WP.27
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

Topic (v): New and emerging methods

GEOSPATIAL EDITING

Supporting Paper

Prepared by Olivia Blum & Rinat Calvo[1], Central Bureau of Statistics, ISRAEL

I. INTRODUCTION

1. Many multi-dimensional phenomena happen and interact in space, and therefore, space is relevant for their understanding (Calder, 2000). Spatial data have already been used in the production of spatial statistics; however, the technological developments of the last decade have supplied a platform and a boost for geospatial analysis, using statistical analysis tools within geographic information systems (GIS). The strength of this type of analysis is in the inclusion of spatial relationships in models, and in the use of spatial traits, like distance and volume, to better understand patterns and processes.

2. The military industry has initiated the technological development of GIS in order to improve orientation in space and to facilitate the perception of objects in spatial dimensions of interest. Civilian use of GIS has followed suit, especially in earth sciences, where the spatial dimension defines the core nature of the discipline. Characterizing patterns of air pollution, predicting natural disasters like earthquake or snow avalanche, analysing land use and its impact on the atmosphere, these are all examples of geospatial research and analysis overtime (Brabec, 2000). Graphic presentation of images and basic statistical characteristics are embedded in the geospatial systems and as such, provide the infrastructure for geospatial statistical analysis. Yet, their use is not widely spread and the benefits are not fully utilized.

3. Censuses, by their nature, are geography-dependent. The new censuses around the world are even more so: The US, France, the Netherlands and Germany rely on integration of information, at one point of time or overtime, in order to get small area census estimates (Durr, 2005; Schulte-Nordholt, 2005; Szenzenstein, 2005; Waite & Reist, 2005;). So is the case in the 2008 census of population and housing in Israel. It is basically an administrative census, based on the central population register (CPR). Two sources of information are used to correct it, administrative records and field evaluation surveys: Administrative records support the definition of the population frame on the individual record level. The field evaluation surveys supply the estimates for under- and over-coverage of the administrative files, on local and global levels. Since census estimates are of the total population in the census area, in localities and in small areas, the geospatial system is an ideal platform for most processes of data processing. Not only census processes are geographically based, but also the common denominator of the heterogeneous data sources is the geographic reference of each record. These geo-references enable data integration (Blum & Calvo, 2001; Calder, 2000). Editing and imputation (E&I) of census data can and should enjoy the advantages offered by the technological tools and by the spatial perception of statistical phenomena

[1] Email: blum@cbs.gov.il; rinatc@cbs.gov.il

on three levels: Graphic presentation of data, enhancement of data with spatial attributes, and geospatial statistical analysis.

4. In this paper there is a brief description of spatial statistics in geographic information systems and examples of deterministic spatial imputations. It does not present the more significant advantage of the method for spatial imputation, based on geostatistical analysis, yet, it gives the reader an idea about the potential use of the geospatial environment for E&I processes.

II. SPATIAL STATISTICS IN GIS[2]

5. The geospatial environment is multidimensional, comprising layers of geo-referenced information and time indicators. Different layers may carry different spatial entities, which can represent a. stable physical entities like buildings, roads, and city-blocks, b. virtual but stable entities, districted by administrative boundaries, like states, localities, statistical areas, c. ad-hoc virtual entities, comprised of a combination of other entities, whole or partial, like a polygon whose vertexes are road junctions with fatal car accidents within a cloud of emissions from industrial factories, d. other objects, like people and households, that can be linked to a spatial entity or can represent a physical spatial entity in constant motion. These spatial entities have four dimensions; x-y-z coordinates (location in space) and time to which this location corresponds. They can be characterized by non-spatial attributes like a colour of a building or an education of a person. Statistical analysis of these spatial entities and of non-spatial objects and attributes, refer to variability, correlations and trends within and between layers of information.

6. Statistical analysis in a geospatial environment can use statistical tools developed for that matter, or it can use the SAS software, since it has been integrated with the GIS. The statistical functions these tools provide are of descriptive statistics that include location, spread and shape of a distribution, trend analysis and prediction. The challenge lies in the need to work with spatial polygons as the computation units, rather than point values. It is answered by creating a Voronoi map of distance parameters, within a defined neighbourhood of polygons, in which a chosen point, the one that all locations within the polygon are closer to, represents the polygon. Neighbouring polygons are defined as such by the proximity of their point to the chosen point. The computation of statistics becomes familiar since now it relates to point values within a neighbourhood. It is done in iterations, like the calculation of a mean and a mean of means, when the relevant population is located in several neighbourhoods, within or across spatial layers (ESRI, 2005).

