluid, permanent or temporary, created in a deterministic process or in a statistical one. As a result, the reliability and cohesiveness of a complex entity is not a constant value but rather changing.



In a population census, the pivot of the complex entity, in the above illustration, would be the apartment (where the people reside in). Its profile would be comprised of the four different spatial identifiers: dwelling address (locality and street code, house and apartment number), water meter number, mail address (like mail box) and xy-coordinates. Linking them all to one entity is not a straightforward operation. At times, the same water meter would be linked to a neighboring entity and some of the mailing addresses wouldn't be qualified to be linked to any spatial entity. It means that the complex entities created should have a reliability index. The lower the reliability the higher the probability to change its composition.

The suggested components of the *reliability index* are:

1- the quality of the data sources;
2- the existence or absence of corroborating information;
3- stability of the data. Is it stable like the dwelling address or changing like the mail address;
4- frequency of updating and use. Water meter is considered very reliable in the local authorities databases because of the frequent use of water and because of it being a reference item for charging for water consumption;
5- uniqueness of the profile. Are there other complex entities that share one or more identifiers?

This index is an important decision making tool in an integrated census, where the information for a single person or for a single variable does not always derive from one source of information. Moreover, in a statistical census, where not all the people are actually enumerated in a direct or indirect interaction, the probability to err while creating complex entities and links is higher. There are some loose ends that have to be left open-ended so that micro data is not very accurate, but the macro level picture is not distorted.

2.2 Infrastructure

In population censuses, the boundaries of the area to be enumerated are defined beforehand and so is the geographic infrastructure with its discrete features (roads, houses). However, these features have additional role in an integrated census comparing with the conventional one, in the process of covering the defined area in population terms.

The traditional enumeration process involves a systematic inquiry of the populated area. In many cases it is a face-to-face interaction and therefore, an actual visit of census takers in the potentially populated houses. Census units and items, collected in a fieldwork operation, feedback to the original geographic infrastructure, complete it and form the census database.

Some of the countries that perform traditional censuses do not have a direct contact with the population at the first phase of the process. They use mail-out mail-in mechanism. This remote contact is possible only if all physical dwelling units within the census geographic boundaries are identified and carrying unique ID numbers. The bureau of the census in the US, for example, invests great efforts to improve the addresses file in order to have a complete and intact population coverage. This sort of operation implies that if the infrastructure is not built during the enumeration process, it should be based on a relevant updated database like the addresses file.

However, addresses for themselves are not enough if there is no mechanism to ensure full coverage of the census area. In a traditional census, there is a follow-up stage, where enumerators do visit problematic areas. In an integrated census, where most of the census population is not contacted at all, the spatial dimension stipulates the ability to cover it. The evaluation of the census has to be made against a ruler and the administrative sources cannot serve as such. They may be partial or full and not always with the metadata required to understand who and what are included and who or what are not. It means that the sample surveys needed to supplement and correct administrative data have to be drawn from an unknown frame. The body that represents the whole is either missing or not defined. The GIS can be considered as a geospatial-ruler because of its relatively final and stable spatial infrastructure items. It enables a spatial analysis whose results are *synthetic entities* within the area boundaries, to be used for spatial sampling and enumeration follow-up process.

2.2.1 Census Frame and Sampling

A conventional census is the only data collection process where the frame is built as an integral part of the process itself. Censuses have traditionally supplied a frame of population, households, dwelling units and other physical and analytical units for the following statistical activities. In an integrated census, data are first collected from administrative registers and if the target population is not defined as the population found in the administrative registers there are two possible modes of action; to detect the missing people and attributes in a field operation and to cover them all, or to use sample surveys to enable the imputation of missing data. The first option is costly and may be suitable in extreme cases when the missing units and items are well identified and located and when their scope is fairly limited (like the nomad population in the southern part of Israel). Otherwise, a traditional census may appear as a better choice, resources and reliability wise. The second option seems to be more reasonable when the administrative files are not as dependable as desired but are good enough to enable the identification of missing or biased population profiles. For example, if the registers show too many dwelling units comparing with people in a financially supported area, sampling the error prone areas or sampling people registered in these areas for correction survey, may provide corrected estimates of the population.