A. E&I of Spatial Phenomenon

7. Spatial analysis serves E&I by introducing variables of spatial location and variables of different sources that can be integrated to better specify the models. Yet, the more remarkable fact is that the ideology of E&I is embedded in the perception of the surrounding as phenomena in space. Spatial phenomenon relates to a continuous surface, like a wave or a cloud, which encompasses the whole population with the relevant attributes. Interpolation and extrapolation are an integral part of its characterization and creation and therefore, treating outliers and missing values is basically a default process in spatial analysis. The generation of a continuous surface, based on a sample, is parallel to estimating the population parameters from a sample. E&I are required to minimize the distance between the real surface and the generated (predicted) one.

8. Detecting global outliers has no advantage in geospatial systems; however, detecting local outliers benefits from the spatial dimension when using cluster and entropy methods, or a semivariogram in Voronoi map:
- Cluster method marks clusters of similar polygons and identifies those that cross dissimilarity threshold.
- Entropy methods mark polygons that are dissimilar to their neighbouring polygons. Areas with high entropy are possible local outliers.
- Semivariogram (half of the distance squared) is used to fit a model of spatial correlation. It can be used to identify outliers under the assumption that neighbours in space are more alike than those at a distance. All neighbouring pairs are checked and the high values of semivariogram are singled out.

[2] The examples of spatial statistics are based on applications developed by ESRI, since the GIS infrastructure in the CBS is their product.

9. As for imputation, spatial analysis uses deterministic and geostatistical solutions:
Deterministic techniques are based on parameters that control the extent of similarities of the values, or the degree of smoothing in the surface. Deterministic imputation is done simultaneously with the creation of homogeneous surface, when characterizing a spatial phenomenon. Discontinuities are treated by interpolation within the surface, according to its shape. Since it is a multi-dimensional environment, imputation into a surface can be of profiles and not only of single variables, given that there are enough variables that enable the reference of the record to the surface. Error is introduced when the parameters that control similarity and smoothing are erroneous: The record is ascribed to the wrong surface or the imputation within the surface creates wrong pairs of neighbours and alters the shape of the surface.
Geostatistical techniques are based on random processes with dependence. The resulting statistical models take into account the distribution of entities in spatial cells (hot spots, which are highly populated points, cold spots and outliers), the pattern of the distribution (uniform or random, clustered or dispersed) and the trends (local or global). The statistical models also calculate spatial correlations between multiple data sets and autocorrelations, over-time. As in non-spatial environment, the imputation solutions are based on the best-fit model; yet, the spatial presentation of a line with the lowest semivariogram values, or a predicted surface with the same trait. Errors are associated with the predicted values and can be presented as maps of standard errors attached to spatial entities.

B. Prerequisites

10. The two prerequisites for E&I in a geospatial system are the completeness of the geo-infrastructure, and the data integrity as presented by the entire set of geo-referenced data-items (Blum, 2005): The geospatial infrastructure has to be built and updated continuously, in a register-based approach. All reference polygons, like streets and buildings that the data have to be ascribed to, have to be complete and updated, otherwise the spatial environment is not fully defined. As for data integrity, all data items have to carry a geo-reference in order to be anchored to a spatial entity, otherwise, the spatial analysis will generate flawed results because of missing values.

11. There is a fundamental difference between the two prerequisites; the infrastructure is of relatively fixed spatial entities, while the spatial data items are of changing phenomena and therefore, they are temporary and relate to a point of time. Fixed spatial entities are completed and updated by a designated data collection in the field, and usually not by statistical imputation. E&I is useful for temporary entities.

12. The first process of imputation required is of spatial indicators. Geography matters when trying to analyse or predict socio-economic attributes. For example, the education of a person living in a remote village is probably different from that of a person living in the metropolitan core, since accessibility to education is also geography-dependent. As a result, the identification of one’s spatial location improves the imputation of his education. Spatial reasoning is not only instrumental, but also conceptual; the existential envelope of people is spatial and is influenced by surrounding processes. Indicators of personal characteristics and potential skills and qualifications are embedded in the spatial location.