However, although registers provide indicators for problematic populations and areas, they still do not draw the census frame (again, unless the registers population is defined as the target population). At this stage of the census process the frame is expected to be identified for sampling and the surveys are expected to serve the goal of a full coverage database. Developing the census frame for the 2006 census of population in Israel is going to be based on spatial integration of information by spatial overlay, geocoding and analysis. This process of anchoring data in space and forming complex entities within spatial boundaries, define discontinuities of the boundaries and within them. In order to have an independent macro perspective, data generated *on different aggregate levels* should be introduced. As for the boundaries themselves, an external macro level input of the census area is the routine procedure. Countries usually define the census boundaries by the country borders. As for the discontinuities within, the complex entities created will serve as life indicators, pointing at a potential census population or census area (micro data viewpoint). A combination of micro and macro spatial analysis is a preferred way to draw and define an integrated census frame.

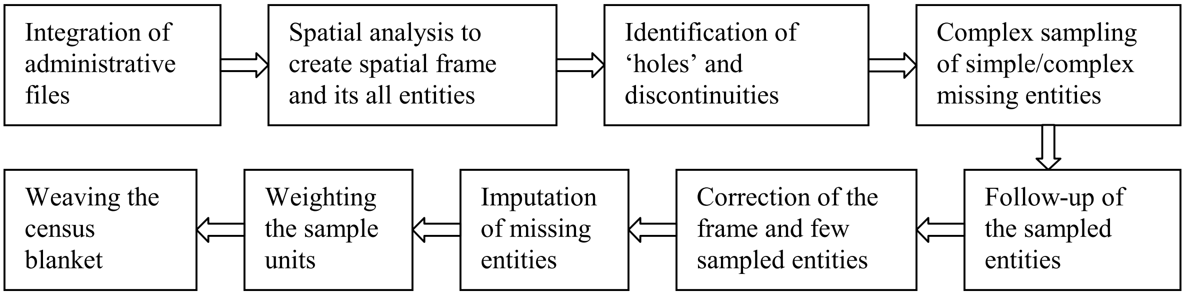
The spatial infrastructure, created for and during the integrated census, is the most tangible frame to relate to when completing and correcting administrative data by sample surveys. Sampling units can be of diversified nature, they can be of a raw unit, like people or buildings or of a complex entity, they can be based on a one-dimensional attribute, like a street, or on spatial one, like area. The spatial infrastructure does not stipulate the continuance of the process to *spatial sampling;* on the contrary, it introduces additional sampling ability to be carried out in combination with other methods, which could not be activated properly without the spatial frame.

2.2.2 Enumeration Follow-up

One of the most challenging tasks when conducting a census is its management and control. When the infrastructure is basically ready and the frame is well defined, the challenge derives from the size and complexity of the operation. However, in a multisource data collection approach, when the infrastructure is built during the process, when the frame is not a final list of units but rather a statistical product of spatial analysis and when the spatial entities are synthetic, management and control can become tricky.

The central issue is the enumeration follow-up of units that are not always defined and whose mere existence is unknown or in question. If the composition of a complex entity, and therefore its boundaries, are vague, it is not so clear what can be considered as covering the unit. For example, a situation where one of the entities is comprised of data collected from a record whose identifier is a parcel number and data from a record whose identifier is a regular address, while a second entity shares the parcel number but has a water meter number which is not nested in the regular address, and a third entity shares the water meter but not any other identifier. It seems that the best strategy is to enable the fieldwork done during the surveys to change the structure and the boundaries of the entities whose reliability index is relatively low, while keeping stable those

whose reliability index is high. Consequently, the goal of the spatial analysis, performed to define the frame, is to maximize the rate of fixed entities.
The suggested process within an integrated census is as follows:


Integration of administrative files
Spatial analysis to create spatial frame and its all entities
Identification of 'holes' and discontinuities
Complex sampling of simple/complex missing entities
Follow-up of the sampled entities
Correction of the frame and few sampled entities
Imputation of missing entities
Weighting the sample units
Weaving the census blanket

Since the entities and the frame are of spatial nature, coverage follow-up can be graphically presented. The census headquarters can ask for two or three-dimensional presentation of the enumeration progress, either triggered by a query or as an automatic process. Technological solutions will be needed for the interactive follow-up of changing entities.