III. DETERMINISTIC SPATIAL IMPUTATION

13. The following examples refer to imputation of spatial location within administrative records of the census-population in Israel. It is one of the first processes of data processing, since the following census processes depend on it, and as explained above, it is a prerequisite for any analysis in a spatial environment.

A. The Problem

14. In the Israeli 2008 integrated census, the estimation of the population is based on corrected counts of the CPR. This census is a geography-based process, in which an address is assigned to each individual record and, based on the results of the evaluation surveys, a census weight is attached to it. The address of an individual in his administrative record is a key factor. The better the initial address is, the better the small area estimates are. Moreover, the sample size of the over-coverage evaluation survey depends on the number of people who have not been found in their administrative initial address in the under-coverage evaluation survey, i.e. the better the initial address is, the lower the census costs are.

15. The problem is presented by incomplete address information in the CPR; not all records carry a detailed address (locality, street and house number) that can be referred to a building polygon on the map (the smallest reference geographic unit). However, judging by objectives, the required geo-reference should allow for population and demographic estimates on aggregated levels of sampling-cell and statistical area (3-5 thousand people), and not necessarily on a building level. Good estimates of the population in the sampling-cells are needed for sample size control and for the organization of the fieldwork, while estimates of the population in the statistical areas are defined as the census goal. The implication is that some of the imperfect addresses can still be imputed to a useful spatial surface, and latter can be geographically referred to the functional aggregates.

B. Spatial Imputation Processes[3]

16. There are small localities in Israel, like the Kibbutzim, that have no street addresses. The register records of their residents carry a locality name only. However, since they comprise one statistical area, no further treatment is required. The evaluation surveys try to answer the question if these people live in the locality or out of it.

17. There are large Arab localities in Israel that are divided into several statistical areas; yet, they have no street addresses for most of the locality, if at all. This is not a problem in a traditional census, when the enumerator goes door to door, however, it is a major problem in this unique census, that relies on geographic samples (of area cells) in order to assign coverage weight to the administrative records. The imputation conducted is a *cold-deck spatial imputation* of synthetic addresses given in the previous (traditional) census. It is a reasonable process due to low migration rates of the Arab population in Israel. It cannot be a simple cold deck imputation since the spatial entities of the localities have been changed during the last decade. The 1995 statistical areas, streets and buildings layers, are not identical to the updated ones, and both are needed for the imputation process together with the population records of the 1995 census and the relevant records in the updated CPR. The required data integration is of geographic infrastructure entities and statistical entities. The spatial imputation process is as follows:

- An individual record in the CPR is linked to an individual record in the 1995 census.
- The synthetic address in the record of the 1995 census is geo-referenced to a building polygon in the 1995 buildings layer.
- The building polygon in the 1995 layer is geo-linked to a building polygon in the updated layer.
- The synthetic address (street and house number) of the building polygon in the updated layer is imputed to the CPR record.

18. When the streets and houses in the locality carry identifiers but the addresses in the CPR records lack part of the information, it is possible to use spatial cold-deck imputation in a similar manner as in localities without address. The possibilities can be extended by geo-linking of different spatial entities, like a building polygon and a land parcel polygon or a point representing the location of an electricity consumption meter. If other sources of information have the record of the individual whose address is to be imputed, and they are geo-referenced to any spatial entity, *drilling down through the layers* enables the geo-linking of the entities and as a result, the matching of individual records, which stipulates cold-deck imputation.

19. When the cold-deck imputation process is exhausted, a *deterministic hot-deck spatial imputation* is a possible solution. The most straightforward imputation by interpolation is the case of a missing house number, not in the CPR record, but in the GIS infrastructure. Looking for neighbours is a spatial process; the neighbour of house number 36, for example, can be 37 or it can be 38 while 37 is on the other side of the street. Moreover, direction matters; the house numbers can go up in one direction or the other. Interpolation is possible once the direction and the location of the neighbours are identified. In this case, an ad-hoc spatial entity is created to represent the imputed address.

20. Imputation of an *aggregated spatial entity* is used when it can be totally ascribed to a statistical area. When a house number is missing but the full length of the street is within the boundaries of the statistical area, the imputation is of a location in a street polygon. If this polygon represents the

[3] See illustrations in attachments I-III.

population density along the street, the imputed address can be the geographic center-point, or, the population gravity-center.

21. A two-stage imputation enabled by the geospatial system is an *imputation of spatial intersection* of polygons. When two sources of information provide addresses that can be geo-referenced to spatial entities, which are not fully contained in one statistical area, the spatial intersection of the two may deliver the required results. It can be the case when a CPR record has the street but not the house number, and the street crosses boundaries of several statistical areas. If another file, of users of heating gas for example, carry a geographic identifier that relate to the centers they are connected to, the residential area of the users connected to a specific center can be represented by a new spatial entity. A cold-deck imputation of this entity to the record of the CPR, and a second hot-deck imputation in the CPR file, using this entity and the street data, may provide a group of possible locations within one statistical area.

IV. CONCLUDING REMARKS

22. The perception of our surrounding as a system of processes in space implies that the correction of a partial or a distorted picture of a section of this surrounding has to take into account the spatial elements, so a more accurate picture is reconstructed

23. In this paper we have presented the spatial environment and some of the spatial E&I processes enabled by the system, while creating and using spatial entities. Statistical analysis in this realm leans on measurements of distance and identification of neighborhoods, which are used extensively in non-spatial E&I. However, the spatial proximity indices incorporate a flexible geographic reference, which brings about new and heterogeneous definitions of neighborhood, and moreover, they integrate variables from contexts of other disciplines that are distributed in space and interact with the statistical data to be edited.

24. This is only a glimpse to geospatial statistical analysis, and to its potential use for editing and imputation processes. The existing technological applications provide us with sophisticated statistical analysis tools that should be used to improve the quality of the edited data.

V. REFERENCES

Blum, O. and R. Calvo (2001). “Geospatial Data Collection and Analysis as Crucial Processes in an Integrated Census”. 2001 FCSM Research Conference, Washington DC.

Blum, O. (2005). “The Geospatial Perception and Its Impact on The Content and Processes of a Multi-Source Data Collection. 2005 FCSM Research Conference, Washington DC.

Brabec, B. (2000). “A Nearest Neighbour Model for Regional Avalanche Forecasting”. UNECE Work Session on Methodological Issues Involving the Integration of Statistics and Geography. WP26. Neuchatel, Switzerland.

Calder, A. (2000). “Spatial Analysis – Why Bother?”. UNECE Work Session on Methodological Issues Involving the Integration of Statistics and Geography. WP12. Neuchatel, Switzerland.

Durr, J. M. (2005). “The French New Rolling Census”. Statistical Journal of the UNECE. Vol. 22 no. 1.

ESRI (2005). “ArcGIS 9: The Principles of Geostatistical Analysis”; “ArcGIS 9: Exploratory Spatial Data Analysis”. ESRI Software Documentation Library.

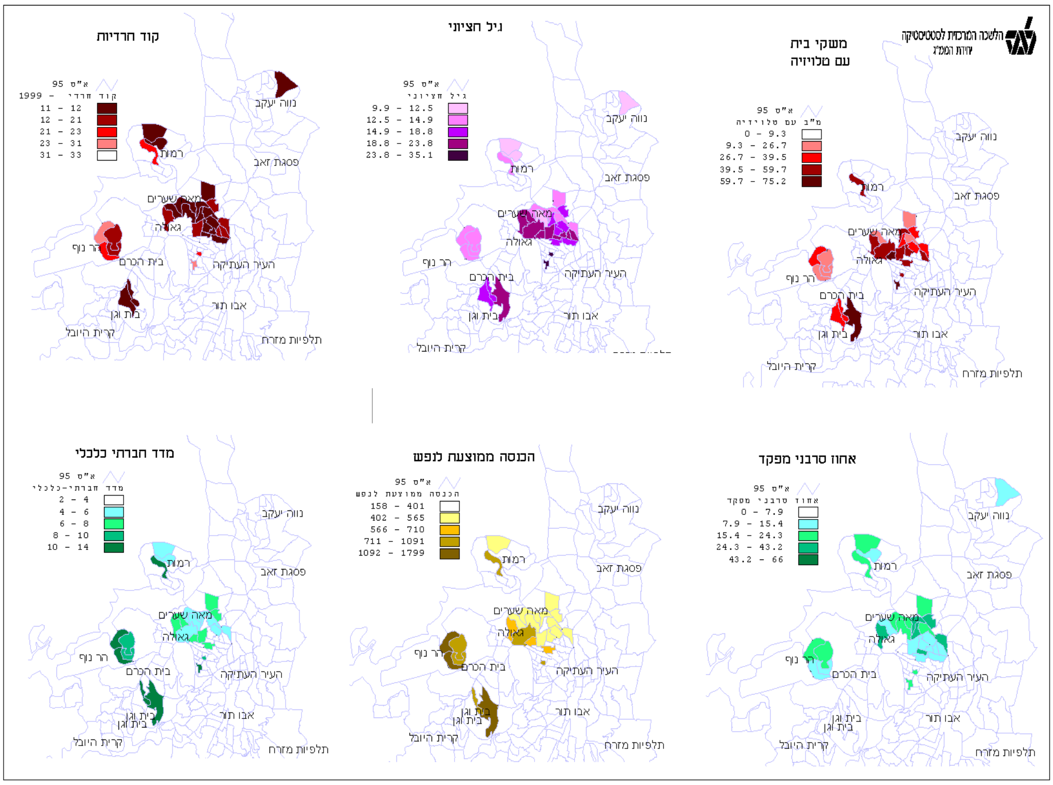
Schulte Nordholt, E. (2005). “The Dutch Virtual Census 2001: A New Approach Combining Different Sources”. Statistical Journal of the UNECE. Vol. 22 no. 1.