If the census turns to be an ongoing process, coverage follow-up will begin with the capture of administrative data. The raw and the complex entities are already defined in the integrated file (the IAF) and their reliability is rated. The direct data collection in the field continues to serve as an input for a correction phase. Along time, less and less fieldwork will be needed and more stable the entities will be. Under the assumption that the goal wouldn't be altered, the complexity of the follow-up is expected to be reduced.

2.3 A Source of Independent Information

The technological ability of geographic information systems, using geospatial analysis and complex entities, creates added value beyond the infrastructure information and the collected data. Three census processes and GIS features produces information: spatial analysis, spatial imputation and integration of data into open spatial system.

The graphic presentation and the 3-D features imply the measurement of distances, areas and volumes not otherwise obtained. For example, the distance between a house and water hose, the total roofs area in a defined geographic perimeter and the volume of industrial space available in a given area. Moreover, information is generated with no pre-planning, by the mere interaction between alphanumeric and geographic information. It means that the integration of spatial and non-spatial data, coming from heterogeneous sources, ignites interactions that, in turn, produce new data.

As for the imputation capability, the ability to identify missing entities renders the ability to impute them. Nearest neighbor imputation, for example, will take into account the spatial dimensions and attributes of potential donors. However, taking advantage of the working environment means that instead of imitating the mere statistical imputation process, one should be engaged in analyzing spatial pattern of a phenomenon, regardless of any known boundary.

Administrative borders of statistical area or a locality shouldn't play a role, but rather an automatic process of drawing boundary lines of a phenomenon is to define the nearest neighbors.

Finally, since it is an open system and any data with spatial anchors can be integrated into it, the spatial information is expected to be richer, improved and more reliable**.**

3 Concluding Remarks

The scope of spatial analysis in processes of integrated census is yet to be explored. The spatial dimension and the physical infrastructures within the geographic information systems seem to be a natural choice for anchoring data in any practiced method of census taking. However, the integrated census has to lean heavier on geospatial analysis. Its administrative and survey data are integrated by geospatial analysis, imputation relates to patterns and attributes discovered by spatial analysis, and on top of it all, the basic census frame and the entities within it are defined and produced by spatial analysis. GIS, because of its capabilities to store, analyze, produce and present spatial and non-spatial information, provides a common denominator for the infrastructure, processes and products of the integrated census.

This approach of recruiting geospatial analysis as a fundamental and, to some extent, a critical component of a census, contributes to the census quality and expands its horizon. In previous censuses, technological development responded to and created new methodological needs. In the next census a third party is playing a principal part, the geographic information system that although embraces technology and methodology within it, can also interact in an evolving process with both. Some of the key concepts that are going to be introduced to census takers are: *complex spatial entities*, created for multi-source data integration, *flexible spatial entities* that enables the discovery of spatial patterns of phenomena, *reliability index* of an entity and *spatial sampling*, needed when the frame is missing.

Bibliography

Antenucci, John C. 1999 "The Coming Quantum Change in GIS Consulting and Systems Integration Services" in GIS 2000: the next millennium. ESRI.

Ben-Moshe, Eliahu 1997 "Integration of a National GIS Project within the Planning and Implementation of a Population Census" in Euro-Mediterranean Workshop on New Technologies for the 2000 Census Round. Israel

Blum, Olivia 2000 "Combining Register-Based and Traditional Census Processes as a Pre-defined Strategy in Census Planning". Statistical Policy: Working paper 30. 1999 FCSM Research Conference.

Calvo, Rinat 1998 "Redistricting Enumeration Areas and Defining the Organizational Structure of the 1995 Census" in the Proceedings of the Eighteenth Annual ESRI Users Conference. Redlands, California.

Dangermond, Jack 1999 "GIS in the Next Millenium" in GIS 2000: the next millennium. ESRI.

Goodchild, Michael F. 1999 "GIS in Science" " in GIS 20: the next millennium. ESRI.

Schaefer, Mark 1999 "GIS and Emerging Jeffersonian Technologies" in GIS 20: the next millennium. ESRI.

United Nations 2000 Handbook on GIS and Digital Mapping, UN, NY.