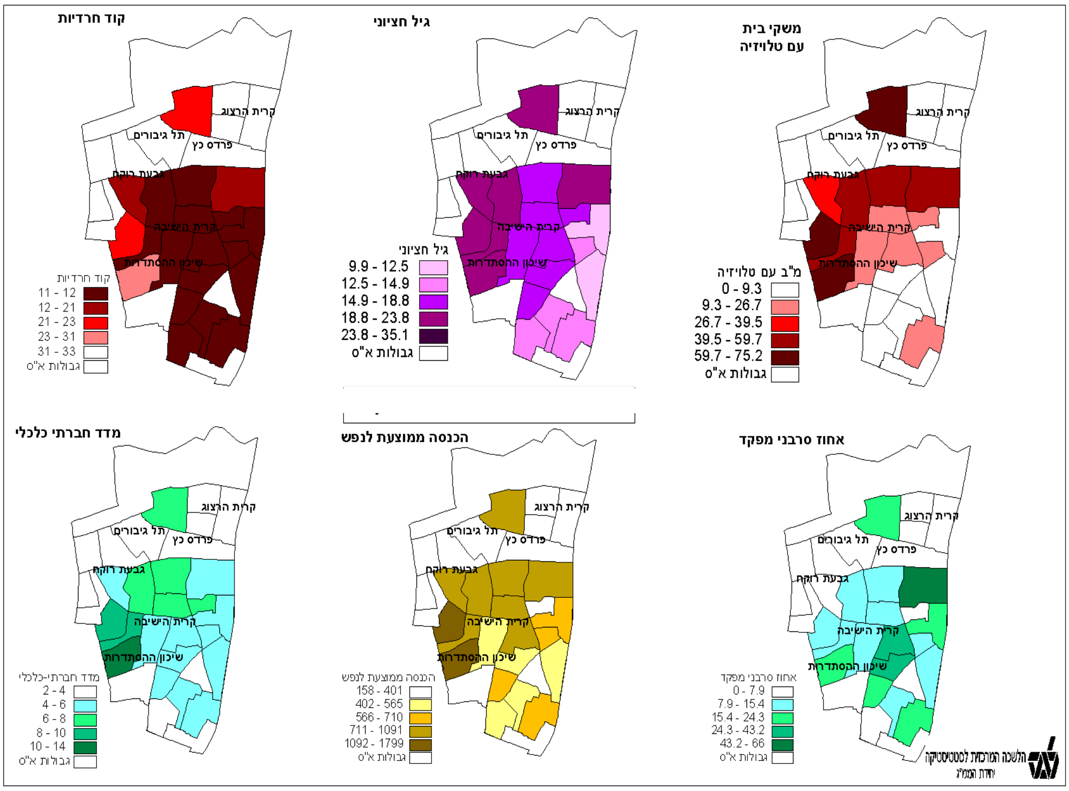
Szenzenstein, J. (2005). “The New Method of the Next German Population Census”. Statistical Journal of the UNECE. Vol. 22 no. 1.

Waite, P. J. & B. H. Reist (2005). “Reengineering the census of population and housing in the United States”. Statistical Journal of the UNECE. Vol. 22 no. 1.

ATTACHMENT I

The following illustrations are of two groups of six thematic maps characterizing the population in specific statistical areas of two cities in Israel (See explanation bellow).

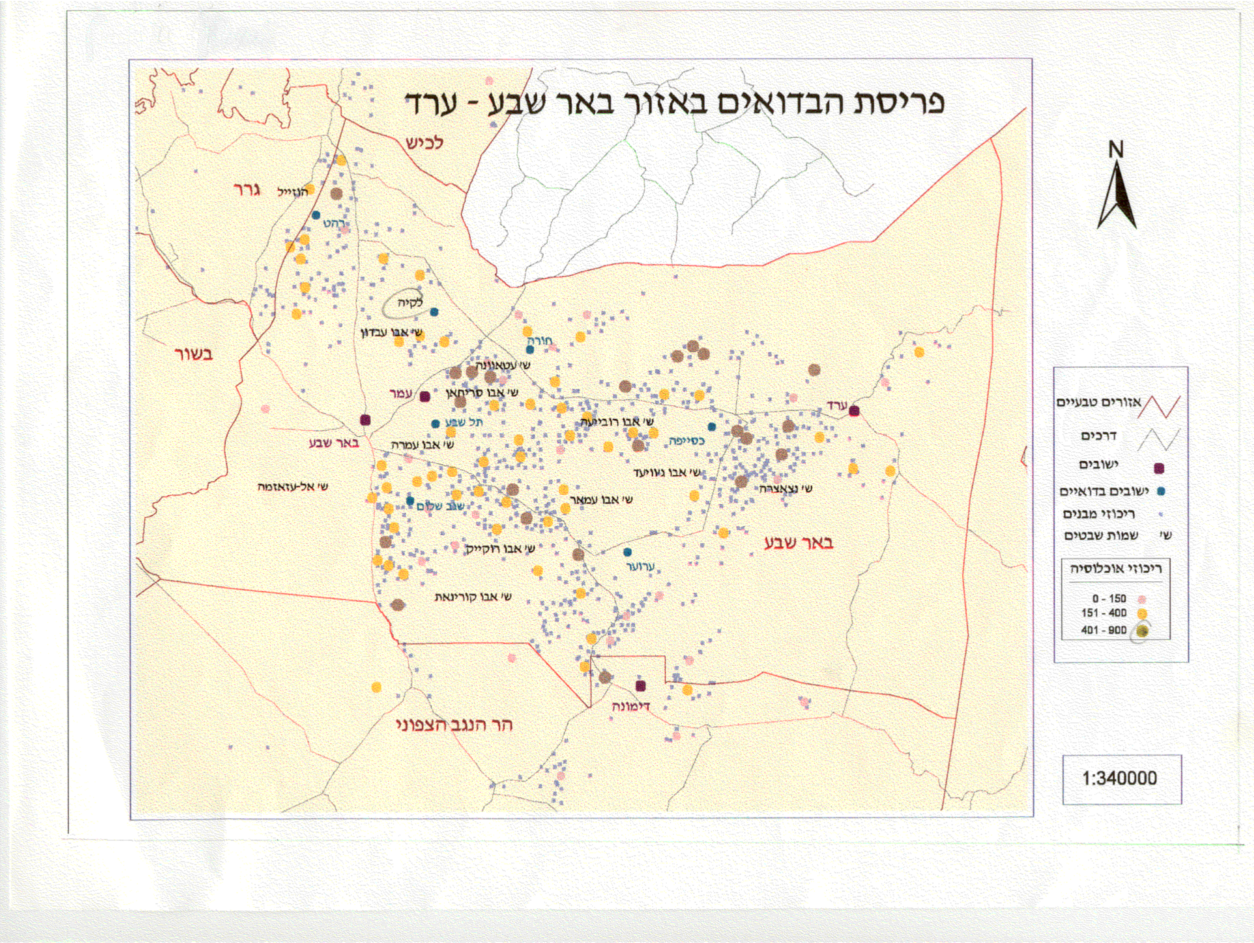




Drilling down through the different maps, in the same x-y coordinates or through the aggregates of the same statistical areas, provides geography-depend profiles. The different profiles represent the characteristics of different locations and geographic aggregates, and therefore, enable imputation of the nearest neighbor, for example.

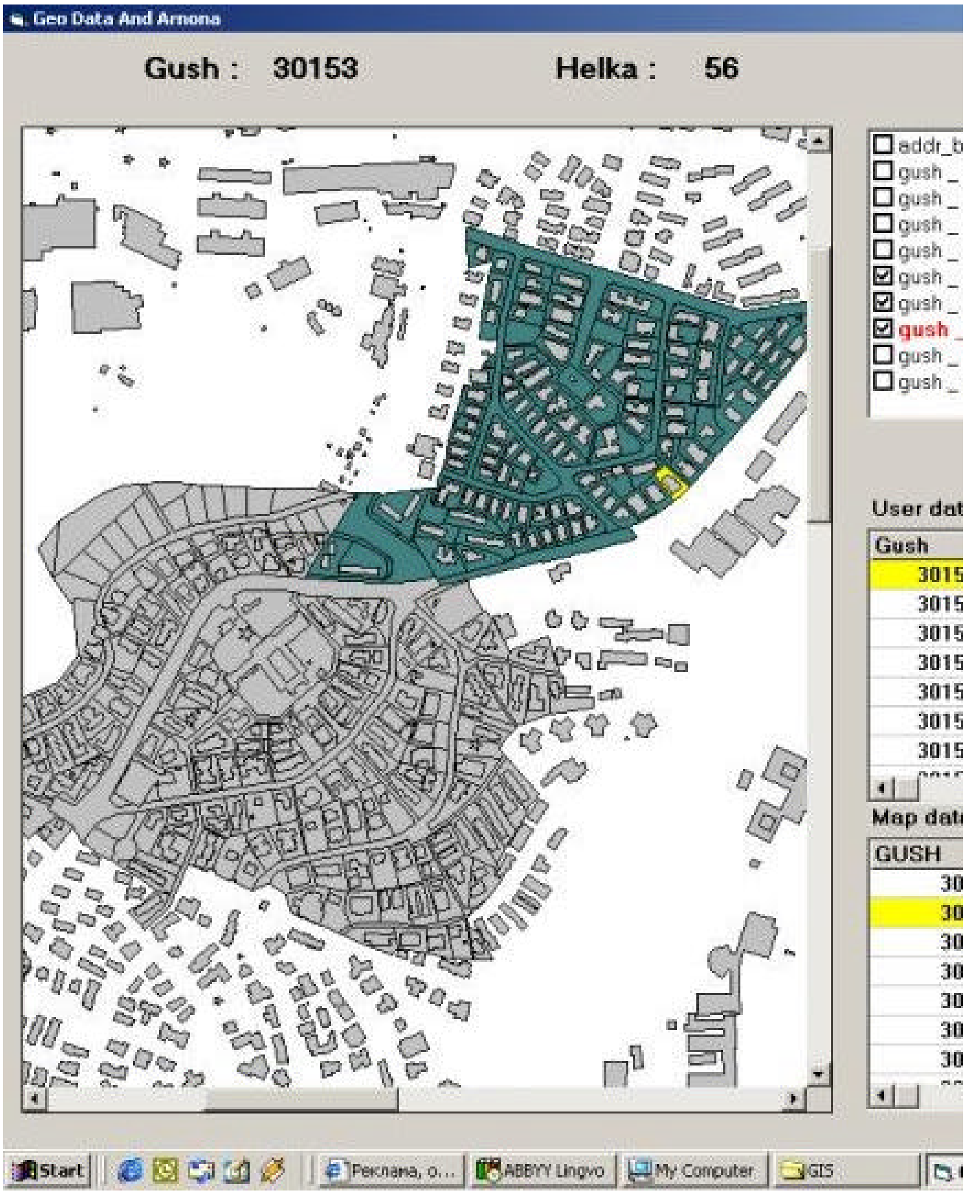
ATTACHMENT II

The following illustration is of spatial distribution of the nomadic population in the Negev Desert. One of the results of distance analysis is that this population is always within three kilometres from a main road. Any imputation of location should take it into account.



ATTACHMENT III

The following illustration displays the yellow building as a result of intersects of three spatial entities, buildings, streets and land parcels. The missing location of a household is the street address of this building.



Gush : 30153 Helka : 56

□ addr_b
□ gush _
□ gush _
□ gush _
□ gush _
☑ gush _
☑ gush _
☑ gush _
□ gush _
□ gush _

User dat

Gush
3015
3015
3015
3015
3015
3015
3015

Map dat

GUSH
30
30
30
30
30
30
30